Combined proceedings of the

Nordic Sound and Music Computing Conference 2019

and the

Interactive Sonification Workshop 2019

18-20 November, 2019

Stockholm, Sweden

Editors:

Andre Holzapfel and Sandra Pauletto

Title Proceedings of the Nordic Sound and Music Computing Conference 2019 (NSMC2019) and the Interactive Sonification Workshop 2019 (ISON2019)

Editors Andre Holzapfel and Sandra Pauletto

Publishers KTH Royal Institute of Technology

Scientific Committee for 1st Nordic SMC

- Roberto Bresin, KTH Royal Institute of Technology, EECS School of Electrical Engineering and Computer Science, Stockholm, Sweden
- Gerhard Eckel, IEM, University of Music and Performing Arts Graz, Graz, Austria
- Henrik Frisk, Royal College of Music, Stockholm
- Kjetil Falkenberg Hansen, KTH Royal Institute of Technology, EECS School of Electrical Engineering and Computer Science, Stockholm, Sweden
- Andre Holzapfel, KTH Royal Institute of Technology, EECS School of Electrical Engineering and Computer Science, Stockholm, Sweden
- Sandra Pauletto, KTH Royal Institute of Technology, EECS School of Electrical Engineering and Computer Science, Stockholm, Sweden
- Bob Sturm, KTH Royal Institute of Technology, EECS School of Electrical Engineering and Computer Science, Stockholm, Sweden
- Stefania Serafin, Aalborg University, Copenhagen, Denmark
- Sofia Dahl, Aalborg University, Copenhagen, Denmark
- Vesa Välimäki, Aalto University, Espoo, Finland
- Alexander Refsum Jensenius, University of Oslo, Oslo, Norway
- Stefan Östersjö, Luleå University of Technology, Luleå, Sweden
- Rúnar Unnþórsson, University of Iceland, Reykjavík, Iceland

Scientific Committee for ISon 2019

- Roberto Bresin, KTH Royal Institute of Technology, EECS School of Electrical Engineering and Computer Science, Stockholm, Sweden
- Henrik Frisk, Royal College of Music, Stockholm
- Kjetil Falkenberg Hansen, KTH Royal Institute of Technology, EECS School of Electrical Engineering and Computer Science, Stockholm, Sweden
- Thomas Hermann, CITEC, Bielefeld University, Bielefeld, Germany
- Sandra Pauletto, KTH Royal Institute of Technology, EECS School of Electrical Engineering and Computer Science, Stockholm, Sweden

Preface

The Interactive Sonification Workshop and the 1st Nordic Sound and Music Computing Conference were held from 18 to 20 September 2019 in Stockholm, Sweden.

These two events gather researchers and practitioners from different fields in the intersectional area between Art, Technology and Design, stimulating on open discussion and exchange of ideas that will hopefully encourage new research, innovation, creativity, and the development of a sustainable society in the fields of Sound and Music Computing and Sonification.

We would like to thank the local organizing team in Stockholm, the reviewers who supported our conferences, as well as our sponsors, KTH Royal Institute of technology, Nordic SMC Network, NAVET, NordForsk, KMA Royal Academy of Music, and KMH Royal College of Music who made the organization of these two events possible.

Stockholm, November 2019.

The organizers
Roberto Bresin, Kjetil Falkenberg Hansen, Thomas Hermann, André Holzapfel, Sandra Pauletto

Contents

Chapter 1

Keynotes

Kia Höök - Professor at Media Technology and Interaction Design, KTH Royal Institute of Technology

Soma Design – intertwining aesthetics, ethics and movement

I will discuss soma design —a process that allows designers to examine and improve on connections between sensation, feeling, emotion, subjective understanding and values. Soma design builds on pragmatics and in particular on somaesthetics by Shusterman — combining soma as in our first person sensual experience of the world, with aesthetics as in deepening our knowledge of our sensory experiences to live a better life. Soma design engages with bodily rhythms, touch, proprioception, bodily playfulness, but also with our values, meaning-making processes, emotions, ethics and ways of engaging with the world. Soma design also provides methods for orchestration of the 'whole', emptying the digital and physical materials of all their potential, thereby providing fertile grounds for meaning-making and engagement. Soma design is imbued with ideals of what a better life might be.

In my talk, I will discuss how aesthetics and ethics are enacted in a soma design process. Our cultural practices and digitally-enabled objects enforce a form of sedimented, agreed-upon movements, enabling variation, but with certain prescribed ways to act, feel and think. This leaves designers with a great responsibility as these become the movements that we invite our end-users to engage with, in turning shaping them, their movements, their bodies, their feelings and thoughts. I will argue that by engaging in a soma design process we can better probe which movements lead to deepened somatic awareness; social awareness of others in the environment and how they are affected by the technology-human assemblage; enactments of bodily freedoms rather than limitations; and aesthetic experience and expression.

Ulf Olausson - Foleyworks AB, IMDB page: https://tinyurl.com/yfsjhwqc

Making Film Sound Real

In this talk I will describe the work of an Foley Artist and show a couple of examples how we create sounds on movies and Tv series. This will show how much sound work is made in the Postproduction.

Eleonora Maria Irene Oreggia - xname.cc, Queen Mary, University of London

Audiovisual performance composition using electromagnetic waves

The enigma of the electromagnetic field is explored from a perceptual perspective, transforming it into a sensor system for contactless interaction and rendering it musical material that can be present to the mind. The relation between the human body, the perpetual capability of the senses and the subsequent image of reality is questioned: can technology expand the senses and become a method to discover new forms of perception? The result of this imaginative and speculative approach to engineering is REBUS, a novel, fully programmable musical instrument that innovates the 100 years old Theremin technique, allowing a new dimension for interactivity and expressivity in new media art and electronic music.

Chapter 2

Papers presented at Nordic SMC 2019

SONIC FOOTPRINTS OF WEB TRACKING

Otto Hans-Martin Lutz
Weizenbaum Institute for the Networked Society
Fraunhofer FOKUS
otto.lutz@fokus.fraunhofer.de

Manfred Hauswirth
Weizenbaum Institute for the Networked Society
Fraunhofer FOKUS, TU Berlin
manfred.hauswirth@fokus.fraunhofer.de

ABSTRACT

Web tracking is found on 90 % of common websites. It facilitates online behavior analysis which can reveal insights into sensitive personal data of an individual. Most users are not aware of the amount of web tracking happening in the background. We present an interactive demonstration of web tracker sonification which is designed to promote awareness for online privacy issues by disclosing the amount of web tracking through sound. While the user is browsing the web, data exchanged with web tracking hosts is transformed into sonic events.

In the demonstration, users can try out different websites testing several browsers, with and without ad blocking add-ons, and compare the different sonic experiences. Additionally, we provide a Wi-Fi access point to which users can connect with their mobile devices so they can listen to the data transferred to third-party tracking providers from their own devices. This demonstration aims to spark discussion not only regarding the sonification design, but to widen the discourse into a critical reflection of online privacy and surveillance issues.

1. INTRODUCTION

Web tracking is the collection of information about a particular user's activity on the World Wide Web. Web tracking technologies are found in 90 % of common websites and in 60 % of websites with highly privacy-critical content [1]. A person's browsing behavior can reveal insights into his or her personality, habits and sensitive aspects such as financial and medical situation or political views. Hence, web tracking may constitute a serious privacy threat [2]. Although complex and very diverse, the ecosystem of web trackers is dominated by a small number of companies, notably by Google and Facebook, who are inconspicuously present as third-party data collectors on many websites [3]. The average user has no sufficient awareness of the extent and possibilities of web tracking [4].

We use sonification of hidden web tracking as a means of raising awareness, creating interest and stimulating reflection on web tracking and privacy issues. The data exchanged with web tracking providers is transformed to direct auditory feedback while browsing. Using sound for disclosing the tracking activity allows the user to reflect on web tracking and privacy issues without having to interrupt the interaction on the web in the visual modality.

2. CONCEPT AND FRAMEWORK

Our framework for live web tracking analysis monitors network traffic, filters connections to known web trackers, extracts tracking-related events, and sends these to a sound generator via OSC [1] (Figure 1). The software runs in the background while the user is browsing the web.

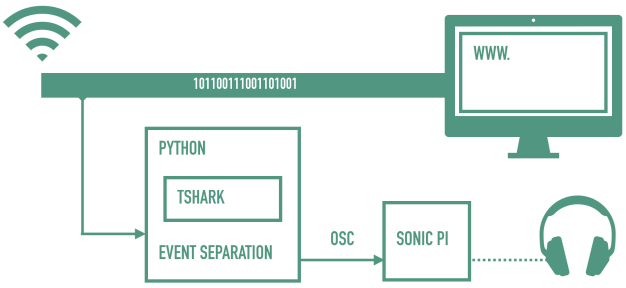


Figure 1. System overview

The framework is implemented in Python. We monitor network traffic with several TShark [2] processes. These processes listen to traffic on the selected network connection on ports 80 (HTTP) and 443 (HTTPS). They are configured with filter lists of web tracker IP addresses, so only traffic to these hosts is analyzed. Two types of events are extracted: Initial connection to a tracking host (SYN) and data exchange with that host (GET and TLS Application Data). Data exchange is sonified by short bleeps consisting of a layered sound (low mallets and high shaped noise). Each tracker IP address is mapped to a certain note. If data is exchanged with the same IP again, this results in the same sound. Although, due to the vast amount of trackers, several IPs are mapped to the same note. An exemplary video is hosted on YouTube [3].

To raise awareness for the prevalence of the top 10 most widespread tracking companies, an audio recording of the company's name is played in a whispered manner each time an initial connection is made to the respective host. Some of the companies are well known to users, others are not (e.g., ComScore). The whispered names are supposed to stimulate questions about these companies as well.

[1] Open Sound Control: *http://opensoundcontrol.org*
[2] text-based version of wireshark: *https://www.wireshark.org*
[3] *https://www.youtube.com/watch?v=ug3GfEe801k*

Further details on our framework for web tracker sonification are given in our ICAD 2019 paper [5]. Compared to the framework presented there, key enhancements of the system used for this demonstration include optimized country-specific filter lists, choice of the sound generator (open source Sonic Pi [6] in place of Ableton Live) and the ability to demonstrate the system not only with a local browser, but to test it with any Wi-Fi enabled device, e.g., laptops or smartphones. We open a secondary Wi-Fi network where our computer acts as an access point. Users can connect to this network with their own device and acoustically explore the data transfer from these devices to web tracking hosts. The sounds are generated on the host computer. In addition to monitoring tracking on websites, this enables listening to exchanges of data between other applications (e.g., mobile games) and web tracking companies as well.

3. DEMONSTRATION

We present two interactive demonstrations of our web tracker sonification. In the first demonstration, users can explore websites of their choice with two browsers (Mozilla Firefox with activated ad blocking plug-ins and native Google Chrome). This aims to raise awareness of web tracking and data-driven economies. Allegedly "free" websites and online services are often connected to a variety of web trackers as part of their business model.

For the second demonstration, we open a secondary Wi-Fi network which users can connect their own personal devices to. This facilitates sonification of traffic to web trackers from user devices and comparison of applications in terms of privacy. The audience can explore the sonic variations between different smartphones and apps. By using their own devices, there is a personal connection to the tracking events and to the data exchanged. This is aimed to spark discussions about tracking, privacy and surveillance on a personal level.

4. RELATED WORK

Sonification of network traffic monitoring to achieve higher situational awareness has been researched widely (see the overview in [7]). The scope of our approach, however, focuses on the average user who, in contrast to network operations professionals, is often unaware of the extent of web tracking [4]. Here, the term "awareness" refers to a general consciousness of the prevalence of web tracking.

Soundbeam by Hutchins et al. [8] is a related project. It sonifies third-party connections extracted by Mozilla Lightbeam, a plug-in for the Mozilla Firefox browser. Our approach of monitoring the internet traffic itself instead of using a browser plug-in extends the potential by supporting all browsers and applications, different web tracker lists, and the capability to monitor traffic of any device connected to a Wi-Fi spawned by the host computer.

Listening Back by Guffond [9] aims to raise awareness for online surveillance and privacy issues. It is a plug-in for the Chrome browser that sonifies access to browser cookies in real time during browsing and aims to be a "mediator of the invisible", analogous to our approach.

5. CONCLUSION

We demonstrate two applications of our web tracker sonification: First, trying out different websites, browsers and ad blocking add-ons and second, providing a Wi-Fi network that visitors can connect to with their mobile devices to experience the data transferred to tracking providers by their own devices and applications. This provides tangible access to the relationship between usage of online services and tracking data collection. The demonstration is aimed to spark discussion in three domains: regarding the technical implementation of the framework, the sonification and sound design choices, and addressing online tracking, surveillance and privacy issues in a broader scope.

Acknowledgments

This work has been funded (in part) by the Federal Ministry of Education and Research of Germany (BMBF) under grant no. 16DII111 ("Deutsches Internet-Institut").

6. REFERENCES

[1] S. Schelter and J. Kunegis, "On the Ubiquity of Web Tracking: Insights from a Billion-Page Web Crawl," *Journal of Web Science*, no. 4, pp. 53–66, 2018.

[2] A. Acquisti, L. Brandimarte, and G. Loewenstein, "Privacy and human behavior in the age of information," *Science*, vol. 347, no. 6221, pp. 509–514, 2015.

[3] S. Macbeth, "Tracking the Trackers: Analysing the global tracking landscape with GhostRank," Cliqz GmbH, Tech. Rep., 2017.

[4] W. Thode, J. Griesbaum, and T. Mandl, ""I would have never allowed it": User Perception of Third-party Tracking and Implications for Display Advertising," in *Proc. International Symposium on Information Science*, 2015.

[5] O. H.-M. Lutz, J. L. Kröger, M. Schneiderbauer, and M. Hauswirth, "Surfing in Sound: Sonification of hidden web tracking," in *Proc. International Conference on Auditory Display*, 2019, pp. 306–309.

[6] https://sonic-pi.net, [Accessed 13.08.2019].

[7] L. Axon, S. Creese, M. Goldsmith, and J. R. C. Nurse, "Reflecting on the Use of Sonification for Network Monitoring," *Proc. SECURWARE 2016*, pp. 254–261, 2016.

[8] C. Hutchins, H. Ballweg, S. Knotts, J. Hummel, and A. Roberts, "Soundbeam: A Platform for Sonyfing Web Tracking," *Proc. International Conference on New Interfaces for Musical Expression*, pp. 497–498, 2014.

[9] J. Guffond, "Listening Back," in *Proc. International Conference on Auditory Display*, 2019, pp. 351–354.

ESTIMATION OF PINNA NOTCH FREQUENCY FROM ANTHROPOMETRY: AN IMPROVED LINEAR MODEL BASED ON PRINCIPAL COMPONENT ANALYSIS AND FEATURE SELECTION

Riccardo Miccini
Aalborg University
rmicci18@student.aau.dk

Simone Spagnol
Aalborg University
ssp@create.aau.dk

ABSTRACT

In this paper, anthropometric data from a database of Head-Related Transfer Functions (HRTFs) is used to estimate the frequency of the first pinna notch in the frontal part of the median plane. Given the presence of high correlations between some of the anthropometric features, as well as repeated values for the same subject observations, we propose the introduction of Principal Component Analysis (PCA) to project the features onto a space where they are more separated. We then construct a regression model employing forward step-wise feature selection to choose the principal components most capable of predicting notch frequencies. Our results show that by using a linear regression model with as few as three principal components, we can predict notch frequencies with a cross-validation mean absolute error of just about 600 Hz.

1. INTRODUCTION

Binaural sound rendering can be achieved by incorporating the acoustic effects of the human head into a given sound, so as to simulate the pressure at the entrance of the ear canals. The set of functions used to perform this are called Head-Related Transfer Functions (HRTFs), and consist of digital filters characterizing sounds coming from a specific point in space. Unfortunately, obtaining personal HRTFs is only possible with expensive equipment and invasive recording procedures. For this reason, non-individual HRTFs are often preferred in practice, with the drawback of being prone to systematic localization errors such as front/back reversals, wrong elevation perception, and inside-the-head localization [1].

The most relevant differences between the HRTFs of two subjects are due to the different shapes, sizes, and orientations of the pinnae. The pinna has a key role in shaping HRTFs because of the reflections and resonances occurring in its rims and cavities, which can be seen in the HRTF as sequences of notches and peaks, respectively. The spectral location of peaks and notches represents a pivotal cue to the characterization of the sound source's spatial position, and in particular of its elevation. Despite the availability of various recent research works targeted at predicting the HRTF or some of its features from pinna anthropometry (see for instance [2–4]), we are still far from a complete understanding of the underlying relationships.

The relationship between the center frequencies of the three main pinna notches (known as N_1, N_2, and N_3) in a set of frontal median-plane HRTFs and 13 different anthropometric features of the pinna was explored in [5] with linear regression models. The anthropometric feature set included global pinna measurements (e.g. pinna height, concha width) as well as measurements that vary with the elevation angle of the sound source (distances between the ear canal and pinna edges). The results of that work showed that while the considered features are not able to approximate with sufficient accuracy neither the N_2 nor the N_3 frequency, eight of them are sufficient for modeling the frequency of N_1 within an acceptable margin of error, and that distances between the ear canal and the outer helix border are the most important features for predicting N_1.

In this work, we take a step forward by further investigating the model presented in [5] on the same input data and considering linear transformations and selection within the feature space in order to improve N_1 prediction. Given the presence of high correlation among features as well as repeated anthropometric parameters for each record pertaining a certain subject, we propose the introduction of Principal Component Analysis (PCA) to project the features onto a space where they are more separated. Subsequently, we apply feature selection on the regression model in order to preserve the components with higher predictive power, thereby reducing overfitting.

2. METHODS

2.1 HRTF feature extraction

The raw dataset consists of measured Head-Related Impulse Responses (HRIRs) for the 33 subjects from the CIPIC database [6] for which full anthropometric data (records and single-ear pictures) is available. We consider HRIRs measured in the frontal half of the median plane, with elevation ranging from $\phi = -45°$ to $\phi = 45°$ at 5.625-degree steps (17 HRIRs per subject). Elevations higher than $45°$ were discarded because of the general lack of spectral notches in the corresponding HRTFs [7].

Pinna notch frequencies in each HRIR are extracted with the *ad-hoc* signal processing algorithm by Raykar *et al.* [8].

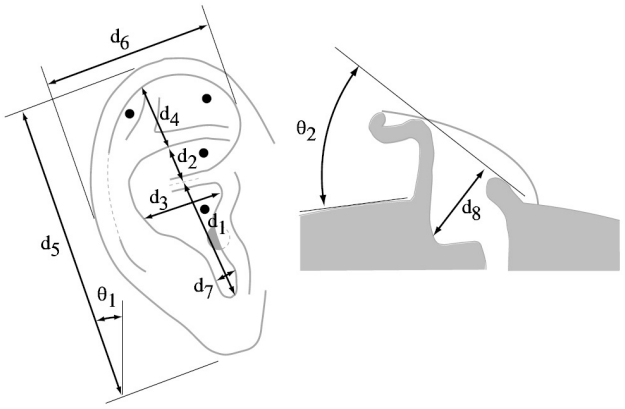


Figure 1. The 10 pinna parameters included in the CIPIC database (figure reproduced from [6]).

Then, for each available elevation, the extracted notches are grouped in frequency tracks along adjacent elevations with the McAulay-Quatieri partial tracking algorithm [9], with a matching interval of $\Delta = 1\,\text{kHz}$ [10]. When available, the three longest tracks are labeled as N_1, N_2, and N_3 in increasing order of average frequency; if a subject lacks a notch track, labels are assigned according to the closest notch track frequency median [5]. Given the previously reported low correlation between anthropometric parameters and the two notches N_2 and N_3, we focus on N_1 as prediction target, for which we have a total of 367 different observations belonging to 29 different subjects.

2.2 Anthropometric feature extraction

In addition to the 10 global pinna features contained in the CIPIC database and reported in Fig. 1, we extract 3 elevation-dependent features from scaled individual pinna images according to the following ray-tracing procedure. The three contours corresponding to the outer helix border, the inner helix border, and the concha border/antitragus (C_1, C_2, and C_3 respectively) are traced by hand and stored as sequences of pixels. Then, as can be seen in Fig. 2, the point of maximum protrusion of the tragus is chosen as the reference ear-canal point for the computation of distances. For each elevation $\phi \in [-45, 45]$, we compute distances in centimeters between the reference point and the point intersecting each pinna contour along the ray originating from the reference point with slope $-\phi$, and store them as $r_k(\phi)$, where $k \in \{1, 2, 3\}$ refers to contour C_k.

We assume that the $r_k(\phi)$ features are, together with the 10 individual global pinna features, good predictors of the N_1 frequency in the HRTF measured at elevation ϕ. This assumption is due to the results of a previous work [11] that highlighted a qualitatively reciprocal linear relationship between distance from the ear canal to the hypothesized pinna reflection points and pinna notch frequencies.

2.3 Dimensionality reduction

The so created dataset is composed of 13 features (d_i, $i = 1 \dots 8$, θ_j, $j = 1 \dots 2$, $r_k(\phi)$, $k = 1 \dots 3$, $\phi \in [-45, 45]$) and 367 observations for N_1. As the focus of this work

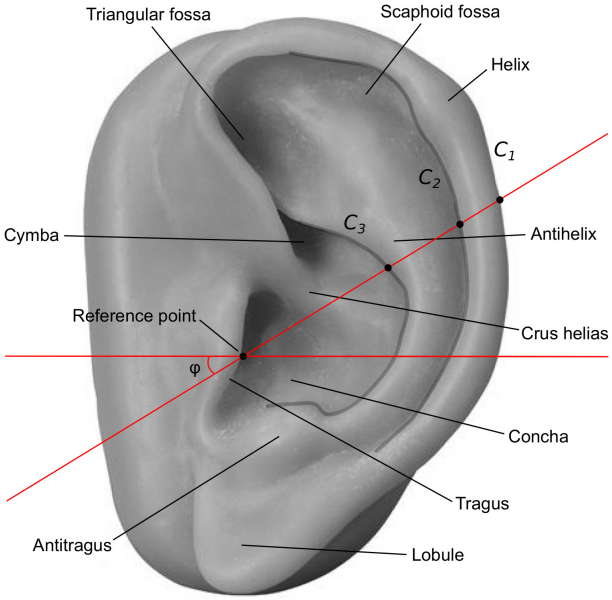


Figure 2. Pinna anatomy and extraction of the three elevation-dependent features.

is on anthropometric features, the elevation angle ϕ is not considered as regressor. However, since our sample comprises only 29 unique subjects, the global pinna features — which do not depend on elevation — are repeated for each specimen. Moreover, the features are mutually correlated, with an average Pearson correlation coefficient of 0.24 and a maximum of 0.95 across the 78 feature pairs.

In order to untangle the data and reduce its dimensionality, we apply a PCA to its features. This technique is used to find a new orthogonal coordinate system in the original data space, which best represents the variance expressed by the data. We therefore obtain 13 new features, called *principal components*, which are largely uncorrelated (average Pearson coefficient of $1.9\text{e}{-16}$) and ordered by decreasing eigenvalue. It is important to point out that the original features have been preemptively normalized into 0-mean and unit variance, so as to avoid features with large magnitudes dominating the results.

2.4 Regression model

Finally, multiple linear regression with forward step-wise feature selection is performed on the principal components using all the 367 data records. The feature selection step improves the generalization capabilities of the model by discarding predictors which are irrelevant and may instead cause overfitting. The procedure consists in instantiating a regression model for each of the available predictors, evaluating their performances using the most appropriate metric, then selecting the predictor resulting in the best performances and repeating the previous steps with models composed of the previously selected features along with any of the remaining ones, until the desired number of predictors is reached. In this case, we settled on 3 principal components, which is the minimum amount required to explain

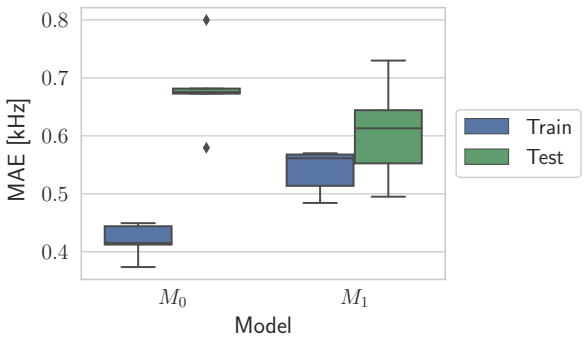


Figure 3. Box-and-whisker plot showing the mean absolute error (expressed in kHz) of models M_0 (baseline) and M_1, for training and test sets respectively.

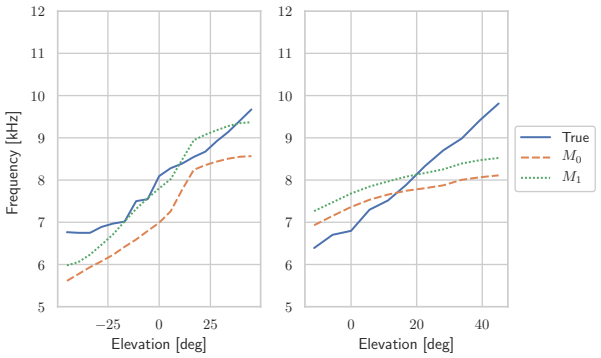


Figure 4. Sample plots of notch frequency over elevation, with both true and predicted values, taken from two subjects in the test set.

more than 50 % of the variance in the data.

The metric used to determine the best-performing predictors is the *mean absolute error*, calculated as

$$\text{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j| \tag{1}$$

where n is the number of observations, y_j is the true value, and $\hat{y}_j$ is the predicted value. This metric was preferred over root-mean-square error because it provides an intuitive representation of the average residual, and because it is robust to outliers and large errors.

The resulting model, M_1, is validated using a 5-fold cross-validation scheme. The pool of 29 subjects is divided into 5 approximately equal-sized subsets. During each iteration, one of the subsets is set aside and used to validate the model, whereas the remaining ones constitute the training set. Therefore, the ratio between training and test data is approximately 20 %, depending on the exact number of observations available per subject. While this scheme does not guarantee a constant training-test ratio, it ensures that no test subject appears in the training set, which would otherwise greatly simplify the prediction task.

The model M_1 described above is then compared against the baseline M_0, a multiple linear regression model comprising all the original features.

3. RESULTS

Figure 3 shows the performances of models M_0 (baseline) and M_1 in terms of their mean absolute error, aggregated over all the cross-validation folds. It is interesting to notice how the baseline model performs better than the custom one on the training set (average MAE equal to 419 Hz for M_0 and 539 Hz for M_1), while the opposite is true for validation data (average MAE equal to 682 Hz for M_0 and 607 Hz for M_1, representing an 11 % average decrease in error). This means that some of the variance expressed by the anthropometric features is not useful for generalizing on unseen data. Therefore, our feature selection process renders M_1 more resilient to overfitting.

Figure 4 shows two example instances of true and predicted notch frequency from the two models under consideration, computed during the validation step. Despite both models being capable of modeling the mostly monotonic relation between frequency and elevation (and by proxy, $r_k(\phi)$ features), the baseline model presents a larger bias, a clear consequence of the aforementioned overfitting.

PC_i	PC_1	PC_2	PC_3	PC_4	PC_5
ρ	-0.43	0.50	0.13	-0.43	-0.13

Table 1. Pearson correlation coefficients between each of the first 5 principal components and target frequencies.

We also found that principal components PC_i with $i \in \{1, 2, 4\}$ are consistently selected across folds, revealing their predictive potential. Despite explaining 11 % of the data variance on average, PC_3 is never selected as a predictor; while this may seem counterintuitive, it can be explained by looking at the Pearson correlation coefficients between said principal components and target notch frequencies, as shown in Table 1. In this case, it is clear how PC_3 does not manifest enough correlation for it to positively impact the performances of the model. In terms of the role of the selected principal components, the matrix of loadings reveals how the component with the most predictive power mainly codes elevation-dependent features $r_k(\phi)$, whereas the second and third ones present a mixture of elevation-dependent and global features.

When evaluating the performances of the models in terms of their psychoacoustical implications, it is desirable to consider whether the predicted notch frequency lies within 10 % of the real one. Indeed, for spectral notches in the high-frequency range, differences lower than said threshold are, on average, indistinguishable [12]. Since every observation is used once and only once throughout cross-validation, it is possible to determine the percentage of *psychoacoustically valid* predictions by counting how many fall within the threshold, and normalizing by the overall number of observations. Therefore, when considering test data only, the percentage of psychoacoustically valid predictions for the baseline and the custom models is 56.3 % and 61.2 % respectively, constituting a modest 8.75 % improvement in perceptually noticeable performances.

4. CONCLUSIONS

The results of this work show that N_1 frequencies can be predicted from anthropometric data within a certain degree of accuracy. However, our regression model was built using a limited amount of training data from a single HRTF database. More recent and documented databases such as HUTUBS [13] will be used in future works in order to carry out larger data analyses, possibly using state-of-the-art feature extraction and nonlinear regression algorithms.

It has to be noted that HRTF data collected on a human population implies issues related to microphone position and head movements that pose critical challenges when merging different datasets. The authors will shortly expand the recently collected Viking HRTF dataset [14], designed to guarantee reproducible measurements on a mannequin with different interchangeable ears, through new acquisitions on a larger ear sample in a controlled environment. These measurements will serve as a solid basis for accurate investigations on the relation between HRTFs and anthropometric data, the final objective being an effective tuning of low-order structural HRTF models [11,15,16]. Applications of these models are expected to range from personal entertainment to assistive technologies [17, 18].

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 797850, and from NordForsk's Nordic University Hubs programme under grant agreement No. 86892.

5. REFERENCES

[1] H. Møller, M. F. Sørensen, C. B. Jensen, and D. Hammershøi, "Binaural technique: Do we need individual recordings?" *J. Audio Eng. Soc.*, vol. 44, no. 6, pp. 451–469, 1996.

[2] K. Iida, Y. Ishii, and S. Nishioka, "Personalization of head-related transfer functions in the median plane based on the anthropometry of the listener's pinnae," *J. Acoust. Soc. Am.*, vol. 136, no. 1, pp. 317–333, 2014.

[3] P. Mokhtari, H. Takemoto, R. Nishimura, and H. Kato, "Frequency and amplitude estimation of the first peak of head-related transfer functions from individual pinna anthropometry," *J. Acoust. Soc. Am.*, vol. 137, no. 2, pp. 690–701, 2015.

[4] K. Iida, H. Shimazaki, and M. Oota, "Generation of the amplitude spectra of the individual head-related transfer functions in the upper median plane based on the anthropometry of the listener's pinnae," *Appl. Acoust.*, vol. 155, pp. 280–285, 2019.

[5] S. Spagnol and F. Avanzini, "Frequency estimation of the first pinna notch in head-related transfer functions with a linear anthropometric model," in *Proc. 18th Int. Conf. Digital Audio Effects (DAFx-15)*, Trondheim, Norway, 2015, pp. 231–236.

[6] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *Proc. IEEE Work. Appl. Signal Process., Audio, Acoust.*, New Paltz, New York, USA, 2001, pp. 1–4.

[7] S. Spagnol, M. Hiipakka, and V. Pulkki, "A single-azimuth pinna-related transfer function database," in *Proc. 14th Int. Conf. Digital Audio Effects (DAFx-11)*, Paris, France, 2011, pp. 209–212.

[8] V. C. Raykar, R. Duraiswami, and B. Yegnanarayana, "Extracting the frequencies of the pinna spectral notches in measured head related impulse responses," *J. Acoust. Soc. Am.*, vol. 118, no. 1, pp. 364–374, 2005.

[9] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, no. 4, pp. 744–754, 1986.

[10] S. Spagnol, "On distance dependence of pinna spectral patterns in head-related transfer functions," *J. Acoust. Soc. Am.*, vol. 137, no. 1, pp. EL58–EL64, 2015.

[11] S. Spagnol, M. Geronazzo, and F. Avanzini, "On the relation between pinna reflection patterns and head-related transfer function features," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 3, pp. 508–519, 2013.

[12] B. C. J. Moore, S. R. Oldfield, and G. J. Dooley, "Detection and discrimination of spectral peaks and notches at 1 and 8 kHz," *J. Acoust. Soc. Am.*, vol. 85, no. 2, pp. 820–836, 1989.

[13] F. Brinkmann, M. Dinakaran, R. Pelzer, J. J. Wohlgemuth, F. Seipl, and S. Weinzierl, "The HUTUBS HRTF database," 2019, DOI: 10.14279/depositonce-8487.

[14] S. Spagnol, K. B. Purkhús, S. K. Björnsson, and R. Unnthórsson, "The Viking HRTF dataset," in *Proc. 16th Int. Conf. Sound and Music Computing (SMC 2019)*, Malaga, Spain, 2019, pp. 55–60.

[15] C. P. Brown and R. O. Duda, "A structural model for binaural sound synthesis," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 5, pp. 476–488, 1998.

[16] S. Spagnol, E. Tavazzi, and F. Avanzini, "Distance rendering and perception of nearby virtual sound sources with a near-field filter model," *Appl. Acoust.*, vol. 115, pp. 61–73, 2017.

[17] F. Avanzini, S. Spagnol, A. Rodá, and A. De Götzen, "Designing interactive sound for motor rehabilitation tasks," in *Sonic Interaction Design*, K. Franinovic and S. Serafin, Eds. Cambridge, MA, USA: MIT Press, 2013, ch. 12, pp. 273–283.

[18] S. Spagnol, G. Wersényi, M. Bujacz, O. Balan, M. Herrera Martínez, A. Moldoveanu, and R. Unnthórsson, "Current use and future perspectives of spatial audio technologies in electronic travel aids," *Wireless Comm. Mob. Comput.*, vol. 2018, p. 17 pp., 2018.

PERFORMANCE OF PIANO TRILLS: EFFECTS OF HANDS, FINGERS, NOTES AND EMOTIONS

Xu Han
KTH Royal Institute of Technology
xuhan@kth.se

Roberto Bresin
KTH Royal Institute of Technology
roberto@kth.se

ABSTRACT

Trill is a type of musical ornament. In automatic playback of piano music scores, trills are usually synthesised as a sequence of repeated notes with equal duration and dynamic level. This is not how trills are performed by pianists. In this study, trills were performed by three pianists on a Yamaha Disklavier and recorded as both audio and MIDI files. Then note duration, inter-onset interval (IOI) and key velocity for each note were extracted from Midi files and analyzed in relation to hands, notes and emotions. Four significant effects were found; 1) hand effect: trills on right hand were in average performed with a faster rate, shorter note duration, longer off duration and faster key velocity, 2) finger effect: within the two notes forming a trill, notes with lower fingering number were performed with shorter off duration, while keeping note duration and key velocity close, 3) emotion effect: emotion mainly contributed to dynamic level, 4) crescendo effect: when crescendo happened, note duration and off duration compensated with each other and kept IOI at a almost constant value.

1. INTRODUCTION

With the development of acoustic analysis technology, researchers started to explore the characteristics of musical performance quantitatively. Various parameters (e.g. inter-onset intervals, loudness levels, note duration, etc.) are used to model different aspects of expressive musical performance. Some of the models are more descriptive than quantitative, and some other models focus more on specific aspects [1]. This project mainly focuses on descriptive piano trills modeling.

Trill is a type of musical ornaments, which "requires one of the fastest alternating movements of which the hand is capable" [2] in piano performance. It is an expressive component in piano performance which has been derived from *tremolo* in the XVII century. However, it is still a problem in synthesizing piano trills for music score playback. Depending on the capability of hand as well as mass and dimension difference between each finger, trill is not a combination of fast notes with equal duration and dynamic level, but contains many variations. There are similar patterns of ornaments in other forms of musical performance, for example stringed instruments and singing, which is called vibrato or trill. Both qualitative and quantitative characteristics of vibrato have been studied for decades. In 1990, [3] assessed trill rates of nine professional early music singers and found slow trill rates ranged from 2.0 to 6.9 Hz and fast trill rates ranged from 7.5 to 12.4 Hz. In 1994, [4] measured vibrato rate for ten singers singing Schubert's Ave Maria. He found the vibrato rate increased at the end of each tone, which was also verified to exist in stringed instruments as well. In 2004, [5] analyzed *tremolo*, trill and vibrato rate for 56 string players. He found ornaments on low notes were performed slower and player with more advanced skills can perform faster. Besides, mean rates for slow, medium and fast tempo are 6.4, 9.3 and 12.1 Hz respectively. These studies above mainly focus on vibrato rate and other qualitative parameters, and there are also other work developing comprehensive computation model for vibrato performance. Schoonderwaldt and Friberg [6] modeled violin vibrato with rate, extent and sound level envelopes, using the method of analysis-by-synthesis approach.

Different from most of the other studies which focused on the temporal aspects of trills, Moore [2] analyzed both on-off timing and dynamic level of notes in piano trills. In his study, trills were recorded by using one of the first Yamaha Disklavier models (a Grand Piano Model C3 equipped with a MIDI interface). Fingers movements and electromyographic (EMG) were monitored at the same time. Four subjects were required to perform a trill exercise consisting of D4 and E4, and also a passage of musical passage from the first movement of the Beethoven Piano Concerto No. 1, which included trills with note pairs D5/E5 and D5#/E5. As a result, average trill rate was 12 Hz; note duration differed by less than 1 ms for D4 and E4; average dynamic level differed by less than 2 MIDI units between D4 and E4, which was less than variation within note D4 or E4; and the greatest performance difference was the gap between offset to onset of next note, in which average gap between offset of D4 to onset of E4 was 17 ms while the average gap from E4 to D4 was 24 ms; also it was observed that increases in key velocity led to increases in note duration. Similar results were observed for note pairs D5/E5 and D5#/E5.

Besides the playing techniques, also the emotional expression used by pianists can have an effect on the performance of trills. Emotion in music can be communicated in terms of variations of a number of parameters such as note height, intensity, duration, articulation, sound level, attach

speed [7,8]. Therefore, since note duration and sound level are two of the main expressive parameters in piano performance, in this study we will also investigate how emotional expression can influence the performance of trills. This study focuses on both temporal and dynamic aspects of piano trills in relation to hands, pitches and emotions. The trills considered in this study are continuous trills that extend over several bars.

2. EXPERIMENT

Three right-handed young pianists (2F, 1M; average age 21.7, SD 3.5), from the master programme in Classical Music Performance at the Royal College of Music in Stockholm, were recruited as participants for the experiment. They have being playing piano for 13 years in average (SD 5.2). The experiment consisted of 3 parts: 1) performance practice, 2) performance recording, and 3) post-interview about the performances. JS Bach composition Invention in D minor BWV 775 [1] was chosen to be the experiment material because of its relatively long trills to be played by both hands (three bars for right hand, and five bars for left hand).

In part 1, the experimenter sent out the score to the participants so the participants could get enough time to practise and be able to perform the piece fluently. The instructions and consent forms were also sent together with the score. In part 2, the participants were asked to come to the experiment location and produce performances in the following 6 conditions:

1. Please perform the piece in your own interpretation as in a concert.
2. Please perform the piece in the same way as in task 1 without trills, which means hold the beginning note of trill until the end of trill.
3. Please perform the piece in exaggerated happy emotion.
4. Please perform the piece in the same emotion as in task 3 without trills, which means hold the beginning note of trill until the end of trill.
5. Please perform the piece in exaggerated sad emotion.
6. Please perform the piece in the same emotion as in task 5 without trills, which means hold the beginning note of trill until the end of trill.

The participants were allowed to play in each condition as many times as they wanted, which resulted in a set of performance recordings, until they felt satisfied with the recording. The pianists did not make use of the piano pedals in their performances. After finishing performance under each condition, participants were asked to indicate the best performance recording, which would be used later in data analysis. This means that for each pianist only one of the performances made with the same expressive intention was kept for the analysis.

This study mainly focused on trill performance, so only data collected under condition 1, 3 and 5 were analyzed later. The performances were played on a YAMAHA Disklavier E3, and automatically stored in MIDI format using the control unit available on the Disklavier. Stereo audio recordings of the performances were also made using a sound card (RME Babyface) connected to a MacBook Pro laptop computer (all recorded data are freely available online [2]). In part 3 of the experiment, participants were interviewed about their technique when playing trills in different conditions, and about their overall experience during the experiment.

3. RESULTS

The MIDI Toolbox 1.1 [9] was used for analizing the MIDI recordings. Onset time, duration, dynamic and note number of each performed note were extracted. In MIDI format, dynamic is given by MIDI velocity (0..127) and note is given by MIDI note number (0..127). The MIDI note numbers of the trill in the score were respectively, 52 (E3, 164.81 Hz) and 53 (F3, 174.61 Hz) for the trill played by the left hand, and 72 (C5, 523.25 Hz) and 74 (D5, 587.33 Hz) for the right hand. All trills were extracted. As described in the experiment instructions, participants were asked to play the piece in three difference expressions, including sad emotion, happy emotion and the emotional expression they would use when performing the piece in concert, which is called "concert emotion" in the following. Trill rate, inter-onset interval (IOI), note duration, off-duration and MIDI velocity were taken as dependent variables. Trill rate refers to number of notes played per second. IOI refers to the time between the onset time of one note and that of the next note. Trill rate was calculated by dividing the number of notes by the duration of the trill passage. Off-duration refers to the time interval between the time when a key is released and the onset time of the next note. Larger off-duration values indicate longer break (also called Key-Detached Time, KDT, see [10]) between finger switch in trills, which corresponds to a more *staccato* trill. On the contrary, shorter off-duration time value indicates shorter break, and when it has a negative values, adjacent notes are overlapping (also called Key-Overlap-Time, KOT, see [10]), resulting in a more *legato* trill. Hand (left or right), note and emotion were taken as independent variables. Hand refers to which hand played the trill. After Kolmogorov-Smirnov normality test, trill rate was shown to be approximately normally distributed, while IOI, note duration, off-duration and MIDI velocity failed the test. Therefore, when examining correlations, Pearson's Product-Moment Correlation test was used for trill rate and Spearman Rank-Order Correlation test was used for other variables.

[1] https://ndmusicedition.files.wordpress.com/2011/01/inventions-largerspacing-4.pdf

[2] https://kth.box.com/v/nsmc2019trills

3.1 Hand Effect

3.1.1 Trill Rate

The mean for trill rate was 10.1 note/s and the standard deviation was 1.5 note/s regardless of hand. Results showed that trill rate and hand were significantly correlated ($p = .002 < 0.05$) and right hand had higher trill rate (see Table 1).

Table 1: Means and standard deviations of trill rate on each hand

Hand	M (note/s)	SD(note/s)
Left	9.1	0.4
Right	11.1	1.4

Table 2: Means and standard deviations of IOI on each hand

Hand	M (ms)	SD (ms)
Left	108.56	17.34
Right	93.44	28.08

3.1.2 Note Duration

There was a negative correlation between note duration and hand, which was statistically significant ($p = 0.000 < 0.05$) (see and Table 3).

Table 3: Means and standard deviations of note duration on each hand

Hand	M (ms)	SD (ms)
Left	89.94	22.90
Right	64.96	26.60

3.1.3 Off-Duration

Hand and off.duration were significantly correlated ($p = 0.000 < 0.05$). Notes in trills played by the right hand had longer off-duration than thoes performed by the left hand (see Table 4).

Table 4: Means and standard deviations of off duration on each hand

Hand	M (ms)	SD (ms)
Left	18.62	25.31
Right	28.49	18.63

3.1.4 Key Velocity

A significant positive correlation was found between Key velocity and hand ($p = 0.000 < 0.05$) (see Table 5).

3.2 Note Effect

Since hand effect was significant to all dependent variables, data were grouped by hand in the following analysis.

Table 5: Means and standard deviations of key velocity on each hand. Values are expressed in MIDI velocity values, which vary between $0 = silentnote$ and $127 = loudestsoundlevelpossible$

Hand	M	SD
Left	55.84	6.99
Right	61.33	6.40

3.2.1 IOI

For left hand, IOI and note were significantly correlated ($p = 0.000 < 0.05$) and IOI from lower note to higher note was longer than the opposite (see Table 6).

Table 6: Means and standard deviations of IOI for each note on left hand

Note	M (ms)	SD (ms)
E3	116.70	13.40
F3	100.11	16.95

For right hand, IOI and note were also significantly correlated ($p = 0.000 < 0.05$) and IOI from higher notes to lower notes was longer than the opposite (see Table 7).

Table 7: Means and standard deviations of IOI for each note on right hand

Note	M (ms)	SD (ms)
C5	83.94	29.21
D5	102.81	23.50

3.2.2 Note Duration

A significant positive correlation between note duration and note was found for the left hand ($p = 0.007 < 0.05$), but not for the right hand (see Table 8).

Table 8: Means and standard deviations of note duration for each notes on left hand

Note	M (ms)	SD (ms)
E3	86.66	22.48
F3	93.35	22.90

3.2.3 Off-Duration

For left hand, when checking correlation between off duration and note, trill data showed significant correlation between off-duration and note ($p = .000$). Off-duration was longer when playing from lower to higher note than from higher to lower note (see Table 9).

For right hand, the correlation between note and off duration was also significant ($p = 0.000 < 0.05$) and it took longer to switch from higher notes to lower notes (see and Table 10).

Table 9: Means and standard deviations of off duration for each pitch on left hand

Note	M (ms)	SD (ms)
E3	30.04	22.50
F3	6.76	22.50

Table 10: Means and standard deviations of off duration for each note on right hand

Note	M (ms)	SD (ms)
C5	19.18	16.58
D5	37.66	15.93

3.2.4 Key velocity

For the right hand, note and key velocity show a positive significant correlation ($p = 0.000 < 0.05$) (see Table 11), but not for the left hand.

Table 11: Means and standard deviations of key velocity for each note on right hand

Note	M	SD
C5	59.70	5.68
D5	62.93	6.68

3.3 Emotion Effect

3.3.1 Trill Rate

Trill rate and emotion are not significantly correlated for either hand ($p = .535 > 0.05, p = .874 > 0.05$) (see Tables 12 and 13). Still it is possible to observe a tendency for slower rate for the sad performances compared to the faster rates of the happy and concert ones.

Table 12: Means and standard deviations of trill rate for each emotion on left hand

Emotion	M (note/s)	SD (note/s)
Sad	8.99	0.42
Concert	9.21	0.56
Happy	9.23	0.46

Table 13: Means and standard deviations of trill rate for each emotion on right hand

Emotion	M (note/s)	SD (note/s)
Sad	10.70	1.67
Concert	11.62	1.28
Happy	10.91	1.88

3.3.2 IOI

Since IOI was found to be significantly correlated to hand and note, the data were grouped by hand and pitch for variable control. For left hand, IOI shows a significant correlation with emotion for the left hand and note E3 only ($p = 0.042 < 0.05$) (see Table 14). No significant correlation was observed for the notes played by the right hand.

Table 14: Means and standard deviations of IOI for each emotion on note E3

Emotion	M (ms)	SD (ms)
Sad	120.31	13.66
Concert	113.99	13.05
Happy	115.79	12.82

For right hand, IOI had no significant correlation with emotion on either note.

3.3.3 Off-Duration

Since off duration was found to be significantly correlated with note for both hands, off duration data was grouped by pitch during analysis. For the right hand, the correlation between off duration and emotion was found significant only for note D5 ($p = 0.019 < 0.05$) (see Table 15).

Table 15: Means and standard deviations of off duration for each emotion on pitch D5

Emotion	M (ms)	SD (ms)
Sad	41.06	13.56
Concert	38.51	12.47
Happy	33.28	20.05

3.3.4 Key Velocity

Significant positive correlations between emotion and key velocity was found for all the notes played by both left ($p = 0.005 < 0.05$) (see Table 16) and right hands ($p = 0.001 < 0.05, p = 0.000 < 0.05$) (see Tables 17 and 18).

Table 16: Means and standard deviations of key velocity for each emotion on left hand

Emotion	M	SD
Sad	54.85	7.25
Concert	55.53	6.57
Happy	57.20	6.97

Table 17: Means and standard deviations of key velocity for each emotion on note C5

Emotion	M	SD
Sad	58.04	5.90
Concert	59.28	4.54
Happy	61.88	5.95

3.4 Crescendo Effect

From the post-experiment interview, it emerged that pianist 2 intended to do a crescendo in trills. Therefore the effect of crescendo was analyzed for all three pianists.

Table 18: Means and standard deviations of key velocity for each emotion on note D5

Emotion	M	SD
Sad	59.85	7.41
Concert	62.41	5.47
Happy	66.64	5.04

3.4.1 Key Velocity

From the analysis of key velocity values of the trills performed by the three pianists, we found that pianists 1 and 2 performed a crescendo with their left hand (see Figure 1).

3.4.2 IOI

IOI did not show any obvious tendency when plotted with time (See Figure 2), and it was kept almost constant throughout the trill.

3.4.3 Note Duration and Off-Duration

Note duration increased (See Figure 3) during the performance of trills, and as a consequence off-duration simmetrically decreased with time (See Figure 4).

4. DISCUSSION

4.1 Hand Effect

Humans have usually a dominant hand which can perform a better, faster and more precise task than the other hand. Non-dominant hand can also be trained to improve, however, skills asymmetry still exists for professional musicians [11]. In this study, handing effect was significant to trill rate, IOI, note duration, off-duration and key velocity. Compared to left hand, right hand had higher trill rate. The difference of trill rate between left hand and right hand was approximately 2 note/s. Right hand also had shorter IOI and note duration in average, which is reasonable because these enabled a faster trill rate. However, right hand had longer off-duration. This indicates that pianists performed on average more *staccato* with the right hand than with left hand, and that they had more time for preparing for the next note to be played in the trill. Meanwhile, right hand had higher key velocity, made possible by a more *legato* performance.

To sum it up, trills on right hand had faster trill rate, shorter off note duration, longer off duration and higher key velocity while keeping note duration as consistent as possible.

4.2 Note Effect

Note effect was significant to IOI, off duration for both hands, note duration for left hand and key velocity for right hand. Note effect can also be treated as fingering effect. For left hand, higher note was played by a finger with lower fingering number (lower finger). On the contrary, for right hand higher note is played by a finger with higher fingering number (higher finger). In the results, higher note for left hand and lower note for right hand had shorter IOI, which also indicated that lower finger for both hands had shorter IOI. With the same analysis, it was easy to find that lower finger also had shorter off duration. This finding was corresponding with Moore's study [2]. He found that the greatest performance difference between fingers was "the gap between the end of one note and the onset of next". Although note effect was significant to note duration for left hand, the mean difference between notes was approximately 7 ms, which was much smaller than standard deviation values of note duration. Similarly, the mean difference of key velocity between notes was approximately 3, which was also smaller than standard deviation values.

In the analysis above, IOI and off duration were shorter for lower finger. At the same time, note duration was either not significantly affected by note or the mean value difference between notes was very small. Since IOI is the sum of note duration and off-duration, off-duration is possibly the main contribution factor to the difference of IOI between notes. Also, pianists 1 and 2 started from the upper note in the trill performed by the right hand, while pianist 3 from the lower note.

In summary, within 2 notes forming the trill, notes with lower fingering were performed with shorter off-duration, while keeping note duration and key velocity almost at constant values.

4.3 Emotion Effect

Emotion effect was significant to key velocity. For note E3 on left hand, emotion effect was also significant. However the mean differences between different emotions was smaller than the standard deviation values. This result indicated that emotion effect mainly contributed to dynamic level but not temporal aspect of the trill passages and happy emotion resulted in higher dynamic level. This is reasonable since because of the high speed at which notes is trills are performed, it is easier for the pianist to control the sound level than the duration of the already very short notes.

4.4 Crescendo Effect

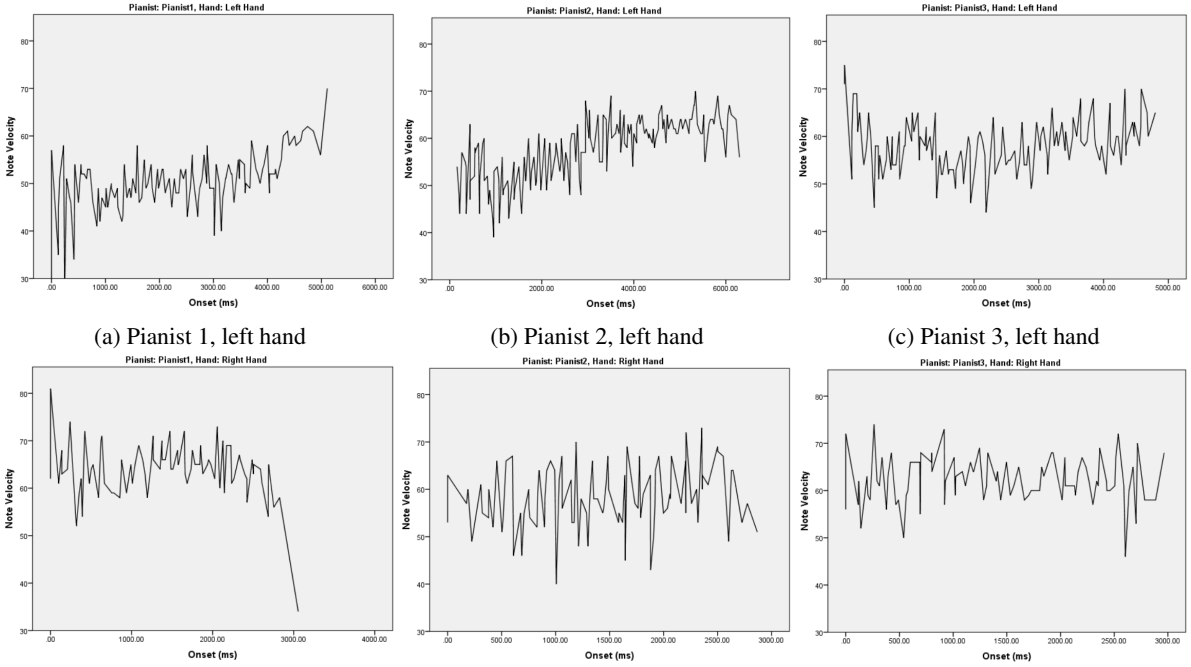
With *crescendo* in trill passage, key velocity and note duration increased with time, while off-duration decreased. Besides, IOI values did not show any obvious tendency. In conclusion, when *crescendo* happened, note duration and off-duration compensated with each other and kept IOI at almost constant value.

5. CONCLUSION

By implementing the experiment and analyzing recordings from the experiment, the main findings are listed below:

1. Hand effect: Trills on right hand have faster trill rate, shorter note duration, longer off duration and higher key velocity while keeping note duration almost constant.

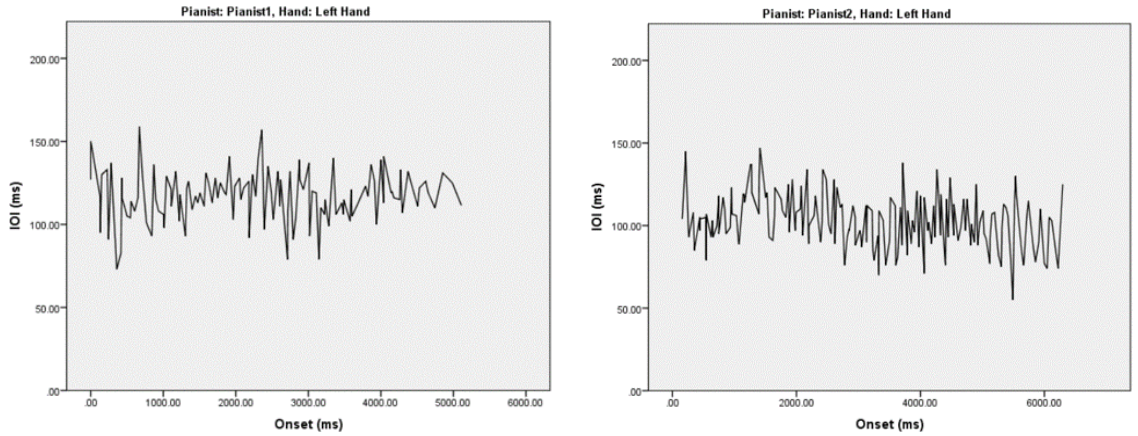
2. Note effect: Within the two notes forming the trill, notes with lower fingering have shorter off-duration, while keeping note duration and key velocity close.



(a) Pianist 1, left hand (b) Pianist 2, left hand (c) Pianist 3, left hand

(d) Pianist 1, right hand (e) Pianist 2, right hand (f) Pianist 3, right hand

Figure 1: Plot of key velocity vs onset time for each note in the trills played by the three pianists



(a) Pianist 1, left hand (b) Pianist 2, left hand

Figure 2: Plot of IOI for the trill played with the left hand by pianists 1 and 2.

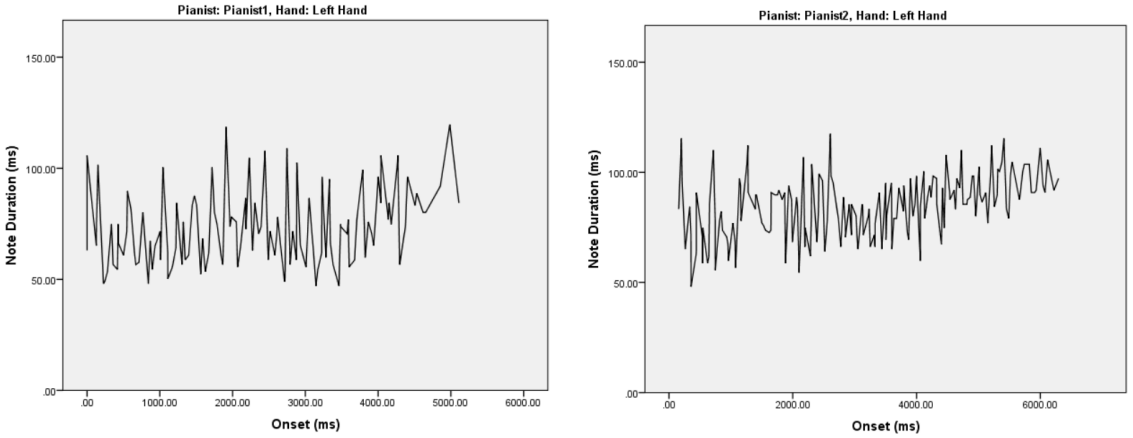
3. Emotion effect: Emotion mainly contributed to dynamic level and happy emotion resulted in higher dynamic level, compared to sad emotion

4. Crescendo effect: In trills with crescendo, note duration and off-duration compensate each other and IOI is therefore kept at almost constant value.

These findings cover both temporal aspects and dynamic level of a trill. They can provide qualitative constraints for trill performance modelling. However, there are also limitations in this experiment. Firstly, there were only three pianists that participated in the experiments which was a small sample size. Secondly, the piano skills to these three pianists as well as their performance style varied which lead to slightly different performances. Nevertheless it was possible to observe significant effects in the performance of trills, probably because of the high speed at which they are played. For future study, it will be interesting to verify the conclusions with more pianists, different compositions and tempi. Also, analysis of commercial recordings of the same and similar pieces will be conducted. The final results will hopefully lead to even more strong conclusions, which will help to implement a rule for the automatic expressive performance of piano trills at least of the Baroque repertoire.

Acknowledgments

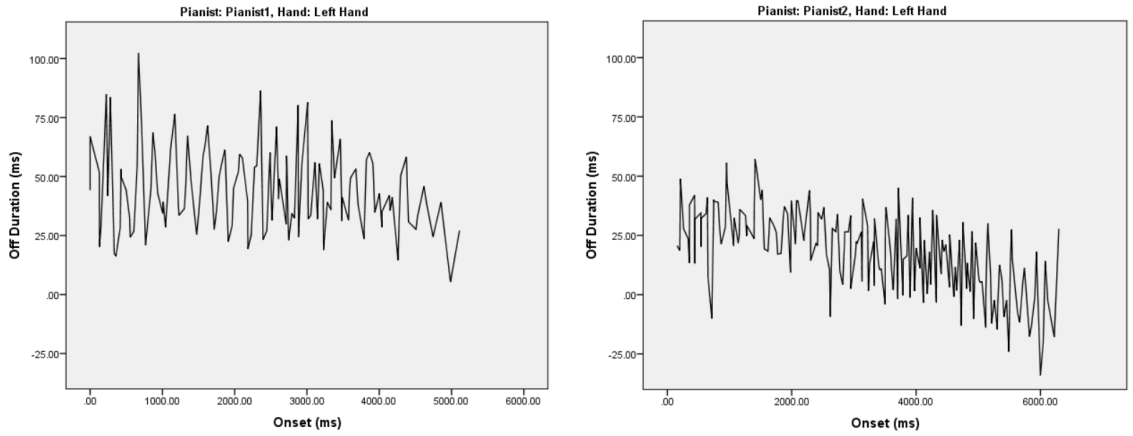
We want to thank the three anonymous pianists from the KMH Royal College of Music who took part in the experiment. This project was partially funded by ”NAVET - A hub to navigate to unexplored regions between art, tech-



(a) Pianist 1, left hand

(b) Pianist 2, left hand

Figure 3: Plot of note duration to onset time for each pianist



(a) Pianist 1, left hand

(b) Pianist 2, left hand

Figure 4: Plot of off duration to onset time for each pianist

nology and design" [3] , a KTH centre.

6. REFERENCES

[1] G. Widmer and W. Goebl, "Computational models of expressive music performance: The state of the art," *Journal of New Music Research*, vol. 33, no. 3, pp. 203–216, 2004.

[2] G. P. Moore, "Piano trills," *Music Perception: An Interdisciplinary Journal*, vol. 9, no. 3, pp. 351–359, 1992.

[3] J. Hakes, E. T. Doherty, and T. Shipp, "Trillo rates exhibited by professional early music singers," *Journal of Voice*, vol. 4, no. 4, pp. 305–308, 1990.

[4] E. Prame, "Measurements of the vibrato rate of ten singers," *The journal of the Acoustical Society of America*, vol. 96, no. 4, pp. 1979–1984, 1994.

[5] D. Moelants, "Temporal aspects of instrumentalists' performance of tremolo, trills, and vibrato," in *Proceedings of the International Symposium on Musical Acoustics (ISMA'04)*, 2004, pp. 281–284.

[3] `http://www.kth.se/navet`

[6] E. Schoonderwaldt and A. Friberg, "Towards a rule-based model for violin vibrato," in *Workshop on Current Research Directions in Computer Music, Pompeu Fabra University, Audiovisual Institute, Barcelona, Spain.*, 2001, pp. 61–64.

[7] C. E. Seashore, "Measurements on the expression of emotion in music," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 9, no. 9, p. 323, 1923.

[8] R. Bresin and A. Friberg, "Emotion rendering in music : Range and characteristic values of seven musical variables," *Cortex*, vol. 47, no. 9, pp. 1068–1081, 2011.

[9] T. Eerola and P. Toiviainen, *MIDI toolbox: MATLAB tools for music research*. Department of Music, University of Jyväskylä, 2004.

[10] R. Bresin, "Articulation rules for automatic music performance," in *Proceedings of ICMC 2001, Havana, Cuba*, 2001, pp. 294–297.

[11] C. Palmer, "Anatomy of a performance: Sources of musical expression," *Music perception: An interdisciplinary journal*, vol. 13, no. 3, pp. 433–453, 1996.

IMMERSIVE AND INTERACTIVE MUSIC FOR EVERYONE

Hans Lindetorp
Royal College of Music in Stockholm, KMH
Royal Institute of Technology in Stockholm, KTH
hans.lindetorp@kmh.se

ABSTRACT

This study seeks to understand how new and accessible technology can be used and developed to include producers of standard music into making immersive, interactive, music experiences. Through observations during a student project and an analysis of the participant's reflections it argues that even if the technology worked well, there are still many opportunities for improvements. The result shows that the repeated, non-creative, tasks like exporting and naming files can reduce musical inspiration for students with little interest in technology and that further development and studies potentially could make interactive music accessible even for them. The aspect of the project that caused most positive response was producing and and mixing for super-surround which led the students to new insights and ideas for their everyday music production. Finally the result indicates that even if there would have been no technical barriers interactive music production might not appeal to everyone. Interactive music should maybe be seen as a separate discipline and students with a linear approach to composition will not necessarily find it interesting.

1. INTRODUCTION

The development and design of digital technology over the last decades have brought us new tools for creativity that are both cheaper and easier to use than previous products. Applications that previously were expensive and technically advanced are now available free or for a low cost for people with little or no prior experience. In the music production field where expensive, analog equipment used to be the only option, Digital Audio Workstations (DAW) [1] now have made music production inexpensive and accessible for everyone with a computer [2]. Interactive and immersive environments like spatialization of musical sound [3] are increasingly important research fields. Music production for these environments is a more complex task [4] and is not yet accessible for most music producers. This study adds to earlier work on composition and arrangement techniques for interactive music [5] seeking to get a better understanding of current challenges and potential solutions for the making of immersive, interactive music. It hopefully contributes to making these music formats accessible for music producers with little or no prior programming skills.

1.1 Earlier experiences

The Royal College of Music (KMH) [1] was founded 1771 and has a wide range of educations for musicians, composers and music teachers in different genres. Since 2001 there is also a bachelor's program in music production. Most of those students have a background in pop and rock music but through different courses they get to try music production for different targets. This could typically involve producing music for film, computer games, web pages and interactive installations. The students are normally mixing for stereo but during the education they also learn how to mix music for other formats like 5.1 and multichannel. Every year since 2013, different student groups have produced multi-channel, interactive music installations for different venues including Kulturhuset, [2] Stockholm City Museum, [3] Nobel Creations [6] at Nobel Price Museum [4] and The Sound Forest [7] at the Museum of Performing Arts [5].

Figure 1. Student interacting with her music production in Klangkupolen.

[1] http://www.kmh.se
[2] http://kulturhusetstadsteatern.se
[3] http://stadsmuseet.stockholm.se
[4] http://nobelprizemuseum.se
[5] http://scenkonstmuseet.se

1.2 The project

This study was conducted alongside a course-based project at KMH 2019. The students were introduced to general concepts and the technical framework through seminars, tutorials and workshops to prepare for the project. The course description defined the project to be a music production for "Klangkupolen" [8] - a multi-channel, super-surround speaker setup - where the music playback should be controlled by the audience using two or more smartphones (see fig.1). The technical framework was built upon the source code from earlier works and adjusted for this particular case.

2. TECHNICAL SYSTEM

The technical system is built around iMusic [6] - a javascript framework for organizing and playing back audio in interactive environments. Beside iMusic, smaller components have been developed to add features like socket communication, webcam control and synchronization. iMusic is the result of an artistic participatory design process [9] at KMH, involving students and teachers in interactive music projects.

The development of the technology (except Klangkupolen itself) follows three design goals: No cost, no installation requirements and easy-to-use. These goals lead to the use of smartphones as sensors (while all participants already have one), an available Mac-Book as the main computer (but could have been any standard computer) and open source, javascript libraries (as they are free and the only language the students know). The following sections describe the three software components (client, server and master) forming the technical framework for this project (see fig.2). All source-code is available for download [7].

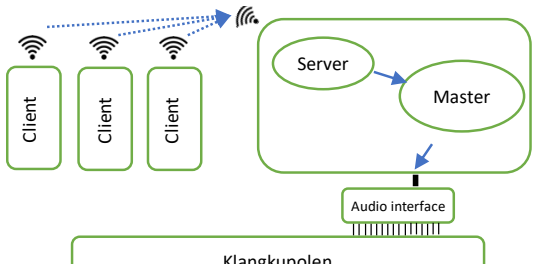


Figure 2. Technical system for the project.

2.1 Client

The client application was developed as a mobile web page and uses the browser's built-in Device Orientation API [8] to record the mobile's movement around the y and z axis. The client's only task is to capture the participants' hand movements and pass the information on to a server. The communication between the client and server runs on a standard WiFi-connection and uses a framework called socket.io [9].

To optimize performance without overloading networks and processors, a refresh rate of 100Hz was chosen. In addition to this streaming data, the library shake.js [10] was used to notify the server when the mobiles were shaken. The client was successfully tested on both iOS / Safari and Android / Chrome in current versions and also on older mobiles such as iPhone 5 with iOS 10. The biggest challenge was to prevent the phone from automatically activate sleep-mode when not touched for a while and made it clear that smartphones primarily are built for touch interactions and visual feedback.

2.2 Server

The server's function is to receive data from the clients and send it to the master application. Inspired by solutions like Soundworks [10] from the CoSiMa project, the server is a web server developed on the node.js [11] platform with express.js and socket.io to manage the communication between the various devices. The application was developed generically to be easily reused in future projects with one single function for clients to pass on data to each other or, as in this case, to a master. Even if node.js is free to download, it requires an installation. The server does therefore not fully meet the design goals.

2.3 Master

The master application controls the audio and is built with three main layers dealing with audio playback, music mapping and communication. The audio playback is controlled by the iMusic framework which is built upon a standard web API called Web Audio API and handles loading and buffering of audio data, synchronization of audio files, looping, randomization and real-time modulation of sounds using parameters like volume, panning, filter, delay, reverb and more. iMusic runs in all browsers supporting Web Audio API, but it varies to what degree they conform to the specification. In this case Google Chrome was chosen as it supports 32 separate output channels. During this project the following features were added to iMusic:

- Mapping a control value to an audio parameter which made it possible for a student with little programming experience to map any sensor data to any audio parameter in the system.
- Multi-channel panning which let the students specify a list of audio channels between which the sensor data could control the position of a sound.
- Addressing any internal audio bus to any number of output channels which made it possible to send reverb for a particular sound to a specified set of speakers.
- Envelope curves controlling audio parameters which was used in this project to connect an event from a client to trigger a volume change on a reverb bus.

[6] http://github.com/hanslindetorp/imusic
[7] http://hans.arapoviclindetorp.se/klangkupolen2019
[8] https://www.w3.org/TR/orientation-event/
[9] https://socket.io
[10] https://github.com/alexgibson/shake.js
[11] https://nodejs.org/en/

The music mapping is separated into a script written by the students. In this script they define the musical structure including organizing the files into iMusic's predefined musical objects; looped tracks, parts, motifs and variations through a file name convention system. (see fig.3)

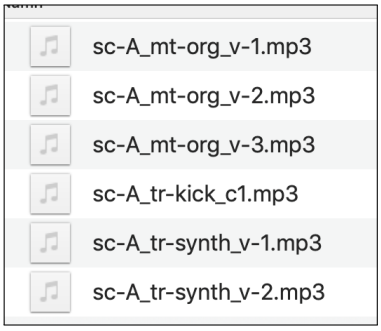


Figure 3. Example of the file naming convention

Another feature of the music mapping is to determine how the incoming sensor data should control the music. In interactive music production, this is crucial to the final result and sometime as important as the composition and recording. Most students wanted to connect one or several incoming data streams to control an audio parameter. This could typically be to connect the tilting of a smartphone to control the volume of a sound. They also connected shake events to trigger the playback of a certain sound.

3. METHOD

This study was conducted alongside a student project 2019 at KMH. Twelve students - seven female and five male students between the ages of 21-36 - took part in the project and contributed to the study with feedback during the supervised sessions and through written reflections after the project. The reflections should answer the following questions and the students gave their permission for their texts to be analysed and quoted anonymously for research purposes. After the texts were analyzed, quotes were collected and sorted to find common themes.

- What was your responsibility and what skills did you develop through the project?
- Which parts of the course have been important for the project and how have you gained from them?
- What parts of the result are you extra happy with and what would you recommend students in coming projects to do in a similar way?
- What would you recommend students in coming projects to do in a different way?

4. RESULT

The text analyses pointed out three themes; "non-linear music", "interactivity" and "multi-channel". The following sections presents these three themes plus a few separate thoughts and ideas formulated by the students.

4.1 Non-linear music

Many students described that the biggest challenge in this project was to produce non-linear music (music without a predetermined form). Like the participants in previous studies [11], they recognized that it was a bit like "telling a story to someone without ever coming to the end". Especially students with a singer / songwriter background who usually create music based on their voice and an instrument described that this step required "a completely different way of thinking" or that they "did not know how to think". They felt that their music became static, that they were struggling with being too much in control and that it was difficult not to get stuck in a particular thinking. One student also stated that "with interactive music, an important parameter in the narrative technique disappears, namely the time the audience is exposed to a certain part of the piece."

The students who usually produce loop-based music more easily related to the non-linear form and expressed that "Every song you write is at some stage non-linear until you record it" and "I'm used to working with loops and like to have DAW projects with several small beats I cut and move (also live) which made me feel comfortable ". One student experienced this part of the project as the most stimulating and wrote "The most fun I thought was to play with rhythms, to insert short rhythmic loops that played randomly and in different constellations with other random rhythm loops".

When the students were asked how to take on the task if they were to do a similar project again, they emphasized starting with a musical idea that reflects a feeling rather than one from the technology, to release the need for control and to choose softer sounds and to use sounds that are difficult to get tired of.

4.2 Interactivity

More than half of the students described that a large part of the production time was used to deal with the technical challenges relating to interactivity. They formulated e.g. "Having to name the audio files in a certain way and cutting them up in different parts for it to work, killed my creative process along the way." And "In the interactive world with our technical knowledge, this step takes an incredible amount of energy and the result can be like that." Several students described how they had to lower their musical ambitions because the technical work took so much time. When they got over the technical barriers, they also experienced challenges related to interactivity and music and described e.g. "Creating something user-friendly requires a lot of testing and exploration." And that "in order for the interaction to work in a practical way, you have to use extremes to make the listener perceive the difference." The challenge of getting a final, musical result that they like when listeners are controlling the playback was also highlighted as well as the difficulty of using vocal tracks in an enjoyable way without a nagging and annoying result.

The students who expressed themselves more positively about their results were particularly satisfied with the interactions that were easy for the participants to understand

and sounds that invited them to play and interact with. According to the students, complex solutions were not always preferable to simple ones and one student described that "A simple sinewave with pitch-bend controlled by tilting the phone became a successful instrument. It was incredibly responsive and fun to play with and also contributed to the composition".

The students contributed with recommendations for producing music for interactivity and mentioned "to start from a musical idea rather than a technical", "decide what would be controlled by the producer and what would be influenced by the participants", "make sure that all parts are fun to play with", "make the background layer minimalistic so it doesn't collide with the participants interactions" and "use short samples with fast attack for the interactions".

4.3 Multi-channel

Many of the students was euphoric about the possibility of producing music for Klangkupolen and used adjectives like "crazy", "funny", "special" and that they were "completely sold on the idea". One student described the experience in the following words:

> *"Being in such an open area and surrounded by music was really wonderful. I couldn't help but dance to the faster, more rhythmic songs while the quieter, more meditative compositions got me and some classmates to lay down on the ground to take in the music."*

This quote both confirms the spatial concept with super-surround and reveals that when the listener is laying down on the floor, the immersive perspective is removed and Klangkupolen is turned into a multi-channel system in front of the listener. The challenges the students highlighted were that it is a huge difference to mix for a 29.4 system compared to stereo and that they needed more time to test their music and adjust their mixes to achieve better results. One student stated that when mixing in stereo one never thinks about how it should sound behind the listener because there is no "behind" in a stereo mix. On the other hand, there is no specified "front" in the Klangkupolen and the listener can experience the music faced in any direction.

Further thoughts that were highlighted were that Klangkupolen invites you to "compose the piece for the room rather than writing from a musical flow", that Klangkupolen can add a new dimension to the meaning of a song and that the sounds "thrown between different sides of the room" can be linked to the meaning of the lyrics. One student thought it was, in one way, easier to create a meaningful musical experience for Klangkupolen than for stereo because panning and levels become more important than details in the music. Several students also expressed how Klangkupolens visual factor invites both to create powerful and "big" music and that those who come to listen to music there will experience something cool just by seeing the room.

4.4 Other thoughts and ideas

Some students further reflected on new experiences and that will benefit their regular music production. One insight was to use a more minimalistic approach without adding instruments that will not be heard in the final mix. One student reasoned that the technical knowledge required in interaction design and programming prevent many music producers from exploring music production for interactive environments. Another student was excited to have produced something that became a creative place where people could meet and create something new together. Finally, some got ideas for new applications and suggested a music format for i.e. Spotify where a part of a song could be interactive and someone else could see an interactive music installation in Klangkupolen which was like a game. Finally, a quote that captures both the opportunities and challenges of creating interactive music and also problematizes the boundary between the composer and the participants:

> *"It should not only sound good but also be fun to participate in. If you give too little control to the audience, it loses its point of being interactive. If you, on the other hand, give them too much control, it may lose the point of composing it at all in advance."*

5. DISCUSSION

This project included several new challenges for the students, each of which would be interesting to study separately. It was obvious that students were euphoric about the immersive sound in Klangkupolen and that the project gave them important ideas for their traditional productions. Therefore, probably a super-surround project would be valuable in itself, without the interactive elements. More than half of the group expressed their frustration with making music for interactions in a non-linear structure, mostly expressed by students with a typical singer / songwriter background. The students with a more loop-based approach in their normal music production was in contrast more positive and engaged more in learning the possibilities with the technical framework. This observation points towards viewing interactive music production as a different discipline compared to standard music production. Maybe technology will never solve the gap between a songwriter and a non-linear context simply because the songwriter wants to express a linear story.

The other group, the students who managed to pass the technical and practical barriers, expressed that this project had given them new perspectives on how their music can work when it becomes a platform for interaction and communication between people. This was said by students with no earlier experience of producing music for interactions and it motivates more attempts to make the technology even more easy to use for those who have no prior experience in programming and interactivity. A challenge to look further into in future studies is how this type of technology can be designed to be easier to use without limiting

the creative possibilities for the music producer. The evaluation of the technical platform also led to a few important observations:

- Open source, installation free, web-based technology performs well enough to work as a platform for projects like this.
- There is huge community developing for web technologies and there will probably be a continuous increase in libraries for sensor detection.
- The web API's are still quite new and not that well established. It's important to constantly read about the latest specification changes to make an application sustainable when the users browser auto-updates.
- The web is primarily made for visual interfaces with touch control and this becomes a problem when using it for a different purpose. In this project especially the auto-play blocking policy and the automatic sleep-mode was causing problems to the interactions.

6. CONCLUSION

Using mobile phones, web technology and Klangkupolen to make immersive music interaction accessible for students with no prior experience was in many ways successful and also gave insights into some challenges. Even if some of the technical barriers were removed, the workflow could be even easier and more fluent. The result also indicates that the artistic challenge to produce music for interactivity differs a lot from standard music production and might not appeal to all. The immersive experience offered by Klangkupolen gave the students valuable insights and an increased interest in trying new approaches in music production. KMH's unique technical infrastructure together with the ongoing development of everyday technology such as mobiles, smart watches and web technology can potentially make projects like this even more accessible to many generations of students in the future.

Acknowledgments

Thanks to all students that contributed with reflections and ideas, to KMH for the funding and not the least to Bill Brunson for designing the technical playground for research and artistic practice at KMH.

7. REFERENCES

[1] A. Bell, *Dawn of the DAW: The Studio as Musical Instrument*, 02 2018.

[2] P. Lewis, "Plug in, turn on, tune up." *Fortune*, vol. 149, no. 4, pp. 52 – 54, 2004.

[3] J. Dashow, "On spatialization," *Computer Music Journal*, vol. 37, no. 3, pp. 4–6, 2013.

[4] T. Redhead, "The emerging role of the dynamic music producer." The Australasian Computer Music Association, 2018.

[5] A. Berndt, K. Hartmann, N. Röber, and M. Masuch, "Composition and arrangement techniques for music in interactive immersive environments," 10 2006.

[6] J.-O. Gullö, I. Höglund, J. Jonas, H. Lindetorp, A. Näslund, J. Persson, and P. Schyborger, "Nobel creations : Producing infinite music for an exhibition," *Dansk Musikforskning Online*, pp. 63–80, 2015.

[7] E. Frid, H. Lindetorp, K. F. Hansen, L. Elblaus, and R. Bresin, "Sound forest - evaluation of an accessible multisensory music installation," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems :*, ser. CHI '19. ACM, 2019, pp. 1–12, qC 20190625.

[8] H. Frisk and W. Brunson, "Building for the Future - research and innovation in KMH's new facilities," in *SMC Sweden 2014: Sound and Music Computing: Bridging science, art, and industry*, R. Bresin, Ed., 2014, pp. 10–11.

[9] D. Schuler and A. Namioka, *Participatory design: Principles and practices*. CRC Press, 1993.

[10] N. Schnell and S. Robaszkiewicz, "Soundworks – A playground for artists and developers to create collaborative mobile web performances," in *'Proceedings of the Web Audio Conference (WAC'15)*, Paris, France, 2015.

[11] H. Lindetorp, "Musik utan slut: Erfarenheter från musikinstallationen i Nobel Creations," in *Elva studier om kreativitet i musikproduktion*. Stockholm: Gullö, Jan-Olof, 2017, p. 150.

BLESync: Wireless Synchronization Between Computers and Tap Tempo Effect Pedals

Juan Pablo Carrascal
Microsoft / Spacebarman
jp@spacebarman.com

ABSTRACT

Electric guitar players often rely on a disparate combination of pedals, whose effects are an essential aspect of the performance language. Many of these effects feature time-based parameters such as delay time or modulation frequency. In a performance incorporating computers and guitars, these parameters should ideally be synchronized with the song tempo as determined by software. While some modern effect pedals incorporate MIDI inputs that can be used for synchronization, a large amount of them require tempo to be "tapped" by stepping on an additional footswitch. In this demo, I present BLESync, a simple MIDI Bluetooth LE device that facilitates the synchronization between a computer and multiple tap tempo-based devices. A wireless connection allows better physical on-stage independence between the computer and the guitar performer.

1. INTRODUCTION

The electric guitar is an extremely versatile music instrument. The combination of elements in its sound chain (the instrument itself, the amplifier, effect units, etc.) give the performer access to an endless tonal palette. By definition, the most radical sound transformations are achieved with effect devices, most commonly implemented in a pedal form factor. Musicians mix and match units to customize their sound, becoming an defining aspect of their performance language [1][2].

Many popular effects depend on time-based parameters. Some examples are the time parameter in echo or delay pedals, and the low frequency oscillator (LFO) in modulation effects such as tremolo, flanger and phaser. By adjusting delay time and LFO frequencies it is possible—and most often desirable—to synchronize the effect to the current song tempo. While MIDI Clock [3] is a frequently used method for synchronizing digital instruments and effects, many effects pedals do not feature MIDI capabilities, hence tempo synchronization is done by means of a "tap tempo" footswitch. While this is a reasonable solution when performing with a band of human musicians, accurate synchronization of an effects pedal with a digital system using tap tempo is difficult.

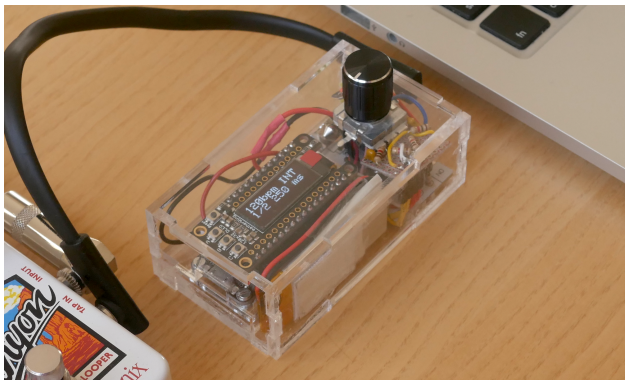


Figure 1. BLESync

I introduce *BLESync* (Figure 1), a Bluetooth LE (BLE) MIDI device that allows wireless synchronization between a computer and tap tempo effect pedals. BLESync's output connects to effects pedals' tap tempo input and simulates tapping by means of a relay switch. Its internal tempo is updated whenever the tempo in the master computer changes, and three taps are sent to its output whenever the internal tempo is updated. The result is that the connected pedal stays synchronized with the master computer, while sparing the need of running additional cables across the stage.

2. DESIGN AND IMPLEMENTATION

2.1 Hardware

Wang et al. [3] evaluated the performance and compatibility of BLE MIDI and concluded that it has potential for replacing wired MIDI interfaces. With this in mind, I built BLESync around the Adafruit Feather 32u4 Bluefruit microcontroller [6], which features a Bluetooth radio compatible with BLE MIDI. This allows the microcontroller to be recognized by computers, mobiles and tablets as a wireless MIDI device. The Feather offers other advantages, including a built-in LiPO battery connector and charger and the possibility to connect a diversity of peripherals, called "FeatherWings". I used a FeatherWing OLED display to visualize the current parameters of the device, including current tempo, tempo signature of the delay, delay time (in milliseconds) and whether the tempo is set internally or externally, i.e., for switching between setting the tempo manually or from a computer via Bluetooth.

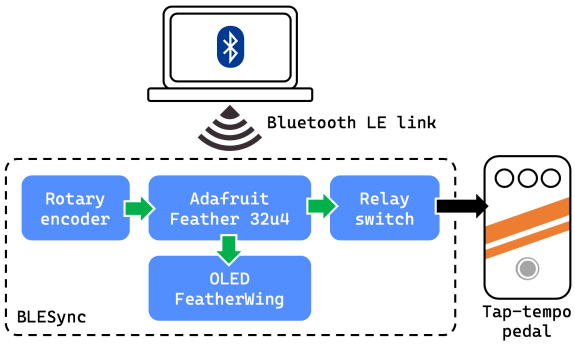


Figure 2. Block diagram

A rotary encoder with a built-in push button, connected to GPIO pins of the Feather, allows the following operations:

- Rotating the encoder manually adjusts tempo when not slaving to a computer.
- By briefly pushing the encoder, the tempo signature switches between different fractions of a beat (1/1, 1/4, 3/4). More values could easily be added with simple firmware code changes.
- If the encoder is pressed for more than 2 seconds, BLESync will switch between tempo sources, i.e. Bluetooth (*EXT*), manual (*INT*) or both (*I+E*).

Another GPIO pin from the Feather controls a relay whose output contacts are connected to a TS jack in the case of the device. A simple patch cable can be used to connect this jack to the tap tempo input of the pedal to be controlled. By opening and closing the relay in quick succession, a tap tempo action is simulated hence controlling the internal tempo of the pedal.

The device is enclosed in a custom-made, laser-cut methacrylate box.

2.2 Software

An Arduino/C++ sketch running on the Feather manages the Bluetooth connection, OLED screen, user input and relay control. The sketch requires several Adafruit's libraries for Bluetooth LE, Bluetooth MIDI and OLED display.

Tempo encoding

MIDI clock is the preferred synchronization method for MIDI devices, thus it would be a natural choice for this project. However, MIDI clock is not part of the Bluetooth MIDI specification [3]. The alternative is to transmit tempo changes as a continuous controller (CC) values. As single CC value are limited to the 0-127 range, tempo can be encoded as a MSB and LSB combination using CC 16 and 48 (general purpose CC, as per MIDI specs [3]) to allow a wide range of tempos. The encoding is simple:

```
tempoMSB = bpm / 10
tempoLSB = bpm % 10
```

Where Both `tempoMSB` and `tempoLSB` are integers, and '`%`' is the modulo (division reminder) operator. Decoding at the receiving end is also very simple:

```
bpm = tempoMSB * 10 + tempoLSB
```

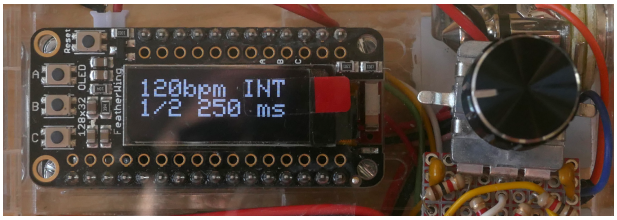


Figure 3. Interface showing current tempo, internally set (i.e. manual, using the encoder), with a signature of 1/2 beat. The period (e.g. delay time) is also shown in milliseconds.

Computer side

In this demo, I use Ableton Live running on MacOS as a master tempo source to control BLESync. As with any BLE MIDI device, a manual connection step is required to make the MIDI port active and available for MIDI software. On MacOS this is done with the Audio MIDI Setup utility. A Max for Live patch obtains the current song tempo from Ableton Live's real time playback information (by means of the `[plugsync~]` object) and applies the encoding described earlier. The patch is inserted in a MIDI channel, whose output is set to be the BLESync MIDI port. The result is that, whenever Live's tempo changes, the new value is transmitted wirelessly to BLESync and its internal BPM value is adjusted accordingly. Please note that, although I use a combination of Ableton Live and Max for the purpose of this demo, the tempo encoding can be implemented with any piece of software that supports MIDI CC messages.

The Arduino sketch and Max4Live patch are available for download at https://github.com/jpcarrascal/BLESync.

Acknowledgements

Dr. Andrea Rosales for letting me borrow her tools.

3. REFERENCES

[1] Darling, T. "A Study of The Edge's (U2) Guitar Delay". http://www.amnesta.net/edge_delay/

[2] Discmakers. "Guitar effects pedals and the evolution of music" (parts 1 and 2): http://bit.ly/2kXT8aR, http://bit.ly/2ms401d

[3] Wang, J., Mulder, A., & Wanderley, M. M. "Practical Considerations for MIDI over Bluetooth Low Energy as a Wireless Interface". In Proceedings of NIME'19, 2019, pp. 25-30.

[4] MIDI 1.0 Specification: http://bit.ly/MIDIspecs

[5] Bluetooth LE MIDI Specification: http://bit.ly/BLEMIDIspecs

[6] Adafruit Feather 32u4 Bluefruit: https://learn.adafruit.com/adafruit-feather-32u4-bluefruit-le/

[7] Max for Live: https://www.ableton.com/en/live/max-for-live/

A LIGHTWEIGHT FRAMEWORK FOR MELODIC INFORMATION ENCODING AND REAL-TIME REPRODUCTION FOR INTERACTIVE SONIFICATION APPLICATIONS

Prithvi Ravi Kantan
Aalborg University Copenhagen
pkanta18@student.aau.dk

ABSTRACT

The use of music in interactive sonification applications is a promising alternative to the simple and often aesthetically unappealing sonic textures used in a considerable proportion of designs. Using synthesized music can be advantageous over pre-recorded music, as the system may exert complete control over individual sonic elements to yield sonically diverse and engaging interactions. The real-time generation of interesting, user-tailored music in an auditory display is a complex process but may be significantly streamlined. The proposed demo showcases a twofold system for rapid and lightweight encoding of polyphonic musical information as well as real-time reproduction of this data format in the user's choice of sonic textures. The system was originally designed for use in a gait sonification system in neurorehabilitation, but may be adapted for multiple data types.

1. BACKGROUND

Despite the promising results that interactive sonification has shown in laboratory settings, auditory guidance at large struggles to find its place in commercial applications. Parseihian et. al. [1] suggest the lack of aesthetic considerations and general formalization of the sonification design process as explanations for this limited adoption. The type of sounds generally used may induce fatigue, or not cater to the taste of the user. Effectiveness and efficiency of auditory guidance have been well-investigated, but the notion of user satisfaction has been absent from most research in this area [1]. Indeed, user preferences vary widely and unpredictably, and the need for user-customizable sonification strategies is inevitable.

I focus on key application areas of interactive sonification such as motor (re)learning and neurological rehabilitation [2] [3], where designers desire that a user engages with the technology repeatedly and frequently in situations that demand high endurance and perseverence. While even the most basic sonic textures such as test signals or white noise can be manipulated to provide effective auditory guidance [1], such designs fail to fully exploit the potential sonic interaction possibilities. Maes et. al. [2] advocate the use of music in movement sonification applications, for its capability to motivate, monitor and modify movement by enhancing self-awareness through extension of the sensory domain. Musical sonification has been applied in multiple applications, such as the D-Jogger [4], and moBeat [5], and multiple studies have shown that feelings of agency through real sonic interaction with music reduce perceived exertion [6] and enhance mood [7].

However, the use of music brings its own set of practical obstacles to the sonification design process. Pre-recorded music (aside from introducing copyright concerns) affords limited interactive signal manipulation without the introduction of unpleasant sonic artifacts. Real-time synthesized music on the other hand allows far more creative interaction possibilities, through spontaneous alterations to musical structure, audio effects and timbre. Moreover, modern computers are more than capable of synthesizing and processing multiple audio tracks while capturing and computing sonification control data. While the design possibilities are no doubt attractive, realizing a system capable of catering to diverse user groups is a multi-faceted challenge. User familiarity has been shown to be a significant factor in the movement-music connection [8].

From a system standpoint, these requirements would necessitate a protocol for flexible encoding of musical information, and a system to decode this information and synthesize user-chosen musical passages in a selection of musical styles, to which target-based sonification metaphors may be applied [1]. It is acknowledged that MIDI is certainly capable of detailed encoding of melodic information. While music-MIDI transcription models exist [9], the process is not trivial and the direct use of existing MIDI files is relatively unreliable in sonification applications. Usable MIDI files for desired music may not always be available, and unpredictable inconsistencies in sound reproduction quality may be introduced by the lack of standardization exhibited by MIDI file creators - for example in terms of octave choices, note velocities and voicing. Furthermore, the rigorous evaluation of sonification strategies would entail a sonic consistency relatively robust to the user's music choice.

The proposed system addresses the aesthetic problem by allowing real-time synthesis of user-chosen music, while opening possibilities for interesting sonification techniques through audio effects and music manipulations that would

be difficult to achieve with pre-recorded music (e.g. D-Jogger). Such an interactive system has the potential to improve adherence in a therapeutic setting, as well as general user retention in other applications. It also streamlines the music structure encoding and decoding processes by standardizing data properties like octaves, polyphony control, voicing and articulation for discrete sonic styles. A robust evaluation procedure inspired by [10] is also planned.

2. DESIGN AND IMPLEMENTATION

The proposed demo aims to first showcase a lightweight polyphonic representation scheme capable of rapidly encoding simplified versions of short melodic passages as integer codes, complete with real-time previewing and editing. These structures are then decoded and reproduced in real-time by a sequencer with added instrumentation in an array of musical styles. The scheme was designed to assist the testing of a gait rehabilitation application, and correspondingly the time signature is fixed at 4/4 with binary patterns suited to human walking [3]. Regardless, similar schemes can streamline the iterative development of user-flexible interactive systems in various applications.

Both the encoder and decoder are written in JUCE (C++) and audio synthesis is carried out in FAUST. The encoder scheme represents melody passages as a combination of a chord pattern and a monophonic melody. Each passage is a four bar (measure) segment, and scale, pitch, velocity, note octave and chord type information is encoded at a sixteenth-note temporal resolution. Thus, an eight-digit integer code corresponds to two beats, and a passage requires eight such codes. The information is represented in a compressed format, for example velocity only has ten levels (0-9) and pitch is stored as scale degrees (1-9), allowing a compact representation. For example, a velocity code of '90009000' represents a quarter note pattern while one of '90909090' is an eighth note. The encoder application allows a user to create up to five passages, monitor the music in real-time and save the passages in CSV format.

The decoder reads the stored melody and chord structures from the CSV and reproduces the passages in the user's choice of styles at any tempo. Timekeeping is handled by a master clock that generates a sixteenth note impulse train, and a sequencer tracks pulses to update musical time counters and fetch melodic information from the encoded format in real-time. Musical evolution over time is defined within the sequencer, but may be altered to suit the situation. The decoder also has a library of style-appropriate percussion patterns represented as velocity codes, and the exact chosen percussion pattern is randomized each time. The musical information is mapped to the FAUST synthesis program at every clock interval, where the information is ultimately converted into audio signals and played back.

3. DEMO DESCRIPTION

The rapid real-time encoding process will be showcased, and the sonic possibilities will be auditioned and assessed by participants. As this is part of ongoing work, the sonification system may be in need of further testing but suggestions and ideas will be welcomed, as well as ideas for gestural mapping. Subjective ratings of the synthesized music will also be instrumental in honing the design of the synthesis algorithms.

4. REFERENCES

[1] G. Parseihian, S. Ystad, M. Aramaki, and R. Kronland-Martinet, "The Process of Sonification Design for Guidance Tasks," *Wi:Journal of Mobile Media*, vol. 9, no. 2, 2015. [Online]. Available: https://halshs.archives-ouvertes.fr/halshs-01230638

[2] P.-J. Maes, J. Buhmann, and M. Leman, "3mo: A Model for Music-Based Biofeedback," *Frontiers in Neuroscience*, vol. 1, 12 2016.

[3] N. Schaffert, T. B. Janzen, K. Mattes, and M. H. Thaut, "A Review on the Relationship Between Sound and Movement in Sports and Rehabilitation," *Frontiers in Psychology*, vol. 10, p. 244, 2019.

[4] B. Moens, C. Muller, L. van Noorden, M. Frank, B. Celie, J. Boone, J. Bourgois, and M. Leman, "Encouraging Spontaneous Synchronisation with D-Jogger, an Adaptive Music Player that Aligns Movement and Music," *PLOS ONE*, vol. 9, 12 2014.

[5] B. Vlist, C. Bartneck, and S. Mueler, "Mobeat: Using Interactive Music to Guide and Motivate Users During Aerobic Exercising," *Applied Psychophysiology and Biofeedback*, vol. 36, pp. 135–45, 03 2011.

[6] T. Fritz, S. Hardikar, M. Demoucron, M. Niessen, M. Demey, O. Giot, Y. Li, J.-D. Haynes, A. Villringer, and M. Leman, "Musical Agency Reduces Perceived Exertion During Strenuous Physical Performance," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 110, 10 2013.

[7] T. Fritz, J. Halfpaap, S. Grahl, A. Kirkland, and A. Villringer, "Musical Feedback During Exercise Machine Workout Enhances Mood," *Frontiers in Psychology*, vol. 4, p. 921, 12 2013.

[8] K. S. Park, C. Hass, B. Fawver, H. Lee, and C. Janelle, "Emotional States Influence Forward Gait During Music Listening Based on Familiarity with Music Selections," *Human Movement Science*, vol. 66, pp. 53–62, 03 2019.

[9] E. Benetos and S. Dixon, "A Shift-Invariant Latent Variable Model for Automatic Music Transcription," *Computer Music Journal*, vol. 36, no. 4, pp. 81–94, 2012. [Online]. Available: http://www.jstor.org/stable/41819549

[10] G. Parseihian, C. Gondre, M. Aramaki, R. Kronland Martinet, and S. Ystad, "Exploring the usability of sound strategies for guiding task: toward a generalization of sonification design," in *Proc. of the 10th International Symposium on Computer Music Multidisciplinary Research*, Marseille, France, Oct. 2013.

BASS AS AN INDICATOR OF QUALITY

The Relation Between Bass Levels and Quality Perception in Headphones

Martin Linder Nilsson
KTH, Royal Institute of Technology
The Department of Media Technology and Interaction Design
hmni@kth.se

Johannes Loor
KTH, Royal Institute of Technology
The Department of Media Technology and Interaction Design
loor@kth.se

ABSTRACT

Bass is a key component in music and its use in modern genres has resulted in lower frequencies generally taking a more prominent role in the frequency spectrum. The preferred level of bass and what is considered as high quality sound is however subjective. This leads to several questions of interest regarding the way we perceive sound quality and how this perception relates to bass levels. With the ambition to explore this, a study was conducted mapping the connection between bass levels and a perception of quality while listening through headphones. Three audio files, representing two different genres of music and one recorded audio book, was edited beforehand with three different levels of bass and then listened to by 41 test subjects in a blind test setting. Test subjects did not get to see or touch the headphones in between versions, which created the illusion of listening to several different headphones when in fact the same headphones were used for all versions and levels of bass. This method, of leading the subjects into believing that they evaluated the headphones instead of different bass levels, changes the scope from only evaluating the bass to instead focusing on the correlation between bass and an overall sense of quality. The results show a positive correlation between high bass levels and perceived quality, where test subjects rated audio with a higher bass level as having higher quality than audio with lower bass levels.

Keywords

Bass; sound quality; quality perception; headphones; frequency response; music

1. INTRODUCTION

Bass frequencies are important to us. In music bass makes us want to move [6], low frequencies are used to create a sense of presence in movies and in many cultures around the world bass has been used as a conveyer of rhythm in the form of drums and other instruments. Bass frequencies constitutes the lowest part of the frequency spectrum, a popular definition being the range from C0 to C4 in the western chromatic scale, which translates to frequencies between 16Hz and 256Hz. Frequency response in music however, is highly preferential. The amount of bass or treble one prefers may vary substantially from one person to another [7]. In the case of headphones, bass levels are a particularly well discussed topic [9,10], both in regard to the amount of bass reproduced but also how far down the frequency spectrum it is being rendered. Headphones with excessive bass levels may in some cases prove problematic as these frequencies tend to obstruct the audibility of others. The effect being a misrepresentation of the original sound material, assuming that the composer's intention is for all frequencies to be audible. Hence, the aim of this study is to investigate whether high levels of bass has become synonymous with high quality. Specifically, how the perceived sound quality of a piece of music or speech recording changes at different bass levels.

2. THESIS AND PREVIOUS RESEARCH

2.1 Previous Research

Musical instruments that deliver low-frequency sound have been used to create rhythm in music in many cultures through the ages. Why the bass has been given the role of rhythm bearer to such an extent has therefore been the subject of several studies. A recent study [6] shows that this culturally widespread practice might utilize a neurophysiological mechanism in us humans, whereby low-frequency sounds, to a greater extent than high-frequency, affect how we perceive rhythm. This effect is also strongly linked to our motor skills and our ability to move.

A study by Varlet et al [13] researching the relationship between motor skills and auditory and visual rhythm

showed that people had easier synchronization of movements to low-frequency sound than to high-frequency sound. This phenomenon is further investigated in another study [12] where subjects danced to club music with varying volumes on the bass drum. Their movements were measured using motion capture technology showing that with increased volume of the bass drum, the subjects not only danced with larger, more energetic movements, but also had an easier time connecting to the tempo of the music.

Dynamic compression has a long history of use in the world of music. The reason for this is to even out the dynamics of the audio, adapting it to for instance radio or listening in the car [4]. With compression follows the opportunity to subsequently amplify the music, thus creating a piece of music that is perceived as louder than the original track and is more rich in energy, a value usually calculated as RMS energy (Root mean square). This way, the dynamic range is reduced while increasing the perceived volume, also called "loudness". In a study, Hove, Vuust and Stupacher [4] show that hit songs between 1955-2016 have seen an increase in RMS energy, where bass frequencies constitutes the biggest contributing factor, while dynamic range has decreased. This gradual increase in volume, usually referred to as the "loudness war", aims to describe the desire of various record companies and artists to make music as loud as possible, thus having an edge in a music industry characterized by fierce competition. However, new technology such as automatic loudness-normalisation facilities in both broadcast and consumer playback systems, might reduce or even remove loudness as a factor[11].

A not as commonly discussed topic is the effect that the loudness war has had on the frequency reproduction in music, where compression has led to rising levels of both bass and treble. Modern genres such as rock, pop and schlager generally have less dynamic range than older genres such as orchestral music, choral pieces and chamber music, while newer genres contain a greater amount of bass and treble than the older ones [5]. This is something that was taken into consideration in the process of choosing genres for this study.

Several studies [2,3] have explored the impact of frequency response on the experience of music. Gabrielsson et al [2] conducted a study where the perception of music and speech recordings were evaluated after passing through various filters. A test group was asked to evaluate these filtered sounds according to subjective attributes such as clarity, fullness, spaciousness and brightness. A low-shelf filter that increased the level of bass frequencies, much like the one used in our study, made subjects experience sounds as more full and muffled and at the same time less clear, spacious and bright. While Gabrielsson focuses on how these perceived attributes change, the basis for our study is how the frequency range around the bass affects the overall quality experience.

A study [1] shows no correlation between price range and the frequency response of headphones. The test was done on headphones of the type in-ear, supra-aural and circumaural. This was taken into account when choosing headphones for our study.

2.2 Thesis

The main aim of this research report is to examine the effect bass levels have on the perceived sound quality of headphones, hence the main research questions being:

Will an increase in bass be perceived as a marker of high quality, making a piece of music or a pair of headphones sound better or feel more expensive?

Thus, the purpose of this study is to help provide some clarity in the psychology of bass levels. Furthermore exploring whether results vary through different genres of music or in speech recordings.

Our thesis is that sounds more rich in bass will generally be perceived as more qualitative than sounds containing less bass.

3. GATHERING OF DATA

3.1 Method

A study involving 41 test subjects was conducted on April 1-3 2019. Test subjects were divided into three groups, two of which were assigned a genre of music each and the third a speech recording. Genres, pop and classical, were chosen to be as different as possible, thus broadening the scope of the experiment and making it possible to choose songs where the amount of compression varies substantially. The song representing the pop genre was “Anyone out there” by Iselin[1] and the chosen classical piece was “Piano trio no. 3 in C-minor, Op. 101: Andante Grazioso” by Johannes Brahms for piano, cello and violin[2]. Each genre was prepared in three different versions: one with the original bass level and the other two with reduced and enhanced bass levels respectively. Preparations were done in Cubase software using a low-shelf filter, Waves Audio Q10 [13], which increased or decreased bass frequencies by 5dB respectively, with the cut-off point at 200 Hz. The pop-song was the track that varied the most in regard to the amount of bass between versions, and also the one being the most dynamically compressed. The speech recording was a male voice reading of an audiobook played in mp3-format, 154 kBit/s, and all versions of the music tracks were played in mp3-format, 320 kBit/s. All clips used in the study were about 20 seconds long.

The study was conducted with two pairs of headphones, Sennheiser HD600 and AKG K240. The HD600s are considered as being open headphones and the K240s are semi-open which was an important factor since the bass

[1] https://open.spotify.com/track/48wEohztw5FeQboMzIb2LB?si=fkBnyy39Q2GlXbTWne8iGA

[2] https://open.spotify.com/track/08JqZxMgLW6Y8vyGYC6WJC?si=JhyfuU_4RNOxAKgTku6C-Q

loss of headphones that do not fit correctly is greater with closed headphones than with open headphones. The ease of putting the headphones on was also of importance as well as a varied price point. At the time of the study the Sennheiser headphones had a retail price of 3300 swedish kr (SEK) and the AKG:s cost 700 SEK.

3.2 Preparatory Test

With the purpose of providing the same perceived volume for all versions of the sound clips used in the main study, a preparatory test was conducted. Seven test participants, not partaking in the main study, listened to several different versions of each of the sound clips in an A/B test-setting and evaluated the perceived volume. The result of the test determined which tracks were then used in the main survey.

3.3 Main Study

During a blind test, participants listened to the same 20-second audio clip in three versions, one of which being the original and the other two were edited in advance with different amounts of bass. Between each version an illusion of changing to a different pair of headphones was created by not letting the participants see or touch the headphones as they were put on and taken off. This aspect, that the participants believe that they evaluate different headphones instead of different bass levels, is important to be able to determine how the level of bass affects the perceived sense of quality in a pair of headphones. Hence shifting the focus from only evaluating the amount of bass to comparing the connection between bass and an overall sense of quality. The use of the same pair of headphones, for all three versions of the track, also minimizes potential factors of error concerning the fit of the headphones and the varying frequency response of headphones. Before each audio clip the interviewer asked if the headphones fit properly around the ear.

After listening to each audio clip, the subjects were asked to fill out a questionnaire with the following four questions:

- How did you experience the sound quality?
- How did you experience the bass levels?
- How did you experience the treble levels?
- How much do you think the headphones cost?

For the first three questions, the subject answered on a scale from one to ten. The ends of the scale, 1 and 10, were described as “Very poor” and “Very high” for the first question and "Very weak" and "Very strong" for the second and third question. For the final question three options were given: Less than 1000 SEK, between 1000 and 2000 SEK and more than 2000 SEK.

3.4 Test Group

The test group consisted of 41 college engineering students, all of which majored in Media Technology at the Royal Institute of Technology in Stockholm. Subjects were between 19 and 29 years old, all of them perceived themselves to have normal hearing and their experience of listening to music and speech in headphones varied.

The test group was divided into three subgroups, each being assigned a genre. The pop song and the audiobook was listened to by fourteen people each and thirteen people listened to the classical piece. Half of the test group started with the Sennheiser headphones and the other half with the AKG:s. In order to minimize possible sources of error that may arise from the audio tracks being played in a certain order, the tests were arranged so that all possible orders had equal representation. Adding the option to play either the AKG headphones or the Sennheiser headphones first, result in the following calculation:

3! possible orders * 2 headphones = 12 different orders

Hence the audio clips were played in a total of 12 different orders.

4. RESULT

Test data from the different series were compared and analyzed focusing on differences in the mean values between the original audio clip and the other two clips. Results regarding the perceived sound quality and cost are thoroughly reviewed, while answers from the bass and treble questions are compared briefly with the other data. To test the significance, Anova tests were used.

4.1 Analysis of Perceived Sound Quality

	Reduced bass (-5dB)	Original	Increased bass (+5dB)
AKG	6,27	6,66	7,24
Sennheiser	6,32	7,22	7,41
All answers	6,29	6,94	7,33

Table 1. Mean value table for the question "How did you experience the sound quality?", Where possible answers range from 1 (very poor) to 10 (very high). Divided by headphones and all answers.

4.1.1 *All Responses*

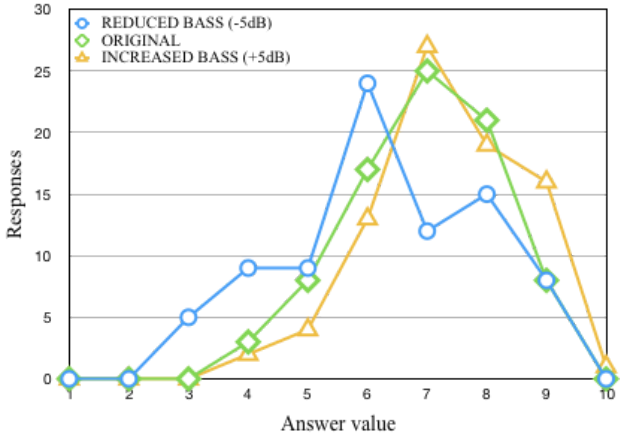


Figure 1. The result of all answers to the question "How did you experience the sound quality?" for the different audio clips. Answer value 1 corresponds to (very poor) and 10 to (very high).

The result of all 82 responses, across all three genres and in both headphones, shows that the soundtracks with enhanced bass levels produced slightly higher values than the originals. As shown in Figure 1, both versions had the largest amount of responses at value 7. The bass enhanced version however generated more of the higher values on the scale producing the higher mean of 7.33 compared to 6.94 for the original tracks, as shown in Table 1. A significant difference can be established at 5% between the originals and the tracks with the enhanced bass ($p = 0.0497$).

The tracks with reduced bass levels generally show lower values than the other two, which is reflected in the lower mean of 6.29. A test of significance as the one above, between the original and the reduced bass tracks, also show a significant difference ($p = 0.0059$).

Following the analysis of variance, we can therefore determine a significant difference in the experience of sound quality, at a significance level of 5%. This is true in both the comparison between the bass-reinforced audio clips and the originals, as well as between the originals and clips with reduced bass levels.

When looking at the results from the two kinds of headphones used in the study separately, shown in Table 1, it's apparent that they follow the same trend as for all answers combined. Statistical significance can be established at 5% between all series except when comparing the bass reduced tracks with the originals in the case of the AKG:s, and in the comparison between the originals and the bass-enhanced tracks in the case of the Sennheisers.

4.1.2 *Responses Divided by Genre*

The results regarding the classical piece, presented in Table 2 and Figure 2, show that listening to the song with more prominent bass generally received higher response values than the remaining versions. The average value for this is 7.38, while the original and the reduced bass sound clip received an average of 6.81 and 6.85 respectively. However, comparison of these test series does not meet the requirement for significance at the level of 5% ($p > 0.05$).

	Reduced bass (-5dB)	Original	Increased bass (+5dB)
Classical	6,85	6,81	7,38
Pop	5,36	6,71	7,00
Audio book	6,71	7,29	7,61
All answers	6,29	6,94	7,33

Table 2. Mean value table for the question "How did you experience the sound quality?", where possible answers range from 1 (very poor) to 10 (very high). Divided by genre and all answers.

When looking at the results from the question about sound quality regarding the generally more bass rich pop-genre, it is apparent that the differences in perception of sound quality between the different versions are greater. As shown in Table 2, the bass amplified sound clip gets the highest average of 7.00 followed by the original at 6.71. However, this difference can not be viewed as significant ($p > 0.05$). The version with reduced bass levels is experienced by the test group as much poorer in sound quality and receives an average value of 5.36. The difference between the test group's experience of this audio clip versus both the original and the clip with increased bass levels is significant at 5% ($p < 0.05$).

Mean values for the perception of the sound quality regarding the audio book reading, also shown in Table 2, follow the same pattern as most other series in the experiment, ie the track with enhanced bass levels gets a higher average value than the original track and the one with reduced bass. The bass-enhanced track receives an average of 7.61, while the original and the bass-reduced audio clip receive an average of 7.29 and 6.71 respectively. Significance tests between these different versions show that there neither is significant difference between the lowered bass version and the original, nor between the original and the reinforced version ($p > 0.05$). A statistical comparison between the track with reduced bass and the one with enhanced bass, however, shows a difference in experience, at significance level 5% ($p < 0.05$).

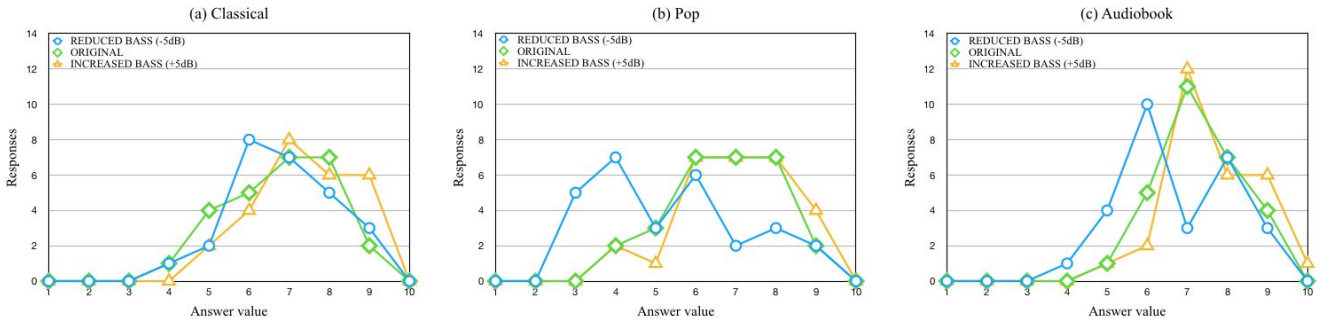


Figure 2: The result of answers to the question "How did you experience the sound quality?" for the different audio clips when listening to the genres classical (a), pop (b) and audiobook (c). Answer value 1 corresponds to "Very poor" and 10 "Very high".

4.2 Analysis of Presumed Price Range

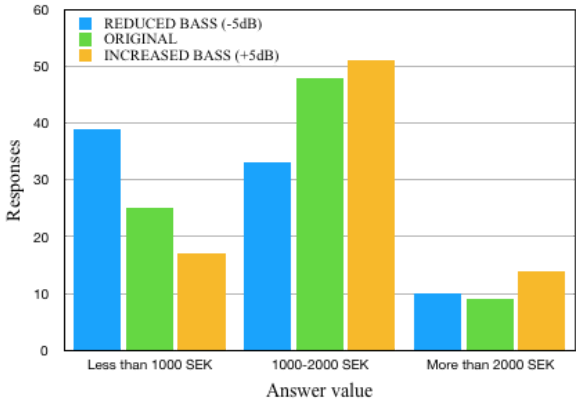


Figure 3. The result of all answers to the question "How much do you think the headphones cost?" for the different audio clips.

4.2.1 All Responses

For the question "How much do you think the headphones cost?" there were three response options, "Less than 1000 SEK", "1000-2000 SEK" and "More than 2000 SEK". The results show that both headphones used in the test are perceived as being a bit more expensive by the test group while listening to sound clips with boosted bass, followed by the original versions and finally those with reduced bass. The difference is slight however. As shown in Figure 3, tracks with lowered bass generated the most results on response option 1 “Less than 1000 SEK”, while the original versions and those with increased bass received most answers on alternative 2 “1000-2000”. However, versions with increased bass received more answers to option 3 “More than 2000 SEK” than the originals. If these response options are labeled as 1, 2 and 3 as described above, the tracks with reduced bass receive an average of 1.65 for both headphones combined. The original versions receive an average of 1.80 and those with increased bass receive an average of 1.96. However, an Anova test shows that the only difference that can be considered statistically significant ($p < 0.05$) is the one between the tracks with reduced bass and those with boosted bass levels. For the others, no statistical difference was seen ($p > 0.05$).

4.2.2 Responses Divided by Genre

When looking at the results from the question about the presumed price range separated by genre, one finds that all genres follow the same trend as the results for the entire test group. Thus, the headphones are perceived as the most expensive while playing the bass-enhanced audio tracks, a little cheaper when playing the tracks in the original version and even cheaper while playing the bass-reduced tracks. The biggest difference is seen in the pop genre. Using the same scale as above, naming the categories 1, 2 and 3, gives an average of 1.43 for the reduced bass track, 1.75 for the original version and 1.96 for the amplified bass track. However, an Anova test shows that there is no statistically significant difference between the versions at a significance level of 5%, in any other case than between the bass-reduced track and the bass-enhanced.

4.3 Bass and Treble Levels

	Reduced bass (-5dB)	Original	Increased bass (+5dB)
Bass level	4,82	5,94	6,66
Treble level	6,23	5,95	5,85

Table 3. Mean value table for experienced bass and treble levels.

Between each audio clip the test subjects were asked to answer the questions “How did you experience the bass level? (Low sounds)" and "How did you experience the treble level? (High sounds)”. In both cases, options

ranged from 1 to 10, where 1 corresponded to very weak and 10 to very strong.

As shown in Table 3, analyzing the answers to the question about bass level shows that the mean value of the perceived bass level varies according to the prepared variation of the track. Worth noting is that the result from the question about the bass experience follows the same trend as the one about sound quality. As for the question about the perceived treble, the mean level of the perceived level decreases as the bass level increases.

5. DISCUSSION

With the purpose of contributing to the research regarding bass and how the perceived sound quality varies when changing it, we considered the question: "How do bass levels affect the experience of quality when listening through headphones?". The overall result from the study reinforces our thesis, that increasing bass levels leads to an increase in perceived quality. How results vary across different genres and how bass level correlates with an assumed price were also explored to form a more holistic view.

To gain a more definite result, a larger test group would have been prefered. In our opinion this is problematic because when analyzing the answers by different factors, by genre for example, we could see a clear trend in the mean values that increased amounts of bass led to a perception of better sound quality. However, these trends could not always be statistically proven, a contributing factor being the low number of responses.

Although we were able to show the positive effects increased bass levels have on the perceived sound quality when analyzing all answers, the same can not be said for all genres separately. The track in which the perceived sound quality varied the most between the different bass versions was the pop song, which is interesting considering it is the track already containing the lowest frequencies of bass. Raising or lowering the bass levels therefore gives a more noticeable effect than in other genres, as more information is available at low frequencies. The classical piece lacks deep bass, when compared to the pop song, and therefore the effect of different bass levels is harder to perceive. This is also reflected in the results, where we see that the classic genre is the one varying the least in perceived sound quality between the different versions of the song.

We also examined the difference between how bass levels affect the quality in music in contrast to in speech. Although not statistically determined, both the music and speech tracks follow the same trend as previously mentioned where an increase in bass gave a higher assessment of the sound quality.

In order to provide a more generally applicable result, rather than for only one pair of headphones, tests were conducted using two pairs of different headphones. A more comprehensive result could have been achieved with a larger number of headphones. This would, however, have required a substantially larger test group, which unfortunately was not possible to acquire within the framework of this study.

The question about the assumed price only had three relatively large categories, which may have made the process of obtaining a well-substantiated result more difficult, as well as making findings problematic to prove statistically. With a larger set of response options, nuances in the presumed price difference could have been better accounted for. Categories in higher price ranges may also have been a good addition, to cover the high end range of headphones. As is, no conclusion can be drawn from the result of this question, other than a statistical difference between the bass-reduced tracks and those with increased bass levels, where the latter are perceived as more expensive. This fact is also true if you look at the results from the headphones separately. With a larger test group and, as mentioned above, more response options, one might be able to establish statistically significant results from the trends found through analysis of the mean values. Consequently, regarding how the headphones price was estimated, our results are in line with previous research [1], in that no direct correlation between bass level and cost was found.

Results from the question about how the bass level was perceived was consistent with the actual change between the audio clips, which indicates that subjects were aware of the changes in bass level. A problem discussed in the introduction is that an overemphasized bass tends to overshadow other parts of the frequency spectrum. If bass frequencies constitutes too large a part of the total frequency reproduction in headphones, musical works might not receive the auditory representation the author intended, as certain frequencies are not represented. The results of the question about how the level of treble was perceived shows a mean value decrease as bass level (both perceived and actual) increases. This gives an indication that an oversized bass does have a tendency to take over the soundscape, making it harder to perceive other frequencies to the same extent, in this case specifically the treble.

It should be borne in mind that the test group used in this study should not be considered representative of an entire population, since the interest of sound and music can be assumed to be somewhat higher than average in media technology students. In terms of listening experience, however, there is research to support that experienced and inexperienced listeners generally prefer the same speakers in a blind test [8]. The age range of the test group, 19-29 years, can also be considered a source of error, as research [9] shows that age is an important factor to bass and treble preference. Younger people, according to the study, generally appreciate a greater amount of bass and treble than the older people. So although the results of the study therefore cannot be assumed to be generally applicable, it can however, give a strong indication of the effect of bass levels on the perceived sound quality.

So what does this result, that we appreciate bass-heavy sounds more than others, mean? It may have something

to do with being indoctrinated by developments during the second half of the 20th century and onwards, where modern genres have gained an ever more pronounced bass [4], creating in us a preference for a more prominent low-end. Genetic reasons might also be speculated in, as studies [6,13] show that the brain responds to bass frequencies in a different way than to other frequencies, especially when it comes to rhythm and motor skills.

6. CONCLUSIONS

Through the results of our study we see a positive correlation between bass levels and a perception of quality when listening through headphones. This correlation applies to a total result across music genres pop and classical music as well as an audio book speech recording. However, results are unclear in terms of the genres individually, largely due to the size of the test groups. The design of the test group should also be taken into account, as results can be considered valid only for the demographic group represented in the study, ie college students. No clear link between bass levels and the presumed price range of the headphones could be established by the result.

7. REFERENCES

1. Jeroen Breebaart. 2017. No correlation between headphone frequency response and retail price. *The Journal of the Acoustical Society of America* 141, 6: EL526–EL530.
2. Alf Gabrielsson, Björn Hagerman, Tommy Bech‐Kristensen, and Göran Lundberg. 1990. Perceived sound quality of reproductions with different frequency responses and sound levels. *The Journal of the Acoustical Society of America* 88, 3: 1359–1366.
3. Alf Gabrielsson, Björn Lindström, and Ove Till. 1991. Loudspeaker frequency response and perceived sound quality. *The Journal of the Acoustical Society of America* 90, 2: 707–719.
4. Michael J Hove, Peter Vuust, and Jan Stupacher. 2018. Increased levels of bass over time in popular music recordings and their relation to loudness. *The Journal of the Acoustical Society of America* 145, 4: 2247.
5. Martin Kirchberger and Frank A. Russo. 2016. Dynamic Range Across Music Genres and the Perception of Dynamic Compression in Hearing-Impaired Listeners. *Trends in Hearing* 20: 2331216516630549.
6. Tomas Lenc, Peter E. Keller, Manuel Varlet, and Sylvie Nozaradan. 2018. Neural tracking of the musical beat is enhanced by low-frequency sounds. *Proceedings of the National Academy of Sciences* 115, 32: 8221.
7. William McCown, Ross Keiser, Shea Mulhearn, and David Williamson. 1997. The role of personality and gender in preference for exaggerated bass in music. *Personality and Individual Differences* 23, 4: 543–547.
8. Sean E Olive. 2003. Differences in performance and preference of trained versus untrained listeners in loudspeaker tests: A case study. *Journal of the Audio Engineering Society* 51, 9: 806–825.
9. Sean E Olive and Todd Welti. 2015. Factors that influence listeners' preferred bass and treble balance in headphones. *Proceedings of the 139th AES Convention*.
10. Sean Olive and Todd Welti. 2015. Factors That Influence Listeners' Preferred Bass and Treble Levels in Headphones. *Audio Engineering Society Convention 139*.
11. Hugh Robjohns. The End Of The Loudness War? Retrieved October 28, 2019 from https://www.soundonsound.com/techniques/end-loudness-war.
12. Edith Van Dyck, Dirk Moelants, Michiel Demey, Alexander Deweppe, Pieter Coussement, and Marc Leman. 2013. The Impact of the Bass Drum on Human Dance Movement. *Music Perception: An Interdisciplinary Journal* 30, 4: 349.
13. Manuel Varlet, Rohan Williams, and Peter E. Keller. 2018. Effects of pitch and tempo of auditory rhythms on spontaneous movement entrainment and stabilisation. *Psychological Research*.
14. Q10 Equalizer user guide. Retrieved October 29, 2019 from https://www.waves.com/1lib/pdf/plugins/q10-equalizer.pdf.

Design of an In-Lab Experimentation Rack System for the ACUTE Multi-Channel Tactile System

Elvar Atli Ævarsson
University of iceland
eae19@hi.is

Mathias Damgård
University of Iceland
mld8@hi.is

Árni Kristjánsson
University of Iceland
ak@hi.is

Runar Unnthorsson
University of Iceland
runson@hi.is

ABSTRACT

Listening to music, as an ensemble of audible acoustic frequencies carefully organized over a time period, is an experience that can generate pleasure and enhance the mood but many take for granted. However, many people with cochlear implants (CI) describe the experience as not so pleasant as the implants alter the frequency perception. Using tactile feedback to aid the hard of hearing is a methodology that has been studied in recent years. Building on previous work from the H2020 funded Sound of Vision project (No: 643636), we propose to modify the tactile belt developed in that project by replacing the motors with voice coils in order to make it a better suitable tactile display for conveying musical features. The aim is to help the CI recipients to better enjoy music.

This paper will outline the ideology of the project as well as describe the in-lab system used for experimentation while designing the ACUTE Multi-Channel Tactile System (Multi-CaTS). With 64 channels of audio to drive the array of voice coils comprising the tactile display, the in-lab rack system will be used to design methods for mapping music to multi-channel tactile feedback.

1. INTRODUCTION

Cochlear implants can vastly improve the quality of life of people with severe hearing impairment. Being able to hear and understand speech is an untold luxury to people who have spent years or decades in silence. For the purpose of one-on-one communication in a quiet environment, the cochlear implant has shown great results. However, with more complex acoustic signals, such as music, the results can be disappointing as crucial musical elements like pitch and timbre are lost [1], which greatly affects the perception of melody, harmony and tone quality.

In the Sound of Vision (SoV) project, a cross-disciplinary research group developed a system for assisting visually impaired people with orientation and navigation [2–5]. The tactile belt, shown in figure 1, is an important part of that system. Consisting of a 6×10 array of eccentric rotating mass vibration (ERM) motors forming a tactile display, the SoV Tactile Belt receives signals from a computer which maps images from a 3-D camera (worn on the subject's forehead) to localized vibrations on the tactile display, creating an image of potential obstructions in the near environment.

Figure 1. The Sound of Vision Tactile Belt comprised of a 6×10 array of ERM motors.

Our goal is to use the general idea of the SoV Tactile Belt to further increase the music enjoyment of cochlear implant users. In doing so, we modify the belt by replacing the ERM motors with specially designed voice coils. While ERM motors worked well in the SoV project, they have the disadvantage of working at a fixed frequency depending on operating voltage and load [2]. Any change in pressure applied on the motors by the body will thus have an effect on the vibrational frequency. Voice coils on the other hand directly react to audio signals. Substituting ERM motors with voice coils, driven by low-power audio amplifiers, will therefore greatly simplify the control of both vibrational frequency and amplitude.

Since the new tactile display will contain 60 voice coils, with each controlled individually, the system will need to include 60 independent audio channels with the required amplification for each coil. Making it a compact and portable system will be a challenging task. The first step will however be to design and build a proof-of-concept system that can be used for lab experiments where the use of voice coils will be compared to the old SoV Tactile Belt, where

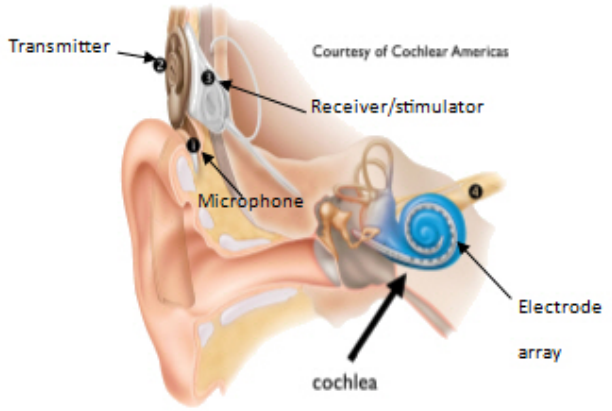


Figure 2. Diagram of the cochlear implant. Electrodes surgically implanted inside the cochlea receive information of stimulation level and directly trigger the auditory nerves. [6]

different audio-tactile mapping techniques will be compared, etc. This paper describes the hardware system setup for this work and further development.

2. COCHLEAR IMPLANTS

The basic structure of a human ear with a cochlear implant can be seen in figure 2. A microphone placed by the outer ear picks up acoustic signals which a processor then analyzes in terms of frequency and amplitude and divides into frequency bands, with each band assigned to a single electrode in an array of electrodes surgically implanted in the cochlea. A transmitter sends radio signals containing information on the stimulation level assigned to each electrode. The electrodes then activate the hair cells similarly positioned in the cochlea, thus directly triggering the auditory nerves and enabling the person to hear.

Since the number of hair cells in the inner ear is in tens of thousands, implanting one electrode per hair cell for a full resolution is impossible. Instead space limitation restricts the number of electrodes down to 16-22 (depending on the manufacturer), thus greatly affecting the precision of sound. This heavily distorts the perception of music.

2.1 Cross-modal plasticity in CI users

The brain has a way of adapting to sensory loss. With one sense lost, the area of the brain devoted to that sense gets a new role dedicated to other senses. For example, blind people tend to learn how to use auditory cues to a greater effect. The above mentioned SoV project showed good results in using the tactile belt as pathway between the somatosensory and visual systems.

Recent studies show that children with CI are generally susceptive to cross-modal reorganization between the somatosensory and auditory systems for speech recognition [7]. Furthermore, music has been presented to the deaf and hard of hearing using tactile displays, for example in form of the Model Human Cochlea [8] with a 2×8 array of voice coils arranged in the back of a chair. However, almost all tactile solutions have been low resolution tactile devices [8–10]. In this project we are developing a high-resolution tactile display (60 actuators) with actuators that cover a wide frequency range (see section 3.4).

With this project, our goal is to design a system that can in some way replace the musical elements lost to CI users, by means of localized vibration patterns applied alongside music signals.

2.2 Experimenting with audio-tactile mappings for musical pleasure

There are many issues to consider when expressing music using tactile stimulation. Mapping will play an integral part in the overall experience, as it ultimately determines the way the music is perceived by the CI user. Two types of audio-tactile mappings have been proposed and are currently being tested in our initial study using the SoV Tactile Belt with ERM motors. The lowest part (up to 300Hz) of the audio signal is divided onto six frequency bands, 5-50Hz, 50-100Hz, 100-150Hz, 150-200Hz, 200-250Hz and 250-300Hz. The reason for using six frequency bands is so that each row of the 6×10 array can represent a single frequency band. Each row of 10 motors is split up in the middle into two rows of five motors with the left side representing the left channel of the stereo signal and the right side representing the right channel. The first audio-tactile mapping treats the motors in each row as an indicator of intensity, with the vibration moving closer to the middle depending on the magnitude of the frequency band. The second mapping also utilizes the same frequency bands but activates all the motors in each row at the same time.

Experimenting with different mappings will be crucial. Frequency scaling, from the frequencies perceptible by the ear (20-20kHz) to the the frequencies perceptible by the skin (5-1kHz) and melody extraction is also something to look at. This calls for a reliable hardware system where the experiments can be conducted.

3. THE IN-LAB EXPERIMENTATION RACK SYSTEM

The hardware system comprises of an audio interface that supports the Multichannel Audio Digital Interface (MADI) standard, digital/analogue (D/A) converters, custom made multi-channel low-power amplifiers and an actuator array (see figure 3), set up as an in-lab rack system.

3.1 Audio interface

The system uses the RME MADIface TX audio interface. MADI is a digital audio standard that supports bit transmission of up to 64 channels over a single fibre-optic cable. Since the MADIface TX has two MADI I/O optical ports, this setup allows for up to 128 channels of audio. For the purpose of this project only one port will be used, as a single MADI port will provide more than enough audio channels for the 6×10 tactile display.

A laptop personal computer (PC) connects to the audio interface via USB. A stereo signal from an audio source (which may be the PC itself or an external source) is fed

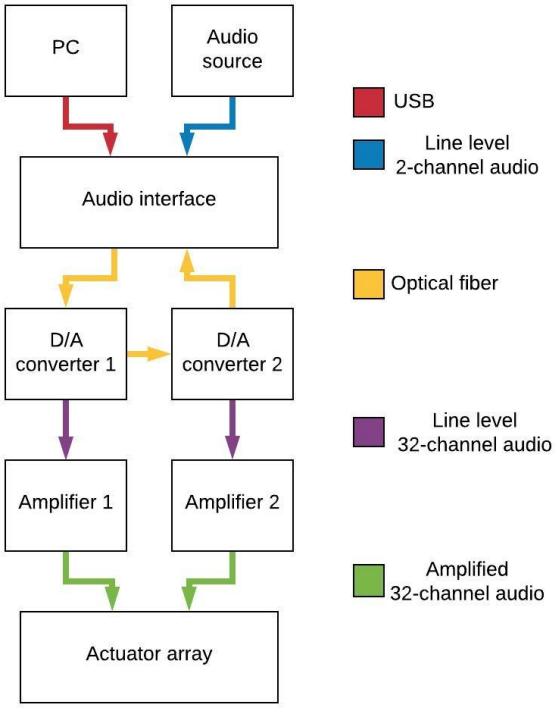


Figure 3. System flowchart. The in-lab rack system receives a stereo signal from an audio source and delivers 64 independent audio channels to a tactile display.

to the MADIface TX via two 6.3mm / $1/4$" TRS connectors. Software on the PC then analyzes the signal, filters it via bandpass filters and routes it according to the selected mapping, to the outputs of the TotalMix FX digital real-time mixer included in the MADIface TX interface. Each output is thus assigned a selected pass-band.

3.2 Digital-to-analog conversion

The 64 channels of audio now need to be converted back to analog as digital audio will not affect the voice coils. The system uses two Ferrofish A32 A/D-D/A converters. The A32 has a MADI I/O port and 32 analog outputs via four Dsub25 / TASCAM connectors. By daisy chaining the two converters, TotalMix FX can route to both converters thus getting the full 64 channels of analog audio.

Each MADI I/O port induces slight latency (3 samples), which could cause the output signals of the two devices to be incoherent. The A32 offers the option of defining the device as primary or secondary in the setup menu. By setting the first A32 as primary and the second A32 as secondary, both devices will be in sync with the same delay. Further latency can be expected due to the D/A conversion process, depending on the sample rate (0.1625ms at 48kHz, 0.05625ms at 96kHz and 0.034375ms at 192kHz). Although the overall latency is minimal, it can be made up for by delaying the playback audio accordingly, if necessary.

Figure 4. Lofelt L5 actuator. When flat side is put up against the body it results in parallel vibration on the skin [11].

3.3 Amplification

To effectively drive the voice coil actuator (L5 from Lofelt GmbH, see section 3.4), the line-level signal needs roughly a 20dB gain. A custom made 32-channel low-power audio amplifier (E32) for this purpose is currently under development by the authors. The in-lab experimentation rack system incorporates two E32's. Each amplifier will receive audio signals from a Ferrofish A32 via four Dsub25 / TASCAM connectors and deliver amplified signals to the ouputs via 32 RCA / phono plugs. Combined, the two E32's will thus deliver more than enough audio channels for the 6×10 array. The E32 is housed in a 2U 19" rack mount enclosure.

The E32 operates with the Texas Instruments TPA2025D1 integrated circuit (IC) amplifier. TPA2025D1 is a Class-D low-power audio amplifier IC that delivers 1.9W of power to an 8Ω load at a 3.6V supply (1% THD+N) and has a fixed gain of 20dB. With an efficiency of around 85%, heat dissipation is minimal which will be an advantage when implementing it to a wearable solution. It comes in a 1.53×1.98mm space saving 12-ball BGA package (0.5mm pitch).

The TPA2025D1 has a built in battery tracking automatic gain control (AGC) which limits battery current consumption and extends the battery life. This will also be of benefit when implementing to a wearable solution. For the moment, however, the TPA2025D1 IC's will be powered by the E32's 3.6V / 2A power supply.

3.4 Actuators

L5 actuators (see figure 4) were chosen for the new tactile display. Designed by the Lofelt GmbH company, the L5 actuator is a voice coil packaged in a plastic case, 17×20.5mm in length and width and 6.2mm in height. It can be driven by standard low-power audio amplifiers as an 8Ω load and has an acceleration response of a minimum of 0.5G for the frequency range of 35Hz to 1kHz. The L5 draws an average current of 57mA at maximum volume which means that the E32 will be able to supply enough current with all of its outputs in active mode.

By applying an audio signal to the voice coil, an alternating current (AC) runs through the winding, producing a magnetic field and interacting with a permanent magnet in the middle, pushing it either up or down (referring to figure 4). Every time the AC changes polarity the permanent magnet is pushed in the other direction, creating a vibration at the same frequency as the audio signal. Putting the L5's flat side up against the body will therefore result in a parallel vibration on the skin.

The ACUTE tactile display is estimated to contain 60 L5 actuators in an array of 6×10 like the SoV Tactile Belt used as frame of reference. However, experimentation will be conducted concerning the number of voice coils necessary for the full effect. In experimenting, various numbers of L5 coils are mounted on approximately 15cm thick layer of foam sponge which straps around the waist. A similar device (referred to as vibro-sponge) was used to measure relative vibrotactile spacial acuity in the SoV project [4,5]. Each L5 has an audio cable pair connected to it through pin sockets soldered to the coil wires and each pair has an RCA connector soldered to the other end. The cables form a tail with numbered RCA connectors that can be connected to the E32 amplifier outputs in a sequence befitting the vibro-sponge array.

By altering the software, adjusting the mapping and possibly changing the frequency bands and number of band-pass filters, many possible combinations of tactile feedback to represent music are available with the in-lab experimentation rack system. With a total of 64 independent channels, tactile stimulation can be presented to various body parts at once. For example, a 6×8 array could be placed on the torso while a strip of 8 coils can be placed on each arm.

4. CONCLUSION

We have presented the ACUTE Multi-CaTS proof-of-concept system, an in-lab rack mount solution for the purpose of evaluating different methods of audio-tactile mapping. The system will represent music using tactile stimulation to enhance the musical enjoyment of cochlear implant users. It will be used while experimenting with audio-tactile mapping as well as the number and placement of actuators. We will resume work, aiming towards a compact and wearable solution.

Acknowledgments

The work of the first author was funded by the Icelandic Technology Fund (Project number: 176713-0611) and the work of the second author was financed by the NorForsk-funded Nordic Sound and Music Computing Network (Project number: 86892).

5. REFERENCES

[1] C. J. Limb, "Cochlear implant-mediated perception of music," *Current Opinion in Otolaryngology and Head and Neck Surgery*, vol. 14, no. 5, pp. 337–340, 2006.

[2] I. Jóhannesson, R. Hoffmann, V. V. Valgeirsdóttir, R. Unnþórsson, A. Moldoveanu, and Kristjánsson, "Relative vibrotactile spatial acuity of the torso," *Experimental Brain Research*, vol. 235, no. 11, pp. 3505–3515, 2017.

[3] I. Jóhannesson, O. Balan, R. Unnthorsson, A. Moldoveanu, and Kristjánsson, "The sound of vision project: On the feasibility of an audio-haptic representation of the environment, for the visually impaired," *Brain Sciences*, vol. 6, no. 3, pp. 1–20, 2016.

[4] R. Hoffmann, V. V. Valgeirsdóttir, I. Jóhannesson, R. Unnthorsson, and Kristjánsson, "Measuring relative vibrotactile spatial acuity: effects of tactor type, anchor points and tactile anisotropy," *Experimental Brain Research*, vol. 236, no. 12, pp. 3405–3416, 2018. [Online]. Available: http://dx.doi.org/10.1007/s00221-018-5387-z

[5] R. Hoffmann, M. Brinkhuis, R. Unnþórsson, and Kristjánsson, "The Intensity Order Illusion: Temporal Order of Different Vibrotactile Intensity Causes Systematic Localization Errors," *Journal of Neurophysiology*, 2019.

[6] "Cochlear Implants | California State University, Northridge." [Online]. Available: https://www.csun.edu/ncod/cochlear-implants

[7] G. Cardon and A. Sharma, "Somatosensory Cross-Modal Reorganization in Children With Cochlear Implants," *Frontiers of Neuroscience*, vol. 13, no. June, 2019.

[8] M. Karam, F. A. Russo, and D. I. Fels, "Designing the model human cochlea: An ambient crossmodal audio-tactile display," *IEEE Transactions on Haptics*, vol. 2, no. 3, pp. 160–169, 2009.

[9] D. Eagleman, "Plenary talks: A vibrotactile sensory substitution device for the deaf and profoundly hearing impaired," *2014 IEEE Haptics Symposium (HAPTICS)*, no. Ci, pp. xvii–xvii, 2014. [Online]. Available: http://ieeexplore.ieee.org/document/6775419/

[10] J. B. van Erp, "Presenting directions with a vibrotactile torso display," *Ergonomics*, vol. 48, no. 3, pp. 302–313, 2005.

[11] "L5 Actuator." [Online]. Available: https://lofelt.com/technology

STUDENT INVOLVEMENT IN SOUND AND MUSIC COMPUTING RESEARCH: CURRENT PRACTICES AT KTH AND KMH

Kjetil Falkenberg Hansen
Roberto Bresin
Andre Holzapfel
Sandra Pauletto
KTH Royal Institute of Technology
Sound and Music Computing Group
kjetil@kth.se

Torbjörn Gulz[1]
Hans Lindetorp[2]
Olof Misgeld[3]
Mattias Sköld[4]
KMH Royal College of Music
[1]Department of Jazz
[2]Department of Music and Media Production
[3]Department of Folk Music
[4]Dept. of Composition, Conducting and Music Theory

ABSTRACT

To engage students in and beyond course activities has been a working practice both at KTH Sound and Music Computing group and at KMH Royal College of Music since many years. This paper collects experiences of involving students in research conducted within the two institutions.

We describe how students attending our courses are given the possibility to be involved in our research activities, and we argue that their involvement both contributes to develop new research and benefits the students in the short and long term. Among the assignments, activities, and tasks we offer in our education programs are pilot experiments, prototype development, public exhibitions, performing, composing, data collection, analysis challenges, and bachelor and master thesis projects that lead to academic publications.

1. INTRODUCTION

At both KTH Royal Institute of technology and KMH Royal College of Music we conduct teaching and research in Sound and Music Computing, and to engage students in research tasks is at the centre of the offered learning activities. Undergraduate research participation has been a pedagogical practice at KTH since many years back, and similarly, with the recent increase in artistic-based research at KMH, the foundation for constructively engaging students of music in research is well established. In this paper we aim to argue for why this is a practice that benefits both under- and postgraduate students as well as faculty members, and how we plan to continue developing our pedagogical approach.

There are well-documented benefits of including undergraduate students in research, both for enhancing research skills [1], general study skills [2], and building a community [3]. Thus, we know that students' participation in research activities is rewarding. Many of our courses have project work components, and the description and instructions which set the framework for these are typically inspired by how research projects are normally carried out. Roberts and Allen [4] measured research participation cost-benefit ratio along seven items, and they found that the students' perceived educational value of research participation was higher than the drawbacks or costs they experienced in research participation. The most prominent factor was that this participation complements teaching, while the expected influence on own future research was lower.

Several research tasks ca be expensive in terms of workload, for instance data collection in listening experiments, and students can often take on such tasks efficiently and with a positive effect on learning. However, many studies show the potential risks of student research involvement. Considering that a major driver for the student is that research experience complements teaching, it has been argued [5] that the ambiguous role of the teacher/researcher can confuse the student with regards to divergent interests in learning and in researching.

To let students carry out a faculty researcher's study can potentially lead to ethical problems: First, there are few standards or guidelines for involving students to do research, and even participate as research subjects [6]. Second, the power differential between faculty and students makes students a vulnerable population and research activities should not be placed on them just because of them being easily accessible [7]. Also, power inequities are known to cause the stronger to act in the best interest of the weaker, which can harm the trust relationship and thus the quality of the work [8]. A more tenable approach may be to encourage and allow students to formulate and conduct research with a higher degree of independence. Still, senior researchers have the important role of warning students to avoid common pitfalls, given the short time available for many of the projects.

It is quite common that postgraduates supervise undergraduates, and a study on research-based teaching activities presents five tensions between faculty, postgraduate and undergraduate students [9]. Of particular interest is that while undergraduates can increase research productiv-

ity, they also require more time for training, and postgraduates vary in their mentoring abilities. Furthermore, the undergraduate and postgraduate relationship was found to generate hierarchical structures that may have a negative effect on their work environment. As such, postgraduates' involvement should be considered carefully.

Even if undergraduates need more time for mentoring and training, they are an important asset. To attract students to participate and engage in research activities, we need to advertise their involvement: lack of visibility has been identified as a barrier to student participation, both in terms of awareness of ongoing research, and of results dissemination [10]. As incentive to motivate students to participate to research projects with both advise and assist students in submitting their works for publication and presentation in international conferences, such as the SMC conference, the Nordic SMC Conference and SMC Sweden. [1] .

Interestingly, teachers with larger course loads are slightly more likely to engage in undergraduate research than their colleagues [11]. This contrasts the above mentioned potential problems and risks and indicates that even for faculty, the benefits of having students active in research are greater than the perceived costs.

In the following, we describe the backgrounds of KTH and KMH in terms of teaching and researching in sound and music computing, and we provide some examples of current and recent undergraduate work.

2. BACKGROUND

In 2016, an agreement between KTH and KMH was signed that declared the intention to intensify collaboration in both research and education. Since then, external funding was obtained to develop a framework for master level work, including thesis, in the wider area of music technology, which aims at the conduction projects in collaboration with companies and organisations involving students from both KTH and KMH. Recently (June 2019), the research centre NAVET [2] was established, with the aim of exploring new research direction in the intersectional area Art, Technology and Design. This research centre involves currently four of the main higher education institutions in Stockholm, among them KTH and KMH. This dense network of established collaboration creates the environment to further develop strategies to combine research and education at KMH and KTH.

The Sound and Music Computing research group at KTH originates from the Speech, Music and Hearing lab. The Music Acoustics group, led by Johan Sundberg, started early to publish master theses in the Quarterly Progress and Status Report series, for instance by Erik Jansson in 1966. [3] This started a long tradition within the field of SMC at KTH, currently lead by the SMC group [4] , characterized by an ongoing exchange of research activity between undergraduate students and faculty.

At KMH, research has become an increasingly important part of the faculty and students' activities. This was further emphasised by the move into the new campus in 2016, where an advanced digital infrastructure and a concert hall with a loudspeaker dome with up to 45 speakers were built with music research in mind. Today, research at KMH is conducted in a variety of music-related subjects such as artistic research, music education, music and health and music technology.

The joint organization of the conference *Music and and Music Science* conference in 2004 [5] , chaired by Johan Sundberg and Bill Brunson, made it clear that KTH and KMH had to collaborate in both research and teaching. In the current close collaboration between KMH and KTH, in particular with four KMH teachers also enrolled at KTH as doctoral students, we have activities across the institutions. These four teachers cover the topics of music production, jazz and folk music performance, as well as composition. Their simultaneous affiliation as researchers and teachers enables them to integrate projects of their research into the teaching practice at KMH. This way, students are introduced within their artistic education to research practice in SMC, which diminishes observed or forseeable barriers between the different research and academic traditions.

3. ACTIVITIES

In the following we present a selection of examples from courses, degree projects, exhibitions, and research projects where undergraduate students are involved with research. Typically, the results are disseminated publicly—either as presentations, publications, concerts, or exhibitions.

3.1 Course activities at KTH

In the framework of the *DT2213 Musical Communication and Music Technology* course (7.5 ECTS, second cycle), students work in a small project as one part of the course. In previous years, this study was suggested and designed by the students, but since 2019 all students get the same research challenge and are divided into groups based on competence matching. Course evaluations confirm that the new approach is more appreciated. To increase motivation, the study is set in a realistic environment and is disseminated through a public presentation, currently held at the Swedish Museum of Performing Arts. [6] In this way, the students can collectively contribute to ongoing research.

In spring 2019 a new bachelor course *DM1579/2579 Media Production* (6 ECTS first cycle/7.5 ECTS, second cycle) was developed at KTH. The main project work for this course, to be carried out in teams, is the development, production and postproduction of an interactive documentary [12]. Interactive documentaries are a new form of nonfiction narrative that uses at its core action, choice, and immersion to create storytelling. Through this project the students are challenged to engage with the most current

[1] http://smcsweden.se/

[2] https://www.kth.se/navet

[3] https://www.speech.kth.se/qpsr/show_by_author.php?author=Jansson%2C+E.

[4] https://www.kth.se/mid/research/smc

[5] http://www.speech.kth.se/music/music_and_music_science_2004.pdf

[6] https://scenkonstmuseet.se/

issues in media and sound production (non-linear storytelling, audience interaction, and object-based media). Although only one cohort so far has taken this new course, it is clear that some of these students productions will contribute to research in media and sound production. Additionally, through this course bachelor students come into contact with research as often they choose to select a research topic or a research team as the subject of their production.

Further examples of second-cycle courses that provide students with the possibilities to conduct research projects as part of the course examination are *DT2300 Sound in Interaction*, *DM2350 Human Perception for Information Technology*, and *DT2470 Music Informatics* (all 7.5 ECTS, second cycle). In all these cases, students are provided with a list of project proposals to choose from, which they then develop, supported by faculty staff, by shaping a detailed project plan, implementing this plan, and documenting the outcome in form of public project presentations that resemble settings of scientific conferences. In the current version of the *DT2300 Sound in Interaction* course, students work on a theme; for example for 2020 the theme is *The soundscape of the future*, connected to one of the current research projects run by the SMC group at KTH. In this and other courses students can work with technology such as permanent installations [13], sensor technology (i.e. motion capture systems), 3D printing, and laser cutting.

In the *DT22157DT2216 Advanced Individual Course in Music Communication* courses (6.0 and 9.0 ECTS, second cycle) students can work on individual research projects assigned by researchers and that are usually associated to ongoing research projects. Often this research results in international publications (see for example [14, 15].

In *DM2799 Advanced Project Course in Media Technology* (7.5 ECTS, second cycle), student groups choose and work on a project suggested and supervised by faculty researchers and teachers. With this approach, faculty have an outspoken interest in proposing good projects as they lead to results relevant for research and possibly attracting further engagement in Master's theses.

In summary, the above mentioned courses create a large diversity of research areas in which the students are stimulated to integrate their learning process with scientifically novel research projects. These areas comprise among other, the design of new music instruments and interfaces, sonification, music perception, expressive body motion, music performance, and music information retrieval. With this background, students come well prepared to do a rewarding master thesis work.

3.2 Course activities at KMH

A large part of the research conducted by doctoral students enrolled in the KTH/KMH joint programme could be applied within the area of educational tools. Such tools and methods can be used to develop the learning environment within KMH. In *DG1067 Sonology and Studio Technology* at KMH, composition students have the last couple of years been involved in working with experimental music notation providing data for research in composition and analysis. While their participation was rather as test groups than performing the actual research, the fact that the notation system they used was the subject of current research made them more invested in the work than previous students, who had been provided with a complete and unchanging system (see [16] for more information).

Since 2013 all music production students at KMH are introduced to interactive music production. Students in three courses—*EG1005 Media Production*, *EG1043 Media Production 3*, and *EG1073 Media Production 4*—have been a part of an iterative development process of the interactive music playback framework *iMusic* which is written and maintained by teacher Hans Lindetorp. [7] Artistic ideas expressed by the students have been the driving factor for new features and the framework and documentation have been tested in projects run by the students.

The Sound Forest [13] is a large-scale Digital Musical Instrument (DMI) situated at the Swedish Museum of Performing Arts. It was designed by researchers from KTH and the experiments on its first prototype were conducted in the framework of a master thesis presented at SMC 2016 [17]. Sound Forest has been used as an experimental platform for several interactive music productions by master students from KMH. The students combine working on their own research questions with participating as informants for research conducted by PhD students at KTH [18].

In the course *EA2001 Music Production 1* during 2014–2016, master students from KMH composed interactive musical soundscapes for the Nobel Price Museum. The exhibition was a multidisciplinary, artistic interpretation of the different prices where students from different artistic schools collaborated to make an interactive installation with audio, video, fashion design, product design, photography and textile design. The project was documented in an article [19].

3.3 Examples of joint course activities

In *DM2350 Human Perception for Information Technology* at KTH, ten students from the jazz department and ten students from the folk-music department at KMH took part in a study on tempo perception in music. The study was planned and conducted by students from the Interactive Media technology master programme at KTH. In addition to results useful in ongoing research, the project also gives students from different school environments an opportunity to create inspiring contact areas for possible later projects. The experimental setting and the measurable results is a new experience for the KMH students, and the KTH students encounter situations and challenges to solve that are hard to replicate in a KTH setting. The experiment is based on identifying the pulse in 20 jazz samples of varying difficulty and identifying the pulse in 20 samples from Swedish folk music by a tapping exercise. These kind of experiments are addressing central themes in Folk Music Theory and the results can also be relevant to course content development at KMH.

[7] `https://github.com/hanslindetorp/imusic`

In the research on strategies in jazz improvisation, for instance, students from the jazz institution have participated in pilot studies. Apart from the main work that is developing methods, the experiment also helped the students to discover parts of their improvisational ability. They also had the opportunity to reflect on their playing together with a senior teacher. One of the contributing reasons for making the strategies more visible was the technical analysis tools that were being developed and tested within KTH [20].

In the research on the interaction between dancers and musicians in Swedish Folk music students from the folk music department and from the folk dance education at DOCH, School of Dance and Circus at Uniarts participated in Motion Capture pilot studies in the PMIL Performance and Multimodal Interaction Lab at KTH. These experiments relate closely to how these interactions are addressed within the educations and invites students to reflect on research strategies and their own practice.

3.4 Bachelor and Master degree projects

Bachelor and Master's projects at KTH are typically conducted within one semester (15 and 30 ECTS, respectively), with the students making the choice of their thesis subject. Within the SMC team, we provide a list of project topics for the students to choose from, but also encourage them to reshape these or even formulate their own research topic. This process of deciding on the thesis subject is well-supported through the project experiences the students had in previous courses.

In recent years, the outcomes of Master and Bachelor projects are published in a thesis that follows the ACM format with a maximum length of 10 pages. This way, students are prepared and encouraged to use the final thesis as a basis for the submission to a peer-reviewed conference. Before, when the format of the written report was a more traditional one (40 to 100 pages long), the ratio of theses submitted to conferences and the total number of presented theses was usually low. However, we believe that connecting Master theses to ongoing research—both by proposing relevant subjects and by using formats commonly used in research—creates an environment that immediately supports and mutually benefits research and education. Moreover, the conference submissions that actually emerged led to an exposure of the student to a wider community, enabling students to profit from their thesis projects in ways that are relevant for their future employment.

Examples of a master theses that resulted in a peer-reviewed research publication include the work by Emma Frid [21] (best paper award at ICMC - SMC 2014) on the design of tactile display for a live-electronics notification system and that of Lilia Jap [22] who implemented an interactive system for dancers, in which body movements of dancers are tracked in real-time and the playback speed of the audio track is adjusted in interaction with these measurements. In the area of artificial intelligence, another master thesis focused on the analysis of broadcast audio content using deep neural networks, and the work was presented at ISMIR 2019 [23].

Another example was a collaboration between master student Ludvig Elblaus and opera composer and singer Carl Unander-Scharin who was teaching at the University College of Opera. Elblaus' master thesis [24] was awarded best thesis in Sweden in acoustics, published in the SMC conference [25], and invited to Journal of New Music Research [26]. Following this, both Elblaus and Unander-Scharin have completed post-graduate studies at KTH [27, 28].

The Bachelor thesis projects conducted at KTH employ a similar framework as the Master thesis projects. However, the goal of the Bachelor projects is regarded rather as preparation for more elaborate work within a Master thesis, and as a first structured approach to planning, conducting, and presenting research. Despite the limited experience of the Bachelor students, the exposure to a framework that combines research and education has resulted in a number of theses that were presented publicly within conferences and exhibitions [29,30], including accepted submissions to this year's NordicSMC [31, 32].

At the heart of the thesis work for music students at KMH is the artistic output. Concerts and scores form essential and necessary parts of their work. But reflective and exploratory work in dialogue with current research has become increasingly important over the past 10 years in the theses. A trend observed among composer students is that of continuing towards a PhD degree after their master. The students themselves decide which aspects of their artistic work they write their theses on. While some describe their creative process in close dialogue with the artistic work, others bring up topics at a more general level though still in relation to their own artistic activities in a form of inside perspective.

The integration of the different approaches to Master thesis work is a challenge in the current collaboration between KMH and KTH, which we approach within the ongoing development of a common framework outlined above. One strategy is the formulation of Master thesis studies at KTH that promote close collaboration with KMH students in their longer projects.

4. DISCUSSION

While it is common that students formulate research questions themselves and design the studies without constraining supervision, an approach where the study is already defined by the teacher has proven to be both efficient and appreciated. Among the reactions in evaluations were that less time was wasted on agreeing on an idea and that they could instead explore a solution space within the framed proposal.

Another finding from course evaluations is that students are in general very positive to be included in ongoing projects and current research as compared to more traditional course work. The tasks become significant, and motivation to study becomes intrinsic. We believe that research in sound and music computing is particularly applicable because most, if not all students, have a personal relationship to the field, it is more intuitive to grasp, and because of its multidisciplinarity involving different areas such as multimodality,

physical interaction, design, computer science, HCI, perception, and musicology.

With the above illustrated variety of courses with research experience activities included in the intended learning outcomes and course goals, we have as a group built up a strong foundation for undergraduate involvement. Thus, we can in our daily teaching competently offer different opportunities and appropriate challenges for being engaged in research—both in projects with freedom and with more steered tasks.

To conclude, we list the benefits we consider this pedagogical approach will give students in their learning, and faculty in research. These benefits together form our common strategy for including KTH and KMH students in such activities.

4.1 Benefits for students

Students that have partaken in research activities get a first hand experience of how research work can be conducted and therefore a better understanding of what an academic does—something they might then decide to pursue themselves. They learn both to produce results and report these in a format that can lead to publications, that they can add to their portfolio, CV and list of merits for future careers. Naturally, the needs for tutoring the writing process vary depending on a range of factors, but a well-proven and rewarding method for supporting student independence is peer-instruction (among students), which is widely used in our courses.

In terms of the learning outcome, it is known that intrinsic motivation is better for deep learning than grades; it has also been found that a moderate extrinsic motivation enhances the effect of the intrinsic motivation [33, 34]. With regards to our course activities, we typically let the examination take the form of an academic report and a project presentation. When students feel they own the project and do a meaningful task, intrinsic motivation increases, and because a published paper gets more visibility than grades, even the extrinsic motivation increases.

4.2 Benefits for research

The group's quantitative research output in the form of peer-reviewed publications cannot be considered to have increased through the involvement of students in research projects. A setting that would aim at this aspect would have senior researchers design experiments and compile the written reports, and students supporting data collection. But, as detailed in the introduction section above, such an approach would be in conflict with various ethical considerations. Instead, we see the major benefit for research in the increased diversity of the research outcomes. Since students are provided the freedom to shape their projects under guidance by experienced researchers, new lines of research are established that go beyond the main research lines of the senior researchers.

4.3 Authorship and ownership

Students and researchers at KTH are protected by policies for intellectual properties (IP). Regulations state that *the basic assumption is that rights to IP created as a result of teaching and research shall vest in the inventor. Inventors can be students as well as employees at KTH.* [8] , and that *if you proposed or own the assignment for the degree project and nothing else was agreed, you own the intellectual property rights that you generate during the assignment.* [9]

In a similar way, the students own the copyrights to all music they create in course-based projects. Therefore it is important to sign agreements on how the music can be used outside the projects, or issue the works with an appropriate licence such as Creative Commons. [10] This involves contracts between students and venues to possibly continue to use the music after a project is finished. It also applies agreements between students and researchers for publishing audio and video resources as references from publications. One positive side effect for the students is that they learn how to handle licences, intellectual property, and contracts as a natural part of their education.

4.4 Strategy

In our dual roles as teachers and researchers, we are aware of the power inequality with regards to students taking our courses and taking part in our ongoing research projects. We intend to continue our tradition of promoting possibilities and providing means for exposure to our academic communities, where students are encouraged to design studies, conduct experiments, and disseminate results.

We need to ensure that course activities are always designed aiming at their intended learning outcomes and supporting incremental development of students' curricula. Researchers should keep within the realm where students can make inspiring learning experiences instead of meeting a closed office door. We must make students aware that their creative and theoretical outputs can substantially contribute to research. In summary we believe that our efforts in engaging students in our research at different stages of their studies have produced very positive results of which both students and researchers are having benefit.

5. REFERENCES

[1] C. M. Kardash, "Evaluation of undergraduate research experience: Perceptions of undergraduate interns and their faculty mentors." *Journal of Educational Psychology*, vol. 92, no. 1, pp. 191–201, 2000.

[2] R. E. Landrum and L. R. Nelsen, "The undergraduate research assistantship: An analysis of the benefits," *Teaching of Psychology*, vol. 29, no. 1, pp. 15–19, jan 2002.

[8] Policy for management of intellectual property created at KTH

[9] https://www.kth.se/en/samverkan/exjobb/studenter/riktlinjer-1.293804

[10] https://creativecommons.org/

[3] H. Thiry and S. L. Laursen, "The role of student-advisor interactions in apprenticing undergraduate researchers into a scientific community of practice," *Journal of Science Education and Technology*, vol. 20, no. 6, pp. 771–784, jan 2011.

[4] L. D. Roberts and P. J. Allen, "A brief measure of student perceptions of the educational value of research participation," *Australian Journal of Psychology*, vol. 65, no. 1, pp. 22–29, feb 2013.

[5] T. Wulf-Andersen, P. Hjort-Madsen, and K. H. Mogensen, "Research learning – how students and researchers learn from collaborative research," in *The Roskilde Model: Problem-Oriented Learning and Project Work*. Springer International Publishing, 2015, pp. 211–231.

[6] A. F. Leentjens and J. L. Levenson, "Ethical issues concerning the recruitment of university students as research subjects," *Journal of Psychosomatic Research*, vol. 75, no. 4, pp. 394–398, oct 2013.

[7] M. Cleary, G. Walter, and D. Jackson, "Editorial," *Contemporary Nurse*, vol. 49, no. 1, pp. 93–95, dec 2014.

[8] L. M. Ferguson, O. Yonge, and F. Myrick, "Students' involvement in faculty research: Ethical and methodological issues," *International Journal of Qualitative Methods*, vol. 3, no. 4, pp. 56–68, dec 2004.

[9] E. L. Dolan and D. Johnson, "The undergraduate–postgraduate–faculty triad: Unique functions and tensions associated with undergraduate research experiences at research universities," *CBE—Life Sciences Education*, vol. 9, no. 4, pp. 543–553, dec 2010.

[10] H. A. Wayment and K. L. Dickson, "Increasing student participation in undergraduate research benefits students, faculty, and department," *Teaching of Psychology*, vol. 35, no. 3, pp. 194–197, jul 2008.

[11] K. L. Webber, T. F. N. Laird, and A. M. BrckaLorenz, "Student and faculty member engagement in undergraduate research," *Research in Higher Education*, vol. 54, no. 2, pp. 227–249, dec 2012.

[12] J. Aston and S. Gaudenzi, "Interactive documentary: setting the field," *Studies in Documentary Film*, vol. 6, no. 2, pp. 125–139, 2012.

[13] R. Bresin, L. Elblaus, E. Frid, F. Favero, L. Annersten, D. Berner, and F. Morreale, "Sound Forest/Ljudskogen: A large-scale string-based interactive musical instrument," in *Sound and Music Computing*, 2016, pp. 79–84.

[14] A. Elowsson, R. Schön, M. Höglund, E. Zea, and A. Friberg, "Estimation of vocal duration in monaural mixtures," in *ICMC*, 2014.

[15] X. Han and R. Bresin, "Performance of piano trills: effects of hands, fingers, notes and emotions," in *Proceedings of the 1st Nordic Sound and Music Computing Conference*. KTH Royal Institute of Technology, 2019.

[16] M. Sköld, "Combining sound-and pitch-based notation for teaching and composition," in *TENOR'18–Fourth International Conference on Technologies for Music Notation and Representation*, 2018, pp. 1–6.

[17] J. Paloranta, A. Lundström, L. Elblaus, R. Bresin, and E. Frid, "Interaction with a large sized augmented string instrument intended for a public setting," in *Sound & Music Computing Conference*. SMC Sound & Music Computing Network, 2016, pp. 388–395.

[18] E. Frid, H. Lindetorp, K. F. Hansen, L. Elblaus, and R. Bresin, "Sound Forest – evaluation of an accessible multisensory music installation," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI 19*. ACM Press, 2019, article ID 677.

[19] J.-O. Gullö, I. Höglund, J. Jonas, H. Lindetorp, A. Näslund, J. Persson, and P. Schyborger, "Nobel creations : Producing infinite music for an exhibition," *Dansk Musikforskning Online*, pp. 63–80, 2015. [Online]. Available: http://www.danishmusicologyonline.dk/arkiv/arkiv_dmo/dmo_saernummer_2015/dmo_saernummer_2015_lyd_musikproduktion_04.pdf

[20] T. Gulz, A. Holzapfel, and A. Friberg, "Developing a method for identifying improvisation strategies in jazz duos," in *Proc. of the 14th International Symposium on CMMR :*, 2019, pp. 482–489, qC 20191029. [Online]. Available: https://cmmr2019.prism.cnrs.fr/

[21] E. Frid, M. Giordano, M. M. Schumacher, and M. M. Wanderley, "Physical and perceptual characterization of a tactile display for a live-electronics notification system," in *ICMC SMC 2014*. McGill University, 2014.

[22] L. Jap and A. Holzapfel, "Real-time mapping of periodic dance movements to control tempo in electronic dance music," in *Sound & Music Computing Conference*. Zenodo, May 2019.

[23] Q. Lemaire and A. Holzapfel, "Temporal convolutional networks for speech and music detection in radio broadcast," in *Proceedings of ISMIR*, 2019, in print.

[24] L. Elblaus, "Exploring a design space by prototyping a gesture-controlled augmentation of an operatic lead part," Master's thesis, KTH Computer Science and Communication, Stockholm, 2012.

[25] L. Elblaus, K. F. Hansen, and C. Unander-Scharin, "Exploring the design space: Prototyping "the throat v3" for the elephant man opera," in *Proceedings of the Sound and Music Computing Conference*, S. Zanolla, F. Avanzini, S. Canazza, and A. de Götzen, Eds. Padova, Italy: Padova University Press, jul 2011, pp. 141–147.

[26] ——, "Artistically directed prototyping in development and in practice," *Journal of New Music Research*, vol. 41, no. 4, pp. 377–387, 2012.

[27] L. Elblaus, "Crafting experience : Designing digital musical instruments for long-term use in artistic practice," Ph.D. dissertation, KTH, Media Technology and Interaction Design, MID, 2018, qC 20180507.

[28] C. Unander-Scharin, "Extending opera - artist-led explorations in operatic practice through interactivity and electronics," Ph.D. dissertation, KTH, Media Technology and Interaction Design, MID, 2015, qC 20150119.

[29] P. Mattei, S. Stolica, and K. F. Hansen, "SoundCubes: Providing new stimulating auditory training for people with hearing impairments," in *Sound and Music Computing Conference Sweden*, 2017.

[30] M. Herrera, G. Schierbeck, and K. F. Hansen, "Disadvantages of using non-linear video in shallow learning situations – a critical perspective on current trends," in *KTH Scholarship of Teaching and Learning*. Stockholm: KTH ECE, March 2017, p. 14. [Online]. Available: https://www.kth.se/polopoly_fs/1.712153!/20170314_abstracts_B.pdf

[31] M. Svahn and J. Hölling, "Rhythm as sensorimotor support for gait disturbance caused by neurological disease," in *Proceedings of the Nordic SMC*, 2019.

[32] M. L. Nilsson and J. Loor, "Bass as an indicator of quality: The relation between bass levels and quality perception in headphones," in *Proceedings of the Nordic SMC*, 2019.

[33] Y.-G. Lin, W. J. McKeachie, and Y. C. Kim, "College student intrinsic and/or extrinsic motivation and learning," *Learning and Individual Differences*, vol. 13, no. 3, pp. 251 – 258, 2003.

[34] K. Chamberlin, M. Yasué, and I.-C. A. Chiang, "The impact of grades on student motivation," *Active Learning in Higher Education*, p. 146978741881972, dec 2018.

Evaluation of Two Music Tactile Display Encodings for Cochlear Implant Recipients

Mathias Lyneborg Damgård
University of iceland
mld8@hi.is

Elvar Atli Ævarsson
University of Iceland
eae19@hi.is

Árni Kristjánsson
University of Iceland
ak@hi.is

Runar Unnthorsson
University of Iceland
runson@hi.is

ABSTRACT

This paper investigates the effect of using haptic feedback in conjunction with music using a tactile display worn as a belt. Studies have shown that Cochlear Implant (CI) users have limited ability to perceive music. This is partly due to the fact that CIs have been developed and optimized for speech and not for delivering the acoustical properties of music. The idea behind this work is to use tactile stimulation to enhance the musical experience of CI recipients. Using a tactile display from a previous project, music was encoded and mapped to the display to allow for more senses to be involved during the listening experience. To test if this had a positive impact, we compared two encodings with music filtered through a cochlear implant simulator and just the filtered music as control. Our results suggest that using tactile feedback can positively affect the experience of listening to music.

1. INTRODUCTION

Cochlear implants have improved speech recognition significantly [1]. However, living with a cochlear implant(CI) also presents a number of obstacles. This paper will discuss the impact CI's can have on the experience of listening to music, and a possible way of improving their listening experience, to increase the users' quality of life. CI's work very well for speech signals in quiet environments and for rhythm perception, but are severely lacking in other features such as timbre and pitch [1]. The focus of the CI processing strategy is often to remove temporal fine-structures and preserve temporal envelopes to improve speech perception. This unfortunately also removes much of the spectral information [2]. Even with further advancement in CI technology, obtaining a new implant is costly and as such, a different approach could be useful. Even though CI users face poor music perception studies have shown that there might not be a relationship between music perception and music enjoyment [1].

This initial study reports on the first steps taken towards the development of a tactile display representation, of music for cochlear implant recipients. It is partly based on previous work in the Sound of Vision project [3–6] where a tactile display in the form of a belt was developed with an array of 60 motors to assist navigation for vision impaired people. Specifically the belt is reused for this initial study. The belt (see fig 2) is reused to save development time and to quickly create an underlying foundation for launching other studies on new mappings, placement of tactors, choice of songs and so on.

Other haptic designs have been made to further the experience for the hearing impaired such as the the emotichair [7] and Eagleman's vibratory vest [8] but neither focus on tactile stimulation in conjunction with CI's.

The aim of this study was to gauge if introducing the somatosensory modality through tactors with music, would at all change the enjoyment of listening to music for CI users. Using the tactile display previously mentioned, two different ways of mapping music to the tactile display were developed. A cochlear implant simulator was used to filter music to stand in for actual CI users. The two mappings in conjunction with simulated CI music, were compared to each other and to the experience of just listening to the simulated version of the music. The result show a significant preference for the belt as opposed to just the simulated music by itself.

2. COCHLEAR IMPLANTS

As opposed to normal hearing aids, the cochlear implant is, as the name suggests, an implant. The implant has an electrode array surgically placed into the cochlea (see fig. 1). On the outside, a microphone picks up sound which a processor can then analyse to determine the level of stimulation each electrode is sent [9]. This information is sent via radio transmission. The implant then sends electrical signals to the electrodes placed in the cochlea. The signals activates the auditory nerves which is then projected to hearing related regions of the central nervous system, same as any normal hearing person. Many factors determine what the end experience is for a CI user. The most important one is the amount of channels the CI has. The implant divides the sound into channels which drive the aforementioned electrodes. Because of the size of the cochlea organ,

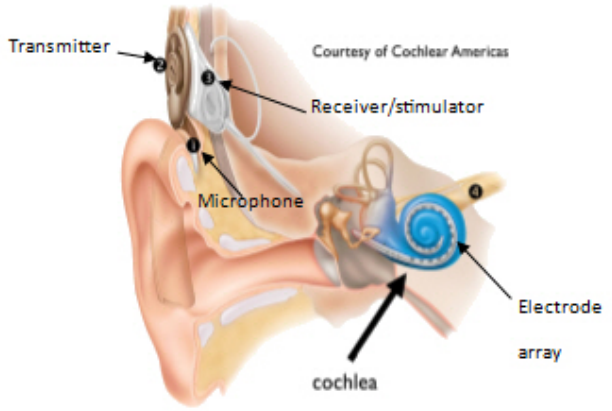


Figure 1. diagram of the hearing organ with CI [10]

the amount of electrodes and wires you can fit is limited. Since the sound is divided into a limited amount of channels a lot of fidelity is lost, especially in pitch perception. While this technology has enabled people to hear again, it leaves a lot to be desired in the way of music enjoyment.

2.1 CI simulation

It is very hard to determine the exact experience of a CI user. A multitude of factors are involved: The technological features of the implant itself, surgical factors, physiological factors of the user, factors concerning the users hearing impairment: duration, age of onset, auditory training and so on [9]. For accurate simulation; testimonies, indirect observations and analysis of audio signal transformations can be investigated. The Cochlear Implant Simulation(CIS) [9] uses these to approximate the listening experience of a CI user. For this initial study of the influence of haptic stimulation on music experience of CI users, we used this simulator with people with normal hearing, instead of testing actual CI users. There are other, more recent simulations available such as the SPIRAL [11, 12] but it was decided that the CIS was sufficient for this project, especially since it has not been shown whether the SPIRAL provides better music sound quality.

3. THE BELT

The tactile display that is used in this study was designed and built in the H2020 project Sound of Vision (no: 643636). It consists of an array of 60 motors in a 6*10 grid (see fig. 2). It is meant to be worn around your waist with the motors on your belly. It uses parallel rotating ERM motors that are controlled with an arduino microcontroller connected to a computer.

4. MAPPING MUSIC TO AN ARRAY OF TACTORS

The question of what kind of mapping provides the best experience is not an easy one to answer. Designing tactile encoding for good music appreciation is not trivial, as there are number of factors that need to be considered: focus on providing stimulation to differentiate pitches or focus on

Figure 2. Belt consisting of 60 motors

the beat and percussion, focus on low frequencies, mids, highs or the entire spectrum, which type of song do we use, is the beats per minute(BPM) and important factor and what hardware are we dealing with.

CI users can perceive rhythm and transients relatively well compared to pitch. With this in mind, a mapping strategy focused on the beat and percussion would not provide as much new information as a strategy focused on pitch. However, Prevoteau et al. [1] suggests making music more percussive and/or enhancing the lower frequency content may "enhance music enjoyment for CI users."

The type of music to test is also important. We could choose a song that fits a CI user more, such as a genre with a strong steady beat and a relatively simple melody without many layers. This could come from the electronic genre or perhaps pop music. The other direction is something that would be difficult for a CI user to listen to, such as a genre with a lot of layers and complexity [1, 13]. With a genre that doesn't fit a CI user, the addition of a new modality could have a strong effect on the enjoyment of music, however, in choosing a song that does not fit, the experience as a whole could be so bad even engaging other senses would not help it. Since we are taking our first steps in to the development, we chose to go with the more simple option and start with a song that fits CI users.

A 30 second sample of the chorus of Mika - Grace Kelly was chosen. Its beat and bass line is relatively steady. The vocal track is well known for its high variance in pitch which could help CI users detect the change in pitch.

Since CI users can perceive rhythm well, the BPM is not expected to be an important factor. The biggest influence of our choice here involves the limitations of the technology used for this particular initial study. Our base knowledge is also a limiting factor as some decisions require more information to be taken into account.

The hardware used is also an important factor. The tactors used in this study were simple vibrating motors. They vibrate at a certain frequency which changes with the load you put on them. As such, pressing them against a human body, for example, will change the output frequency. This makes it very hard to control. Given this difficulty the intensity was not a variable we decided to map for any of the encodings.

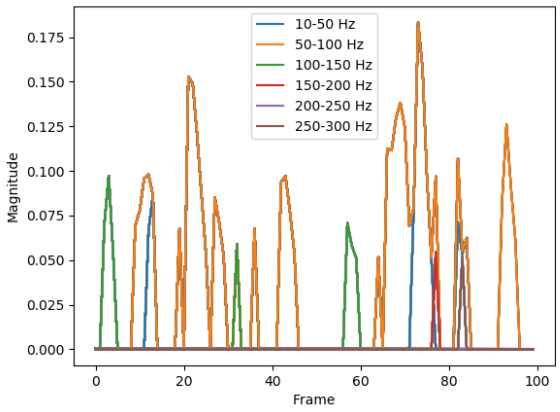


Figure 3. Graph of a 10 second sample after analysis and dividing the sample into bands of 50 Hz. Every frame a corresponding magnitude for a band is found and can be mapped to the belt.

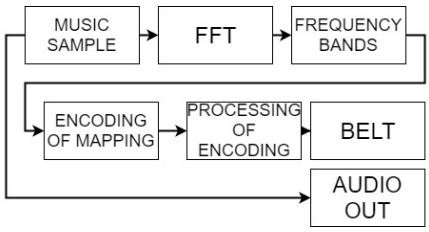


Figure 4. Flowchart diagram outlining the processing. A music sample is sent through an FFT and divided into frequency bands that are used to encode the mapping. This encoding is then processed and sent to the belt.

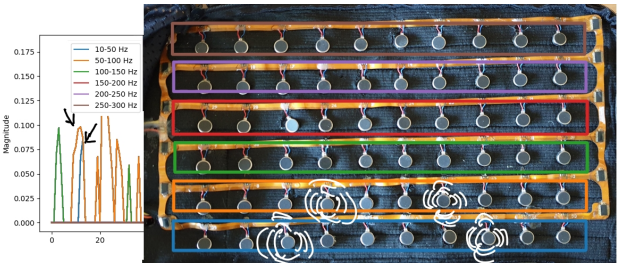


Figure 5. When the software detects activity in a certain band it activates according to the intensity. In this case the second highest in the 50-100 Hz range and third highest in the 10-50 Hz range.

4.1 Mappings

In order to transform music into data that the belt (see fig 2) can interpret, two strategies must be developed. One to determine the specific data that is sent to the belt, and the second being the encoding itself. For this, custom software was developed. The software is split into two parts; Encodings for the belt programmed from an analysis of the music sample, and the software controlling the belt that then executes the encoding. The software is written in python with an arduino sending signals to the belt itself.

As the belt was designed to provide navigational aid to blind people, and has a 6*10 array of motors tactile display, it is easy to draw inspiration for mappings from a visual standpoint. Since there has been very little research in this particular area, this initial study is also intended to set a reference point for future studies. New mappings, tactors and placements of tactors and so on will be compared to this study. Some of the software is the same for every different mapping; analysis of the music and the execution of the resulting encoding. What changes is the encoding. The analysis part of the code runs through a fast fourier transformed piece of music within a specific range of frequencies. If the magnitude of the signal is higher than a certain threshold, that magnitude is saved in a list. We do this for every frequency band. See fig (4). The software controlling the belt is coded to run through a simple txt file which consists the motors targeted by the encoding, their intensities and finally a symbol that denotes a pause. Intensity was a constant number throughout the mappings with the option for customization left in for future iterations. The algorithm will look through the current motors and intensities and send those signals to the arduino, if there is a symbol for a pause it will pause for 10 milliseconds and then look at the next range of motors and intensities. As an example, see fig (3). The file processed is a ten second long sample of jazz music featuring a double bass.

4.2 The EQ mapping

The first mapping considered divides the music into six bands from 0 to 300Hz, each band containing 50Hz. Each row of motors consists of a band of frequencies. The top row has the highest frequencies and the bottom one the lowest. Each row is also split up into two halves representing the left and right audio channel. The position on the row represents the magnitude of the frequency detected in that band, with the middle one the highest. The left and right extremes are the lowest (before 0) values. For example; if a frequency in the range of 50-100 was detected, the appropriate motors in the second last row would activate (see fig. 5).

4.3 The line mapping

The second mapping considered discards some complexity in favor of a clearer experience. Every row is divided into bands again, but the entire row activates if there is any activity in the band (see fig 6). Again the bands go from 10-50, 50-100 and so on up to 300.

5. EXPERIMENT

A 30 second sample of Mika - Grace Kelly that was processed through the CI simulator (see section 2.1) was used to test three treatments. One is the sample by itself, the second is the sample with the first encoding method (see section 4.2) and the third is the second encoding method (see 4.3). the UEQ test [14] was used with the goal oriented scales taken out. We focused on three subscales: Attractiveness, Stimulation and Novelty with the most important

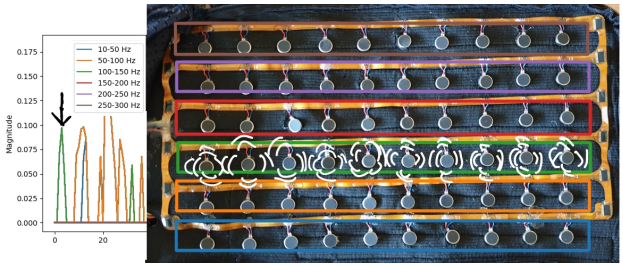


Figure 6. When the software detects activity in a certain band it activates according to the intensity. In this case the entire line of motors in the 100-150 Hz band.

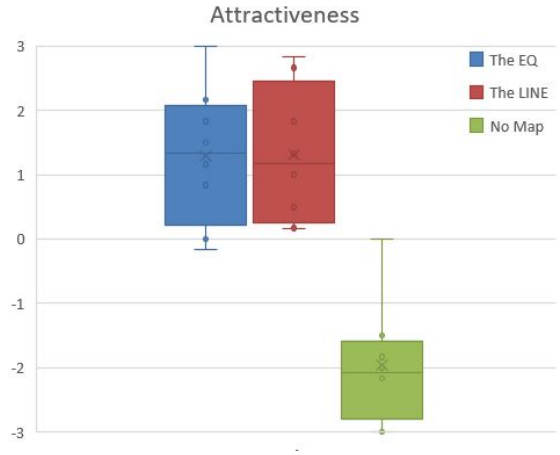


Figure 7. Boxplot showing the spread of data between the three conditions in the Attractiveness scale

being Attractiveness. The first and second treatment were randomized for every test subject with an assessment given after each treatment. Eight subjects were tested, 6 male and 2 female with a mean age of 33,25. The answers from the UEQ test were transformed from the original likert scale of 1-7 to a score between -3 to +3 where -3 represents the most negative value and +3 the most positive. An ANOVA test was used to test if there was a significant difference between the groups followed by a post hoc test to determine interaction between the treatments.

5.1 Results

There was a significant difference between the treatments on all three subscales. (see fig. 7,8 and 9)

The post-hoc analysis reveals that there was no significant difference between The EQ mapping and The LINE mapping.(t(14) = -0,039, P one-tail = 0,485) There was, however, a significant difference between the No Mapping and the EQ mapping (t(14) = 6,14, P one tail = 8,03E-06) and between No Mapping and the LINE mapping (t(14) = 6,513, P one tail = 6,86E-06).

6. DISCUSSION

Our results suggest that providing haptic stimulation in conjunction with music played through a cochlear implant simulator improves the musical experience of listeners. This is promising in terms of improving the experiences of CI

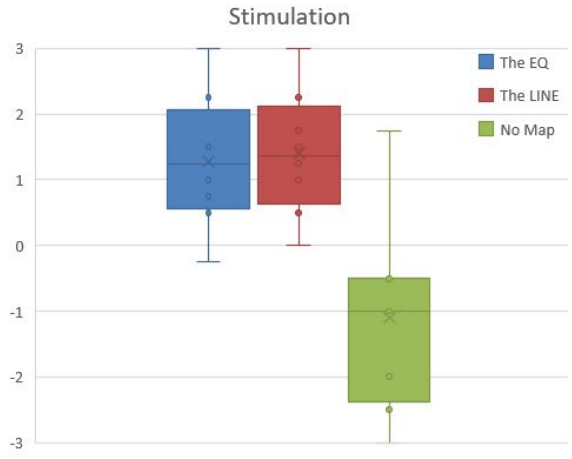


Figure 8. Boxplot showing the spread of data between the three conditions in the Stimulation scale scale

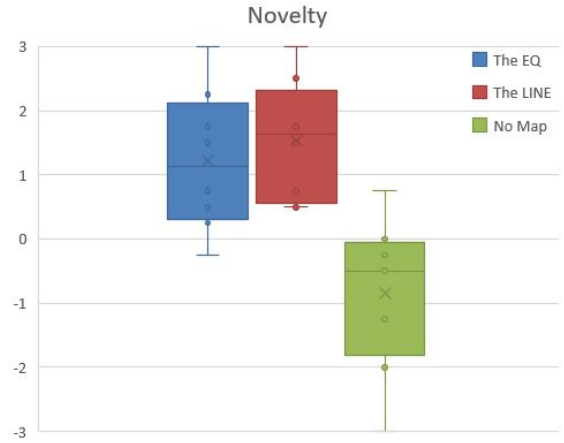


Figure 9. Boxplot showing the spread of data between the three conditions in the Novelty scale scale

users. But as the experiment was not performed on actual CI users but rather music presented through a CI simulator song, we can not say for certain if we would get the same results if the participants were CI users. While we can be fairly certain the simulator is a good approximation, the experimental participants were all exposed to it for the first time. A CI user, on the other hand, could be used to it to a certain degree and thus the scores might be more neutral. For these reasons we find this path very promising for further studies.

7. CONCLUSION

The experimental results suggest that there is a difference between the experience of listening to CI simulated music with and without the belt keeping in mind there were eight participants which could be considered low. While the scores for both mappings were not especially high, neither mapping was very low. These results are very promising for future iterations, and we will move onward with focusing on the placement of tactors and the choice of tactors themselves.

Acknowledgments

The work of the first author was financed by the NorForsk-funded Nordic Sound and Music Computing Network (Project number: 86892) and the work of the second author was funded by the Icelandic Technology Fund (Project number: 176713-0611)

8. REFERENCES

[1] C. Prevoteau, S. Y. Chen, and A. K. Lalwani, "Music enjoyment with cochlear implantation," *Auris Nasus Larynx*, vol. 45, no. 5, pp. 895–902, 2018. [Online]. Available: https://doi.org/10.1016/j.anl.2017.11.008

[2] K. E. Gfeller, C. Olszewski, C. Turner, B. Gantz, and J. Oleson, "Music perception with cochlear implants and residual hearing," *Audiology and Neurotology*, vol. 11, no. SUPPL. 1, pp. 12–15, 2006.

[3] I. Jóhannesson, R. Hoffmann, V. V. Valgeirsdóttir, R. Unnþórsson, A. Moldoveanu, and Kristjánsson, "Relative vibrotactile spatial acuity of the torso," *Experimental Brain Research*, vol. 235, no. 11, pp. 3505–3515, 2017.

[4] I. Jóhannesson, O. Balan, R. Unnthorsson, A. Moldoveanu, and Kristjánsson, "The sound of vision project: On the feasibility of an audio-haptic representation of the environment, for the visually impaired," *Brain Sciences*, vol. 6, no. 3, pp. 1–20, 2016.

[5] R. Hoffmann, V. V. Valgeirsdóttir, I. Jóhannesson, R. Unnthorsson, and Kristjánsson, "Measuring relative vibrotactile spatial acuity: effects of tactor type, anchor points and tactile anisotropy," *Experimental Brain Research*, vol. 236, no. 12, pp. 3405–3416, 2018. [Online]. Available: http://dx.doi.org/10.1007/s00221-018-5387-z

[6] R. Hoffmann, M. Brinkhuis, R. Unnþórsson, and Kristjánsson, "The Intensity Order Illusion: Temporal Order of Different Vibrotactile Intensity Causes Systematic Localization Errors," *Journal of Neurophysiology*, 2019.

[7] M. Karam, C. Branje, G. Nespoli, N. Thompson, F. Russo, and D. Fels, "The emoti-chair: An interactive tactile music exhibit," *Conference on Human Factors in Computing Systems - Proceedings*, pp. 3069–3074, 2010.

[8] D. Eagleman, "Plenary talks: A vibrotactile sensory substitution device for the deaf and profoundly hearing impaired," *2014 IEEE Haptics Symposium (HAPTICS)*, no. Ci, pp. xvii–xvii, 2014. [Online]. Available: http://ieeexplore.ieee.org/document/6775419/

[9] de la Torre Vega, M. B. Martí, R. de la Torre Vega, and M. S. Quevedo, "Cochlear Implant Simulation version 2.0: Description and usage of the program," *University of Granada, Spain*, 2004.

[10] CSUN, "Cochlear Implants | California State University, Northridge." [Online]. Available: https://www.csun.edu/ncod/cochlear-implants

[11] J. A. Grange, J. F. Culling, N. S. L. Harris, and S. Bergfeld, "Cochlear implant simulator with independent representation of the full spiral ganglion," *The Journal of the Acoustical Society of America*, vol. 142, no. 5, pp. EL484–EL489, 2017. [Online]. Available: http://dx.doi.org/10.1121/1.5009602

[12] M. D. Fletcher, S. R. Mills, and T. Goehring, "Vibro-Tactile Enhancement of Speech Intelligibility in Multi-talker Noise for Simulated Cochlear Implant Listening," *Trends in Hearing*, vol. 22, pp. 1–11, 2018.

[13] C. Fuller, "Early Deafened , Late Implanted Cochlear Implant Users Appreciate Music More Than and Identify Music as Well as Postlingual Users," vol. 13, no. October, 2019.

[14] A. Hinderks, M. Schrepp, F. J. Domínguez Mayo, M. J. Escalona, and J. Thomaschewski, "Developing a UX KPI based on the user experience questionnaire," *Computer Standards & Interfaces*, vol. 65, pp. 38–44, 7 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0920548918301247

DISTANT READING STRATEGIES AIDING THE COMPOSITION OF NEW IDIOMATIC MUSIC FOR CLASSICAL GUITAR

Giovanni Albini
Conservatorio "Jacopo Tomadini"
Eesti Muusika- ja Teatriakadeemia
Conservatorio "Franco Vittadini"
mail@giovannialbini.it

Matilde Oppizzi
Conservatorio "Franco Vittadini"
m.oppizzi@icloud.com

ABSTRACT

Idiomatic music for classical guitar - i.e. music that can accommodate and highlight the specific characteristics and features of the instrument - is an easy task for composers-guitarists but could be a real issue for composers with no specific expertise in guitar music, even more so if they aim to compose adhering to a specific difficulty level. In this context, the aim of this paper is to answer the following questions: can distant reading methods help in defining strategies for composing new idiomatic music for guitar? If so, how can they be defined and implemented? To answer the questions a software has been developed to analyze through distant reading strategies different corpora of music for classical guitar in order to extrapolate knowledge about their idiomatic features in terms of left hand behaviors. The results have been consequently examined and structured with the purpose of mapping idiomatic patterns and to set rules for defining new ones accordingly. Their use have then been tested in the composition of new music for guitar, checking how they fostered creativity, granting at the same time novelty and playability.

1. INTRODUCTION

"Idiomatic music reflects what an instrument can and cannot do, what it does willingly and what it does reluctantly" [1]. In fact, "musical passages can be characterized as more or less idiomatic depending on the extent to which the music relies on instrument-specific effects". Moreover, "the mechanics of musical instruments commonly influence how the music itself is organized" [2]. However, idiomatic music for guitar seems to be implicitly referred to the traditional repertoire of the instrument, making the two words almost synonymous [3–5]. This raised the questions: are there any ways of aiding and fostering the composition of new music for classical guitar that a performer could feel at the same time idiomatic and new? Can they be defined so that they also assist composers with no expertise in the technique of classical guitar and therefore enriching the repertoire of the instrument? Accordingly, the goal is to keep the focus at the uncharted possibilities at the core of the idiomatic performance habits of the instrument, leaving out of this research the novelty of the non-conventional extended techniques that are typical, for example, of the avant-garde.

Distant reading is an approach for the studying and understanding of a corpus not by taking into account its texts separately and by having humans reading them, but by aggregating and analyzing them as massive amounts of data with a quantitative approach, usually computerized [6, 7]. It has shown in the last decade its efficacy, mainly for literature and text-based corpora. Moreover, the use of statistics in musicology has recently intensified, especially in studying musical style [8–13], as well as its use for intercepting knowledge of relevance to composers [14, 15].

In this context, the aim of this paper is to answer the following question: can distant reading methods help in defining strategies for composing new idiomatic music for classical guitar? If so, how they can be defined and implemented? To answer it, 1) an algorithm has been developed so to use distant reading methods to intercept in different corpora of scores written for classical guitar some knowledge related to their idiomaticity and difficulty level, 2) the achieved knowledge has been then organized in a way that is accessible and usable by composers for the composition of new music , and finally 3) the effectiveness of the aforesaid algorithm and strategy has been tested qualitatively by one of the two authors, a composer, using it for the composition of a new score, and by the other one, a classical guitarist, who studied and performed it.

2. WHICH IDIOMATICITY?

To define and quantitatively measure idiomaticity, in absolute, is problematic because it is not a binary nor an objective concept. Its perception varies from one person to another, given their unique history and experiences with their instrument and with repertoires and the unique shape of their hands. Furthermore, several elements are involved in the performance of a musical score and in the determination of its idiomaticity and difficulty level, engaging different skills.

For these reasons, this research 1) narrowed down the elements to those that could determine left hand behaviours, i.e. admitted fingerings and transitions between them, and 2) defined an algorithm that takes into account a corpus of scores that is considered idiomatic by a specificic (set of)

guitarist(s), so to satisfy any individual requirement of idiomaticity/difficulty basing it to the repertoire given as an input. The idea is therefore to intercept in the corpus of all the music that is known by and *at the hands* of a performer the admitted and known sequences of left hand fingerings [1] and, from these, to derive new ones to be perceived indeed new and as idiomatic as the original ones.

3. THE SOFTWARE

To implement the algorithm a software has been developed in Python using Music21 [2] and NetworkX [3] libraries. It receives as inputs the corpora of scores as collections of files in MIDI and Music XML formats, it parses the complexity of the multiple parts of the scores to a succession of chords through a *salami slicing* method [4] , and retrieves the transitions between each left-hand fingering, enumerating every couple of consecutive chords, as it is depicted in Fig. 1.

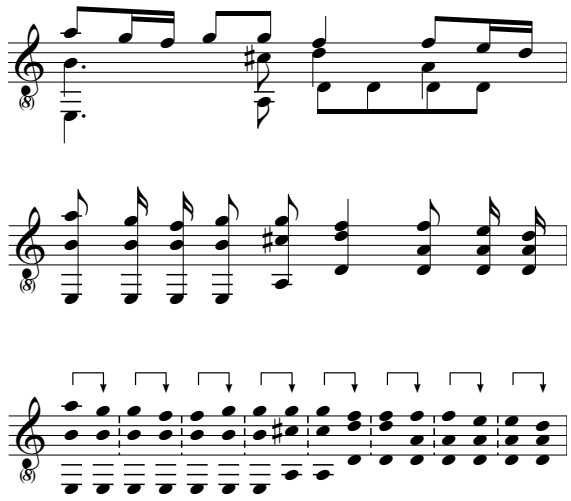


Figure 1. Bar five from Bach's *Fugue* BWV 1000 (first line), the same bar after it has been chordified (second line), and the dictionary of the eight possible transitions between chords, thus different left hand fingerings (third line).

The basic idea is that whatever could be the actual fingerings chosen or needed to put into practice the given chords and their succession in time, they must have been learnt by the guitarist(s) to whom the music will be written and they will be perceived the more idiomatic the more often they appear. Thus, by limiting the grammar of a new composition just to the resulting list of couples of chords (evidently put in any order and eventually considering also rests) the idiomaticity of the new score will be granted.

Such an approach has at least the following benefits: 1) working on very common file formats it is fully automatic, quick and reliable, 2) it can work on huge corpora, mixing authors, styles and genres and then possibly triggering new music ideas, 3) it works on the music itself and it does not need data about the fingerings, and 4) it offers a dictionary of chords and a grammar for using them without the need of specific knowledge about the guitar technique for the composers.

A criticality of the software and the constrained composition system hereby described is that it does not consider that the same chord may have two different fingerings in different couples/transitions. This may then result in difficult - if not impossible - sequences in the new music written abiding by the grammar. To fix it the authors are developing a further algorithm to guess the actual fingerings involved and so to list the admitted transitions between chords not only in terms of the music but also of the fingerings.

Furthermore, each of the two file formats accepted as input has its specific advantages and disadvantages. Since music notation refers to conventions that not always show the actual music content (arpeggios, for instance, keep the same fingering but are usually notated in ways that would be salami-sliced in several chords), occasionally a Music XML format could be not the best choice and MIDI is better. Moreover, the enharmonic pitch representation of MIDI files is irrelevant from the point of view of guitar fingering. However, sometimes MIDI files are written so to mimic a performance, tweaking note durations, then resulting in not reliable data after a chordification. In any case, the wide diffusion of music in both formats lets the user decide case by case the one that fits better.

4. A CASE STUDY

To test the effectiveness of the aforesaid software and constrained composition strategy a single score has been given as the input corpus, Leo Brouwer's *Fugue N. 1* [22]. Despite the shortness of the score, forty-nine bars for less than two minutes and a half of music, the different chord (hence fingering) transitions are totally 334. In table 1 are shown all those with occurrence major than one (ascending), excluding trivial transitions between or including one-note chords. In Fig. 2 all the chords transitions are on the other hand depicted as directed edges on a circular graph where the vertices are the chords. The list given is by itself sufficient for a short constrained new composition.

To ensure a further degree of novelty other chord transitions have been added to the lists: the ones obtained from the 334 translating them horizontally and vertically along the fretboard. This allows an harmonic variety that is also enhanced by the irregularity of the tuning of the strings. In fact, classical guitars are typically tuned in a series of ascending perfect fourths but with a single major third be-

[1] For quantitative approaches to the issue of stringed instrument left-hand fingerings see [16–19].

[2] "Music21 is an object-oriented toolkit for analyzing, searching, and transforming music in symbolic (score-based) forms" started at MIT [20]. https://web.mit.edu/music21/

[3] NetworkX is a Python package for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks. https://networkx.github.io

[4] A method to "represent musical sequences as ordered collections of consecutive pitch slices, each slice having a proper duration. Every time a new pitch starts or stops being played, the current slice ends and a new one begins. The process applies systematically for every note encountered within the score, including embellishments. This segmentation method, widely used in computational musicology, also goes under the name of pitch simultaneities" or chordification [21].

tween the third and the second string [5] . Therefore, any *vertical* translation crossing the tuning kink results in a new sonority.

With the resulting list - and also admitting open string chords and rests in between disconnected transitions - a new score, *Prigioniero I, Op. 49 N. 1b*, of which the first page is shown in Fig. 3, has been composed by one of the authors.

First Chord	Second Chord	Occurrence
D4E4	A3B4	2
G4F5	G4G5	2
G4G5	D4B4G5	2
D4B4G5	F4G4	2
D4C5A5	G4A4	2
F♯3G4A4	G3G4A4	2
A3D5G5	A3B4E5	2
D4D5G5	D4B4E5	2
B3B4E5	C4D5G5	2
C4D5G5	C4B4E5	2
B3G4	A3A4	2
D4A4	E4G4	2
E4G4	E4C5	2
E4C5	D4D5	2
D4D5	G4D5	2
G4D5	A4B4	2
G4F♯5	C5F♯5	2
C5F♯5	D5E5	2
D5E5	D5A5	2
D5A5	A3D5B5	2
C4D4	A3D4	2
F3D4	F3G4	2
F4D5B♭5	F4C5B♭5	2
E3E4	E3F♯4	2
E3F♯4	E3C5	2
E3C5	E3F5	2
E3F5	E3B5	2
E3G♯5	E3A5	2
E3A5	E3B4	2
E3B4	E3D5	2
E3D5	E3E5	2
E3G♯5	E3F5	2
E3E4	E3G♯4	2
E3G♯4	E3D5	2
E3D5	E3G5	2
E3G5	E3C♯6	2
E3A♯5	E3B5	2
E3B5	E3C♯5	2
E3C♯5	E3E5	2
E3E5	E3F♯5	2
E3A♯5	E3G5	2
E3E4	E3B4	2
E3B4	E3F5	2
E3F5	E3B♭5	2
E3B♭5	E3E6	2
E3C♯6	E3D6	2
E3D6	E3E5	2
E3E5	E3G5	2
E3G5	E3A♭5	2
E3E5E6	E3G5G6	2
E4B4E5	E4D5G5	3
B3D4	B3G4	3
D4D5B♭5	G4B♭5	3
G4B♭5	F4D5B♭5	3
A3A4	D4A4	4
A3D5B5	C4F5D6	4
E3B5	E3G♯5	4
E3E6	E3C♯6	4
E4D5G5	E4B4E5	5
E3C♯6	E3A♯5	6

Table 1. All the chord transitions in Brouwer's *Fugue N. 1* with occurrence major than one (ascending), excluding trivial transitions between, or including, one-note chords.

The new score has then been studied by the second author, a guitarist, who had Brouwer's score in her repertoire, and who confirmed the easiness of learning it and the friendliness of the fingering involved. However, as a result of the chords obtained from translations along the fretboard and of newfound sequences, the score has not been experienced by the performer as a sibling of the corpus piece, but as an original new one.

[5] From the lowest string, to the highest: EADGBE.

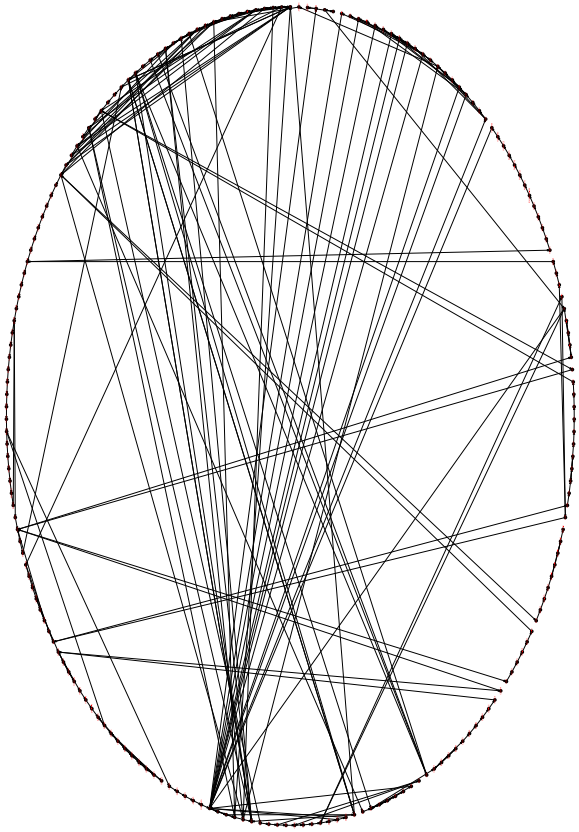


Figure 2. All the chords transitions in Brouwer's *Fugue N. 1* depicted as directed edges on a circular graph where the vertices are the chords.

5. CONCLUSIONS

The experience presented connected the artistic practice of the two authors, a composer and a guitarist, trying to provide answers to artistic issues by the means of technology and through research.

Looking for strategies for composing new idiomatic music for guitar that could be useful also for composers with no experience with the instrument, a software has been developed to analyze through distant reading strategies different corpora of music in order to map idiomatic patterns in transitions between left hand fingerings and to set rules for defining new ones accordingly. Their use has then been tested in the composition of new music for guitar, checking how it fostered creativity, granting at the same time novelty and playability.

Future developments will refine the algorithm recognizing the actual fingerings involved and perfecting the software's usability, so to offer a system that could also automatically provide the composer an enlarged and user-

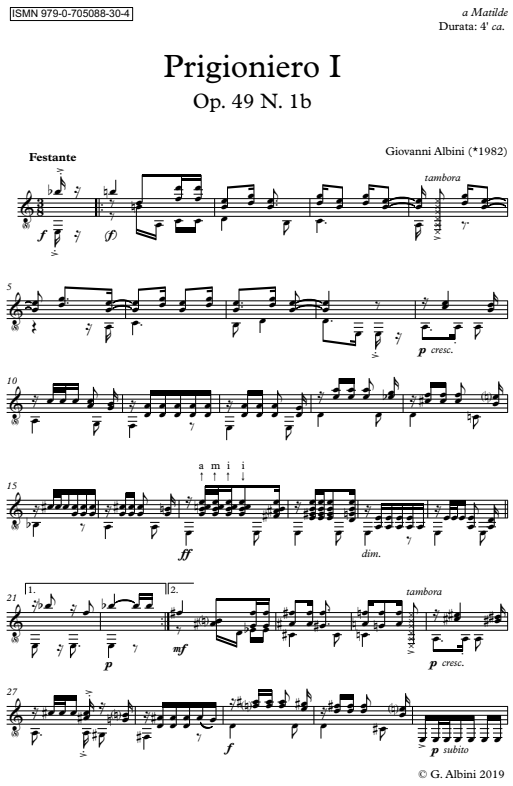


Figure 3. First page of the score composed by one of the authors with the constrained system based on the list of chord transitions obtained from Brouwer's *Fugue N. 1*.

friendly list including all the translations along the fretboard. Moreover, to address the issue formalizing left-hand fingering/position changes only is not without its flaws: to study them taking also into account other parameters - such as for instance tempos and durations - could help in inferring their idiomaticity more effectively and in more detail. Further studies will be conducted then to test the efficiency of the strategy surveying the experience of several different performers and composers.

6. REFERENCES

[1] J. D. Souza, *Music at Hand: Instruments, Bodies, and Cognition*. Oxford Studies in Music Theory, 2017.

[2] D. Huron and J. Berec, "Characterizing idiomatic organization in music: A theory and case study of musical affordances," in *Empirical Musicology Review*, vol. 4/3, 2009, pp. 103–122.

[3] G. Radole, *Liuto, Chitarra e Vihuela*. Edizioni Suvini e Zerboni, Milano, 1997.

[4] G. Wade, *A concise history of the classical guitar*. Mel Bay, Pacific, PO, USA, 2001.

[5] ——, *Traditions of the Classical Guitar*. John Clader, London, UK, 2012.

[6] F. Moretti, *La Letteratura Vista Da Lontano*. Einaudi, 2005.

[7] ——, *Distant Reading*. Verso, London, 2013.

[8] L. Knopoff and W. Hutchinson, "Entropy as a measure of style: The influence of sample length," in *Journal of Music Theory*, vol. 27/1, 1983, pp. 75–97.

[9] L. B. Meyer, *Style and Music: Theory, History, and Ideology*. University of Pennsylvania Press, 1989.

[10] J. Snyder, "Entropy as a measure of musical style: The influence of a priori assumptions," in *Music Theory Spectrum*, vol. 12/1, 1990, pp. 121–160.

[11] D. Huron, *Sweet Anticipation: Music and the Psychology of Expectation*. MIT Press, 2006.

[12] M. Kaliakatsos-Papakostas, M. Epitropakis, and M. Vrahatis, "Musical composer identification through probabilistic and feedforward neural networks," vol. 6025, 04 2010, pp. 411–420.

[13] A. Brinkman, D. Shanahan, and C. Sapp, "Musical stylometry, machine learning, and attribution studies: A semi-supervised approach to the works of josquin," in *Proceedings of the 14th Biennial International Conference on Music Perception and Cognition*, 2016, pp. 91–97.

[14] G. Albini and F. Nastari, "Composing new with the old: Musical stylometry as a means of composing new music for classical guitar," in *Proceedings of the II Dublin Guitar Symposium, Dublin*, 2019.

[15] G. Albini and M. Oppizzi, "Composing idiomatic music for guitar using distant reading strategies," in *21st Century Guitar Conference, Ottawa*, 2019.

[16] J. D. Souza, "Fretboard transformations," in *Journal of Music Theory*, vol. 62/1, 2018, pp. 1–39.

[17] T. Koozin, "Guitar voicing in pop-rock music: A performance-based analytical approach," in *Music Theory Online*, vol. 17/3, 2011.

[18] J. Rockwell, "Banjo transformations and bluegrass rhythm," in *Journal of Music Theory*, vol. 53/1, 2009, pp. 137–162.

[19] D. R. Tuohy and W. D. Potter, "A genetic algorithm for the automatic generation of playable guitar tablature," in *Proceedings of the International Computer Music Conference*, 2019, pp. 499–502.

[20] M. S. Cuthbert and C. Ariza, "music21: A toolkit for computer-aided musicology and symbolic music data," in *11th International Society for Music Information Retrieval Conference (ISMIR 2010)*, 2010, pp. 637–642.

[21] L. Bigo and M. Andreatta, "Filtration of pitch-class sets complexes," in *Mathematics and Computation in Music. MCM 2019*. Springer, Cham, 2019.

[22] L. Brouwer, *Fuga N. 1*. Max Eschig, Paris, 1972.

Efficient Rendering and Perception of Acoustical Environments in Augmented Reality Audio

Alex Baldwin
Dirac Research, Uppsala, Sweden
alexbaldwinmusic@gmail.com

Jonas Holfelt
GN Store Nord, Copenhagen, Denmark
jholfelt@gmail.com

Cumhur Erkut
Aalborg University Copenhagen
cer@create.aau.dk

ABSTRACT

This paper presents an efficient real-time estimation and rendering of an acoustical environment for augmented reality (AR). An AR application obtains information about a given environment based on the detected surfaces with the camera, as well as the positions virtual sound sources and listener. Then an artificial reverberator is calibrated based on these data. An evaluation is carried out both in an actual and in a virtual room (by using pre-rendered impulse responses and convolution). Results indicate a perceptual preference towards simpler artificial reverberators. Taken together, our implementation and perceptual evaluation results contribute to the challenge of real-time interaction in a real, dynamic environment so that the virtual objects are registered in that environment not only visually, but also in the auditory modality.

1. INTRODUCTION

Immersive experiences in virtual environments require realistic but efficient sound rendering algorithms [1]. The sound design pipeline for these experiences usually consist of three major elements: *source modeling, room acoustics modeling* and *listener modeling* [1]. In this paper, we focus on the room acoustics modeling. We implement some efficient delay-based algorithms first on Unity. We then deploy them to a mobile AR device protoype. We finally evaluate immersive sonic experiences.

Our work derives from two implementations. The first one is ScattAR [2]. The application allows the user to scan an environment from which a 3D mesh is generated. First-order reflection points were calculated using the image-source method [3]. The source and reflection points, and listener were connected according to Scattering Delay Network (SDN) technique [4], and reverberation was calculated. Both the listener and sound source can be moved around in space, while reflections were updated in real time. Fig. 1 depicts ScattAR, where a sound source (drone) moves in an actual (left) or virtual (right) room, and the direct path (green) as well as first-order reflections (red paths) are calculated, visualized, and auralized.

Recent work has shown interest in modeling sparsely reflecting scenes and proposed a solution utilizing the Digital Waveguide Web (WGW) [5]. The model is highly flexible, but computational costs increase exponentially with added scattering junctions. WGWs were added by Jonas Holfelt to the ScattAR framework [6]. Computational benchmarks of WGWs in our framework indicates that even the simplest cases using just five reflection points are computationally too demanding for real-time implementation.

The main contributions of this paper are as follows. First, ScattAR is a valuable application in teaching and demonstrating room acoustics, and we hope to maintain the development by porting it to other AR-frameworks. This requires a compact description of previous implementations for reference. Second, besides comparing the simulated environments to the original ones, we clearly need different strategies in conducting interactive listening tests. In ScattAR, an interactive perceptual test on object presence indicated only a marginal difference between the ratings of SDN processed sound and unprocessed anechoic sound, favoring the SDN [7]. Recent tests indicate the difference is clearly perceivable in non-interactive conditions, and require intermediate steps as well as a more accurate model.

This paper is organized as follows. Section 2 gives a brief overview of important fundamental reverberation methods, followed by an overview of the SDN and WGW methods, which form the theory needed for the implementation. Section 3 describes the implementation of SDN and WGW. Section 4 presents an evaluation in a virtual environment. Finally, Section 5 discusses the results and concludes the paper.

2. BACKGROUND

This section presents two fundamental methods for artificial reverberation: *convolution-based algorithms* and *delay networks*. The former method is realistic but not flexible, whereas the latter represents the foundation for the implementation of both SDN and WGW. In this paper, they will be used in tandem, and the *image-source* method will also be utilized.

2.1 Convolution Reverb

Convolution reverb is performed by convolving a *room impulse response* (RIR) with a dry input signal. The RIR can be obtained either by a recording in a physical acoustic space, or by synthesis. The result is a very accurate reverberation model, though it is not flexible [8], plus it is computationally heavy. *Fast convolution techniques* have been proposed, where one solution is to multiply signals after

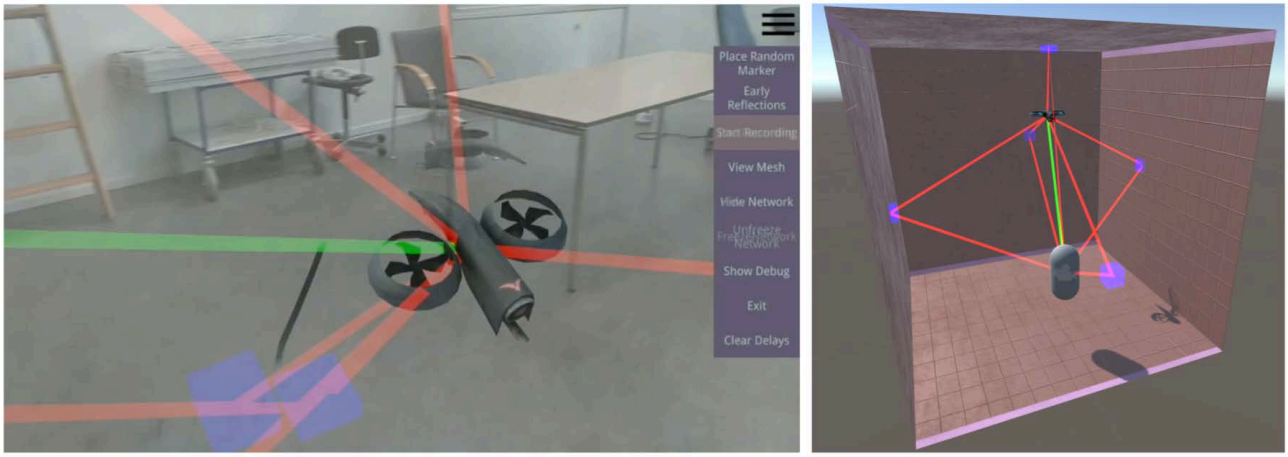


Figure 1. A virtual drone in a real-room on the ScattAR mobile augmented reality audio application. a) Left: Real-time path overlays of direct path (green) and first-order reflections (red). b) Right: Simulation on Unity.

FFT has been performed [8], thus simplifying the computation.

2.2 Image-Source Method

In the image-source method, physical boundaries of a room are replaced by an infinite lattice of image sources [9]. The first-order reflections are geometrically determined by assuming perfectly reflecting walls [8], whereas higher-order reflections are typically modeled by artificial reverberation.

2.3 Delay Networks

Delay networks are fast and efficient, but cannot directly model a given physical space. The most common recursive linear time-invariant reverberator is known as the Feedback Delay Network (FDN). An FDN is build by a number of delay lines in a feedback loop utilizing matrices for attenuation and filtering. Ideally the loop should be lossless, thus energy-preserving [8]. Absorption filters can be designed to obtain a desired reverberation time in various frequency bands [4]. The FDN offers high quality reverberation at an efficient cost, and with the right delay-lines and filter matrices precise models can be made. It however lacks the source and positions updates.

2.3.1 Digital Waveguide Network

A digital waveguide is a bidirectional delay with a port-impedance where traveling waves propagate and scatter if there is a discontunity in the medium [10]. *Digital Waveguide Networks* (DWNs) are an example of a *delay network*, consisting of a closed network of digital waveguides interconnected by *scattering junctions* [10]. A scattering junction, shown on Fig. 2 is placed where the traveling sound waves hit e.g., the walls of a room (blue boxes on Fig. 1). The sound pressure $p_J(n)$ at such a junction J at the time instant n is calculated as:

$$p_J(n) = \frac{2}{\sum_{i=1}^{N} \Gamma_i} \sum_{i=1}^{N} \Gamma_i p_i^+(n), \qquad (1)$$

where N is the number of waveguides connected at the scattering junction, Γ_i is the admittance of waveguide i, and p_i^+ is the incoming traveling pressure wave to the junction from waveguide i.

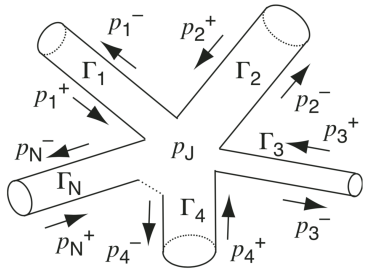


Figure 2. A scattering junction, where the junction pressure is P_J.

These junctions are lossless so that the signal power is conserved [8]. The outgoing pressure wave p_i^+ from the junction is then calculated by

$$p_i^-(n) = p_J(n) - p_i^+(n). \qquad (2)$$

The operations in Eqs. 1 and 2 can be combined into a single matrix operation so that the reflected waves are calculated by multiplication of scattering matrix $\mathbf{S}$ with a vector of incoming waves. The network topology and waveguide impedances can be determined from geometrical analysis of a given environment to be simulated based on a path-tracing algorithm [10].

2.3.2 Scattering Delay Network

The scattering delay network extends the digital waveguide network [4] and provides a computationally efficient approximation to geometric ray tracing [11]. The method is conceptually similar to feedback delay networks (FDN) (see above). The SDN involves, however, discrete nodes instead of a circulant matrix. These nodes are connected by a set of waveguides and represent the first-order reflection points of a given space. The nodes are also referred to

as scattering junctions and are interconnected via bidirectional waveguides.

A detailed description of SDC can be found in [4] and [5]. The SDN is an efficient and effective reverberator for acoustic room simulation, providing accurate first-order reflections and rich higher order reflections [4]. For more complex systems involving objects such as rigid cylinders rather than just planes, it requires directional filtering [5]. In addition, the importance of precise first-order and second-order reflections in the modelling of outdoor acoustic spaces is indicated in a previous study [12]. The Waveguide Web offers this type of modelling and will be described in the next section.

2.3.3 The Waveguide Web

The WGW is, in essence, very similar to SDN, but it offers additional accuracy. Like the SDN, the modeled space consists of wall-nodes connected to each other with bidirectional delay lines. The source and receiver are also connected to the wall-nodes by unidirectional delay lines. ppThe main difference between the two algorithms is the scattering action at each node [5]. SDN only allows for one filtering operation at each node. WGW allows for directionally dependent filtering to be implemented, which means that each node can provide different filtering according to where the incoming pressure wave signal is coming from. This can be done to second-order reflection precision. The nodes can be placed at any 3D position, just like SDN.

Full details of WGW computation can be found in [5], and a MATLAB implementation of the alogrithm is available online at [13].

2.3.4 Evaluation of the SDN and WGW

A comparison between SDN and WGW for the same space (a shoebox room of 9x7x4 meters) was performed in [5]. The results proved that the two algorithms produced identical first-order reflections and similar reverberation tails, with a noticeable difference due to the different characterization of the higher-order reflections. An evaluation of the computational requirements of WGW was also performed. For the simplest case of 5 nodes, the simulation required 4.36 seconds and 5.65 MB of memory to produce 1 second of audio at a sample rate of 48 kHz. For higher numbers of nodes, the computational requirements increased exponentially, where a case of 30 nodes would require 667.23 seconds and 102 404.16 MB of memory. This exponential increase is due to the N^2 at each node, which means an increase of $N(N^2) = N^3$ filters for the entire system [5]. To our knowledge, no perceptual evaluation of SDN or WGW have been carried out previously.

3. IMPLEMENTATION

The SDN and WGW were implemented on Unity. For prototyping, a virtual room was created. It contains the following essential objects, which can be all idenfied in Fig. 1b : 1) A sound source, to which reverberator and reflection finder scripts, and an Audio Source component are attached (a drone in Fig. 1b). 2) A listener, with a rigid body and a mesh collider attached to it, so that it can be hit by rays. The Main Camera is snapped to the position and orientation of the listener object. 3) Boundary objects, which set the colliders for the surfaces that are to reflect the emitted rays to the listener. In Fig. 1b the boundaries are simply the walls of a room. 4) Scattering Junctions are the blue boxes shown at the points of first-order reflections. The scattering junctions are instantiated, when the first-order reflection points have been calculated.

The reflection path (the red lines on Fig. 1) holds the audio source position, destination position (the listener position), a ray list for every ray segment, and a float list containing the lengths of these segments. For each first-order reflection, there are two segments: one from source to wall-node, and one from wall-node to receiver. The reflection finder is a component on the source object. The direct path between the source and the listener is easy to find by utilizing the built-in ray tracing system in Unity. The component finds a pre-specified number of reflections, which is generally 6 for first-order reflections in an empty rectangular room. To find the reflections for each surface of the room model the spherical Fibonacci point set algorithm [14] is utilized. The main WGW component receives the list of paths and instantiates the WGW node objects at the positions of the first-order reflections points. These nodes are depicted by the blue boxes on Figure 1.

The WGW nodes hold the scattering position in the environment, a list of connections to the other nodes, two delay lines for input and output, N^2 absorptive IIR filters, the reflection path which the node is placed upon, and finally the scattering values which are applied through a for-loop.

The WGWnode class is based on the node structure proposed by [5]. It holds a class definition for delay lines, a function for propagating through the network, a class definition for the WGW connections (connections between all the other nodes), a function for performing the scattering operation and the initializing of the audio streaming thread. When a number of WGW nodes have been instantiated, they are all connected with bidirectional delay lines. Filters and coefficients for second-order reflections were defined according to [5].

3.1 Performance: Running the WGW in real-time

The WGW was not able to produce output in real-time in our Unity application, even when we set all the filters to constant coefficients, The algorithm was verified by letting the WGW run and record the output offline. For six nodes, the system spent averagely 100 ms for each sample, which is far from real-time. Therefore, to be able to make a real-time simulation for interactive tests, we generated an impulse response and performed convolution-based reverberation, with Unity's Native Audio Plugin SDK [15]. This process of generating an impulse response can took up to 4-5 seconds, which fits the results from [5], where the computational benchmark found the WGW to generate 1 second of audio in 4.35 seconds. This means that upon finding a new reflection path (when either source position or listener position is updated), it will take 4-5 seconds before the audio output from the convolution reverb will ac-

tually be based on the current position. Therefore we have designed an evaluation on the fixed source and listener positions.

4. EVALUATION

Before the perceptual tests, we first compared the impulse responses generated with SDN and WGW and validated our implementation. In a virtual room in Unity (see Fig. 5, scaled to a 9x7x4 meter shoebox room modeled and evaluated in [5], the wall reflectance coefficient was set to 0.97, and the filters in both simulations were designed as allpass filters. Both the impulse response and the T30 plots gave similar results; so we proceeded to the perceptual tests, by asking the participants if they prefer one rendering over the other. In addition to the sounds processed with SDN and WGW, an unprocessed anechoic sound was also used as a control condition.

4.1 The virtual environment and stimuli

The room was designed in Unity, based on geometries identical to the signal comparison. These geometries produce a room that would be considered quite large in a real-world setting. Since it would be difficult to assess the size of the room without any relative elements, some objects were implemented as seen in Figure 3. The room needed to remain empty, for the modeled reverberation would not consider diffusion of furniture and other objects. Therefore only a door and windows were added to the room, since they are flat and would not alter the room reflections. The door is scaled relative to the 4 meter tall wall. Buildings, grass and road objects are added to the background, for a parallax effect when looking through the windows, to obtain a greater sense of depth. To make the room more relatable, textures were added to floor, wall and ceiling as seen in Figure 3.

Figure 3. A virtual environment that represents a large room (9x7x4 m).

All decisions regarding the size of objects, textures, lighting and materials have been made without precise measurements or references. Neither has the environment been tested for verification of its perceived size. Therefore, there was a risk that the participants will not be able to perceive the actual size of the room. Nevertheless, it should not affect the end results, since the condition will be the same for all cases.

The input sounds for the perceptual test were recorded in an anechoic room. The sounds were the impact of a ball hitting a surface and a small speech excerpt. These two sounds served as the input for the sound source in two different conditions. One condition shows the character speaking (see Figure 3) while the other condition shows a ball bouncing. The sounds were pre-rendered for the experimental test. The test was designed to take the participant through three different positions in the same room.Therefore a recording for each sound (ball and speech) and reverberation type, was recorded at each of the three positions, producing 12 audio files. A wall reflectance coefficient of 0.97 was empirically chosen for the recordings. Both reverberators had identical filters (allpass). Similar settings were used for the original evaluation of WGW [5]. The two original anechoic recordings were used as the control condition.

All sounds were normalized, and played over a laptop with headphones to the participants. A 13-inch Macbook Pro was used to display the room and record the ratings.

4.2 Experimental Design

The design of this test follows the method proposed by [16] to evaluate the audio quality of synthetic sound effects. In this method, the participant is asked to evaluate a number of sound samples on a scale from *very unrealistic* to *very realistic*. Different adjectives were used in our test on the scale for the specific context in the virtual room, since *realistic* is too general in this case. The interest is to know how realistic the audio output is in the given context of the room and sound source position, therefore the scale is formulated as in Figure 4.

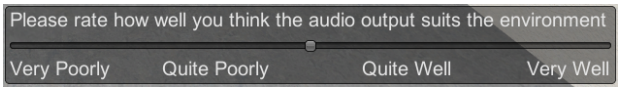


Figure 4. The scale in discrete steps in our test.

The participants were asked to evaluate a total of 18 cases, consisting of combinations between three different room positions, two different sound sources and three different types of processing (SDN, WGW and anechoic). The test started with a quick tour of the room, where the camera moved around in a fluent motion in the room to show the surrounding space, ensuring that the participant was well-aware of the room dimensions and their position in the room. Thereafter the source and listener positions were initialized to one of the three test positions and the test began. The first half of the test instantiated one sound source, and the second half instantiated the other sound source. Half of the participants was first introduced to the speaking character, while the other half was first introduced to the bouncing ball. The order of position and reverberation combinations were completely random. When the participants had completed the test, they were asked to answer a

short questionnaire, where they were allowed to comment on the environment, sound, and perceived room size.

4.3 Results

20 participants completed the tests. Their audio quality ratings were recorded together with the labels for sound, position, and the algorithm used. The minimum rating corresponding to *very poorly* was 0, and the maximum rating corresponding to *very well* was 10.

Table 1 presents the descriptive statistics of the experimental conditions. The top part shows the results derived from the overall ratings. These have been labeled by WGW, SDN and Anechoic, and is not dependent on the sound source or position. Anechoic was a control condition and was expected to have a low mean, which it definitely had compared to the two other ratings. The ratings for SDN were generally slightly higher than WGW.

The two other parts in Table 1 show the ratings depending on sound source and position respectively. These ratings are independent of algorithm type, and merely represent the total ratings for the given conditions. This is done to evaluate whether some of the conditions were biased.

Algorithm	Mean	STD
WGW	5.76	1.15
SDN	6.32	1.30
Anechoic	3.31	1.76
Sound	**Mean**	**STD**
Ball	5.30	1.16
Speech	4.96	0.97
Position	**Mean**	**STD**
Pos 1	5.28	1.06
Pos 2	4.79	1.15
Pos 3	5.32	1.17

Table 1. Statistics of the ratings. Top: General ratings independent of sound source or position. Middle: Overall ratings (independent from algorithm) for the ball and speech respectively. Bottom: Overall ratings (independent from algorithm) for the various positions

Figure 5 visualizes the ratings as a graph with error bars. pnThe ratings between different sources (ball, speech) and different positions (1-3) are similar, and the standard deviations are generally small for all cases. The mean ratings of Algorithm on were tested for normal distribution using the Anderson Darling test. All three ratings were classified as being normally distributed. Therefore, we concluded that the data is parametric, and a parametric statistical test can be used to test for a significant difference. Our null-hypothesis yielded: *There is no difference between the ratings of audio quality in sound processed with respectively SDN and WGW.*

A t-test was performed for three conditions on algorithms; it successfully rejects the null-hypothesis with a probability value of 0.0396 (0.4 percent probability of a type 1 error). Thereby it can be concluded that there is a significant difference between the ratings of SDN and WGW, with a preference towards SDN.

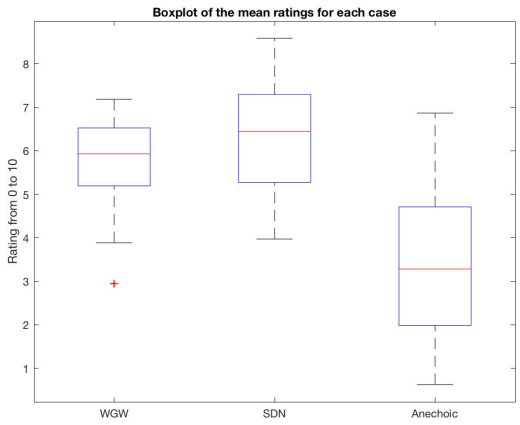


Figure 5. Rating statistics visualized as a graph with error bars.

5. CONCLUSION

This paper has presented an implementation and evaluation of the Scattering Delay Network and the Waveguide Web as acoustical renderers in virtual environments on a popular game engine. Following the related background and the presentation of the object-oriented design structure, studies were presented to evaluate the perceptual differences between the two artificial reverberators. A perceptual test was performed, where participants were asked to rate how well they though the audio output fit into a given environment. The ratings indicate a difference between the reverberators, with a preference of SDN over WGW and anechoic sound. This result is surprising, as it shows that for simple room geometries that contain planar surfaces, a simpler implementation based on SDN should suffice.

Further research and development of WGW would be interesting for more complex environments, where SDN would not be sufficient. The present work posed a compromised solution for running WGW in real-time, by generating an impulse response and performing convolution with the input sound. This limits the system to only update the position of the source and listener at the rate of generating the impulse response. This means, that for WGW to run truly real-time some compromises would have to be made to the complexity, and thereby accuracy of the structure.

The Waveguide Web still offers a wide variety of interesting studies for non-real-time purposes, and by modifying the filter calculation in the directionally dependent filtering structure, new types of scattering could be modeled to provide accurate acoustical models of complex spaces and structures. Recent work presents a novel algorithm for generating virtual acoustic effects from AR, with an extra dimension where it scans both the geometry and the materials of the objects of a room, which are used to determine absorption coefficients and filter responses [17]. This algorithm is not real-time either, works indoors, but further development would consider outdoor areas as well. Therefore, there is a need for an algorithm such as the Waveguide Web, especially if it can be made efficient to run in real-time on mobile processors.

Acknowledgments

We thank Enzo de Sena and Francis Stevens for guiding us through the implementation details of SDN and WGW, respectively.

6. REFERENCES

[1] S. Serafin, M. Geronazzo, C. Erkut, N. C. Nilsson, and R. Nordahl, "Sonic Interactions in Virtual Reality: State of the Art, Current Challenges, and Future Directions," *IEEE Computer Graphics and Applications*, pp. 31–43, 4 2018.

[2] A. Baldwin, S. Serafin, and C. Erkut, "ScatAR: a mobile augmented reality application that uses scattering delay networks for room acoustic synthesis," *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology*, 2017.

[3] L. Savioja and U. P. Svensson, "Overview of geometrical room acoustic modeling techniques," *The Journal of the Acoustical Society of America*, vol. 138, no. 2, pp. 708–730, 2015.

[4] E. De Sena, H. Hacihabiboglu, Z. Cvetkovic, and J. O. Smith, "Efficient Synthesis of Room Acoustics via Scattering Delay Networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 9, pp. 1478–1492, Sep. 2015.

[5] F. Stevens, D. T. Murphy, L. Savioja, and V. Valimaki, "Modeling Sparsely Reflecting Outdoor Acoustic Scenes Using the Waveguide Web," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 8, pp. 1566–1578, Aug. 2017.

[6] J. Holfelt, "Implementing ar in and out," Master's thesis, Aalborg University, Copenhagen, Denmark, 2018. [Online]. Available: http://tinyurl.com/y2tyh8wo

[7] A. Baldwin, C. Erkut, and S. Serafin, "Towards the design and evaluation of delay-based modeling of acousticscenes in mobile augmented reality," *Presented in SIVE Workshop within IEEE VR Conference*, 2018.

[8] V. Välimäki, J. D. Parker, L. Savioja, J. O. Smith, and J. S. Abel, "Fifty Years of Artificial Reverberation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 5, pp. 1421–1448, Jul. 2012.

[9] J. B. Allen and D. A. Berkley, "Image Method for Efficiently Simulating Small-room Acoustics," *The Journal of the Acoustical Society of America*, vol. 95, no. 943, 1979.

[10] M. Karjalainen, "Digital Waveguide Networks for Room Modeling and Auralization," *Proceedings of Forum Acousticum*, 2005.

[11] J. O. Smith, "Scattering delay networks," in *Physical Audio Signal Processing*, online book, 2010 edition ed. `http://ccrma.stanford.edu/~jos/pasp/`, 2010.

[12] D. T. Murphy and F. K. M. Stevens, "Spatial impulse response measurement in an urban environment," *55th International Conference: Spatial Audio*, 2014.

[13] F. Stevens, "Waveguide web example audio," http://www.openairlib.net/auralizationdb/content/waveguide-web-example-audio, 2017, accessed: 2018-05-31.

[14] R. Marques, C. Bouville, M. Ribardire, L. P. Santos, and K. Bouatouch, "Spherical Fibonacci Point Sets for Illumination Integrals: Spherical Fibonacci Point Sets for Illumination Integrals," *Computer Graphics Forum*, vol. 32, no. 8, pp. 134–143, 2013.

[15] Unity, "Unity native audio plugin sdk," https://docs.unity3d.com/500/Documentation/Manual/AudioMixerNativeAudioPlugin.html, 2018, accessed: 2018-05-31.

[16] D. Moffat and J. D. Reiss, "Perceptual Evaluation of Synthesized Sound Effects," *ACM Transactions on Applied Perception*, vol. 15, no. 2, Apr. 2018.

[17] C. Schissler, C. Loftin, and D. Manocha, "Acoustic Classification and Optimization for Multi-Modal Rendering of Real-World Scenes," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 3, pp. 1246–1259, Mar. 2018.

MahlerNet: Unbounded Orchestral Music with Neural Networks

Elias Lousseief and Bob L. T. Sturm
School of Electrical Engineering and Computer Science, Royal Institute of Technology (KTH), Stockholm, Sweden
elias.lousseief@blackwinged-angel.com, bobs@kth.se

ABSTRACT

This paper presents MahlerNet, a deep recurrent neural network that models polyphonic music sequences of arbitrary length with an arbitrary number of instruments. The data representation consists of instrument, pitch, offset and duration, which is motivated by properties inherent in both notated and performed music. It generates units of music (i.e., measures in this work) by sequentially sampling from distributions conditioned on context. This paper details experiments using two established datasets (PIANOMIDI and MUSEDATA), and a new dataset (MAHLER) consisting of all symphonies by Gustav Mahler. The smoothness of the learned latent space is explored by interpolating between two given measures of music. Results show that MahlerNet can generate music resembling its training data in many respects. Long-term structure is present in the form of instrumentation, intensity and rhythm, albeit rarely in the form of longer concrete motives and themes.

1. INTRODUCTION

The involvement of artificial intelligence in Arts practices is becoming a common activity, and music creation is no exception. Notably, AI has recently been used in the creation of popular music [1] and even folk music albums [1], not to mention start-up companies focused on the automatic generation of music for games [2] and soundtracks. [3] Ultimately, results in the field can serve many uses. AI composition tools can be used by composers to develop existing material, give inspiration to fragments of ideas, and to explore new kinds of music and ways of working.

While much research has been published about the modelling and generation of music, most of it is focused on melodies, chord progressions, or single instruments. Important aspects of music, however, are polyphony (multiple simultaneous, autonomous voices or instruments), instrumentation, and the development of musical ideas using multiple voices.

This paper presents *MahlerNet* [2], a polyphonic music model that aims to generate orchestral music with any number of instruments. The music representation of MahlerNet involves parameterizing each event in terms of a pitch, offset, duration, and instrument. MahlerNet models sequences of these events using a sequence-to-sequence network with context conditioning. These models generate new material by sampling sequentially from updated posterior distributions of the four parameters.

The rest of this paper has the following structure. Section 2 briefly reviews polyphonic music modelling and generation. Section 3 presents the representation and architecture of the proposed model. Section 4 describes several experiments with trained models. Section 5 reviews the strengths and weaknesses of the results and 6 describes the contribution of MahlerNet and presents avenues of future work.

2. PREVIOUS WORK

Computational modeling and generation of music has a long history, stretching back decades to experiments in the 1950s [3]. While these first approaches involved expert-based systems, with or without probabilistic components, later approaches take advantage of data for training music artificial intelligence (music AI) [4, 5]. Much research in the domain of music AI – a subdomain of *algorithmic composition* [6] – is concerned with individual musical voices, e.g., melodies or chord progressions. Polyphonic music modeling and generation in contrast can be more difficult, since it is concerned with multiple simultaneous, autonomous voices, perhaps on the same instrument (e.g., piano), or as an ensemble (e.g., choir, chamber ensemble, or orchestra). Though much recent work addresses polyphonic music modelling and generation, many are not concerned with instrumentation at all [7–18]. A few works use four-part choir [19, 20] or ensembles of instruments found in popular music, e.g., bass, guitar, piano and drums [21–24].

Many polyphonic approaches propose different music representations and modelling methods. A common format is "piano roll", which is a matrix of time-ordered column vectors with rows denoting which pitches are on at a given time [7, 9–15, 18, 20–24]. One column represents activity during a fixed amount of time (typically in note length) and is used by a system to generate the activity to follow. This approach is called *time slicing*. This implies that simultaneous events are fed in at one go and that the duration and starting point of each note is implicit, which makes the representation efficient. Problems involve distinguishing a sustained long note from successive shorter ones and in the case of multiple instruments, some extra device has to take care of which instrument plays what.

[1] https://www.helloworldalbum.net
[2] https://melodrive.com
[3] https://www.aiva.ai

An alternative representation, similar to how MIDI works, outputs one event at a time, even if the events are simultaneous [16, 17, 25, 26]. Such a representation needs to express starting time and duration explicitly, even though sometimes, duration is handled by treating the start and end of a note as separate events [16, 25].

These sequences of note events have often been modelled by recurrent approaches, such as recurrent neural networks (RNNs) [7, 9, 13, 14, 16–18]. Restricted Boltzmann machines have also been used, both as an output function of RNNs [7, 9] as well as on their own in deep belief networks (DBNs) [10] or with convolutional neural networks (CNNs) [11]. Other attempts use CNNs [20], generative adversarial networks (GANs) [22], and sigmoid belief networks (SBNs) [8]. The variational autoencoder (VAE) has been used recently [12, 15, 23–25], sometimes in sequence-to-sequence networks [23–25]. These are promising approaches because they learn a structured latent space with several attractive properties, for example the ability to interpolate between musical material and to perform arithmetic, such as imposing some attribute to the output by adding a so-called *attribute vector* to the latent code [24, 25]. Further details about the VAE can be found in [27] and in [28]. Among the state of the art in music generation is MuseNet [26], which is a *transformer* model (an attention-based feedforward neural network). MuseNet seems able to generate impressive, polyphonic compositions in different styles with up to 10 different instruments.

Where instruments are handled, their number is usually restricted to at most four [19–24], with the exception of MusicVAE which handles 8 in one publication [25], and MuseNet which handles 10 [26]. Furthermore, the modelling of instruments is usually hard-coded in the architecture [19–25] and not part of the data representation only, e.g., when an instrument does not play, the network outputs a token that implies silence in that voice. This poses a restriction on the number of instruments that can be used without making changes to the actual model.

3. MAHLERNET

The overall architecture of MahlerNet is inspired by MusicVAE [23–25], and its data representation and conditioning in output layers is inspired by BachProp [17]. MahlerNet is written in python using the deep-learning library Tensorflow, [4] and the Mido [5] library for MIDI input and output. Samples of material generated by MahlerNet can be heard online at `http://www.mahlernet.se`. Attendant code and instructions can be found at the project github repository. [6]

3.1 Data representation

MahlerNet uses a data representation where each played note (an "event") is expressed by four parameters. The *offset* parameter is the duration between the last event and the current one. The *duration* parameter is the duration of the current event. Both of these parameters index into a set of 60 possible values – from the 32nd note to a 32nd note less than five tied quarter notes – plus zero. The zero duration is used in the offset parameter for simultaneous notes, and in the duration parameter to signal an advancement of time only in a non-note event (more on the use of this in Sec. 3.2). The longest duration was chosen from observing how notes are often tied over bar lines in music. The *pitch* parameter is the pitch number of the current event, which indexes a set of 96 MIDI pitches from 17 to 112 (both inclusive). Finally, the *instrument* parameter identifies the instrument playing the current pitch, indexing into a list. The number of instruments modelled by MahlerNet is arbitrary but the current preprocessor instrument plugin class makes use of 23 instruments. The number of instruments is thus not hard-coded into the architecture itself and can be altered by simply altering the plugin class. The representation used by BachProp [17] includes the same parameters except for instrument.

This data representation can be referred to as "sequential polyphony", since simultaneous pitches are modelled sequentially but with offset parameters set to zero. This is in contrast to piano roll format where simultaneous pitches are represented in parallel. The representation used by MahlerNet alleviates problems that come with some other approaches, such as time slicing (e.g., rearticulation), and offers a clear conditioning order among the output notes, from early to later in time and from lower to higher in pitch. The latter order of conditioning is reasonable from a musical perspective where lower notes affect what is played in higher registers rather than the opposite.

Apart from the described data, each sequential step is also fed additional conditioning in the form of active pitches (currently turned on notes) and active instruments (the set of the instruments of the currently turned on notes). With one input representing the start of a piece of music (the context when generating the opening of a piece) and conditioning on position in the underlying metric pulse of eighth notes, the data representation of a piece of music is a sequence of 367-dimensional many-hot binary vectors.

3.2 Preprocessing

For several reasons, MIDI is often not a straightforward conversion from symbolic (written) music to sounding music, which results in both the starting point and the length of each detected event being subject to normalization. The notion of MIDI normalization is used in [17] and implies the definition and use of a set of rules or heuristics with which we interpret the original MIDI event. The normalized MIDI event is the same event but with starting time and duration adjusted with respect to some goal, in the case of MahlerNet what written note it most likely originated from. For example, even note-writing software will make use of existing articulations (e.g., *staccato*, *tenuto*) in the score when writing a MIDI file so that the mapping between the sheet music and the MIDI becomes ambiguous (an eighth note played tenuto might be easy to mistake for a quarter note played staccato). On top of that, many MIDI files on the Internet today are manually recorded from per-

[4] `https://www.tensorflow.org/`
[5] `https://mido.readthedocs.io/en/latest/`
[6] `https://github.com/fast-reflexes/MahlerNet`

formance with MIDI instruments which introduces imprecision in many ways. In orchestral MIDI files, to make the MIDI orchestra sound more like a real orchestra, creators often manipulate the music in many ways, sometimes adding things that are not in the score, or altering things, to make the synthesis sound better.

To deal with these problems, a study of the properties of notated and performed music have been used to create a series of heuristic scoring functions that ultimately decide how to normalize an event in a MIDI file [2]. The goal of this is to pick the most reasonable candidate given a series of candidates. Other heuristics govern which candidates to choose from for a given event. The importance of normalizing MIDI files is further discussed in [17].

To train MahlerNet, a dataset of MIDI files is preprocessed into the data representation described in Sec. 3.1. This entails mapping all MIDI instruments to instrument classes pre-defined in the preprocessor instrument plugin, and then ensuring that all pitches for a given instrument are within the prescribed range ($[17, 112]$). If some pitches are outside this range, they are mapped into the range by octave transposition(s). During the normalization of the offset and duration properties of each input MIDI note event, notes are considered in their context, and not in isolation.

In the preprocessor, each input MIDI file is divided into segments of events where one segment starts with the detection of a time signature MIDI event (or the beginning of the piece) and then contains all the events up until the next time signature (or end of the piece). Fifteen different time signatures are accounted for and segments beginning with other time signatures are ignored. This can cause an interruption in the sequence of consecutive segments, resulting in a single input piece being divided into several different uninterrupted sequences of segments.

Dataset preprocessing creates a collection of data representation files, where each file corresponds to an uninterrupted sequence of segments. During training, the preprocessor reads these files and divides them, on-the-fly, to batches of sequences of either a fixed or variable total duration (hyperparameter). The preprocessor version presented here uses the variable duration of a measure of music (this unit length is variable both because any number of events can actually exist in a measure but primarily because the actual length, in duration, varies depending on time signature). For the rest of this paper, a unit of music is considered a measure.

Input note durations and offsets with normalized durations longer than the longest duration are divided into several smaller, consecutive events in the data representation. First, the full measures that the duration covers are turned into different events. Next, the remaining parts of the original duration, the beginning and end of it, are normalized like all other durations with the exception that the one in the beginning has a fixed end at the end of the measure and vice versa with the remaining piece at the end. When this duration is an offset parameter, this process gives rise to some events that are non-note events that only advance time, and nothing else.

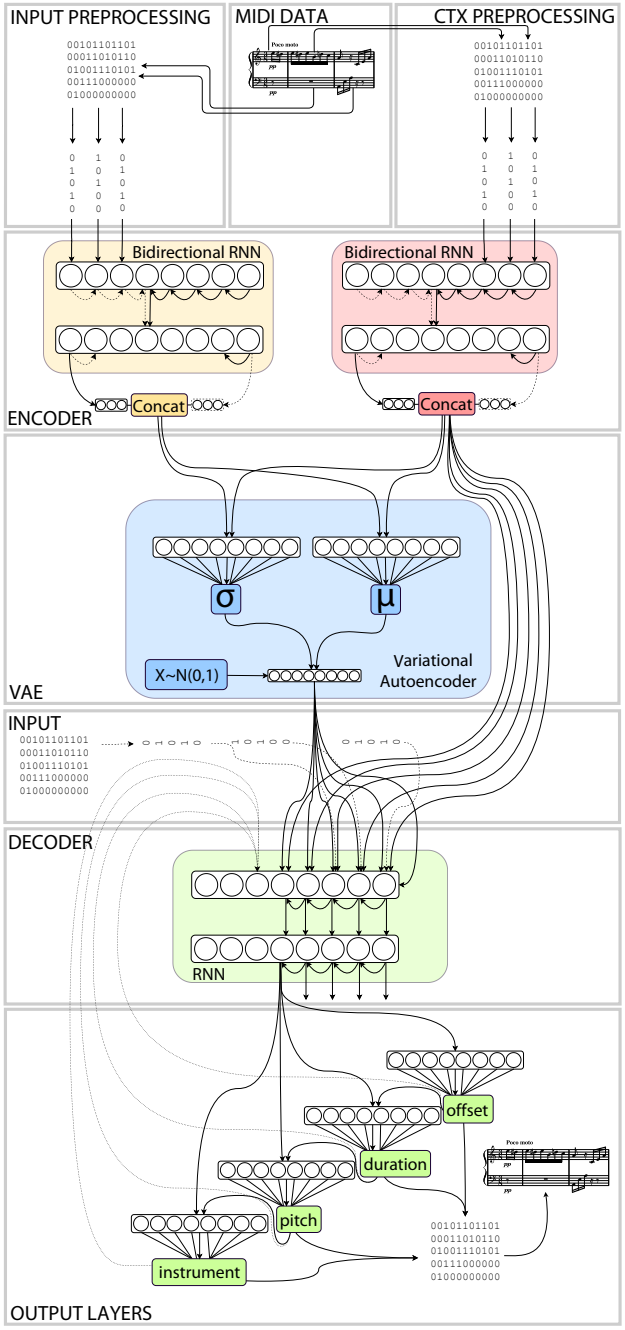


Figure 1. Schematic of the MahlerNet architecture.

3.3 Architecture

MahlerNet is a sequence-to-sequence network using a VAE to model the distribution of the latent state (like MusicVAE [23–25]). As can be seen in the schematic over the MahlerNet architecture in Fig. 1, the encoder side consists of two bidirectional RNNs where the first considers the measure to reconstruct and the second considers a context measure, e.g., the measure before the one to reconstruct. The final RNN states of both these bidirectional RNNs are concatenated and used as an input to the VAE, whose outputs are the parameters of a Gaussian distribution from which to draw a sample $\mathbf{z}$. The decoder then uses the sample $\mathbf{z}$ as a starting state, and the final states $\mathbf{c}$ of the context bidirectional RNN in the encoder as addi-

tional conditioning. This effectively results in the VAE of MahlerNet being a Conditional VAE (C-VAE) [29].

Each time the decoder RNN is advanced, it outputs first the offset parameter, then duration, then pitch and finally instrument. Sampling these parameters is done using distributions created from softmax output layers conditioned on the outputs from the previous output layers, according to the parameter order described above (which is also used in [17]). Denote an observation at state t as $\mathbf{x}_t = (o_t, d_t, p_t, n_t, ap_t, an_t, b_t)$ where o_t, d_t, p_t, n_t, ap_t, an_t, b_t denote offset, duration, pitch, instrument, active pitches, active instruments and metric position, respectively. The softmax layers output the conditional probability distributions of the parameters at the next state according to:

$$P(o_{t+1} \mid \mathbf{z}, \mathbf{c}, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t) \quad (1)$$

$$P(d_{t+1} \mid \mathbf{z}, \mathbf{c}, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t, o_{t+1}) \quad (2)$$

$$P(p_{t+1} \mid \mathbf{z}, \mathbf{c}, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t, o_{t+1}, d_{t+1}) \quad (3)$$

$$P(n_{t+1} \mid \mathbf{z}, \mathbf{c}, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t, o_{t+1}, d_{t+1}, p_{t+1}) \quad (4)$$

Active pitches and instruments, as well as metric position in the measure, are computed beforehand for training data, and are calculated on-the-fly from the sampled parameters during generation. MahlerNet is highly configurable and comes with optional batch normalization and dropout with the possibility to use both the β-*VAE* [30] and *free bits* [31] innovations when calculating the Kullback-Leibler (KL) term of the VAE loss function. These extensions effectively guides the tradeoff between a smooth and well-behaved latent state and high-quality reconstructions.

4. EXPERIMENTS

Three datasets were used for training and testing MahlerNet models: PIANOMIDI,[7] MUSEDATA[8] and a new dataset consisting of most movements of all ten symphonies by Gustav Mahler (MAHLER).[9] . Table 1 gives some statistics of these three datasets. The music in the datasets are used as is with no transposition to a common key. A hyperparameter grid-search was conducted for each dataset whereupon activation function, learning rate and RNN cell type were established.

Trained models were evaluated in two ways: interpolations between given measures from the training datasets, and random or seeded sampling. After sampling from the VAE, the MahlerNet decoder is allowed to decode for as many steps as needed to output events with a total duration less than or equal to a full measure, as dictated by the desired time signature. Each generated measure can be used as context for the next. Teacher forcing (target output is fed as input to update an RNN instead of its actual output) was used in training the decoder.

Training was done on a GeForce GTX1080Ti GPU and took at most two hours for any individual setup and dataset.

[7] http://www.piano-midi.de/

[8] https://musedata.org/ fetched from http://www-etud.iro.umontreal.ca/~boulanni/MuseData.zip

[9] MIDI files for MAHLER were gathered from two online sources: http://gustavmahler.com/midi.html and http://midi-orchestra.xii.jp/ta/mahler.htm

Dataset	PIANOMIDI	MUSEDATA	MAHLER
Measures	47544	37006	17450
Max length	116	158	303
Avg. length	14.16	26.25	29.01

Table 1. Statistics about the three datasets used in this work. The length of a measure is the number of events it contains.

4.1 Model training setup

MahlerNet models were trained separately with and without contextual input, after the initial hyperparameter grid search. Training was done with LSTM (long short-term memory) units in RNN layers, RMSProp optimizer and learning rate set to 0.001. Leaky ReLU (rectified lineary unit) was used as activation function with the exception of the VAE layers which used SoftPlus activation in the log-variance layer and no activation in the layer outputting the distribution mean. A batch size of 128 was maintained for all experiments except for the model trained on the MAHLER dataset with context conditioning; this model used a batch size of 96 due to size. The VAE had a single layer with 256 nodes whereas both encoder and decoder RNNs had two layers with 512 nodes in each. All models were trained with batch normalization in all dense layers, except for in the output layers and in the layers belonging to the VAE. The batch normalization takes place before the activation function but after the multiplication (as opposed to after the activation function as sometimes advocated).

We trained models without contextual input conditioning without dropout in a way similar to MusicVAE [23, 24], annealing the VAE loss function at rate 0.00001 with a β parameter of 0.2, calculated with 48 free bits. These settings favor reconstruction accuracy of training data over the quality of the latent space. By "quality" we mean that the decoder has been exposed to and has had the opportunity to be trained on all (most) points of the latent space, and that the transition between nearby points in the latent space results in gradual and small changes in the output space.

These models are meant to overfit, but to reconstruct input training samples accurately and interpolate between them in the latent space in a meaningful way. Interpolations are done by taking two endpoint measures from the training data and run them through the trained encoder and the VAE to generate a distribution for each. After sampling from both, ten steps of interpolation move from one sampled latent vector to the other. Because of the overfitting objective, no validation set was used and all models were trained for a fixed number of epochs (50).

We trained models using context conditioning with a dropout rate of 0.35 (probability of dropped connection) after the activation and with the VAE loss disregarding beta annealing or free bits. A validation set of 10% of the training data was used to determine when to abort training. These models are meant to create coherent sequences of measures sampling one measure at a time using the last measure as context. They are initialized either with a random vector with the same size as $\mathbf{z}$ sampled from a standard Gaussian

Dataset	Teacher forcing	
	with	*without*
PIANOMIDI	98.29%	80.94%
MUSEDATA	96.11%	45.87%
MAHLER	94.43%	27.20%

Table 2. Comparison of reconstruction accuracy of pitch with and without teacher forcing for models without contextual conditioning.

distribution (random sampling) or with a $\mathbf{z}$ vector sampled from the VAE originating from some chosen input and context vectors (seeded sampling). The context vector is either a "start" measure (only containing one step with only the "start" class set) or with some other measure, e.g., the one preceding the one to compose, for seeded sampling.

4.2 Results

Table 2 summarizes the results of experiments with the three datasets. Models trained without context conditioning were all trained for 50 epochs at which point almost no further improvement was observed. When attempting to reconstruct the training data without teacher forcing, a heavy drop in performance is visible. This is proportional to the average and maximum length of the sequences, as shown in table 1, in the different datasets.

Interpolations with these models seem successful for short sequences (< 30 events), and endpoints are properly reconstructed with a plausible transition between them. An example of this can be seen in figure 2. With longer sequences, the transitions are less plausible and musical deviations from both endpoints take place. Sometimes, not even the endpoint measures are properly reconstructed even though often, the first notes are correct.

We stopped the training of models with context conditioning once the performance on validation sets started to drop. For PIANOMIDI, MUSEDATA and MAHLER, this resulted in the models training for 15, 21 and 31 epochs. With these models, exact reconstruction is worse than with previous models but quite to the contrary, random sampling works better and the decoders produce plausible music output.

We perform interpolations in 10 steps. Generated samples have 10 or 100 bars. We sampled all material at softmax temperatures around 0.9 for models trained on PIANOMIDI and MUSEDATA, and at around 0.6 for the MAHLER model.

5. DISCUSSION

The strongest argument in favour of using MIDI for training data is the availability where thousands of MIDI files can be found online. However, in principle, it is a representation of performed music rather than written. As a result, a very large responsibility is placed on the preprocessor when using this format to model written music.

The reconstruction capability of models trained without context conditioning drastically deteriorates as the number of events increases. In accord with what is said with respect to MusicVAE in [23–25] and brought up elsewhere [32], what is believed to happen here is a phenomenon named "mode collapse": the decoder learns to neglect the latent code and only base its output on the teacher-forced input. Evidently, when not even the endpoints can be properly reconstructed, it affects interpolations in a negative way. Nonetheless, often the measures resulting from these interpolations are plausible as music, however not constituting a smooth and continuous transition between the endpoints as desired.

Random sampled measures (generate a random $\mathbf{z}$ vector and decode it) from these models however do not pass off as music and neither of the output parameters, inspected in sequence, are realistic. This is quite in contrast to the result with seeded samples and it turns out that the lack of regularization of the latent space is the cause for this; the use of a β weight and free bits results in better reconstructions but less smooth latent space. Here it is obvious that it is fairly simple to sample a latent vector that the decoder has never experienced and it is thus uncertain how it will be decoded. However, mode collapse aside, interpolations are still well-behaved and since the MusicVAE is used with interpolations, the slackened regularization of the latent space appears reasonable in that case. For MahlerNet when used with context conditioning however, random sampling is an important factor and so these modifications of the VAE loss must be left out.

At the cost of exact reconstruction, the removal of these components effectively improves the outcome of sampling procedures in the models trained with context conditioning. With consecutive generation of multiple measures, no matter if the decoder is fed a random latent vector or seeded with one coming from a given input and context, the music is coherent in terms of dynamic tension, active instruments, intensity, tonality and rhythm.

We can find clear differences between outputs depending on training data. The model trained on MUSEDATA produces music that sounds Classic and Baroque whereas the model trained on MAHLER generates samples that sound more dissonant and characteristic of music of the Romantic period – both in line with the music found in the datasets. The output from the model trained with PIANOMIDI has almost always piano only whereas instruments such as tuba and trombone, which are quite common in the Romantic era but not in the Classic and Baroque eras, are commonly seen in output from the model trained on MAHLER but not in the other models.

Furthermore, instruments tend to stay within their real ranges and groups of instruments that typically play together during the different eras are active together in generated material as well. Examples of this is the chamber setting with strings, woodwinds and continuo in Baroque music and woodwind sections present in Classic music where strings otherwise dominate. In the 100-bar samples, contrasting parts are present both in terms of instrumentation and homophony versus polyphony. Sections where all instruments play often have brass, horns and timpani added much like in a real scenario with beats 1 and 3 typically

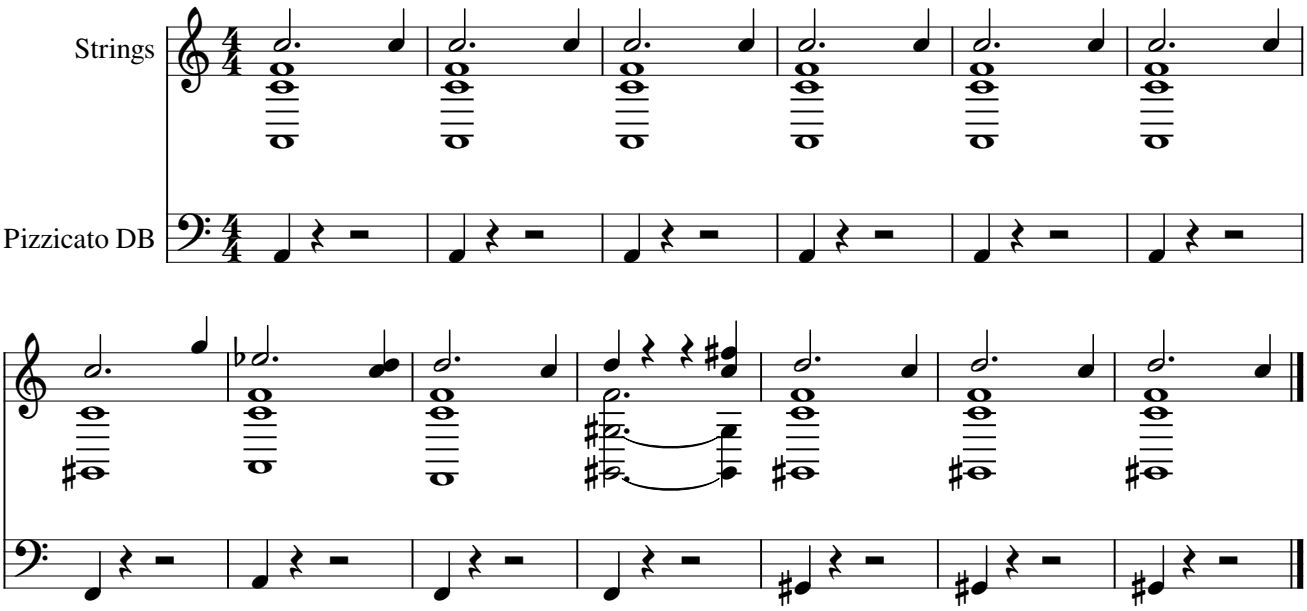


Figure 2. Interpolation (sample 3-1 at `http://www.mahlernet.se`) between two measures from the 4th movement of Mahler's fifth symphony, in a model trained without context conditioning.

emphasized.

Rhythmic consistency is also present and when MahlerNet is seeded with a measure containing particular rhythms (for example triplets) these often prevail for several measures. Finally, phrase endings with V-I motion are common, however not as common as in real Classic music.

Despite many good qualities, MahlerNet does only very occasionally produce music with motivic or thematic content that persist over many bars. This is a common defect in music AI based on neural networks. Nevertheless, in the case of MahlerNet, this tendency is almost expected since the temporal receptive field only spans one measure of history; expanding this context is necessary to improve the conditions.

6. CONCLUSIONS

MahlerNet is a new neural network based on MusicVAE and BachProp with a new data representation and preprocessor based on BachProp. MahlerNet is a first attempt at modelling orchestral music using deep neural networks with an architecture that allows for an unlimited number of instruments. Even with the limit of the current preprocessor, there are, to the authors' knowledge, many more instruments available than in any earlier publication. Music generated by MahlerNet shows coherence and long-term structure in many aspects, not the least with respect to instrumentation. Furthermore musical style is clear and correlates with training data and the music is both plausible and shows qualitative pregnancy. Reoccurring themes and motives, important aspects of music, are however rare findings in samples, perhaps due to the almost naively short context of a measure that is fed as conditioning.

The next step in developing MahlerNet is to increase its temporal receptive field by experimenting with longer contexts. The mode collapse problems in the decoder should be addressed, as well as steering the generation process with more conditioning. Further research on the impact of batch normalization on the latent space is also desirable since there are some indications that the structure of the latent space suffers from the use of it. On a more general level, an interesting direction of the future of music AI is hierarchical and transformer-based architectures.

Acknowledgments

The authors would like to thank the creators of the Mahler symphonies in MIDI format found online. [10] EL wrote most of this paper, and wrote all computer code. BLTS supervised of EL's master's thesis research, and initiated and edited this paper.

7. REFERENCES

[1] B. L. Sturm and O. Ben-Tal, "Let's Have Another Gan Ainm: An experimental album of Irish traditional music and computer-generated tunes," KTH Royal Institute of Technology, Tech. Rep., 2018.

[2] E. Lousseief, "MahlerNet - Unbounded Orchestral Music with Neural Networks," Master's thesis, Royal Institute of Technology, Stockholm, Sweden, 6 2019.

[3] L. Hiller and L. Isaacson, *Experimental Music: Composition with an Electronic Computer*. New York, USA: McGraw-Hill Book Company, 1959.

[4] J. D. Fernández and F. Vico, "AI methods in algorithmic composition: A comprehensive survey," *J. Artificial Intell. Res.*, vol. 48, no. 1, pp. 513–582, Oct. 2013.

[5] J.-P. Briot, G. Hadjeres, and F. Pachet, *Deep learning techniques for music generation*. Springer, 2019.

[10] `http://gustavmahler.com/midi.html` and `http://midi-orchestra.xii.jp/ta/mahler.htm`

[6] R. Dean and A. McLean, Eds., *The Oxford Handbook of Algorithmic Music*. Oxford, UK: Oxford University Press, 2018.

[7] N. Boulanger-Lewandowski, Y. Bengio, and P. Vincent, "Modeling Temporal Dependencies in High-Dimensional Sequences: Application to Polyphonic Music Generation and Transcription," *Proc. Int. Conf. Machine Learning*, 2012.

[8] Z. Gan, C. Li, R. Henao, D. Carlson, and L. Carin, "Deep Temporal Sigmoid Belief Networks for Sequence Modeling," in *Proc. Int. Conf. Neural Information Process. Systems*, 2015, pp. 2467–2475.

[9] Q. Lyu, Z. Wu, J. Zhu, and H. Meng, "Modelling High-dimensional Sequences with LSTM-RTRBM: Application to Polyphonic Music Generation," in *Proc. Int. Conf. Artificial Intell.*, 2015.

[10] F. Sun, "DeepHear - Composing and harmonizing music with neural networks," https://fephsun.github.io/2015/09/01/neural-music.html, 2015, accessed: 2019-10-29.

[11] S. Lattner, M. Grachten, and G. Widmer, "Imposing higher-level Structure in Polyphonic Music Generation using Convolutional Restricted Boltzmann Machines and Constraints," *CoRR*, vol. abs/1612.04742, 2016.

[12] J. A. Hennig, A. Umakantha, and R. C. Williamson, "A Classifying Variational Autoencoder with Application to Polyphonic Music Generation," *CoRR*, vol. abs/1711.07050, 2017.

[13] D. D. Johnson, "Generating Polyphonic Music Using Tied Parallel Networks," in *EvoMusArt2017*, 2017.

[14] F. Liang, M. Gotham, M. Johnson, and J. Shotton, "Automatic Stylistic Composition of Bach Chorales with Deep LSTM," in *Proc. Int. Symp. Music Info. Retrieval*, 2017.

[15] R. Sabathé, E. Coutinho, and B. Schuller, "Deep recurrent music writer: Memory-enhanced variational autoencoder-based musical score composition and an objective measure," in *Proc. Int. Joint Conf. Neural Networks*, May 2017, pp. 3467–3474.

[16] I. Simon and S. Oore, "Performance RNN: Generating Music with Expressive Timing and Dynamics," https://magenta.tensorflow.org/performance-rnn, 2017, accessed: 2019-10-29.

[17] F. Colombo and W. Gerstner, "BachProp: Learning to Compose Music in Multiple Styles," *CoRR*, vol. abs/1802.05162, 2018.

[18] H. H. Mao, T. Shin, and G. W. Cottrell, "DeepJ: Style-Specific Music Generation," *CoRR*, vol. abs/1801.00887, 2018.

[19] G. Hadjeres and F. Pachet, "DeepBach: a Steerable Model for Bach chorales generation," *CoRR*, vol. abs/1612.01010, 2017.

[20] C.-Z. Huang, T. Cooijmans, A. Roberts, A. Courville, and D. Eck, "Counterpoint by Convolution," in *Proc. Int. Symp. Music Infor. Retrieval*, 2017.

[21] H. C. Chu, R. Urtasun, and S. Fidler, "Song From PI: A Musically Plausible Network for Pop Music Generation," *CoRR*, vol. abs/1611.03477, 2016.

[22] H.-W. Dong, W.-Y. Hsiao, L.-C. Yang, and Y.-H. Yang, "MuseGAN: Multi-track Sequential Generative Adversarial Networks for Symbolic Music Generation and Accompaniment," *AAAI 2018*, 2017.

[23] A. Roberts, J. Engel, and D. Eck, "Hierarchical Variational Autoencoders for Music," in *Workshop on Machine Learning for Creativity and Design, NIPS*, 2017.

[24] A. Roberts, J. Engel, C. Raffel, C. Hawthorne, and D. Eck, "A Hierarchical Latent Vector Model for Learning Long-Term Structure in Music," *CoRR*, vol. abs/1803.05428, 2018.

[25] I. Simon, A. Roberts, C. Raffel, J. Engel, C. Hawthorne, and D. Eck, "Learning a Latent Space of Multitrack Measures," *CoRR*, vol. abs/1806.00195, 2018.

[26] C. Payne, "MuseNet," https://openai.com/blog/musenet/, April 2019, accessed: 2019-09-19.

[27] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," *CoRR*, vol. abs/1312.6114, 2013.

[28] D. Jimenez Rezende, S. Mohamed, and D. Wierstra, "Stochastic Backpropagation and Approximate Inference in Deep Generative Models," *CoRR*, vol. abs/1401.4082, 2014.

[29] K. Sohn, H. Lee, and X. Yan, "Learning Structured Output Representation using Deep Conditional Generative Models," in *Proc. Advances in Neural Info. Process. Systems*, 2015, pp. 3483–3491.

[30] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "β-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework," in *Proc. Int. Conf. Learning Representations*, 2017.

[31] D. P. Kingma, T. Salimans, and M. Welling, "Improving Variational Inference with Inverse Autoregressive Flow," *CoRR*, vol. abs/1606.04934, 2016.

[32] S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Józefowicz, and S. Bengio, "Generating Sentences from a Continuous Space," *CoRR*, vol. abs/1511.06349, 2015.

THE VISUAL REPRESENTATION OF SPATIALISATION FOR COMPOSITION AND ANALYSIS

Mattias Sköld
KMH Royal College of Music, Stockholm
KTH Royal Institute of Technology, Stockholm
mattias.skold@kmh.se

ABSTRACT

The motivation for this text is my ongoing research into creating a uniform and comprehensive notation system for music regardless of sound sources, acoustic or electronic. I propose a way to visually represent the positions and movements of sound in composition and analysis of music which in different ways utilises space as a parameter. I address a number of aspects of spatialised music to take into account when defining a notation language for the music. I suggest visually representing the room in different ways depending on how the music relates to the concept of space: as projections from the center of a sphere for more structural work, or as coordinates in a cubic room for works that depict a physical or imagined space. I also show how these descriptions of space are integrated with my existing notation system.

1. INTRODUCTION

Though space has always been an important aspect of music it is with the development of electroacoustic music that spatialisation has become a composition parameter to be fully integrated with the compositional structure of Western music today. In his visionary text, Varèse famously spoke of sound projection becoming the fourth dimension of music [1]. However, if we consider early concepts of music, with music inseparable from dance [2], space and movement as part of the compositional structure of music is an ancient idea. Despite space becoming an increasingly important aspect of electroacoustic music composition, there is little in terms of established notation for the movements and positions of sound.

In this text, I will describe recent developments in the visual representation of spatial sound. I will discuss some concepts that I found useful to consider when deciding on the functionality of the spatialisation notation system to integrate with my existing system for analysis and composition [3]. Finally, with these concepts in mind, I will describe some of the functionality of my own solutions for visually representing spatial sound. What I need is not necessarily a complex system but rather a way of notating sounds that both my students and I will find intuitive to work with. Providing notation tools for spatialisation to composition students, will make it easier for them to formulate and structure composition ideas related to space.

2. BACKGROUND

Even if we disregard the role of dance in the historical development of music, space has always had a strong impact on this art form. The most obvious example is the influence of room acoustics over both timbre and durations. In Messiaen's *Apparition de l'Eglise éternelle* for Organ [4], the large church acoustics usually surrounding its performance make the transition from a dense dissonant chord to a consonant chord a gradual process since each chord lingers in the room long after its release on the keyboard.

When we discuss spatial aspects of music we are faced with a similar problem to that of musique concrète: how to relate to a musical parameter that, like timbre, so clearly refers to phenomena in the physical world. Annette Vande Gorne speaks of the space illusion (L'espace illusion) as a category [5] and Natasha Barrett, when discussing approaches to space in music, speaks of the illusion of a space and the allusion to a space [6]. Marije Baalman, in her text on spatial composition techniques and technologies, notes that spatialisation technologies are usually oriented towards recreating the acoustics of the physical world while composers in their spatialisation techniques are more inclined to create something new, not necessarily relying on what would be considered realistic or not [7]. A good example is how reverb units often specifically name the type of room applied to the sound in their presets even when they are made up of filters rather than recorded impulse responses from actual rooms. So a reverb preset with a long reverb tail is not thought of simply as a reverberant space of a particular size but more specifically a cathedral, suggesting that adding this effect has religious connotations.

Peters, Marentakis and McAdams find in their qualitative and quantitative analysis of spatialisation technologies and techniques that only 38 % of the surveyed electroacoustic composers use notation for spatial aspects of the music [8]. This should come as no surprise considering that a score is not really necessary for the production of electroacoustic music. In fact, that this music is not easily reduced to notation has been highlighted as a strength of the genre [9]. However, in the aforementioned survey several composers cited partly a lack of notation standard as the reason for not using notation for spatialisation [8]. There is indeed no

standard, but if we imagine a connection between the notation of a musical parameter and the graphical user interface used to control the same parameter, the homogeneity of the graphical layouts of interfaces in spatialisation tools would suggest an opportunity to find a notation that should look familiar to composers working with spatialised sound. In other words, the graphical user interfaces found in tools like ambisonics panning software suggest working solutions for visually representing positions and trajectories of spatial sound layers in comprehensible ways. In existing solutions there are indeed connections between notation and the software used for rendering the spatialisation.

3. EXISTING NOTATION SOLUTIONS

While there is no standard for notation of spatial sound, there are some very interesting systems and techniques for the visual representation of space available. The earliest more widely used systems can be found in the world of dance choreography with its long history of notating movements and positions in space along a musical score. The dances in the courts of France and Italy in the 17th Century emphasised the paths of the dancers over the dance floor, giving rise to notation like the systems of Feuillet, Lorin and Landrin representing the dancers' paths as floor patterns where the dancers' body movements are traced over their trajectories in the floor space [10]. Labanotation, perhaps the most famous notation system for dance, also has symbols and methods for visually representing the floor patterns of the dancers [11]. Labanotation and its derivative, Laban Movement Analysis, have been used in various movement related research such as in Abe, Laumond, Salaris, and Levillain's application of labanotation for humanoid robot movements [12].

Many composers work with their own tailor-made solutions for spatial notation, such as Karlheinz Stockhausen's indicators for loudspeakers in the composition process of Kontakte [13] or Pierre Henry's solution in the realisation score of Messiaen's Timbres-Durées [14] where different tape tracks have their own specified spatial identity. I have also throughout the years seen many interesting solutions from my students coping with representing spatial data on paper.

In terms of current research during the last five years there have been some interesting advances in the field of spatial sound notation, all related to technologies for sound projection:

3.1 SVG to OSC Transcoding

Starting from the idea that electroacoustic music is not usually notated but rendered, Rama Gottfried suggests working with graphical software like Adobe's *Illustrator* to create graphical scores with maximum flexibility. The scores are then output as Scalable Vector Graphics (SVG) to be interpreted as OSC to control sound e.g. in Max [15].

The advantages of this system is that there is little compromise in terms of the quality and possibilities of the graphical output, while the data exchange format itself is readable and possible to understand. Also, thanks to XML's tree structure model one can create hierarchies of graphic objects [15]. While I will not go into more detail regarding this solution in this text, this system should be of great interests to composers who continue to rely on their own systems of representation for their work.

3.2 SSMN and the Taxonomy and Notation of Spatialisation

To my knowledge, the most detailed symbolic notation system for spatial sound developed so far is the *Spatialisation Symbolic Music Notation* (SSMN) system from ICST in Zürich. The system builds on a taxonomy of spatialisation that takes into account the different uses and needs from composers working with spatialisation today [16].

The system is based on room descriptors and descriptors of sound sources with their own symbols. The descriptors can have assigned properties such as numeric parameters, trajectories, names and flags for more detail, e.g. when using the system to control the dedicated spatialisation engine for panning over a multi-speaker system using ambisonics. For the first case studies, a version of open source notation software *MuseScore* [1] was modified to handle the notation as well as OSC communication, not otherwise part of the software's functionality.[17]

3.3 SPAT-SCENE

Garcia, Carpentier and Bresson relied on interviews and studies of score drafts of eight composers working with spatial sound to develop new functionality for spatialisation in OpenMusic with focus on aspects such as the possibility of graphical and gestural input, notation and time management for control of the spatial rendering [18].

One key problem they address is the difficulty of defining the temporal aspects of trajectories in the compositional process. Their solution, *SPAT-SCENE*, is a graphical user interface for the spatialisation engine *Spat* [2] in OpenMusic [3] [18]. The visual representation can then be exported as SVG for import in notation software.

3.4 About These Solutions

The SVG to OSC transcoding concept meets the ever present need for composers to feel completely free to go their own ways with their notation and still be able to apply the results to current digital audio projection tools. SSMN meets the need for standardised notation symbols for spatialisation. This is valuable for composition and perhaps even more so for analysis. By subjecting different works to a common notation system, it is possible to analyse and compare works on a deeper level than if the works have different tailor-made systems for the visual representation of their spatial aspects. SPAT-SCENE is not a system of symbolic notation like SSMN, but it caters to the specific needs of electroacoustic composers with regard to input of spatial trajectories, temporal control of the composition,

[1] https://musescore.com
[2] https://forumnet.ircam.fr/product/spat-en/
[3] http://repmus.ircam.fr/openmusic/home

the rendering of the spatialised sound in Spat and the notation possibilities. See Figure 1 for a comparison how trajectories can be visually represented by SSMN and SPAT-SCENE, bearing in mind that both systems are highly flexible and the output of both systems can look quite different depending on how you work with these tools.

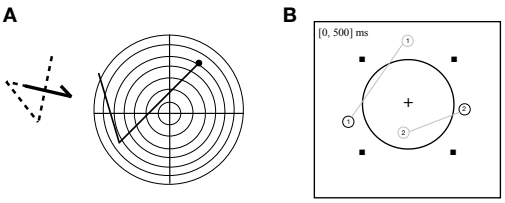


Figure 1: Examples of the kind of displays that can be used in the SSMN system (A) and SPAT-SCENE (B) for visually representing trajectories of sound sources.

4. ASPECTS OF THE VISUAL REPRESENTATION OF SPACE

4.1 Perspective

A major difference between dance notation and spatialisation notation is that of perspective. At the heart of labanotation is the movement of one's body. Composers and music software, on the other hand, relate movement to an ideal listener position, the sweet spot. I believe it would be helpful for the notation of spatial sound to include the possibility of multiple perspectives in the visual representation of music. Just like the dance choreographer must think of what each individual dancer is doing while maintaining control of the overall scene it would make sense to visually represent both the individual movements of sounds while also displaying the composite scene image with all sound layers and components in one frame. For the individual sound layer, we're interested in the start, execution and completion of each individual movement, while the composite image of all movements will help us keep track of how the individual movements affect the overall sound image in each instance. This is important particularly with regard to the concept of balance.

4.2 Balance

One important reason for the need for including a composite image is to keep track of the balance of the spatial audio image. In stereo mixing, maintaining an overall left-right balance is paramount. A recording that as a whole tilts to the left would be considered a mistake. This is also reflected in our behaviour as listeners. If a part of an acoustic concert is performed on the far left side of the stage we turn to face the musician not only to see the performance but also to "balance" the stereo image of our listening experience. In stereo recordings, instruments panned to the left are usually complemented by an equivalent instrument to the right – we can imagine a kind of spatial counterpoint theory where sound appearing on one side of the audio image is commonly complemented with an equal sound on the other side.

The balance of the left-right stereo field is in a way easy to comprehend since the left and right sides of the audio image are equal to our ears. Front-back and up-down shifts are more complicated relationships in terms of balance, since shifts in these fields have hierarchical consequences: A sound right in front of us has greater significance than a similar sound of equal loudness and distance from the rear. Similarly I would argue that a sound in the same height as our ears seems more urgent than something from below or above. While this should be visually apparent by simply looking at the shape of our outer ears, this could also be reflected in the visualisation of the notated composite image both for 2D and 3D displays of the listening field. The notation can otherwise fool the composer into believing that all panning angles are equal. Obviously, for installations and other works without the audience firmly positioned in one direction such a notational feature would not be needed.

4.3 Room and Resonance

Different reverbs with different decay settings on the same recording at the same time is common in stereo mixing today. Also, placing different reverbs in different sets of loudspeaker pairs to simulate the opening of a room inside the room is a suggested spatial technique in Roads and Strawn's Computer Music Tutorial [19]. But when defining the resonance of a sound source for notation it's important to remember that for acoustic music each instrument body is also a small room inside the room acting as a resonating chamber for whatever air or string that has been put in vibration. Since most electric instruments don't have resonating bodies [4] , reverbs are sometimes added not to act as the musical space but as the resonating body of the instrument. There is a reason why a guitar amplifier has its own dedicated mono reverb - without it the guitar risks sounding muted and will not match the sound of any acoustic instruments of the ensemble. Accordingly, the visual representation of reverberation or other resonators should specify whether the effect is that of a resonating body for an individual sound source or as the musical space of the music. There is a grey area here since loud cathedral reverbs are sometimes used as inserts on a single sound source. Also, there are cases when reverb effects act like a synthesis technique rather than a room and should therefore be notated as timbre rather than space–the echo chamber sounds of Stockhausen's *Studie II* is a good example of this [20]. Electroacoustic music is also interesting with regard to the problem of room and resonance in that while electronic music has tone generators that produce sound with no resonating rooms or resonators, the sound material of musique concrète may include both already from start in its material. This of course relfects the contrasting nature of the two fields in terms of their relations to their material.

4.4 Sound or Action

In the existing systems exemplified above, the notation of action and result is one and the same thing, i.e. the nota-

[4] The ondes martenot is a lovely exception

tion also controls the spatialisation rendering. But for the sake of individual sounds and/or performers moving along individual trajectories it can be necessary to extrapolate individual instructions in a way that makes sense for that sound/performer. This has been put to the test in a case study involving the SSMN system where spatial notation was assigned to different performers moving loudspeakers [17]. While the movement of people rhather than sound is not very common, some composers have put this ideas to good use, as in Benedict Mason's 2008 piece *Music for Moderna Museet* [21], where he relied on individual map fragments of the museum building on cards for the individual orchestral members to know where to move during the performance. Also, in performances of Rebecca Saunders' installation piece *Chroma* [22], positions and movements of musicians and audience is a part of the experience.

4.5 Imagined or Projected Space

Methods for panning sound in software and hardware place the sound sources at the centre of the listening space by default, and to position the sound somewhere else means specifically projecting the sound at a position in the possible listening field defined by the loudspeaker system, e.g. in ambisonics panning software, positions are defined by their horizontal (azimuth) and vertical (elevation) angles and their distances from the listener. For much spatialisation work this is practical both from a conceptual and practical viewpoint.

But for compositions where sound objects are imagined and positioned in a physical or imaginary space, when sound objects or indeed musicians are moving across the room, we're dealing with a different kind of spatialisation where each object is not necessarily best defined by its position in relation to the centre. The difference between these approaches is similar to that of Vande Gorne's categories *geometric space* and *illusory space* [5] where the former denotes spatialisation from a structural point of view. One could argue that, from a notational viewpoint, this is a superficial difference considering the fact that values can easily be converted between angles and distances and room coordinates. But if the reason for visually representing the spatial sounds is for us to make sense of a particular composition for the sake of analysis or composition this is an important conceptual difference.

4.6 Distance or Amplitude

When we encounter acoustic sound sources in our surroundings, the difference between dynamics and distance is often evident thanks to our knowledge of how sound loses energy with distance and how sound sources change at low and high sound levels respectively. But for acousmatic music the difference may not be obvious. We are not necessarily familiar with the sounds presented, and even if we are familiar with the sounds, they do not necessarily act as acoustic sources of sound at different volume levels or distances. This becomes a problem for credible three-dimensional panning with respect to distance [23]. And for the purpose of analysing or transcribing acousmatic music we have to rely on our perception to decide when amplitude changes are related to distance.

4.7 Articulation or Change of Position

When defining and notating movements of sound it makes sense to adapt the same approach as with the notation of traditional musical parameters such as pitch and dynamics: Just like there's a structural difference between a vibrato and a glissando or between a tremolo and a crescendo when writing for voices and instruments, there's a difference between sounds that are articulated through movement patterns such as continuous rotation and auto-panning, and sounds moving from one position to another. It is the difference between articulation and a change of position. Even for a heavy vibrato from a singer you won't expect this to be reflected in the sheet music since the pitch from a structural point of view is considered to stay the same regardless of the articulation. (See [24] for a discussion on such notation discrepancies.) Likewise, a sound rotating around the listener can be thought of as a sound in orbit rather than a repositioning of the sound from one place to the other. This should be reflected in the notation.

4.8 The Problem of 3D Represented on Paper

A major problem of visually representing spatial sound as printable notation is the fact that music notation, as well as audio editors, tend to favour one dimensional parameters to be plotted over a time line. This works for displaying simple left-right panning, but adding another dimension for front-back positioning immediately complicates things, not to mention the problem of displaying movement in a 3D space. Many use top view 2D symbols placed above or below the staff, as in the case of both SPAT-SCENE and SSMN which relies on radar style displays as specifiers for descriptor symbols like the *bezier* symbol [17]. But to add further dimensions we may need to consider going beyond spatial dimensions and e.g. incorporate darkness-brightness differences or colour saturation.

5. MY NOTATION SOLUTIONS

The purpose of this text is not in any way to disprove previous research or show how the systems cited here fail in any regard. Also, those systems have possibilities not covered here because they are beyond the scope of this paper. My major concern when formulating my own solutions is to find notation methods that are easy to grasp so that composition students with little or no experience with electroacoustic music will be able to start working with the notation and achieve their desired results. But in order for the system to be of any interest for the students the system must also be versatile and open for custom solutions.

When deciding what spatialisation notation to go with my overall notation system for composition and analysis, as with the previous work [25][3], I'm more interested in finding existing working solutions as starting points rather than inventing a new system from scratch. Of the three existing solutions briefly presented above it is a combination

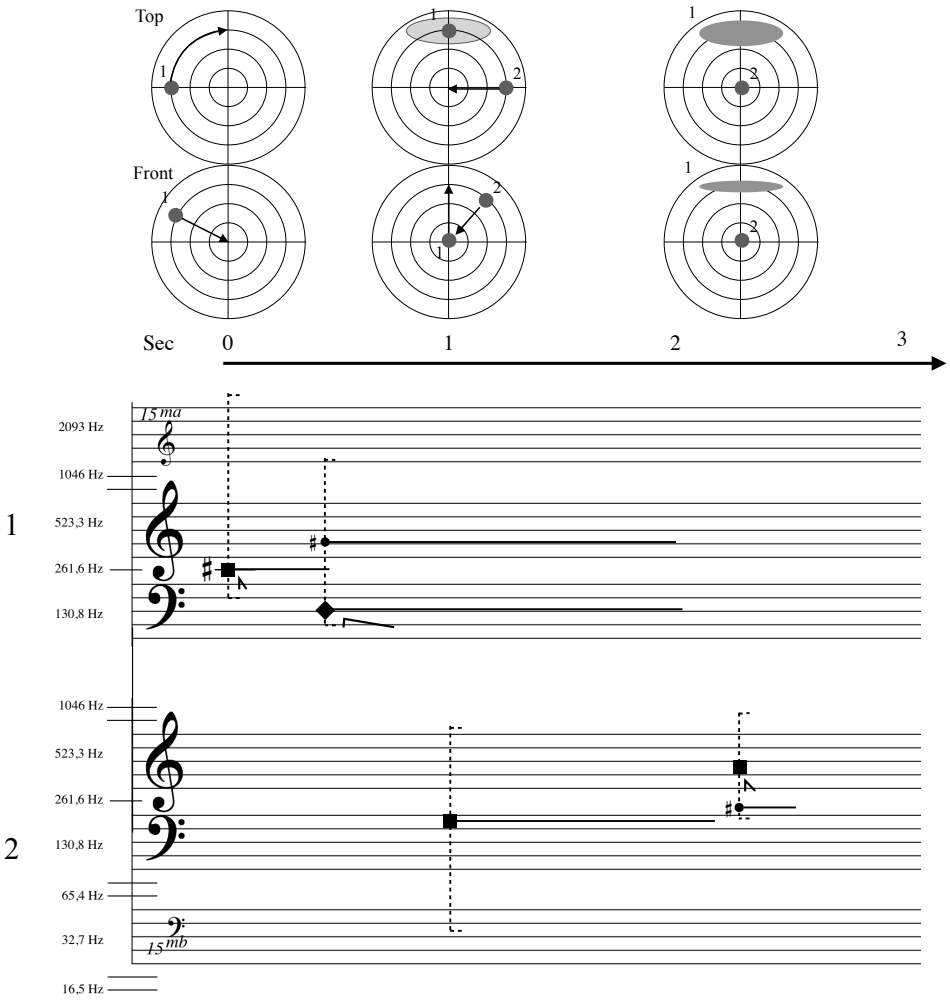


Figure 2: Example of composite displays of positions and movements for staves 1 and 2. In this case, the ambisonics style sphere indicators where used. The second indicator displays a widening of the field occupied by the sound of staff 1, which has happened by the time of the third indicator.

of the SSMN and the SPAT-SCENE that I find most suitable for my purposes. The SVG to OSC solution I find to be more useful for composers who define their own systems and symbols as they go, which remains an important part of electroacoustic music creation among composers [8]. What I will propose therefore is the following:

5.1 Composite Displays of the Sound Image

For key moments or continuously, depending on the composer's wishes, there will be frames that, just like the graphic output of SPAT-SCENE, indicate all sounds' positions at the time. This is particularly important for analysis and composition when exploring or working with the development of the sound image as a whole. For each sound layer indicated their position and their immediate plans for moving and/or changing size and definition are displayed.

It is important how this composite image is displayed and in what format the positions and movements are conveyed. Bearing in mind the discussion of sound projection versus room description above, I introduce the possibility to select one or the other depending on the focus of the music. When visually representing experimental and structural work on spatialisation in works aimed for ambisonics rendering, circular indicators from top view and front view would be used (see Figure 3 A). Positions would then be indicated in the style of ambisonics panning using AED values (Azimuth, Elevation and Distance). On the other hand, pieces that work like a kind of sound choreography, where the physical or imagined positions in the room are in focus, a square-shaped grid would be used to display room positions and values would be indicated using cartesian coordinates (see Figure 3). The need for these two perspectives is of course not a new idea and technologies used for spatialisation often provide both types of displays.

Regarding perspective, both systems would rely on a 0 point in the middle, but while the circular sound projection oriented system, through its design, has a very obvious "bulls eye" center, the cartesian grid style display has a mere "+" sign to indicate the (0,0) coordinates of the image.

In Figure 3 only the first display has a front view below

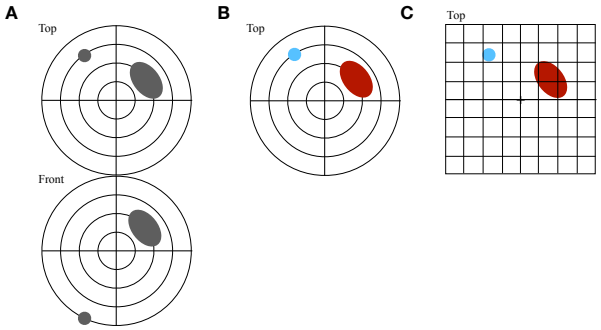


Figure 3: Examples of my solution for visually representing the composite information on positions and movements of the sound layers in the score. Examples A and B are aimed for a sound projection focus, relying on AED (angles and distance) while C is focused on the illusion of the room created by the spatialisation, relying on XYZ (cartesian coordinates). For B and C the elevation of the sound is indicates with a colour scale instead of a separate front view display. From dark to bright red means sounds are moving up from ear level, and from dark to bright blue means sounds are moving down from ear level. The indicator with the smaller size has a more clearly defined position in space, while the larger indicator suggests occupying a larger area of the listening space

to indicate elevation. This is because examples B and C both rely on colour to display elevation or height. I was inspired by the cross-modal correspondence between colour and temperature [26] for selecting red for high positions and blue for low positions. Brighter red indicators suggest sound sources higher up, while brighter blue suggests sound sources lower down. The darker they get, they move closer to ear level. This colour scheme may be subject to change following the case studies with students. An alternative selection of colours could be the ones used in altitude meters on airplanes, where blue represents the sky and brown represents the ground.

5.2 Notating movements

Much advanced spatialisation relies on the programming of trajectories, which is not only a problem to display but also to input in an intuitive way for a rendering engine to understand. For SPAT-SCENE they have introduced an application for mobile devices for tactile trajectory input [27]. To visually represent these movements they show initial positions and target positions over an indicated time interval, while for SSMN there is first the fixed descriptor symbol to indicate type of movement, and then optionally a radar style display to show the specific movement to be performed until the next indicator. Both systems are straight forward and I chose a very similar approach, though unlike SSMN I do not use a separate indicator to define type of movement. See Figure 2 for an example of indicators connected to a short score of four sound objects, notated on two separate staves, one staff for each spatialised musical layer.

5.3 Spatialisation as articulation

In my notation of sound objects, inspired by symbols and ideas from Lasse Thoresen's spectromorphological analysis [28], there's the possibility to notate various paramtetric changes as articulations by providing the appropriate symbols below the main sound object symbol (a circle, square or diamond note head). As with traditional notations of articulation, this should be used for changes that don't appear on a structural level but that provide a certain identity to a particular sound object, like a vibrato or a tremolo. See Figure 4 for example of how modulation is notated in my system. Basic sawtooth waves are indicated here to achieve regular clockwise or anticlockwise rotation. Width and frequency can be added for a more detailed indication. The colon marks after the sawtooth symbols in Figure 4 signify repetition, but one can also define spatial articulations that, like accent marks in traditional music, are carried out only once. If the colon mark in example A of Figure 4 was removed, the corresponding sound object would be expected to complete one rotation during the course of its duration.

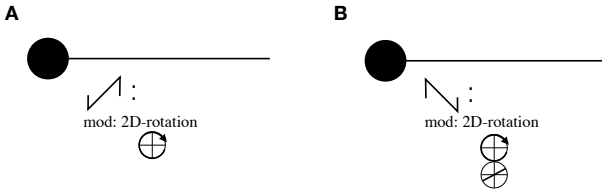


Figure 4: Simple spatialisation attributes notated as articulations of a sound object. Both examples A and B are two-dimentional rotations but example A is a clockwise rotation while example B is a tilted anti-clockwise rotation in a three-dimensional field.

5.4 Room, resonance and reverbs

As mentioned above, for notating the presence of a resonant room or chamber, I suggest first deciding whether it is to be considered *the* space of the music, a space of the music or a resonating room or body for one particular sound or sound layer. For a reverb or a room caught on tape that remains static, there's no reason for continuous notation, only a text or a drawing in the beginning of the score. But for an evolving space, a separate staff line would do for whatever changes that appear. For effects only affecting a particular sound or layer, as in the case of a reverb acting as the resonating body of a sound, these should be notated as a shadow on the extension line of the sound object to indicate the extent of the effect (see Figure 5).

6. CONCLUSIVE REMARKS

In this text I have conveyed my ideas for moving forward with the notation of spatial sound for the purpose of analysis and composition of acoustic and electroacoustic music. The ideas will be put to the test already during the academic year 2019 to 2020, when composition students will

Figure 5: Reverb effects added to specific sound objects, notated as shadows behind the extension lines. Example A shows a softer effect while B is a louder effect.

work with the notation in case studies aimed at exploring new possibilities for teaching composition, analysis and ear training with the notation of sound in focus. Thanks to the establishment of concert halls with large speaker sets for reproducing high-quality three-dimensional sound, there is in the field of electroacoustic music a renewed interest in tools that enable the continued exploration of spatialised sound.

7. REFERENCES

[1] E. Varèse and C. Wen-Chung, "The liberation of sound," *Perspectives of new music*, vol. 5, no. 1, pp. 11–19, 1966.

[2] G. Valkare, *Varifrån kommer musiken?* Gidlunds förlag, 2016.

[3] M. Sköld, "Combining Sound- and Pitch-Based Notation for Teaching and Composition," in *TENOR'18 – Fourth International Conference on Technologies for Music Notation and Representation*, 2018, pp. 1–6.

[4] O. Messiaen, *Apparition de lEglise éternelle*. Paris: Editions Henry Lemoine, 1934.

[5] A. Vande Gorne, "L'interprétation spatiale. essai de formalisation méthodologique'," *Démeter*, 2002.

[6] N. Barrett, "Spatio-musical composition strategies," *Organised Sound*, vol. 7, no. 3, pp. 313–323, 2002.

[7] M. A. Baalman, "Spatial composition techniques and sound spatialisation technologies," *Organised Sound*, vol. 15, no. 3, pp. 209–218, 2010.

[8] N. Peters, G. Marentakis, and S. McAdams, "Current technologies and compositional practices for spatialization: A qualitative and quantitative analysis," *Computer Music Journal*, vol. 35, no. 1, pp. 10–27, 2011.

[9] D. Smalley, "Spectromorphology: explaining sound-shapes," *Organised sound*, vol. 2, no. 2, pp. 107–126, 1997.

[10] A. Hutchinson Guest, *Choreo-Graphics: A comparison of dance notation systems from the fifteenth century to the present*. Routledge, 2014.

[11] ——, *Labanotation: The System of Analyzing and Recording Movement, Fourth Edition*. Routledge, 2005.

[12] N. Abe, J.-P. Laumond, P. Salaris, and F. Levillain, "On the use of dance notation systems to generate movements in humanoid robots: The utility of laban notation in robotics," *Social Science Information*, vol. 56, no. 2, pp. 328–344, 2017.

[13] S. Brandorff and J. La Cour, "Aspects of the serial procedures in karlheinz stockhausen's kontakte," *Electronic Music and Musical Acoustics*, vol. 1, pp. 75–107, 1975.

[14] C. Murray, "A history of" timbres-durées": Understanding olivier messiaen's role in pierre schaeffer's studio," *Revue de musicologie*, pp. 117–129, 2010.

[15] R. Gottfried, "Svg to osc transcoding: Towards a platform for notational praxis and electronic performance," in *Proceedings of the International Conference on Technologies for Notation and Representation*, 2015.

[16] E. B. Ellberger, G. T. Pérez, L. Cavaliero, J. Schütt, G. Zoia, and B. Zimmermann, "Taxonomy and notation of spatialization," in *Proceedings of the International Conference on Technologies for Notation and Representation*, 2016.

[17] E. Ellberger, G. T. Perez, J. Schuett, G. Zoia, and L. Cavaliero, *Spatialization Symbolic Music Notation at ICST*. Ann Arbor, MI: Michigan Publishing, University of Michigan Library, 2014.

[18] J. Garcia, T. Carpentier, and J. Bresson, "Interactive-compositional authoring of sound spatialization," *Journal of New Music Research*, vol. 46, no. 1, pp. 74–86, 2017.

[19] C. Roads, J. Strawn *et al.*, *The computer music tutorial*. Cambride, MA: MIT press, 1996.

[20] J. Harvey, *The music of Stockhausen: an introduction*. Univ of California Press, 1975.

[21] B. Mason, *Music for Moderna Museet*. Stockholm: Moderna Museet, Stockholm, 2008.

[22] R. Saunders, *Chroma*. London: Edition Peters, 2003.

[23] J. Daniel, "Spatial sound encoding including near field effect: Introducing distance coding filters and a viable, new ambisonic format," in *Audio Engineering Society Conference: 23rd International Conference: Signal Processing in Audio Recording and Reproduction*. Audio Engineering Society, 2003.

[24] C. Seeger, "Prescriptive and descriptive music-writing," *The Musical Quarterly*, vol. 44, no. 2, pp. 184–195, 1958.

[25] M. Sköld, "The Harmony of Noise: Constructing a Unified System for Representation of Pitch, Noise and Spatialization," in *CMMR2017 13th International Symposium on Computer Music Multidisciplinary Research*. Les éditions de PRISM, 2017, pp. 550–555.

[26] H.-N. Ho, G. H. Van Doorn, T. Kawabe, J. Watanabe, and C. Spence, "Colour-temperature correspondences: when reactions to thermal stimuli are influenced by colour," *PloS one*, vol. 9, no. 3, p. e91854, 2014.

[27] J. Garcia, X. Favory, and J. Bresson, "Trajectoires: A mobile application for controlling sound spatialization," in *Proceedings of CHI EA '16: ACM Extended Abstracts on Human Factors in Computing Systems*.

[28] L. Thoresen and A. Hedman, "Spectromorphological analysis of sound objects: an adaptation of pierre schaeffer's typomorphology," *Organised Sound*, vol. 12, no. 2, pp. 129–141, 2007.

pya – a Python Library for Audio Processing and Auditory Display

Thomas Hermann
Ambient Intelligence, CITEC,
Bielefeld University, Germany
thermann@techfak.uni-bielefeld.de

Jiajun Yang
Ambient Intelligence, CITEC,
Bielefeld University, Germany
jyang@techfak.uni-bielefeld.de

ABSTRACT

This paper introduces *pya*, a python package for audio processing and sound and music computing (SMC). We focus on the key concepts, the core classes for signals, spectra and spectrograms, furthermore we illustrate the supported interaction types and different use cases. The examples focus on (i) sound synthesis, (ii) multi-channel audio processing, and (iii) sonification, specifically audification and parameter-mapping sonification. We showcase how pya facilitates advanced audio processing. Finally we provide a roadmap for future development.

1. INTRODUCTION

Researchers and developers need tools and software environments to sketch ideas, to prototype, develop, test, and evaluate novel systems, and thereby to advance the state-of-the-art in their discipline. Particularly in the field of sound and music computing (SMC), including auditory display research, sound design, digital audio effects, machine listening, music information retrieval (MIR), to name a few, tools for just-in-time creation of – and interaction with – code, user controls, sound, and algorithms are crucial. At the same time, researchers wish to document their steps, visualize results, and format their system in a reproducible, shareable form.

The Open Source programming language Python has grown for decades into a very mature interactive programming ecosystem, which – supported by powerful packages such as the NumPy [1], SciPy [2], Matplotlib [3], scikit-learn [4]– offers wonderful and highly flexible means for data scientists. Furthermore Jupyter notebooks and Jupyter Lab enable Web-browser-based interactions in a just-in-time manner, enabling to interleave code and documentation. So, data scientists often prefer code-oriented platforms such as MATLAB, Octave, and Python. Particularly for Python, NumPy offers ndarray as a suitable representation for multi-channel audio data, and SciPy features powerful functions for digital signal processing, such as via scipy.signal. Interestingly, research on digital audio signal processing often uses other platforms, such as C/C++, MATLAB. There are also specialized sound synthesis and audio processing languages/tools such as SuperCollider, PureData, Max, Csound, Chuck, Faust, which, however, are more focused on sound design and real-time rendering and less flexible on accessing the data vector for subsequent inspection or analysis. So why not simply use (and add convenient) functions for loading, saving, recording and playing audio and work as such? One reason is that audio signals are not just a data vector or multi-dimensional array: meta data, such as sampling rate, channel names, value range, may vary and need to be specified and provided every time for functions to work properly. This causes substantial coding overhead and makes writing code even for simple audio signal processing tasks less elegant – and in turn, it makes the code harder to read.

For instance, to create 1 second, 2-channel 44100 Hz audio signal and to play its segment from 0.5s and 0.7s in plain Python with NumPy and sounddevice takes

```
sr, nr_chnls, duration = 44100, 2, 1.0
signal = np.zeros((sr*duration, nr_chnls))
sounddevice.play(signal[int(0.5*sr), int(0.7*sr)], fs=sr)
```

whereas pya's Asig and *time slice indexing* condenses the above to

```
a = Asig(1.0, sr=44100, channels=2)[{0.5:0.7}].play()
```

while enabling later attribute access via `a.channels`, `a.sr`, `a.get_duration()`. The advantages become even more obvious as code grows.

There are several existing audio packages within the Python ecosystem. To our knowledge, they mostly fall into two categories: 1) toolkits for audio signal analysis; 2) audio synthesis and I/O. Analytical toolkits such as pyAudioAnalysis, LibROSA, SpeechRecognition and madmon [5–8] all have a strong focus on music information retrieval (MIR) and machine learning. They provide useful functions for audio signal feature extraction, segmentation and classification. These packages gear more toward offline processing of audio arrays and less so on applications in audio I/O. On the second category, there are multiple packages for 'low-level' access to audio hardware and audio files (e.g. PyAudio, playsound, simpleaudio, audiooop, sound), they are more of a barebone package on top of which programmers still need to develop their own synthesis and processing from scratch. An exception is

pyo, a powerful package designed for real-time sound synthesis and signal processing. Its design philosophy is reminiscent of SuperCollider [9] for sound design and composition. Yet to our limited knowledge, pyo's PyoObject is not a static storage for samples but rather a buffer. Thus interactions between a NumPy array and pyo's processing functions are not straightforward. Pydub and Pygame offers easy-to-use audio manipulation but gear more towards application audio such as games, hence they lack more in-depth processing functionalities and support for sample-accurate indexing [10, 11].

In summary, while there are a number of audio packages available, each has certain trade-offs that ultimately led us to developing pya for all our needs, including a clean coding interface, NumPy-focused editing and processing, multi-channel support, sample-precise indexing, easy access to files and audio hardware and builtin methods for plotting. To not reinvent the wheel, pya embraces existing cross-platform low-level audio interfaces, specifically PyAudio, which in turn relies on the C/C++ audio API portaudio.

The paper starts with an introduction to the core pya classes Asig, Aspec and Astft, and to the interface class Aserver. It introduces key concepts to keep an as-clean-as-possible, pythonic interface and providing minimal code examples. In Section 4, we illustrate how pya can be used in selected use cases, namely sound synthesis, multi-channel audio processing, and offline-rendered audifications and parameter-mapping sonifications. We wrap up with a short discussion and conclusion.

2. LIBRARY OVERVIEW AND FEATURES

pya aims to fuse together the advantages of array processing with NumPy and PyAudio to enable users to easily create, analyze, manipulate, record and play multi-channel audio signals. It facilitates method chaining, which is particularly useful when applying multiple processing steps, such as below where multiple processes including signal manipulations, plotting and playback can be daisy-chained together.

```
asig.gain(db=6).fade_out(0.2).pan2(-0.5).iirfilter
    (200).plot('r-').play(rate=0.25)
```

The core idea while designing pya was to offer a set of data representations (Asig, Aspec, Astft) which provide straightforward access to the relevant data structures, and offer conversion methods, e.g. `Asig.to_spec()` or `Asig.to_stft()`, when operations in another reference frame are needed. These classes are accompanied by interface classes `Aserver` and `Arecorder` for playback and recording. `Aserver` features a timed queue to schedule data to be played at a given absolute or relative time. The package aims to minimize dependencies while maximizing flexibility. Particularly, the package features extensions to indexing and broadcasting as explained below.

3. CORE COMPONENTS OF PYA

In the current version (v0.3.1), pya features six classes and a set of helper functions. All classes except the signal generator class Ugen start with the letter 'A', which stands for *audio* just as the 'a' in pya (Python audio). The class diagram in Fig. 2 depicts the structure of the package. It contains the main signal class `Asig`, a spectrum class `Aspec` and a spectrogram class `Astft`. `Ugen` is a subclass of `Asig` that provides quick ways to generate commonly used signals. Then we have `Aserver` and `Arecorder` which in charge of audio playback and record.

3.1 Asig - The audio signal class

`Asig` is the core class of the library. An Asig object contains any audio array in the form of a `numpy.ndarray`, and other relevant information such as sampling rate, an identifier label, channel names, etc. The class offers a broad collection of methods commonly needed for audio signal processing. pya aims at making NumPy users automatically familiar with using pya for signal manipulation and is unique in providing 'numpythonic' indexing extensions.

There are multiple ways to create an `Asig` object, as shown in the code example below (see code comments for description):

```
from pya import Asig
from numpy.random import rand
asig0 = Asig(rand(44100, 2), sr=44100, label='
    noise', cn=['left', 'right'])  # stereo noise
asig1 = Asig(sig="folder/filename.wav", label='
    filename')  # Load an audio file
asig2 = Asig(1.2, sr=1000, channels=5)  # A 1.2
    seconds, 5 channels zeros array at 1kHz
    sampling rate
asig3 = Asig(1024, sr=2048, channels=3)  # A 3
    channels zeros array at 2048Hz sampling rate
    with 1024 sample frames
```

Listing 1. Several ways of constructing an Asig object

Asig's constructor has 5 arguments: the required argument `sig` can be a NumPy array for audio signal, a string as path to a file, a float for creating a silence signal of given duration in seconds and an integer for an integer number of sample frames. The audio array is accessible at `Asig.sig`. `sr` as sampling rate is not required when loading an audio file but recommended for other cases, otherwise it defaults to 44100 Hz. Users can also specify the number of `channels`, yet the shape of a given `sig` takes priority. The `label` serves as a string identifier to the object. `cn`, short for 'channel names', is a list of strings for labelling each channel. This does not affect the signal array, but is useful for more meaningful sub-setting of channels, e.g., users might name a stereo signal's channel via `cn=['left', 'right']`, and index the wished channel as `asig[:, ['right']]`.

Asig offers a growing collection of methods for signal processing. Methods such as `gain()` enables gain

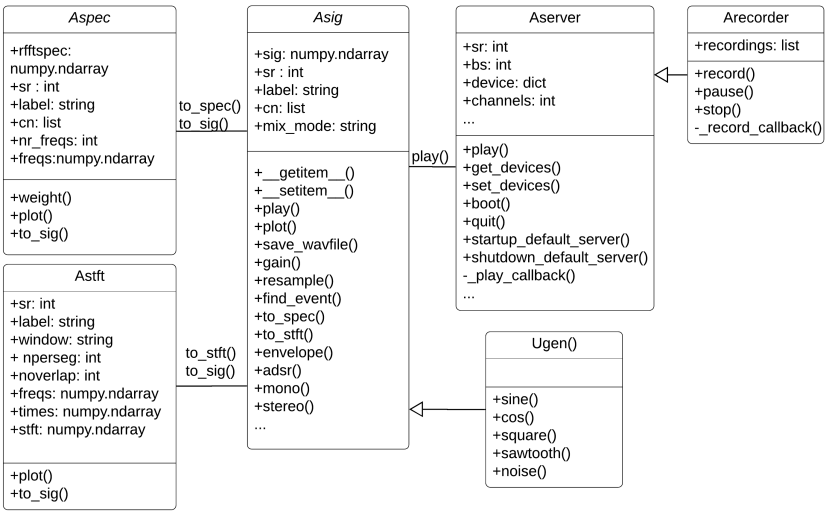


Figure 1. pya Class diagram.

adjustment in amplitude or decibel. `find_event()` enables detection of silence-separated events, `mono()` and `stereo()` convert audio signals to mono or stereo allowing to specify how to mix the original channels. The `plot()` method displays the Asig as time series, even allowing to stack line plots easily via `offset` and `scale` arguments, otherwise forwarding arguments as kwargs to the underlying Matplotlib `plot()`. Asig methods return either a new Asig object or self, which enables method chaining, resulting in a highly compact way to write signal processing paths.

pya extends the already powerful slicing of NumPy arrays to better tailor them for indexing segments of multi-channel audio – with particular benefits for code clarity and conciseness. As this topic is more complex, it will be expanded in the following subsection.

3.1.1 Asig Indexing and Time Slicing

One key feature of Asig is a set of methods to enable indexing, slicing and assignment, realized by defining get_item and set_item in similarity to NumPy, yet extending it for frequent audio use cases.

The pythonic way of sub-setting a list or a NumPy array is via index slicing. Slicing takes the form of `start:stop:step` and this is in analogy implemented for Asigs. However, when specifying audio segments, it is also common to use timestamps rather than the integer sample index. In pya, we introduced *time slicing* in addition to index slicing, allowing to pass timestamps in the form `{start:stop}`. Note that we chose the dictionary type here, as it allows to maintain the colon ':' as familiar separator for slicing. An example is shown in Listing 2 with both the conventional index slicing and our newly introduced time slicing. Note that in analogy to index slicing, negative numbers refer to the position relative to the end, i.e. `a[{-2:-1}]` specifies the second last second of the signal `a`. Finally, note that different from standard slicing syntax, it is not possible to omit the `start` or `stop`. If the last two seconds shall be selected, use `{-2:None}`. The reason is that dictionary syntax always requires key and value. Note that no `step` argument is allowed in time slicing.

```
1 _ = asig[0:150:2]  # Index slicing with step 2
2 _ = asig[{0:1.5}]  # Time slicing in seconds
```

Listing 2. Asig indexing

3.1.2 Asig channel slicing, list slicing

When dealing with multi-channel signals, sometimes we need to subset a subgroup of channels. Similar to the row slicing mentioned previously, column slicing utilizes the same pythonic principle. Specific to pya, we can give a list as column slicing argument. The data type of the list can be of type

- integer for column indices,
- boolean that chooses the columns that are true,
- string or list of strings for channel names from `cn`.

Listing 3 demonstrates the different options for the same indexing of channels. With the use of labeling each channel through `cn`, indexing audio channels can be done in a more linguistic way, which can offer more clarity to users.

```
1 asig = Asig(np.random.rand(5000, 5), cn=['fl', 'fr',
      'cen', 'rl', 'rr']) # front-left to rear-right
2 asig[:, 0::2]  # equiv. to [0,2,4]
3 asig[:, [True, False, True, False, True]]
4 asig[:, ['fl', 'cen', 'rr']]
```

Listing 3. 3 types of column indexing

3.1.3 Broadcasting

We implemented flexible ways of broadcasting signals to a subset of an Asig through `___setitem___()` such as `a1[{1.5:3}] = a2[{4:5.5}]`. Of course, this works

also for simultaneous channel subsetting. In such assignments, as signals are basically NumPy arrays, Asig follows the broadcasting rules of NumPy, i.e., a subset of the array can be reassigned by another. However, the conventional way requires equal shape. When working with audio data, mixing audio signals together, an equal shape is rarely given. This burdens users constantly with assuring in their code that assignments can be executed.

For example, in plain NumPy, mixing a signal `s2` into `s1` from index 300 would require `s1[300:300+s2.shape[1]] += s2`. An error will raise if `s1`'s length is not large enough to accommodate `s2`. It is likely users need to instantiate a new array to accommodate the potential new shaped signal resulting from that mix-shaped operation.

As a solution we invented and introduced into pya (as of now) three different *mix modes* for `___setitem___()`:

- `x` or `extend`. The extend mode automatically expands the destination's row dimension if the right-side operand (which can be either Asig or ndarray) has more rows than available.

- `b` or `bound`. The bound mode automatically fixates the shape of the left-side operand and thus truncates longer right-side operands.

- `o` or `overwrite`. This mode automatically replaces the left-side specified subset of the Asig with the right-side given signal, regardless of their lengths.

In search for the most compact, concise and readable way to enable these modes syntactically, we finally solved it by adding the `@properties x, b, o` to the `Asig` class. In their code, they merely set `self.mix_mode` to the specified mode value ('extend', 'bound', or 'overwrite') and then return `self`. In turn, the assignment `myasig.x[100:1000] = a2` makes sure that at the time the `___setitem___()` function of myasig is executed, the `mix_mode` flag has been properly set, allowing setitem to branch into mode-specific assignment code. Note that before exiting setitem, `mix_mode` is reset in order to avoid any confusion in subsequent/future assignments.

Both the abbreviation (x, b, o) and their corresponding full names can be used when setting the mix mode. Examples of Asig broadcasting are shown in Listing 4.

```
asig1[0:100, :] = asig2[100:200, :]  # By default operands' shapes need to match

asig1.x[-10:] = np.arange(50)  # assign 50 samples starting from last sample-10, i.e. asig1 is extended by 40 samples

asig1.bound[{-1.:}] += asig2[{0:2.0}]  # The first 2 seconds of asig2 will be added to the end of asig1.

asig1.o[{0.5:3.5}] = asig2[{1.0:2.0}]  # The resulting asig1 becomes shorter
```

Listing 4. Examples of Asig mixing modes. Note that here all asigs have the same sampling rate, thus equal duration corresponds to equal sample length.

3.2 Ugen - The unit generator class

`Ugen`, inherited from `Asig`, offers easy generation of several common waveforms, currently available are sine, cosine, square, sawtooth and noise. Each type of signal has its own method that returns an Asig object. All methods use optional arguments, by default they all return a 1s-signal at 44100 Hz with the amplitude of 1. For pitched signal, 440 Hz is the default frequency. An example is shown below:

```
from pya import Ugen
asine = Ugen().sine(freq=440, amp=1.0, dur=1.0, sr=44100, channels=1, cn='s', label='sine')
asquare = Ugen().square(freq=1000, amp=0.5, duty=0.3, dur=3.0, channels=3)
anoise = Ugen().noise(type='pink', amp=0.2, dur=0.5, sr=1000)
```

Listing 5. The unit generator class Ugen used to create sine, square and noise signals.

3.3 Aspec – The spectrum class

Aspec is the spectrum class, using `numpy.fft.rfft()` (Fig. 2). An Aspec object can be constructed directly by giving the required argument `x` as an Asig or a numpy.ndarray. Alternatively, the `to_spec()` method of Asig returns its Aspec counterpart. For multi-channel arrays, the spectrum of each channel is computed independently. Weighting functions can be applied to the spectrum of the signal. A warping function can be specified as argument of `.plot()`.

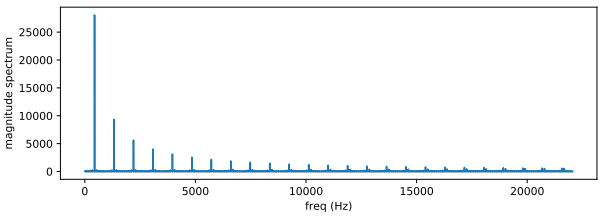


Figure 2. Aspec Spectrum plot, here a square wave via `Ugen().square(440, duty=0.5).to_spec().plot();`

3.4 Astft – The STFT class

`Astft` generates the short-time-Fourier-transform using `scipy.signal.stft()`. Its `plot()` method displays the signal's spectrogram as a mesh plot (Fig. 3). Similar to Aspec, an Astft object can be created by converting an Asig object, e.g. `astft = asig.to_stft()`. Thus it can also be converted back to Asig object i.e. `asig = astft.to_sig()`.

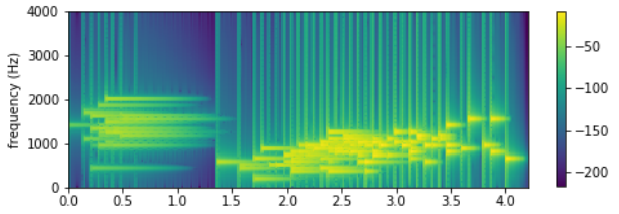


Figure 3. Spectrogram of the sonification from Sec. 4.3.2, plotted via `son[:,1].to_stft().plot(ampdb)`

3.5 Aserver – The audio server class

`Aserver` allows users to boot a threaded audio stream playback server. At Aserver, users can set up the corresponding audio device and other parameters such as the sampling rate. Once the server is booted via `boot()`, audio data is streamed to the dedicated output. Users can not only trigger playback by calling the Asig object's `play()` method, but also schedule a future playback, using the onset argument (relative in seconds or absolute using time.time()). Thus, multiple audio event playback can be orchestrated as needed simultaneously, overlapping or in sequence.

It is possible to construct multiple Aserver instances. In that way, it is possible to have multiple Aserver objects utilizing different audio devices at the same time. Below is an example of playing an Asig via different audio servers at different times.

```
from pya import Aserver, Ugen
as1 = Aserver(device=1).boot()
as2 = Aserver(device=4).boot()
saw = Ugen().sawtooth()
saw.play(server=as1, onset=1)
saw.play(server=as2, onset=2.5)
```

Listing 6. Playing an audio on multiple devices

The current version has not yet implemented sample-accurate synchronous playback to multiple devices. Thus simply calling two play methods does not result in a synchronized playback due to the processing required before the signal being queued to each server. We will work on this feature in the future release.

pya provides a helper function `device_info()` that prints all available devices and their information, from which users can find out the index of each device.

3.6 Arecorder – The audio recording class

Inherited from `Aserver`, `Arecorder` provides recording functionality. Booting a recorder object is similar to Aserver. Once booted, call `record()` to start and resume; `pause()` to pause and `stop()` to stop the recording. Afterwards, the current recording along with meta data will be stored as an Asig object and append to the `recordings` variables as a list of recordings.

At the time of publishing this paper, Arecorder has just been newly introduced to the package, thus various improvements will follow such as an easy way to select one of multiple input channels, or to adjust the individual channel's gain. We plan to introduce these the upcoming release of pya.

3.7 helpers – A collection of useful functions

In addition to the main class components, pya contains a collection of helper functions that can be imported and used, such as the aforementioned `device_info()`. Others include conversions, e.g., between amplitude and decibel (`ampdb` and `dbamp`), or a linear-to-linear mapping (`linlin`). Furthermore there are functions for normalization (`normalize`), clipping (`clip`), conversion between cycle-per-second (cps) and MIDI (`midicps` and `cpsmidi`), etc. Many helper functions are named and defined in analogy to SuperCollider3 functions.

4. EXAMPLES

In this section, we present three examples to demonstrate some of the possible usages of pya.

4.1 Sound Synthesis with pya

In the first example let's create a band-limited square wave using additive synthesis, i.e., by summing few terms of the Fourier series

$$x(t) = \frac{4}{\pi} \sum_{k=1}^{\infty} \frac{1}{2k-1} \sin(2\pi(2k-1)f \cdot t) \qquad (1)$$

As shown in Fig. 4, we zip the vectors of frequencies and amplitudes (odd harmonics and reciprocal amplitudes), summing the sine waves generated by `Ugen()` by looping through the total amount of units.

```
asig = Asig(1.0) # Create a blank canvas
harmonics = (1 + 2 * np.arange(100))
for f, a in zip(3*harmonics, 1/harmonics):
    asig += Ugen().sine(f, a, sr=44100, dur=1.)
```

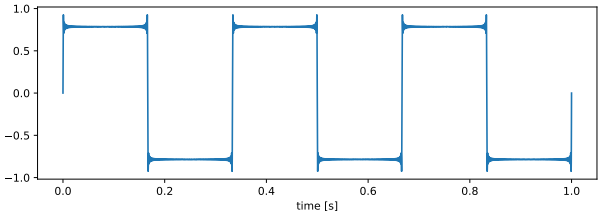


Figure 4. Additive synthesis of 100 sine waves to form a 1 s, 3 Hz band-limited square wave signal.

4.2 Multi-channel audio processing

Due to the flexibility of indexing and broadcasting mentioned in Section 3, pya is highly suited for multi-channel audio editing or playback. Fig. 5 demonstrates arranging a 4 channels mix. First, an Asig with the correct number of channels but only 1 sample is constructed as a placeholder. Using the extend

```
a = Asig(1, sr=1000, channels=4, cn=['a', 'b', 'c', 'd'], label='multichannel')
b = Ugen().sine(freq=50, sr=1000, dur=0.6).fade_in(0.3).fade_out(0.2)
a.x[:, 'a'] = 0.2 * b  # no need to extend as len(src)<len(dest)
a.x[300:, 'b'] = 0.5 * b  # extends a to 0.9 seconds
a.x[1300:, 'c'] = 0.2 * b[::2]  # extends a further, writing beyond end
a.x[1900:, 3] = 0.2 * b[300:]  # note that 3 is 'd'
plt.figure(figsize=(12,3)); a.plot(offset=1, scale=3.5)
```

```
Asig('multichannel'): 4 x 2200 @ 1000Hz = 2.200s cn=['a', 'b', 'c', 'd']
```

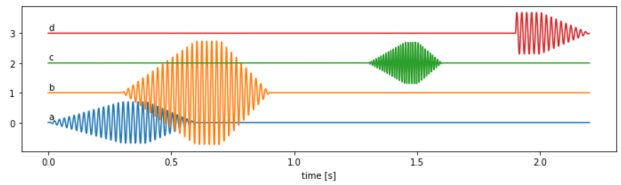


Figure 5. Sequencing signals in a 4-channel audio.

mode `x`, we mix signals in while extending Asig as needed beyond the end. Note that zero-padding is automatically applied to other channels on extension.

For stereo, the same principle as above can be applied. Asig features methods such as `stereo()` for arbitrary blending and `pan2()` for constant power panning to a stereo position in $[-1, 1]$ (left, right).

4.3 Sonification

Sonification is the systematic, reproducible representation of data using sound [12]. There are a variety of sonification techniques to represent complex data, such as audification, parameter-mapping sonification, model-based sonification etc, see the Sonification Handbook [13] for an overview. As data is often processed and analyzed with ipython, it is convenient to have basic sonification methods available without the requirement to interface with other sound synthesis engines such as Supercollider or PureData. Also, there sometimes is the situation that complex (beyond realtime renderability) mappings force offline sonification anyway, and pya lends itself both for simple and fast sonification tasks and complex offline mappings. For the demo we merely feature the example notebooks made available as supplementary material to the paper, available at `http://dx.doi.org/10.4119/unibi/2938412`. Due to limited space, we here focus on audification and parameter-mapping sonification.

4.3.1 Audification

Audification is basically the direct data-as-sound translation for auditory inspection of time series [13]. We demonstrate how pya can help with an example of auditory analysis of EEG data. The example notebook (see supplementary material) loads a provided data set by

```
data = np.loadtxt("data/epileptic-eeg.csv", delimiter=",")
chnls = "FP1 FP2 F3 F4 C3 C4 P3 P4 O1 O2 F7 F8 T7 T8 P7 P8 FZ CZ PZ".split(" ")
aeeg = Asig(data, sr=256.41, cn=chnls)
```

This shows by the way that Asigs are also useful to represent and manipulate non-audio data such as biomedical data. We can now audify and plot for instance channel 5 and 9 at a 10x-speedup in one line of code: `a1=aeeg[:,[5, 9]].norm(1).plot().resample(44100, rate=10).play()` In the provided sound example, we hear the rhythmical (3/sec) epileptic pattern arising and falling back into background signal. Listening at higher speedup (25×) via `a1 = aeeg[:,[5, 9]].norm(1).resample(44100, rate=25).play()`, we discover the ritardando over the seizure. Next, let's select 6 seconds (from 16s–22s), to listen in 'slow motion'. We resample to a modest 8000 Hz audio rate by `a2 = aeeg[{16:22}].norm().plot(scale=0.5, offset=1).resample(8000, rate=1)`.

However, we better modulate the signal – here with a sine wave – since otherwise we would not perceive the low frequencies. So, let's (i) use integer multiples of 90 Hz for the series of channels, (ii) listen to the result in $^1/_2\times$ the original speed, (iii), plot the compiled audio with a thin green line, and (iv), save the audio file in WAV format. This complex task can be performed with 4 lines of code:

```
asum = Asig(a2.get_duration(), sr=a2.sr)
for i in range(a2.channels):
    asum += a2[:, i] * Ugen().sine(90*(i+1), dur=a2.get_duration(), sr=a2.sr)
asum.norm().plot(color="g", lw=0.1).stereo().play(rate=0.5).save_wavfile('eeg.wav')
```

This showcases that a quite complex set of tasks can be easily expressed in short and concise code, which is even shorter than the textual instruction above, yet remains well readable.

4.3.2 Parameter Mapping Sonification

Next we demonstrate pya for discrete parameter-mapping sonification, i.e., data variables are mapped to synthesis parameters and the resulting sound events are mixed to the sound track. Our parameters will be onset, frequency, duration, amplitude and stereo panning. For the demo, we use the well-known Iris data set of 4 geometric features of 150 iris plants [1] . With pya, a 5-parameter synthesizer can be written highly compact as Python function:

```
def mysyn(freq=440, dur=0.34, amp=0.9, att=0.01, pan=0.5):
    return Ugen().sine(freq, dur=float(dur), sr=8000).fade_in(0.01).envelope([0,1,0], [0, att, dur], curve=4).gain(amp).stereo([1-pan, pan])
```

With it, the parameter-mapping becomes

```
son = Asig(1.0, sr=8000, channels=2)
for i, v in enumerate(diris):
    onset = mapcol(v, 2, 0,4) # def: s. .ipynb
    freq = mapcol(v, 1, 200, 2000)
    # some lines skipped here for code brevity
    son.x[{onset:None}] += mysyn(freq, dur, amp, 0.001, pan)
son.norm().plot(offset=1, alpha=0.5).play()
```

Note how the extend mix-mode facilitates superimposing data-driven sound events into the sound canvas `son`.

[1] see https://archive.ics.uci.edu/ml/datasets/iris

5. DISCUSSION AND CONCLUSION

We have introduced pya, a new audio processing package with the aim to make audio coding more pythonic by exploring and extending NumPy and SciPy coding styles, incorporating and thus hiding many tedious coding work into a set of audio classes `Asig` for signals, `Aspec` for spectra, `Astft` for spectrograms and `Aserver`/`Arecorder` as interfaces for real-time and non-blocking playback/recording.

Conceptually new contributions are *time slicing* and broadcasting via *mix_modes*. They alone simplify many audio coding problems, according to our own experience. Being still a novel and young package, some pya interfaces may be changed on deeper understanding or better ideas how to write code even more elegantly. However, the core functionality should only mature or improve.

The examples in this paper demonstrate that pya can be swiftly used for very different tasks, resulting in short, clean and readable code, and we hope that future contributions and extensions help to make it more versatile and useful.

The roadmap for future releases will involve improvements on the freshly added Arecorder class for non-blocking audio recording, additional audio backends such as Web Audio, full documentation and tutorials, custom meta data, introducing more audio processing algorithms and effects, instrument class for designing digital musical instruments with MIDI support, etc.

Pya is hosted under our github organization *Interactive-Sonification* and available on PyPI and Github. It is under MIT license and we are eager to hear comments, feature requests, and to discuss ideas.

Acknowledgments

We thank Alexander Neumann for assistance with package release and continuous integration testing, also for the contribution of the multiple audio backends featuring in the latest release of the package. This work has been partly supported by the BMWI via the ZIM project SoundRefiner, by the BMBF project NeuroCommTrainer, and by the Cluster of Excellence Cognitive Interaction Technology "CITEC" (EXC 277) at Bielefeld University, which is funded by the German Research Foundation (DFG).

6. REFERENCES

[1] T. E. Oliphant, *A guide to NumPy*. Trelgol Publishing USA, 2006, vol. 1.

[2] E. Jones, T. Oliphant, P. Peterson *et al.*, "SciPy: Open source scientific tools for Python," 2001–, [Online; accessed 2019-09-17]. [Online]. Available: http://www.scipy.org/

[3] J. D. Hunter, "Matplotlib: A 2d graphics environment," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007.

[4] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python ," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[5] T. Giannakopoulos, "pyaudioanalysis: An open-source python library for audio signal analysis," *PloS one*, vol. 10, no. 12, 2015.

[6] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8, 2015.

[7] A. Zhang, "Speech recognition (version 3.8) [software]," 2017. [Online]. Available: https://github.com/Uberi/speech_recognition#readme

[8] S. Böck, F. Korzeniowski, J. Schlüter, F. Krebs, and G. Widmer, "madmom: a new Python Audio and Music Signal Processing Library," in *Proceedings of the 24th ACM International Conference on Multimedia*, Amsterdam, The Netherlands, 10 2016, pp. 1174–1178.

[9] S. Wilson, D. Cottle, and N. Collins, *The SuperCollider Book*. The MIT Press, 2011.

[10] J. Robert, "pydub," [Online; accessed 2019-10-22]. [Online]. Available: https://github.com/jiaaro/pydub

[11] W. McGugan, *Beginning game development with Python and Pygame: from novice to professional*. Apress, 2007.

[12] T. Hermann, "Taxonomy and definitions for sonification and auditory display," in *Proc. Int. Conf. Auditory Display (ICAD 2008)*. International Community for Auditory Display, 2008.

[13] T. Hermann, A. Hunt, and J. G. Neuhoff, *The sonification handbook*. Logos Verlag Berlin, Germany, 2011.

RHYTHM AS SENSORIMOTOR SUPPORT: HOW SOUND AFFECTS STRIDE LENGTH AND STEP FREQUENCY

Maria Svahn
EECS, School of Electrical Engineering
and Computer Science, KTH
Lindstedtsvägen 3, 114 28 Stockholm
masvahn@kth.se

Josefine Hölling
EECS, School of Electrical Engineering
and Computer Science, KTH
Lindstedtsvägen 3, 114 28 Stockholm
jholling@kth.se

ABSTRACT

Rhythm therapy has shown to be an effective complement to other methods for treating neurological diseases such as Parkinson's disease or stroke. Several studies have shown improvement in specially the patient's ability to maintain a steady walk with a normal stride length. It has been investigated what actual changes that occur in the brain while listening to music in relation to motor changes, but none of these studies can clearly explain why rhythm therapy works so well for PD patients and what part of the auditory stimulus that affects the patient most. The aim of this article is to investigate how sound is used to treat Parkinson's disease, and through a study of healthy adults create an idea of how different types of auditory stimuli affect stride length and step frequency. The study consisted of 19 participants that walked a distance of 10 meters 6 times each. One time without any auditory stimuli, one time with to the sound of a metronome, three times with different kinds of music that the experiment leaders had chosen and one time with a musical piece of their own choice. The result of this study indicates that the musical piece the participant had chosen for themselves gave the best result.

Keywords
Gait; rhythm; Parkinson's disease; stride length; step frequency; sensorimotor; auditory stimuli

1. INTRODUCTION

Parkinson is a neurologic disease which affects the central nervous system, causing uncontrolled shaking, muscle weakness and deterioration of mobility [1]. The problem is that the cells that make dopamine, which is used to send signal between cells, are degraded. It is not yet known why this degradation of cells occurs. Most patients are over 50 years old and the disease gradually deteriorates to a point where the affected person finds it difficult to walk or do everyday chores. The impaired motor ability also often leads to speech impediments.

The Ronnie Gardeners Method (RGM) started being implemented as rehabilitation for Parkinson in Sweden in 1993, which is a rhythm and music-based physiotherapy used to develop cognitive and sensorimotor control [14]. By following rhythms, the patients are described to be able to talk more fluent, improve their gait, reduce the risk of falling as well as stabilize their body movements in general. The measurements made on the RGM method showed that stride length, gait speed and step frequency (step/second) where improved after training with this kind of sensory motor support [13].

Since the first implementation, there has been a growing interest of seeing this kind of physiotherapeutic methods as treatment instead for or as a complement to medical treatment. Despite this, it is not known exactly why these cognitive and motor improvements occur [8]. Some studies [9] indicates that music and rhythms trigger our reward system which produces dopamine neurons that otherwise, for a person with Parkinson, would have been lost. Other studies have shown the same results with only a metronome as auditory stimuli [2].

2. THEORY AND RELATED RESEARCH

Parkinson patients experience difficulties in walking because of the degradation of dopamine neurons in the cerebrum, just over the brainstem [2]. When the dopamine is destroyed, this part of the brain loses the ability to deliver the type of signals that handles execution of predictable movements and actions [3]. For a person with PD, this results in loss of the built-in intuitive feeling for stride length and the person end up with a shortened stride length that often is compensated with an increase in step frequency [3]. This makes the gait unstable and increases the risk of falling. Several studies have therefore shown that an external cue, that helps the motor cerebral cortex to temporarily adjust and synchronize the stride length, significantly simplifies a PD patient gait ability, including cueing by rhythms and music.

The use of sound as sensorimotor support is established within both treatment and research. Musical activities make special demands on the brain that in many cases has shown to be advantageous in the treatment of motor difficulties in neurological diseases. The parts of the brain that is activated in musical contexts and are connected to motor abilities are the prefrontal cortex, gyrus precentralis and the supplementary motor cortex, which gives us clues about the question that researchers still have not clearly figured out- why it works [4]. It is still unknown what part of the music that affects the disease, for example if it is equally effective to listen to just a steady pace or if it is the other parts of the music combined with a clear pulse that gives effect [9]. Some researchers believe that the biomusicological phenomenon rhythmic entrainment [16], meaning that

particularly humans but also some animals unintentionally synchronize to an external rhythm, has something to do with this.

Even if there is much research done, it is problematic that it does not exist any larger scale studies with many participants or any clearly stated treatment technique. Studies show that music-based movement therapy has a positive effect on motor skills and neutral proof that supports the use of music to increase cognitive function and quality of life [8]. Several studies [5] regarding the matter has a predominantly positive result and shows improvement in measurements of for example stride length and gait speed. There are also examples of studies [6] where the result indicates that it only affects the patient's psychological mood and not the other PD symptoms.

Since a well-coordinated gait is not a natural state for a person with PD might it be counterproductive to focus on synchronizing the gait to an external cue. Because of this, long-term training to music is more effective since the patient gets familiar with the music and the rhythm [3][2]. Although, a short-term auditory stimulus as sensorimotor support has been shown to help the patient to a significantly steadier walk. A study conducted in 2014 [3] examined how healthy persons with varying ability of rhythm sense can synchronize their gait to a rhythm, with the purpose to optimize rhythm therapy for PD patients. The result indicated that for a healthy person with a poor beat perception, the gait became more unstable (increased variations in walking speed and stride length and also longer time standing on two feet) when they tried to synchronize their gait to music that did not have a distinct pace. This since the person's attention must be focused towards finding the rhythm instead of synchronizing their steps. This result is meaningful since Parkinson patients often have reduced capability of performing activities that require multitasking. Therefore, the choice of rhythm for each individual is of great importance.

2.1 Purpose and question at issue

The purpose of this paper is to examine how sound can be used as treatment for neurological diseases, such as Parkinson and stroke. Through a study, we want to create an idea of how the choice of auditory stimuli affects stride length and step frequency when using sound as sensorimotor support. Since there is research to support this type of treatment but it is still unclear why it works, there is room for further studies in the field and a media technological point of view can be valuable.

The question at issue is: *How does the choice of auditory stimuli affect stride length and step frequency when used as sensorimotor support?*

The result would primary benefit the individual being treated and the people in his or her surroundings. But also the healthcare and society in general which benefits from as effective treatment methods as possible. On top of that, it is also valuable to examine how healthy adults is affected by these tests, as it can be an advantage in other research for Parkinson patients.

3. METHOD

3.1 Stimuli

Following stimuli was used in the study:

0. Reference test without auditory stimuli
1. Metronome (120 BPM)
2. Drum Loops by African (drumbeat)[1]
3. 120 BPM by Dead Obies (Instrumental) (hiphop)[2]
4. Tobago by Jonas Rahtsman(pop/electronic/dance)[3]
5. The subjects own favorite song in the tempo interval 115-125 BPM

The study was delimited to try five different kinds of auditory stimuli. All tracks had a tempo of 120 BPM. The part of playback was chosen by what was considered by the experimenters to be a central part of the track. This to avoid start and end parts of the song since these might deviate from the rest. The first sound played was a metronome consisting of a single clicking sound at a rate of 120 beats per second. Song 1, *African Drumloop*, is a drum loop of african drums. Since it was a short repetitive pattern consisting of only drums it had a relatively clear rhythm. Song 2, *120 bpm*, is a funk inspired hip-hop beat. Song 3, *Tobago*, is in the genre of electronic dance music which have a clear rhythm that should be stimulating to apply to the gait [8]. The study was delimited to examine musical pieces with a steady and clear rhythm to give room for other observations of what affects stride length and step frequency.

This study was also delimited to examine short term effects on gait with auditory stimuli, which is the change that became apparent during the test. It was not investigated whether the changes in stride length and step frequency persisted in the long term. The test was performed on healthy adults.

3.2 Participants

19 people, in the ages 19-25 participated in the study. The subjects had varying backgrounds but considered themselves physically healthy.

3.3 Setup

The walking distance of 10 meters was marked with tape on the floor with markings on the start and end to be used on later analysis.

[1] https://jimdooley.net/Drum-Loop/Africanish-Tom-Beat-120-BPM

[2] https://www.youtube.com/watch?v=VbhdhkMx1y4

[3] https://open.spotify.com/track/6H8tEitY5FbP5lEOepM7q8?si=1iA5kZnHRFaDnF6h1G-AYQ

The gait parameters examined was stride length (cm) and step frequency (step/s) as these factors are the most affected for a person with Parkinson's, where stride length is usually shortened and step frequency increases. The stride length was measured by recording the gait with one camera at either end of the walking distance. There were rulers attached to the wall behind the walkway at the same height as the cameras to make exact measurements of where the foot passed the start and end markings.

Figure 1. End markings to measure the subjects exact walking distance. The placement of the last step could be read and used for calculating the difference with the 10 meter marking.

In addition, the entire walk was filmed with a wide-angle camera to count how many steps the subject took in total. Based on that, the stride length was calculated by dividing the number of steps on the exact walking distance. Step frequency was calculated by dividing the number of steps on the total walking time, which could be read from the distance camera.

Sony Vegas Movie Studio HD Platinum 11 was used to analyze the result. To ascertain the statistical significance of the result, one-way ANOVA tests was used at a significance level of $p = 0.05$ [15].

3.4 Procedure

The experiment was conducted on two separate occasions at the Royal Institute of Technology with one subject at a time. Prior to the study, the participant had filled in a Google form where they stated their name, age, length and favorite song within the tempo range of 115-125 BPM (beats per minute).

The subject was asked to walk 10 meter straight with a 3-meter-long start and finish section. The start and finish sections were not accounted during analysis calculations but used for the participants to gain their normal gait pace.

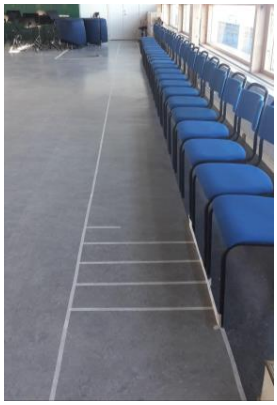


Figure 2. 10 meter walking distance that was used during the study. The participant walked one-way between the two parallel markings.

One of the experimenters showed how to proceed the test by demonstrating the walkway. The participant was asked to start walking when feeling ready. The participant did not know in advance what the study was about, to ensure a, for the person in question, as normal gait as possible.

The result from (0) was used as a reference to measure the changes in the remaining trials. The preselected songs was selected because they are in different genres, have the same tempo (120 BPM) and are relatively unknown, to prevent the participant from recognizing it. Trial (5) was conducted to examine if there was a difference in gait if the subject liked the song used as auditory stimuli. The sounds were played from a computer through the portable speaker Bose Soundlink Mini.

4. RESULTS

4.1 Stride length

Since the stride length depends on a person's height and the study participants vary between 160-199 cm, the change in stride length was measured in percentage relative to the reference value that was measured during test 0.

Observations from the test shows a trend in lengthening the steps when walking with an auditory stimulus compared to without. The 19 participants had an average increase of +2.6 % when they walked to the sound of a metronome. Song 1 resulted in an average increase of +2.3 %. Song 2 resulted in an average increase +2.2 % and song 3 resulted in an average increase of +2.4 %. The auditory stimulus that affected the participants stride length most was their own chosen song, that had an average increase of +4.4 %. The average changes in stride length is illustrated in figure 3.

An observation is that most participants remained constant about if they increased or decreased their stride length independent from the type of auditory stimuli.

Auditory stimulus	Increase (%)
None	1
Metronome	2.6
Song 1	2.3
Song 2	2.2
Song 3	2.4
Self-chosen	4.4

Table 1. Average change of stride length in percentage

4.2 Step frequency

In the step frequency test, number of steps/second was measured and then converted into steps/minute to be able to compare this to the song tempo which is measured in beats per minute (BPM). The metronome as an auditory stimulus resulted in +0.8 steps per minute. Song 1 increased the step frequency with 5.5 steps/minute. Song 2 increased the step frequency with 3.2 steps/minute and song 3 increased the step frequency with 4.9 steps/minute. The self- chosen song increased the average step frequency with 3.5 steps/minute.

Auditory stimuli	Increase (%)
None	1
Metronome	0.8
Song 1	5.5
Song 2	3.2
Song 3	4.9
Self- chosen	3.5

Table 2. Average change of step frequency in percentage

As shown in figure 4, the metronome increased the step frequency least, with an increase of 0.8 steps per minute. When the participants walked to song 1, the step frequency had an average increase of 5.5 steps per minute. Song 2 increased the step frequency with 3.2 steps per minute. The participants reference values (measured without auditory stimuli) were between 98.5 and 129.7 steps per minute. The average value for the participants step frequency without auditory stimuli was 116.5 steps per minute. The participant with the lowest reference step frequency, increased their step frequency with 5.2 steps per minute for song 1, 4.0 steps per minute for song 2 and 3.1 steps per minute for the metronome. The participant with highest reference step frequency increased their step frequency with 0.7 steps per minute for song 1, 9.9 steps per minute for song 2 and 0.2 steps per minute for the metronome.

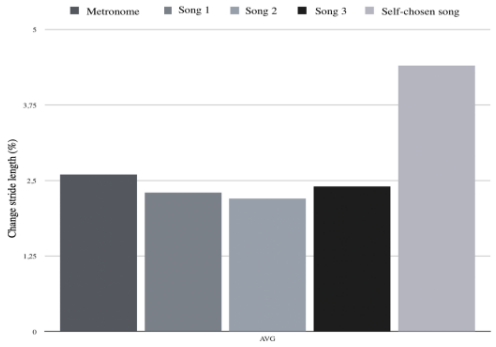


Figure 3. Average change in percentage compared to the reference value regarding stride length to the different auditory stimuli.

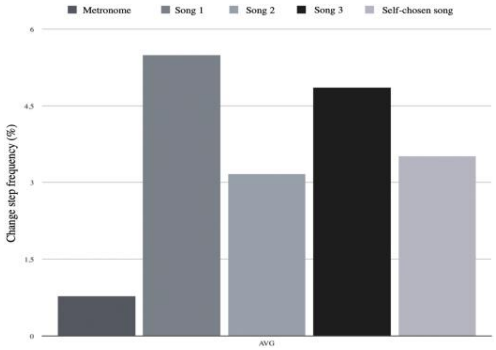


Figure 4. Average change in percentage compared to the reference value regarding step frequency to the different auditory stimuli.

4.4 Statistical significance

To test the results for statistical significance, a one way ANOVA was used. This showed that either stride length nor step frequency was proven to be statistically significant. The resulting changes in stride length had a value of $p = 0.66$ while the change in step frequency resulted in $p = 0.28$.

5. DISCUSSION

In order to discuss the result, it is important to treat a combination of stride length and step frequency, since it is difficult to draw any conclusion about what a low or high step frequency means without having the stride length in consideration. A low step frequency and a long stride length should be the optimal combination for an effective gait in relation to how much energy is needed. The definition of a steady gait is to have a long stride length and a low step frequency, or a long stride length combined with a high step frequency. A short stride length is normally a contributing reason for people with PD having a slow gait speed [12].

Even thought there was no statistically significance, this study shows a trend that auditory stimuli have an effect on stride length since all five tests showed an increase, of which the self-chosen song had the most. Test 1, 2, 3, and 4 had an average increase of 2.6 %, 2.3 %, 2.2 % and 2.4 % while walking to a self-selected song had an average increase of 4.4 %. Test 5 also resulted in the greatest number of people with an increase in stride length, with 15 out of 19. This difference needs to be discussed if it occurs because of the recognition factor of the song and the order of the tests. A previously mentioned study [3] indicated that long-term rhythm therapy had a greater impact on gait since the motor parts of the brain can use an external cue more effectively if it is musically trained and recognizes the rhythm. The result from our study can therefore be interpreted as the fact that a self-selected song has the greatest impact on stride length because it is reminiscent of the effect that comes from long-term rhythm therapy - a musically trained brain where the motor parts recognize the external cue. This also corroborates the result from study [2] which showed that a good beat perception makes it easier to gain a longer stride length because of the advantage in recognizing a rhythm, giving a more effective rhythm therapy and auditory treatment. There were no measurement of beat perception in our study, but since the subjects did not recognize the music in test 2, 3, and 4 it is possible that the lower average of stride length to these auditory stimuli compared to the self-selected occurred because of the varying musical knowledge and beat perception within the participants.

Compared to stride length, test 5 did not result in as big difference of step frequency compared to the other tests. The average increase was 3.5 % compared to test 1-4 that had an increase of 0.8 %, 5.6 %, 3.2 % and 4.9 %. 6 participants had an increased stride length but decreased step frequency which is considered the best scenario. Most participants (9 of 19) had increased both stride length and step frequency in test 5. Because the self-selected songs were of varying genre, rhythm and to some extent BPM, test 5 is difficult to use in the analysis of what *kind* of auditory stimuli affects gait the most. This result could rather be used to draw conclusion about the recognition factor - which seemed to have a greater impact on stride length than step frequency. A possibility is that a healthy person's step frequency already lies relatively close to the tempo of the auditory stimuli which makes it easier to extend the stride rather than changing the ratio of steps.

The song in test 2 had the most tones per second – something that are related to feelings associated with stress, happiness or fear [11]. This might explain why it had the largest increase in step frequency, compared to test 3 that had the same tempo but less tones per second, associated with feelings such as sadness. This was reflected in the result as the stride length only increased by 2.2 % and step frequency by 3.2 % for test 3. An interesting part about test 3 was that the participants performed quite similar. It can be an interesting insight if you are to draw general conclusions about the impact of a certain type of auditory stimuli. The result from test 4 was neither high or low and did not stand out in any specific way, making it less interesting.

The order of the tests need to be discussed, since the order of the songs did not change between the participants. The increase in stride length might therefore be an effect of the test order itself causing a kind of learning curve within the test. This might explain why the metronome resulted in low values for both tests, which is not in line with previous studies. However, the results showed a constant increase in stride length but not in step frequency, which requires further studies about how the participants accustomed to the situation and how the learning factor affects gait in short term.

The overall result indicates a change in stride length and step frequency of healthy young adults when exposed to different auditory stimuli. In order to draw conclusions about how this result might be useful when constructing effective sensorimotor support for people with Parkinson's, it is important to take in account the complications that arise with neurological disorders. A previous mentioned study describes healthy, older, people to increase their stride length and gaining a steadier gait walking to a metronome compared to music. This also applies to most people with Parkinson's because it mainly affects the elderly. Another study [7] showed that a well-known song with a metronome on top of it produced the best result for Parkinson patients. Our study contributes to this research by adding a discussion about the cognitive aspects that are related to why the best result was reached with an auditory stimulus the participant enjoyed listening to. This trigger and rewards the cognitive parts of the brain, rather than just motor. Working cognitively and getting the Parkinson's patient to find pleasure in rehabilitation is a big part of long-term sensory-motor training. A development to this research would therefore be to investigate the change of gait when exposed to a self-chosen song the participant enjoys, with an imposed metronome. In that way, the song itself would activate the cognitive parts of the brain while the metronome helps the motor parts to react to the external cue.

6. CONCLUSION

This study examined the difference between different types of auditory stimuli to contribute to the research regarding gait rehabilitation for people suffering from Parkinson's disease. Since the test results turned out to be non-significant, there cannot be any definite conclusions drawn from this study. However, the result indicates that a song that is known and likeable for the patient has the largest impact on increasing stride length, and an average increase in step frequency compared to the other auditory stimuli. This study shows trends that recognition and learning factors are the most important for manipulating a person's stride length and step frequency among healthy young adults. This results also indicates that is is meaningful to choose auditory stimulus with concern when using it as sensorimotor support, since different stimuli seem to have different advantages.

7. REFERENCES

[1] Tuominen, P. 2018. Parkinsons Sjukdom. 1177 *Vårdguiden*. Hämtad 2 feb 2019 från https://www.1177.se/Stockholm/Fakta-och-rad/Sjukdomar/Parkinsons-sjukdom/

[2] Leow, L-A., Parrott, T., and Grahn, J. 2014. Individual differences in beat perception affect gait responses to low- and high-groove music. *Front. Hum. Neurosci.* 8, (October 2014), 811. DOI:https://doi.org/10.3389/fnhum.2014.00811

[3] E Morris, M., Iansek, R., A Matyas, T., and J Summers, J., 1996. Stride length regulation in Parkinson's disease: Normalization strategies and underlying mechanisms. *Brain* 119, 2 (April 1996), 551–568. DOI:https://doi.org/10.1093/brain/119.2.551

[4] Grau-Sánchez, J., Ramos, N., Duarte, E., Särkämö, T., and Rodríguez-Fornells, A., 2017. Time course of motor gains induced by music-supported therapy after stroke: An exploratory case study. *Neuropsychology* 31, 6 (2017), 624–635. DOI:https://doi.org/10.1037/neu0000355

[5] Sofuwa, O., Nieuwboer, A., Desloovere, K., Willems, A.-M., Chavret, F., & Jonkers, I. (2005). Quantitative Gait Analysis in Parkinson’s Disease: Comparison With a Healthy Control Group. *Archives of Physical Medicine and Rehabilitation*, *86*(5), 1007–1013. https://doi.org/10.1016/j.apmr.2004.08.012

[6] Nombela, C., Rae, C. L., Grahn, J. A., Barker, R. A., Owen, A. M., & Rowe, J. B. (2013). How often does music and rhythm improve patients' perception of motor symptoms in Parkinson's disease. *Journal of Neurology*. https://doi.org/10.1007/s00415-013-6860-z

[7] Benoit, C-E., Dalla Bella, S., Farrugia, N., Obrig, H., Mainka, S., and A Kotz, S. 2014. Musically cued gait-training improves both perceptual and motor timing in Parkinson's disease. *Front. Hum. Neurosci.* 8, (July 2014), 494. DOI:https://doi.org/10.3389/fnhum.2014.00494

[8] Zhang, S., Liu, D., Ye, D., Li, H., & Chen, F. (2017). Can music-based movement therapy improve motor dysfunction in patients with Parkinson's disease? Systematic review and meta-analysis. *Neurological Sciences : Official Journal of the Italian Neurological Society and of the Italian Society of Clinical Neurophysiology*, *38*(9), 1629–1636. https://doi.org/10.1007/s10072-017-3020-8

[9] Raglio, A. (2015). *Music Therapy Interventions in Parkinson's Disease: The State-of-the-Art*. *Frontiers in neurology* (Vol. 6). https://doi.org/10.3389/fneur.2015.00185

[10] L Chen, J., J Zatorre, R., and B Penhune, V. 2008. Listening to Musical Rhythms Recruits Motor Regions of the Brain. *Cereb. Cortex* 18, 12 (April 2008), 2844–2854. DOI:https://doi.org/10.1093/cercor/bhn042 *Neuropsychologia*, *96*, 96–110. https://doi.org/https://doi.org/10.1016/j.neuropsychologia.2017.01.004

[11] Bresin, R., & Friberg, A. (2011). Emotion rendering in music: Range and characteristic values of seven musical variables. *Cortex*, *47*(9), 1068–1081. https://doi.org/https://doi.org/10.1016/j.cortex.2011.05.009

[12] Williams, A. J., Peterson, D. S., & Earhart, G. M. (2013). Gait coordination in Parkinson disease: effects of step length and cadence manipulations. *Gait & Posture*, *38*(2), 340–344. https://doi.org/10.1016/j.gaitpost.2012.12.009

[13] M H Thaut, G C McIntosh, R R Rice, R A Miller, J Rathbun, and J M Brault. 1996. Rhythmic auditory stimulation in gait training for Parkinson's disease patients. *Mov. Disord.* 11, 2 (March 1996), 193–200. DOI:https://doi.org/10.1002/mds.870110213

[14] Pohl, P., Dizdar, N., and Hallert, E. 2013. The Ronnie Gardiner Rhythm and Music Method - a feasibility study in Parkinson's disease. *Disabil. Rehabil.* 35, 26 (2013), 2197–2204. DOI:https://doi.org/10.3109/09638288.2013.774060

[15] ANOVA-test: Definition, Types, Examples. Hämtad 24 april 2019 från: https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/hypothesis-testing/anova/

[16] J. Trost, W., Labbé, C., & Grandjean, D. (2017). Rhythmic entrainment as a musical affect induction mechanism.

Sonification for Process Monitoring in Highly Sensitive Surgical Tasks

[1]Sasan Matinfar, [2]Thomas Hermann, [3]Matthias Seibold,
[3]Philipp Fürnstahl, [3]Mazda Farshad, [1]Nassir Navab
[1] Computer Aided Medical Procedures, Technical University of Munich, Germany
[2] Ambient Intelligence Group, CITEC & Faculty of Technology, Bielefeld University, Germany
[3] Department of Orthopaedics, Balgrist University Hospital, University of Zurich, Switzerland.
sasanmatinfar@gmail.com

ABSTRACT

Surgeons usually have to keep track of many variables during a surgical intervention. This paper introduces three novel sonification approaches for fluid-related process data monitoring in the highly sensitive surgical context. From the instantaneous fluid (creation or expenditure) rate, different feature time series have been computed, including the cumulative fluid volume or filtered signals, which are in turn used for a set of sonification methods that structure a composed soundscape, either natural, musical or hybrid in real-time. We present 3 variations of 3 approaches with introductory example videos, each followed by results of a first user study in search of the preferred/most acceptable auditory representations. The qualitative evaluation of our method shows the potential for further research in this field.

1. INTRODUCTION

The advancements in computer technology have enabled the automatic collection of a vast amount of multi-modal data intra-operatively, i.e., during the course of medical intervention such as surgery. The problem is that the current limitations of display technology for the operating clinician lead to the situation that a considerable part of these data are either missing for the surgical team or cannot be perceived intra-operatively. One reason for this problem is that in the modern operating rooms, visual media have been used as the main approach to presenting the information. Yet sometimes it is difficult or even impossible to use visual monitors for data display. First, because the surgeons would then be forced to constantly switch their focus of attention between patient and monitor; Second, even by using virtual displays in AR, the overlay of many different visual cues occludes the field of view, which can cause the well-known inattentional blindness [1]. These are also the relevant problems in situations where monitoring of real-time data is merely the secondary task and the display persists during the entire time of the surgery.

Sonification, the systematic auditory display of data by using sound [2] offers many possibilities for addressing this problem. The omnidirectional propagation of sound does not require the user to attend to a specific location and they can thus focus on their primary task while perceiving peripheral information. This feature helps us to improve the problem of missing data during the surgery, i.e., information can be processed without requiring to shift the visual focus of attention. However, an important question here is how sonifications for process monitoring should be implemented - both from the aesthetics and individual preferences point of view, so that, particularly in a sensitive situation such as a surgery, it provides an information system which is at the same time ergonomic, informative, and workable. This will be more challenging and critical when the data is multidimensional.

In this paper, we propose three novel approaches to sonify a surgical signal, in particular, for a fluid-related signal. Fluid-related signals are measurements of the expenditure, loss, or consumption of fluid during the intervention. They are typically continuous signals and practical constraints may demand to stay within safe limits of either the instantaneous rate or the integral over the whole intervention. For this paper, we keep our detailed application scenario undisclosed to better generalize for the broader category of these situations. For our sonification methods presented here, we propose different signal processing approaches and generate different signals from a single initial signal to create a richer signal space for an "orchestrated" multi-stream sonification. We applied different kinds of filters, such as the moving average over different periods. Then, we mapped the parameters of the resulting features to sound, using Parameter Mapping Sonification (PMSon). The key idea in this approach is not only providing the original signal but to present additional information streams such as the accumulated signal and moving averages as well. Creating three variants for three approaches we obtained nine sonification types. We evaluated the resulting nine sonifications in a preliminary qualitative user study to determine aesthetics, compatibility with the main task, and to identify the best sonifications designs.

2. RELATED WORK

Sonification approaches for process monitoring are well established in the Auditory Display community. The applications are ranging from complex event-based situations

such as network traffic to production processes [3]. In the following, we provide a brief overview of systems that deal with monitoring as a secondary task.

Hermann et al. [4] proposed an audio-visual system for monitoring, querying and accessing information of different modules or processes in complex systems. They computed the emotional information of the processes and communicated them using sounds according to the emotional status of the process. For this approach, they used Musical Sonification and Model-Based Sonification.

Hildebrandt et al. [5] proposed a multi-modal approach for monitoring business processes using sonification to enhance conventional monitors and information systems which are depending mainly on the visual displays. In [6], they have been analyzed the business process monitoring tasks and the data structure in this field to build a foundation. According to this analysis, the usage of sonification for similar data structure has been studied so they summarized them in a list of recommendations that can be considered as guidelines for sonification of similar data.

Hildebrandt et al. [7] have demonstrated the advantages of continuous sonification: in a subjective experiment, they compared different peripheral process monitoring systems (visual only, visual + auditory alarm, and continuous sonification). Their comparison was based on the reaction time and the error-rate of the users by recognizing the events as they focus on and perform their primary task. Their results show that their continuous sonification for process monitoring, using a data-driven forest soundscape, significantly benefits the timing of attentional shifts towards the monitored system, as they become *anticipation optimal*.

EEG and anesthesia-machine are the established surgical sonification examples which have been already integrated into the surgical routines. Moreover, there have been several works for surgical sonification [8, 9], however, they have mainly focused on navigational tasks.

Matinfar et al. [10, 11], proposed a sonification approach for communicating peripheral information in a surgical task, in particular for ophthalmic surgery. The suggested approach conveys the information about the existence of the surgical tool in three discrete labeled area in the ocular cavity i.e. normal, careful, dangerous. The method can be considered as Event-Based PMSon since it notifies the surgeon as the tool entered a new area through the online modifications of a music stream.

In this context, the main task (performing surgery on a patient) is a critical and highly sensitive situation where the operator is completely focused and might be stressed. Because the result of her/his work directly affects human life, the sonification must be ergonomic, distinctive, and informative to best support the surgeon's work. At the same time, it should not be boring or annoying over long-term tasks, e.g. for an operation that lasts more than 3 hours.

3. APPLICATION SCENARIOS

The application scenario is a medical intervention over an extended time where the use, loss, or consumption or change of properties of one or more fluids is relevant to the success of the intervention [12, 13], we call the general class *fluid-based signal sonification*. The data are mostly continuous signals and practical constraints demand to stay within safe limits of either the instantaneous rate or the integral over the whole intervention. This type of information exists, for instance, in water loss, transpiration, urination, wound water, etc. as fluids; as well the pulse oximeter, for instance, is also concerned with a fluid signal, namely the oxygen saturation in the blood.

Specific to fluid-based signal sonifications is that fluids are usually incompressible, they change with a rate, and accumulate, either in loss or production. So relevant features are the integral over time, the actual rate, the change of the rate over time, the sign of the rate relating to whether the fluid is growing or depleting, and few other features that matter to the specific application scenario. Sometimes it might be interesting to have the integral over a relevant time interval to stay within healthy bounds, e.g. to avoid dehydration, and there could be different intervals that matter to different medical personal (anaesthetist, surgeon, etc.).

As several derived/relevant features are available, we decided, in line with experiences and recommendations from the above-cited related work, to create a multi-stream sonification, i.e., features that influence the sonic shape of a corresponding auditory stream. Specifically, we decided to use natural soundscape sounds and musical streams. The reason is that users must be able to process the sonifications at a low cognitive load since they need to focus on their primary task, so the sonification should only create an awareness of the signals and their temporal progression. The challenge, however, is to combine all the features and constraints into a working auditory display.

4. DATA REPRESENTATION

For this paper we use the variable δ (***delta***) as scalar variable to refer to the instantaneous rate of fluid flow, i.e. the loss or creation as measured by a sensor at a certain rate. For our project, we measured δ at a sampling rate of 1 Hz. Alternatively it can also be determined by subtraction of the current and previous accumulated amount of fluid by

$$\delta(t) = \frac{V(t) - V(t - \Delta t)}{\Delta t} \,. \tag{1}$$

By ***volume*** $V(t)$ we understand the accumulated *delta*, i.e., the integral over *delta* from beginning to time t as

$$\text{volume} = V(t) = \int_0^t \delta(t) dt \,. \tag{2}$$

The *volume* at time $t = n$ is the total accumulated amount of *delta* from $t = 0$ to $t = n$. ***FIR*** or Finite Impulse Response in signal processing characterizes filters whose impulse response is of finite duration, i.e., it settles to zero in a finite time. In this project, we applied FIR filters to the *delta* signal to obtain smooth features which average the instantaneous signal and do not introduce significant noise (s. Fig. 2).

Moving Average (MA), also called moving mean or rolling mean is a type of FIR. Here, we used the simple *MA*

which is simply the unweighted average of a fixed-size subset of a series of numbers. The *MA* at time t for the time period τ is the average over all t values from $t - \tau$ to t divided by the τ:

$$\mathrm{MA}(t) = \frac{1}{\tau}\left(\sum_{k=t-\tau}^{t} \delta(k)\right) \tag{3}$$

where the parameter τ determines the averaging time period.

5. SONIFICATION METHODS

The sonification technique used for this project is PMSon, and in particular in some parts Event-Based PMSon, based on different combinations of the signals as shown in Table 1. The mappings have been introduced in detail in Section 4, and for each signal combination, we proposed a specific approach as described below.

NSS	DDAC	CSSIS
MA 5s	delta	MA 30s
MA 30s	filtered-delta	MA 120s
MA 120s	volume	filtered-delta
volume		volume

Table 1: signal combinations in three sonification categories, see the text for abbreviations. We called *delta* filtered by FIR shortly *filtered-delta*.

We developed our sonifications in three major categories:

- Category 1: Natural Soundscape Sonification (NSS);
- Category 2: Data-driven Algorithmic Composition (DDAC);
- Category 3: Combined Soundscape & Synthetic Information Streams (CSSIS).

The sounds have been chosen according to aesthetic reasons and we cannot prove or claim that they are the best possible sounds for stream segregation or adoption by surgeons. Please see videos V1–V9, available as supplementary material at DOI:10.4119/unibi/2938433, which present and explain these sonifications with synchronized visualization of signal and features.

5.1 Natural Soundscape Sonification

This group included three sonifications, NNS-1–NNS-3, with different mappings (see V1–V3), and all of them have been implemented only based on natural sound samples from environmental soundscapes, such as shaking water, rain sounds, bird songs, seagull songs, etc. The signal combination in this group was the *MA* with different τ values namely 5, 30, 120 seconds, and the *volume* as depicted in Figure 1.

Everybody is familiar with natural sounds from sonic contexts, such as holidays, walks, work, etc. In this work,

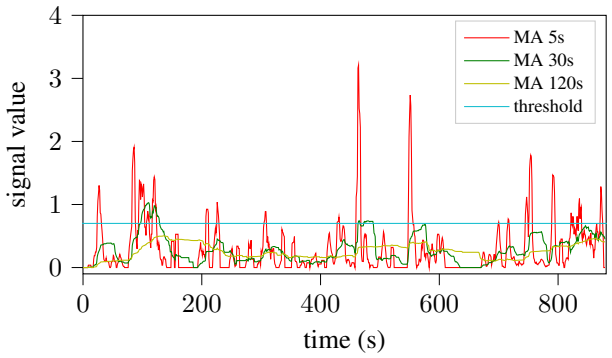


Figure 1: signal setup by NSS, threshold=0.75; x-axis depicts time in seconds, y-axis shows the fluid rate in arb. units. e.g. millilitre.

we want to investigate the impact of such intuitive sounds for process monitoring of highly sensitive surgical tasks.

In all three sonifications of this group, the *MA* of $\tau = 5s$ in the domain of [0, 4] has been mapped to the sound level and the rate of the water shake sample. Note that rate 1Hz refers to the original recording, 0.5Hz and 2Hz are half resp. 2 times the speed, i.e. an octave lower resp. higher.

The other *MA*s, with $\tau = 30s$ within [0, 2] and $\tau = 120s$ within [0, 1], have been mapped in each example differently. In the NSS-1 (V1), the *MA* $\tau = 30s$ and *MA* $\tau = 120s$ have been mapped to the samples of seagulls and motor boat. In NSS-2 and NSS-3, the *MA* $\tau = 30s$ has been mapped to the parameters of the birds sample; whereas, the *MA* with $\tau = 120s$ in NSS-2 has been mapped to a sample of horse, and in NSS-3 to the recorded sounds of rain. The details of the mappings in all nine sonifications is available in the appendix in the supplementary material DOI:10.4119/unibi/2938433. More details on the mapping parameters in NSS-3 illustrated in Table 2.

(a) MA 5s [0, 4]

water shake	level [-14, 0] dB, rate [1, 4] Hz

(b) MA 30s [0, 2]

birds	level [-12, 0] dB, rate [1, 2.5] Hz

(c) MA 120s [0.25, 1]

rain	level [-20, -4.5] dB rate [0.6, 1.2] Hz, pan [1, 0]

Table 2: Mapping details for NNS-3

Besides, we have defined a threshold for the *MA* of $\tau = 30s$, triggering a dedicated sound sample whenever that level is exceeded. Exceeding the threshold in NSS-1 and NSS-3 has been signalized with a thunder sound sample, and in NNS-2 with a herd of sheep. The rising *volume* (i.e., integral of *delta*) has been signalized every 50 ml by church bell strikes in all three sonifications.

The entire source and output values have been clipped to the given ranges, and all the mappings were linear.

5.2 Data-driven Algorithmic Composition

The sonifications of this group, DDAC-1–DDAC-3, have been composed based on the rule-based and stochastic techniques. The compositions have been created based on the synthesized instruments [1] such as a reverb, a musical pad, randomized melodic patterns of the marimba, and numbers of percussive strokes of a bell (V4–V6). The input signals *delta*, *filtered-delta*, and *volume* control a bunch of different parameters of the instruments. The signal mix is illustrated in Figure 2.

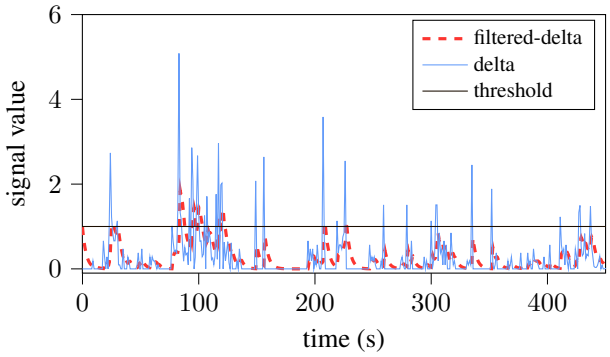


Figure 2: signal setup by DDAC, threshold=1; x-axis depicts time in seconds, y-axis shows the fluid rate in arb. units, e.g., millilitre.

The DDAC examples followed a similar concept of mapping, i.e., in all three variants each signal controls the same parameter, however, each time with different parameter levels. Thus, they create a different intensity in the whole sound texture. *Delta* within [0, 10] ml controlled inter-onset variations in the pad, and parameters such as frequency, detune, cutoff-frequency, and sound level in marimba. On the other hand, *volume* changes within the domain of [0, 250] ml have been reflected in the parameters of the reverb (mix, pre-delay, level, and pre-delay-time) and marimba (attack, sustain, release, reciprocal quality). Finally, exceeding the threshold $= 1$ by *filtered-delta* triggered the bell strikes of different times in each sonification example.

The sequence of marimba notes exhibits rhythmical similarity with the granular texture when a fluid (e.g. rain) drops on a solid surface (e.g. metal surface). Hence, the sonification using musical sounds can activate a metaphoric association to dripping fluid, allowing users, in turn, to infer more easily properties such as the density of grains. As the *delta* value increases, the rhythmical pattern and the tempo get faster and more disordered. The details of the mappings in DDAC-2 are shown in Table 3. All the source and output values have been clipped to the given ranges and the mappings were linear.

5.3 Combined Soundscapes & Synthetic Information Streams

The examples of this group, CSSIS-1–CSSIS-3, built a hybrid combination of previously introduced mapping ideas,

[1] All instruments in this project have been synthesized in SuperCollider, inspired by github.com/elifieldsteel/Supercollider3_tutorials_code

(a) delta [0, 10] in ml/s

pad	inter-onset[random(4.5-5.5), random(0.5-1.5)] s
marimba	inter-onset [random(0.99-1), random(0.1-0.2)] s frequency [1, 3] Hz detune [0, 15] Hz cutoff [1, 5] Hz level [-26, -3] dB pan [random(0-0), (-1-1)]

(b) filtered-delta [1, 4]

bell	3x strikes

(c) volume [0, 250] in ml

marimba	rq* [random(0.005, 0.008), random(0.09, 0.2)] attack [3, 1.5] s sustain [1, 0.5] s release [5, 2.5] s
reverb	reverb-time [1.8, 0.5] s mix [0.5, 0.1] predelay [0.4, 0.1] s level [-2, -14] dB

Table 3: mapping details in DDAC-2; *rq stands for reciprocal quality.

i.e. some of the signals have been mapped to the natural soundscape sounds whereas others have been mapped to music parameters. The general idea of music has emerged by randomizing the melodic patterns based on the Japanese scale *Hirajoshi* [2] . The signal collection in this group consisted of *filtered-delta*, the *MA* with $\tau = 30s$ and $\tau = 120s$ as shown in Figure 3.

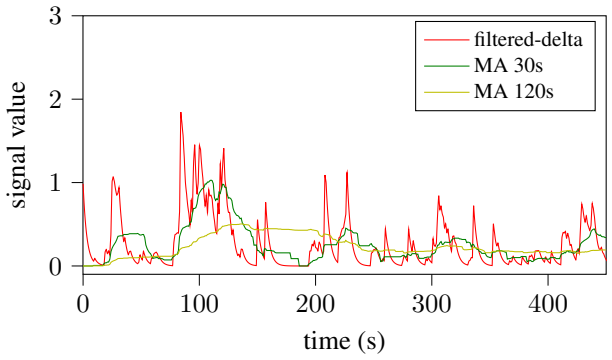


Figure 3: signal setup by CSSIS; x-axis depicts time in seconds, y-axis shows the fluid rate in arb. units, e.g. millilitre.

The *filtered-delta* within [0, 4] has been mapped to the sound level and the rate of water shake sample in CSSIS-1 (V7), birds in CSSIS-2 (V8), and both water shake and

[2] With note C as the root, the scale is C E $F^{\#}$ G B

birds in CSSIS-3 (V9).

The *MA* 30s in the domain of [0, 2] has been mapped, in various combinations, to different musical parameters such as sound level and tempo of a randomized melodic pattern based on the Japanese scale. The melodies have been played by different instrument combinations such as *Guzheng*, *Pipa* [3] , and synthesized marimba. All mapping details are detailed in the supplementary material [4] .

Finally, the mappings of the *MA* with $\tau = 120s$ in the domain of [0, 1] were as follows: in CSSIS-1 to the sound level of the Bamboo flute; in CSSIS-2 to the sound level of the synthesized marimba; and in CSSIS-3 to the sound level and the rate of the rain sample.

The entire source and outputs have been clipped to the given ranges of the mapping functions while all the mappings were linear. The *volume* has been signalized in every 50 ml with the strikes of *Taiko* [5] in CSSIS-1 and CSSIS-2, and with Japanese bell in CSSIS-3. The details of the mappings in CSSIS-1 are illustrated in Table 4.

(a) filtered-delta [0, 4]

water shake	level [-20, 0] dB, rate [1, 3] Hz

(b) MA 30s [0, 2]

melodic pattern	by values > 0.5 : Guzheng: level [-40, -6] dB by values > 0.75 : Pipa: level [-40, -10] dB

(c) MA 120s [0, 1]

Bamboo	by values > 0.45 : level [-26, -14] dB

Table 4: mapping details in CSSIS-1

6. EVALUATION AND RESULTS

We report results of the first step in a two-step quantitative evaluation: step one is the aesthetic and pragmatic evaluation at hand of regular listeners – in order to keep the precious and limited time of the target group of surgeons reserved for step two: the assessment of pre-selected sonification types with the target group.

In the first step, we conducted a qualitative questionnaire study including 9 videos. Each video ($\approx$ 2 minutes) presented one sonification example including the visual presentation of the corresponding signals. The videos consisted of the manually selected highlights of the signal flow. After watching each video, the participants (10 non-experts in music, sound design, or medicine) were asked to express their degree of agreement to 12 statements (resp. questions) using the Likert scale format. The sequence of the videos was reordered anew for each participant. We aimed at identifying the best-rated sonification example in each category for our further optimization and subsequent quantitative studies with the target group. The 12 questions include: 6 questions on how useful, recognizable, pleasant etc. (positive features); 4 questions on how distracting, annoying etc. (negative features) the system was; and the last 2 questions on how long they would estimate to listen to the sonifications in both passive and active modes (questionnaire available in appendix). The participants were asked to give a grade to each question. We gave each response from "Strongly Disagree" to "Strongly Agree" respectively from 1 to 5 points by positive features, and vice versa by negative features. Their estimated time of listening has been pointed from 1 to 5 for a range between "less than 5 minutes" and "more than 60 minutes". The average and statistic error are shown in Table 5.

NSS-1		NSS-2		**NSS-3**	
34.9	2.25	30.8	2.55	**42.7**	**2.41**
DDAC-1		**DDAC-2**		DDAC-3	
39.7	2.38	**40.0**	**2.75**	38.3	2.48
CSSIS-1		CSSIS-2		CSSIS-3	
41.7	2.31	41.6	2.73	38.9	1.95

Table 5: the average (left) and standard error (right) for each example; sonifications with the highest average in each category highlighted.

7. DISCUSSION AND CONCLUSIONS

Although, our preliminary evaluation was not meant to statistically prove the significant effects of our method on recognition of signals and the reaction time, yet they guide us for our future study. The average point in all the examples is above 50% of the maximum possible 60 points, this means that our system has been positively accepted by the participants, especially according to the subjective features such as pleasantness or annoyance (please note that by features such as annoyance, which have negative impact, we inverted the scale so that higher values refer to better impact). This is important since the subjective preferences of the surgical crew affect the acceptance of the system in such cases. Listening to music, e.g. on the radio during surgery is a common practice by many surgeons and in many hospitals. So integrating information into a pleasant, engaging, and not annoying music stream could help bringing valuable information into the surgical routine in a convenient way.

To identify the best candidates in each category, we performed an analysis of variance (ANOVA) for each category separately. The result of ANOVA using $\alpha = 0.05$ fails to reject the null hypothesis for the groups DDAC and CSSIS which means we can't find a significant difference between the examples in any of both groups. However, the test rejects the null hypothesis by NSS and shows there is a candidate with a significant difference (see Table 6). Performing Bonferroni post-hoc for NSS results in that NSS-3 differs significantly from NSS-2 (Table 7 and Figure 4). This suggests NSS-3 as the best possible candidate in its category for further our studies.

[3] *Guzheng* and *Pipa*, both are the traditional Japanese instruments.

[4] DOI:10.4119/unibi/2938433

[5] *Taiko* refers to a broad range of traditional Japanese percussion instruments.

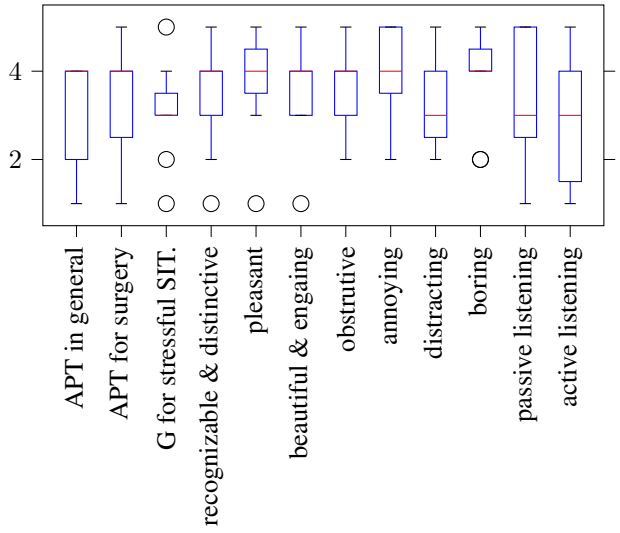
category	F-statistic	PR(>F)
NSS	**6.298162**	**0.005689**
DDAC	0.127451	0.880863
CSSIS	0.456207	0.638478

Table 6: ANOVA shows there is an example by NSS whose average is significantly different from other groups.

compared groups	statistic	p-value
NSS-1 & NSS-2	1.2058	0.2435
NSS-2 & NSS-3	**-3.389**	**0.0033**
NSS-1 & NSS-3	-2.3601	0.0298

Table 7: Bonferroni correction post-hoc comparison. The p-value by NSS-2 & NSS-3 is less than the corrected p-value 0.0166, $\alpha = 0.05$.

Figure 4: boxplot NNS-3



In summary, we presented nine variants of sonification methods for process monitoring of fluid time series as secondary tasks in highly sensitive primary task contexts. Based on our qualitative evaluation, we can assume that our approach of data-driven soundscapes has the potential for further study and quantitative evaluation of the error rates and reaction times in judging variables at random points in time.

8. REFERENCES

[1] B. J. Dixon, M. J. Daly, H. H. Chan, A. Vescan, I. J. Witterick, and J. C. Irish, "Inattentional blindness increased with augmented reality surgical navigation," *American journal of rhinology & allergy*, vol. 28, no. 5, pp. 433–437, 2014.

[2] T. Hermann, "Taxonomy and definitions for sonification and auditory display," in *Proc. Int. Conf. Auditory Display (ICAD 2008)*. International Community for Auditory Display, 2008.

[3] P. Vickers, "Sonification for Process Monitoring," in *The Sonification Handbook*, T. Hermann, A. Hunt, and J. G. Neuhoff, Eds. Berlin, Germany: Logos Publishing House, 2011, pp. 455–491. [Online]. Available: http://sonification.de/handbook/chapters/chapter18/

[4] T. Hermann, C. Niehus, and H. Ritter, "Interactive visualization and sonification for monitoring complex processes," in *Proceedings of the International Conference on Auditory Display*, 2003.

[5] T. Hildebrandt, "Short paper: Towards enhancing business process monitoring with sonification," in *International Conference on Business Process Management*. Springer, 2013, pp. 529–536.

[6] T. Hildebrandt and S. Rinderle-Ma, "Toward a sonification concept for business process monitoring," in *Proceedings of the International Conference on Auditory Display (ICAD 2013)*. Lodz University of Technology Press, 2013.

[7] T. Hildebrandt, T. Hermann, and S. Rinderle-Ma, "Continuous sonification enhances adequacy of interactions in peripheral process monitoring," *International Journal of Human-Computer Studies*, vol. 95, pp. 54–65, 2016.

[8] C. Hansen, D. Black, C. Lange, F. Rieber, W. Lamadé, M. Donati, K. J. Oldhafer, and H. K. Hahn, "Auditory support for resection guidance in navigated liver surgery," *The International Journal of Medical Robotics and Computer Assisted Surgery*, vol. 9, no. 1, pp. 36–43, 2013.

[9] H. Roodaki, N. Navab, A. Eslami, C. Stapleton, and N. Navab, "Sonifeye: Sonification of visual information using physical modeling sound synthesis," *IEEE transactions on visualization and computer graphics*, vol. 23, no. 11, pp. 2366–2371, 2017.

[10] S. Matinfar, M. A. Nasseri, U. Eck, H. Roodaki, N. Navab, C. P. Lohmann, M. Maier, and N. Navab, "Surgical soundtracks: Towards automatic musical augmentation of surgical procedures," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 673–681.

[11] S. Matinfar, M. A. Nasseri, U. Eck, M. Kowalsky, H. Roodaki, N. Navab, C. P. Lohmann, M. Maier, and N. Navab, "Surgical soundtracks: automatic acoustic augmentation of surgical procedures," *International journal of computer assisted radiology and surgery*, vol. 13, no. 9, pp. 1345–1355, 2018.

[12] M. H. Rosenthal, "Intraoperative fluid managementwhat and how much?" *Chest*, vol. 115, no. 5, pp. 106S–112S, 1999.

[13] M. Doherty and D. Buggy, "Intraoperative fluids: how much is too much?" *British journal of anaesthesia*, vol. 109, no. 1, pp. 69–79, 2012.

Chapter 3

Papers presented at ISon 2019

Proceedings of ISon 2019, 6th Interactive Sonification Workshop, KTH Royal Institute of Technology, Sweden, November 18, 2019

TOWARDS INTERACTIVE SONIFICATION IN MONITORING OF DYNAMIC PROCESSES

Niklas Rönnberg

Linköping University,
Division for Media and Information Technology
Linköping, Sweden
niklas.ronnberg@liu.se

ABSTRACT

The modern control room is predominantly made up of visual displays, which might make monitoring different processes a cumbersome and visually cognitively demanding task. Sonification could be used to support the monitoring task. However, it is not clear how the most beneficial sonification should be designed. In this pilot study an experimental setup was developed to explore perception of different sonification approaches. A user experiment was performed to assess perception of the sonification when and where simulated data deviated most from the normal level. It was found that all sonification conditions were generally useful, regardless of the participant's musical experience, shown both in terms of objective and subjective measurements. Stereo panning of the sound was also generally experienced as helpful, but the use of different pitch might not have been experienced to contribute as much for solving the task. The findings in this pilot study will be further used to create new research ideas about sonification for monitoring of dynamic processes.

1. INTRODUCTION

As the foreman arrived at the power plant, the factory or the ironworks in the morning and entered the factory floor, the sound, the sonic ambience or the soundscape of the work place would tell about the night shift, about machines in need of repair or maintenance, about the overall performance and the general status of the plant. For a closer insight individual meters and indicators could be read. As time changed more and more of the surveillance and monitoring were moved to quiet and air-conditioned control rooms (see examples in Figure 1 and 2). These monitoring environments provide more information and one person can easily monitor multiple dynamic processes simultaneously. As a consequence, not only in process control but in monitoring in general, the amount of visual information has increased while auditory information has decreased. The sonic ambience, the peripheral monitoring through the soundscape, has been lost.

If too much information is presented in the visual modality, there is a risk of cognitive overload (see for example discussions in [1, 2, 3, 4]). The consequences might be that information is neglected and ignored, or completely missed (see, for example, discussions in [5, 6, 7]). Not only that, visualization also presents challenges for the visual perception, such as simultaneous brightness contrast [8]. Simultaneous brightness is when a colored area with a set luminance is perceived as brighter when it is surrounded by darker hues compared with when it is surrounded with brighter hues. Another challenge is the Mach band phenomenon [9], which occurs at boundaries between different hues, and a solid hue is perceived as a gradient where it is brighter at the border to a darker hue and darker at the border to a brighter hue. Such challenges might negatively effect the perception of a visualization, where constructs like density levels or amount in data might be encoded as intensity levels. Perceiving differences in intensity levels could

Figure 1: *One example of a modern control room for monitoring of dynamic processes, consisting primarily of displays and visual information. Photo courtesy ABB.*

Figure 2: *One example of a modern remote tower control room, for monitoring and controlling the air traffic at and around an airport, built-up by multiple displays. Photo courtesy LFV.*

be essential for understanding and interpreting visualization correctly, consequently flaws in the visual perception will affect the perception of visual representations negatively. Therefore, it seems justified to argue that sound should be reintroduced in process control and monitoring of dynamic processes.

The challenges with visualization and cognitive load in the visual modality, could be addressed by the use of sonification. Sonification focuses primarily on turning data into sound and could be considered as a complementary modality to the visual modality [10, 11, 12]. Sonification has the ability to provide additional input and further information [13, 14], and the combining of visual and the auditory modalities should be able to present more effective and efficient multimodal visual representations [15]. Also, by adding sound as an additional modality visual cognitive load can be reduced [16]. Sonification can successfully be used in process control and process monitoring [17]. Auditory icons, caricatures of naturally occurring sounds [18], can be used in multiprocessing and collaborative systems for diagnosing problems, monitoring a set of processes as well as individual processes, and providing a shared reference point for collaboration [19].

For sonification to be useful for data exploration, dynamic human interaction is necessary (see discussion in [20, 21]). Therefore, in a monitoring control situation, interactivity is essential for exploration of historical data and for making comparisons between past and present states possible.

Even though some research suggests that natural real-world sounds might be better in a soundscape for monitoring and control [22, 23, 24], sounds can also be designed deliberately with a music-theoretical and aesthetic approach to create a nice sounding sonic ambience. The aim of such a sonic ambience could be to provide a peripheral awareness of the overall status of one or several processes. The use of musical sounds provides design opportunities that are to some degree lacking in sonification approaches based on arbitrary or natural sounds, including musical qualities such as timbre, harmony and tempo. The reason for using a musical approach is somewhat similar to the sonification approach to Barra et al. 2001 and 2002 [25, 26] who used musical sounds or an aesthetic approach to sounds bordering between music and background noise as compared to simple alarm sounds, reasoning that aesthetically designed sounds might minimize fatigue and annoyance in long-term monitoring. Musical structures and compositions have an ability to convey a multitude of information to listeners quickly and intuitively [27], suggesting that the use of musical sounds should be well suited for sonification for monitoring control systems. Music (and a musical approach to sonification) can convey meaning, information, and emotions (see for example discussions in [28, 27]), and sonification with musical sounds also seems to be able to support visual perception [29].

By using sonification for peripheral monitoring it should be possible to provide a sonic ambience that could indicate changes in one or multiple levels of change from the normal status-quo level of a machine, or an operator, or an entire process. Such a soundscape would then provide status information peripherally, creating an awareness of overall system conditions (see an example in Rönnberg et al. 2016 [30]). The notion of peripheral sonification is not new, but the choice of musical sonification of an overwhelmingly visual task introduces some opportunities to broaden the understanding of the concept of peripheral sonification. For instance, continuous soundscapes could provide a base for new and interesting research questions compared with short repetitive approaches or strictly musical treatments [31, 32].

1.1. Aims and objectives

This project is a pilot study, an exploration and analysis of sonification design options. The aim of this research is to develop an experimental setup for sonification of multi-variate time-varying data for a future monitoring setting. However, it is not clear how to design this sonification for monitoring. Therefore, this pilot study aims to examine the following questions:

1. What musical elements are suitable to be used to sonify data from dynamic processes?
2. Can stereo spatial audio be used to support perception of sonification?

An experimental study with an interactive search task was performed to address these questions.

2. METHOD

To examine the use of musical elements and to assess whether a user could distinguish between different levels within each musical element, an interactive search task was designed. In this experimental setup three positions (left, center, right) were sonified (see a screen-shot of the test interface in Figure 3). For one of these positions the sonification changed over time due to the underlying data. The participant's task was to mark the position for which the sonification changed, and when in time the sonification deviated most from the normal state. This is not a typical monitoring situation, but rather a way to assess whether the different sonification design ideas could be useful for future research.

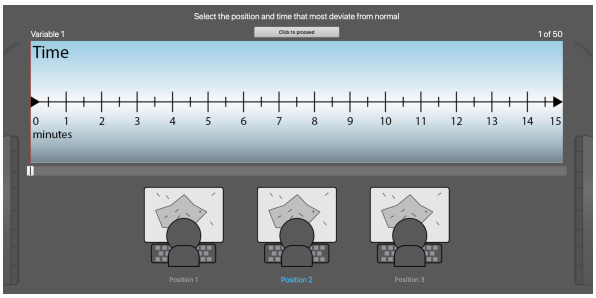


Figure 3: *The user interface used in the experiment. The three positions were selected by clicking on the corresponding icon, the position on the time line was selected by moving the horizontal slider, and the selections were confirmed by clicking on the button (marked "Click to proceed").*

2.1. Simulation of data

The present study uses simulated data, inspired by data that could be obtained in monitoring of dynamic processes. The data was constructed to mimic time-varying continuous data, as well as time-varying discrete data consisting of three levels: 1 - normal levels, 2 - intermediate levels, and 3 - high levels. The continuous data will hereafter be referred to as Data-A, and the discrete data will hereafter be referred to as Data-B. All data was computed using Matlab R_2018b. The data in Data-A could be seen as representative of heart rate, temperature, or stress levels, while data in Data-B could be representative of number of incursions, level of warnings, or severity of conditions.

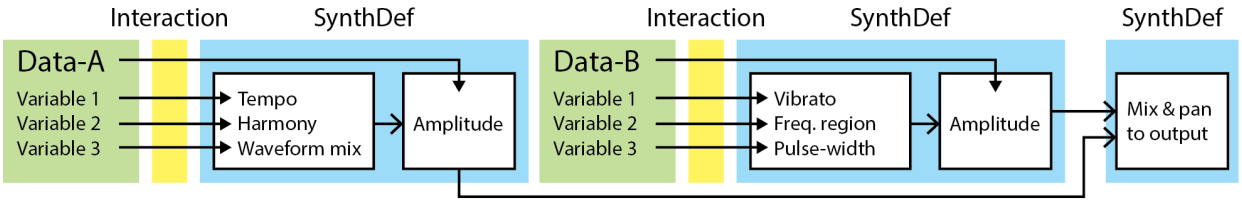


Figure 4: *A model of the sonification implemented in SuperCollider. Showing variables (in green), interaction (in yellow), and synth definitions (in blue).*

2.2. Implementation

The experiment was implemented in SuperCollider 3.10 [33, 34], which is a real-time audio synthesis programming environment. Interaction was implemented using a computer mouse, the participant moved a time bar and the sonification was changed according to the underlying (and invisible to the participant) data (see Figures 3 and 4)). For exploring a sonification of a set of data a static auditory display/graph is not enough, but for a user to be able to compare different positions (left, center, right), interaction is necessary (see further discussions on sonification and interaction in [20, 21]). A questionnaire, printed on paper, was administered to the participants for recording of subjective data.

A short video demonstration can be found here: https://vimeo.com/353209351

SuperCollider was run on a MacBook Pro computer, presenting the user interface on a 21" computer screen and sound through a Universal Audio Apollo X8 sound interface and a pair of AKG K271 MKII headphones. The headphones provided an auditory stimulation of approximately 65 dB SPL. All experiments were conducted in a quiet office. Even if some ambient sounds was present in the background, the experimental environment was deemed quiet enough not to affect the outcome of the experiments.

2.3. Design of the sonification

The sonification was designed to allow the exploration of different musical elements such as sound level, tempo, harmony, timbre by wave form mix, vibrato, frequency region of pink noise, and timbre by variation of pulse-width. Three positions (left, center, right) were used in the experimental setup. These positions could reflect three operators or three machines being monitored. Each position was sonified using one tone each for Data-A, as well as two additional tones each for Data-B. These tones were C, E, G, and thus formed a major C chord [35] (see Table 1).

Table 1: The tones used for the three positions, and for the different types of data, displaying the corresponding note on a piano keyboard, the MIDI note number, and the frequency in Hz.

Position	Type of data	Note	MIDI	Frequency (Hz)
Position 1 (left)	Data-A Data-B	C4 C3, E3	60 48, 52	261.63 130.81, 164.81
Position 2 (center)	Data-A Data-B	E4 E3, G3	64 52, 55	329.63 164.81, 196.00
Position 3 (right)	Data-A Data-B	G4 G3, C4	67 55, 60	392.00 196.00, 261.63

2.4. Musical elements and mapping to data

The sonification was designed to provide information about individual variables, but still work in combination where multiple variables varied simultaneously. The sonifications of Data-A always used the same basic sound, built up by triangle waves, in different pitch for each individual position (left, center, right). The data in Data-A was mapped to clearly distinguishable levels with some variability within each level. This created almost discrete levels in the data (normal levels, intermediate levels, and high levels), somewhat similar to the discrete data in Data-B. For Data-B, the sonification of two variables used the tones, built up by square waves, creating the basic underlying musical sonification (see Table 1). One variable used pink noise. The sonification parameters were designed to be able to both function alone as well as in combination (see Table 2 and Figure 4).

Table 2: Sonification settings for the different simulated data types in the three different levels.

Simulated data	Level 1 - normal	Level 2 - intermediate	Level 3 - high
Data-A	normal sound level	increased sound level	more increased sound level
Variable 1 *Tempo*	60 bpm, 30% sound level of beat	110 bpm, 50% sound level of beat	170 bpm, 80% sound level of beat
Variable 2 *Harmony*	1.5 cent, harmonic, not much beating of frequencies	26 cents, a bit disharmonic, some beating of frequencies	50 cents, disharmonic, beating of frequencies
Variable 3 *Waveform mix*	100% triangle, 0% sawtooth waves	60% triangle, 40% sawtooth waves	10% triangle, 90% sawtooth waves
Data-B	normal sound level, somewhat attenuated high frequencies	increased sound level, more high frequency content	even more increased sound level, most high frequencies
Variable 1 *Vibrato*	no perceivable vibrato	increased vibrato depth	most noticeable vibrato
Variable 2 *Frequency region*	pink noise, 10% output level, BPF cutoff freq. 500 Hz	pink noise, 55% output level, BPF cutoff freq. 1000 Hz	pink noise, 100% output level, BPF cutoff freq. 1500 Hz
Variable 3 *Pulse-width*	50% pulse-width	70% pulse-width	90% pulse-width

Sound level was used in connection with other musical elements for all variables in Data-A. The data level (normal, intermediate, high) in each variable were mapped linearly to exponentially to the amplitude. Thus, as the data increased in level, the amplitude of that specific sonification condition increased as well. A louder sound level might give rise to a higher activity in the listener compared to a lower sound level [36, 37] why sound level should be useful in sonification. No amplitude normalization for

different frequencies was performed.

Tempo was used to sonify the first variable in Data-A. The sonification for all three positions had a basic beat tempo 60 beats per minute (bpm), where an envelope generator was re-triggered for every beat. The tempo of this beat changed in one position according to the data, and was at the intermediate level 110 bpm, and 170 bpm at the highest level. The beat was mixed together with a steady tone to create a continuous signal with a periodic rhythmic pulse. The intensity of the envelope generator increased from 30% sound level in low, to 50% sound level in intermediate, to 80% sound level in high. Consequently, as the data level increased, the tempo of the beat also increased in speed and became more prominent. A faster tempo gives a stronger arousal when listening to music [38] or to sonification, consequently, a faster tempo should suggest an increased level of urgency.

Harmony was used to sonify the second variable in Data-A. Each tone used for Data-A for each position consisted of 5 tones, one at the fundamental frequency and two tones somewhat below respectively above the fundamental frequency. The distance from the fundamental frequency of these harmonics ranged from 1.5 cent at the lowest levels of in the data to 50 cents at the highest levels. Cent is a logarithmic unit, where the interval between each semitone is divided into 100 cent [39]. As harmonic components are further apart in relation to the fundamental frequency the interference between these frequencies creates a beating [40] which is equal to the difference in frequency of the notes that interfere [41, 42]. Thus, as data level increased in one position, the amount of dissonance also increased for that specific position, and the beating created of these frequencies increased.

Waveform mix, i.e. timbre, was used to sonify the third variable in Data-A. Timbre might be described as the "color" of the sound, the tone quality, formed by the different sounds and their inherent characteristics. A softer and more dull timbre might be experienced as more negative compared to brighter timbre [37], and a more complex timbre might be more captivating evoking greater (emotional) responses compared to a simpler timbre [36]. The tones used at each position for Data-A was built up with triangle waves. Triangle waves consists of odd harmonics with quite steep roll off [43] why the triangle wave is perceived as quite soft, a bit round, and without too much high frequency content. As the level in the data increased, sawtooth waves were mixed together with the triangle waves. Sawtooth waves have both even and odd harmonics why the sawtooth is much richer of high frequency components [43] and might therefore be perceived as harsher and sharper. Consequently, as the data level increased, the amount of high frequency content in the sonification increased as well, creating a more distinct and piercing sound.

Sound level and cutoff frequency of a low pass filter (LPF) was used in connection with other musical elements for all variables in Data-B. Similar as for Data-A, the data level (normal, intermediate, high) was mapped linearly to exponentially to the amplitude of the sonification. The cutoff frequency was mapped to be between about 250 to 1200 Hz depending on the level in Data-B as well as the fundamental note used in the sonification. Consequently, as the level in the data increased, the sonification for that data increased in sound level as well as in high frequency content.

Vibrato was used to sonify the first variable in Data-B. Two tones were used together for sonifying the variables in Data-B. The data in the first variable in Data-B was linearly mapped to vibrato depth of the two tones, from not perceivable vibrato at the lowest data level to a vibrato that was +/- a quarter tone of the fundamental frequency. The vibrato speed was set to 4 Hz creating a quite nice sounding vibrato giving a noticeable vibrato effect and, as the data level increased the depth or intensity of the vibrato increased.

Frequency region of pink noise was used to sonify the second variable in Data-B. The pink noise passed through a band-pass filter (BPF). The data level was linearly mapped to the cutoff frequency of the BPF, between 500 (for the low level) to 1500 Hz (for the highest level). The data level was also linearly mapped to the output level of the filter from almost completely attenuated noise at the low level (10% of the output level) to full sound at the highest level. Therefore, as the level in the data increased for one position, the pink noise also increased in sound level and in more high frequency content for that specific position.

Pulse-width variation, i.e. timbre, was used to sonify the third variable in Data-B. The data level was linearly mapped to the pulse-width of the square waves, where the normal level was mapped to 50% pulse-width, intermediate level to 70%, and high to 90% pulse-width. As the pulse-width increases (or decreases) from 50% the amount of harmonics increases [43]. As the harmonics increases the timbre of the sound changes to become richer and more complex. Consequently, as data level increased, the sound quality of the sonification became richer in harmonics and more salient for that position compared to the normal level in the data.

2.5. Panning and sonification level

To explore if stereo panning (left, center, right) could support perception of the sonification, the sonification was used in stereo. The stereophonic sound image always placed position 1 to the left relative the other positions, position 2 in the center, and position 3 to the right (see Figure 5. When a position was selected this position was panned to the center in the stereophonic sound image, and the other positions moved correspondingly. The experienced loudness of a stereo sound is dependent on the panning, and consequently sounds were attenuated appropriately to maintain a good perception of all sounds (see Table 3). The user selected the desired position by clicking on the corresponding operator image, the stereo panning and attenuation was instantly performed and the position of the operator images was also moved accordingly.

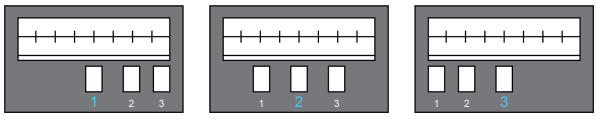


Figure 5: *The panning was positioning the selected position to the center of the stereophonic sound image. Left: position 1 is selected. Middle: position 2 is selected. Right: position 3 is selected.*

Table 3: The panning and attenuation settings for all selected positions.

Selected position	Setting	Position 1	Position 2	Position 3
Position 1 (left)	Panning	0	0.65	1
	Level	0.707	0.84	1
Position 2 (center)	Panning	-0.65	0	0.65
	Level	0.84	0.707	0.84
Position 3 (right)	Panning	-1	-0.65	0
	Level	1	0.84	0.707

2.6. Participants

For the present pilot study, 15 participants were recruited, (8 female) with a median age of 30 (range 23 to 52) with normal, or corrected to normal, vision and self-reported normal hearing. No compensation for participating in the study was provided.

2.7. Experimental procedure and questionnaire

Each test session was initiated by the participant giving a subjective rating of their musical experience, by answering two questions using a 5-point Likert scale. These questions asked whether the participant listens to music from 1 (Not very often) to 5 (All the time), and whether the participant 1 never sang or played an instrument to 5 is playing or singing regularly.

Each sonification condition as well as the user interface was then introduced to the participant for familiarization. The participants task was to mark, using the computer mouse, where on the timeline the sonification changed the most compared to the normal level (which was present in the beginning of the time line), and for which position (left, center, right) this change was connected to (see Figure 3). After the introduction followed either one of the three sonification conditions connected to Data-A or Data-B. The order of these was balanced between participants to avoid order effects. The position that was affected by the increased levels in the data was randomized but overall balanced within each experiment.

After each sonification condition the participant answered two questions in a questionnaire about their subjective experience of the sonification. These questions concerned the experienced difficulty level in finding the position that had a sonification suggesting a deviation from the normal level, as well as the difficulty in finding the time where the sonification differed as most from the normal level. Answer alternatives ranged from 1 (very hard) to 5 (very easy). In total 50 sonification conditions were used in the experiment. After the test the participant answered some final questions in the questionnaire considering to what extent the stereo panning supported in providing answers, as well as to what extent the different tones contributed to providing the answers. Possible answer alternatives ranged from 1 (very little) to 5 (very much).

The experiment yielded subjective accuracy data for how well the participant managed to mark the time and position where the sonification most deviated from the normal, as well as subjective ratings of experience of the sonification.

3. RESULTS

According to Kolmogorov–Smirnov tests the data was not normal distributed, thus non-parametric tests were used. Bonferroni correction for multiple comparisons was applied as appropriate.

Accuracy was measured in terms of the percentage of correct responses given for that sonification condition. Generally the accuracy was high for all six sonification conditions, with a mean accuracy of over 65% in all conditions (see Figure 6). A Friedman test showed no significant differences in accuracy between the six conditions (*tempo, harmony, waveform-mix, vibrato, frequency region, pulse-width*), $\chi^2(5) = 6.81, p = 0.235$. There were no effects of age, gender, or musical experience. However, considering the low number of participants in the present pilot study, studying mean values and 95% confidence intervals might suggest trends in the data. Consequently the accuracy for *harmony* as well as *vibrato* might be less compared to the other sonification conditions.

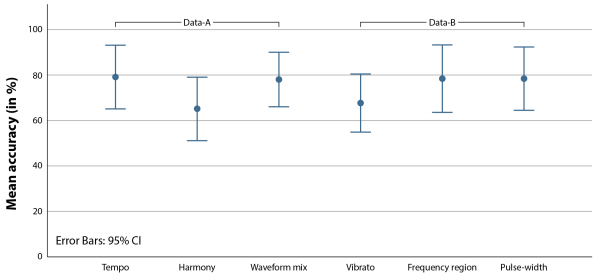


Figure 6: *Error bar graph showing mean accuracy and 95% confidence intervals for the six sonification conditions.*

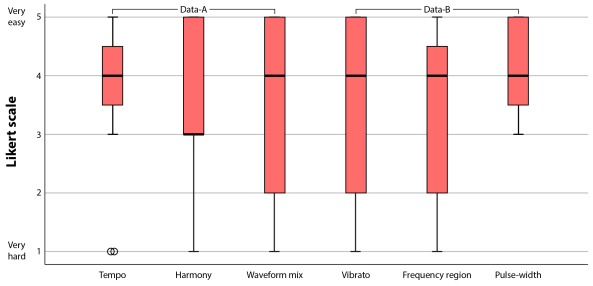


Figure 7: *Box plot showing subjective ratings of experienced difficulty for selecting the correct position.*

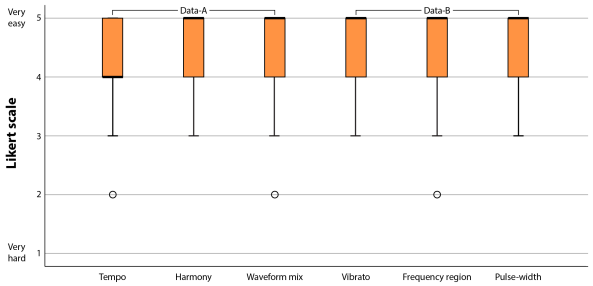


Figure 8: *Box plot showing subjective ratings of experienced difficulty for finding the correct time.*

For subjective measurements, i.e. the difficulty in finding the correct position (left, center, right), and the difficulty in finding the time that most deviated from the normal level, there were no significant differences between sonification conditions (see Figure 7 and Figure 8). Generally, the participants experienced distinguishing between the three positions as fairly easy, and finding the time that most differentiated from the normal level as easy. When comparing the ratings between the position and the time, then maybe finding the time was experienced as somewhat easier than finding the right position (see average rankings in Figure 9). Also, in general the stereo panning was experienced as helpful in as well as the different pitch used for the three positions (see Figure 9).

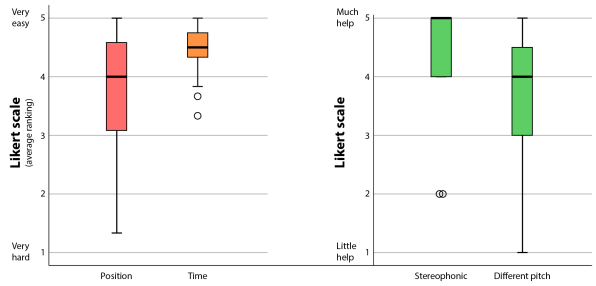


Figure 9: *Box plot showing subjective ratings of difficulty in solving the task (left), and ratings of experienced help (right).*

4. DISCUSSION

It is important to remember that the present pilot study evaluated different sonification design approaches for future studies in sonification for monitoring, but used an interactive search task to assess the perception of the musical elements used to sonify changes in the data. The amount of interaction and the available time for exploring data in a typical monitoring situation is most likely limited compared the present pilot study. The case might be that the operator that monitors a dynamic process has no possibility to interact with the monitored data, but rather needs to attend to changes in the sonification solemnly. The focus of the present study was the use and perception of different musical elements, and therefore the interactive search task was used to assess the sonification. The mean accuracy for finding the highest level in the sonification, was high in all sonification conditions, and the lack of statistically significant differences in accuracy between conditions suggests that all sonification conditions provided enough information for the participants to solve the tasks in the experiment. This result was also supported by the subjective rankings.

The measured accuracy was overall high, which might suggest that the task in the experimental setup was simple, maybe too simple. As the experiment aimed to assess whether the levels in the data was perceivable in the sonification or not, the high accuracy was interpreted as something promising for the sonification approaches. However, response time was not recorded why it is not possible to determine whether all sonification conditions had similar response time or not. It is possible that the same level of accuracy was achieved but with considerably longer response times for some conditions. A future experimental setup where the participants are asked to provide as swift answers as possible, or where there is a time constraint, might reveal differences between the sonification conditions not discernible in the present pilot study.

The number of participants in the present pilot study was low, and consequently the results, and the interpretation of the results, must be considered with this in mind. Despite this, the results give nevertheless some valuable insights for further studies. **Firstly**, *harmony* did not seem to support the participants as much as the other sonification conditions. Maybe the effect of harmony is a more challenging sonification condition, which might put higher demands on the musical experience of the user, even if there were no effects of this found in the present study. This could be due to the low number of participants. Nevertheless, the effect of dissonance could be increased by increasing the range of the mapping in the sonification, and thus create an effect that is more pronounced and noticeable. **Secondly**, there might be a tendency to an effect of *vibrato* with slightly less good accuracy for some participants compared to the other sonification conditions (apart from *harmony*). Also this could be explained by the low number of participants. Nevertheless, in this sonification condition the vibrato depth was altered in relation to the data but the speed of the vibrato, the vibrato frequency, was fixed. It might be hard for the participants to distinguish and perceive the vibrato depth, the amount of vibrato, as implemented in the present study, why an even stronger vibrato could be used in a future study. The vibrato depth could also be mapped to the data together with vibrato speed to create a sonification that might be easier to perceive. **Thirdly**, *tempo* could also be further evolved. The mapping was performed as 60 bpm to 110 bpm to 170 bpm for this highest level, but a mapping such as 60 bpm to 130 bpm to 220 bpm (or similar) could provide more clear and distinguishable steps that would support the users more. **Fourthly**, all three sonification conditions using different methods to alter the *timbre* seemed to provide clear sonification cues for the users, why timbre or the quality of the sounds used in sonification seem important to keep in mind while designing future sonifications. **Fifthly**, there was not an effect of musical experience found for accuracy or subjective ratings for any of the sonification conditions. This is something positive as this suggests that the sonification is useful regardless of musical experience. In a real-life setting, an operator or controller can not be expected to have a music degree to be able to monitor a system and ongoing processes. **Sixthly**, even if the accuracy was overall high, it was not 100%. If sonification would be used in a monitoring setting somewhat similar to the experiment in the present study, the sonification would be present in a context with detailed visual information available as well. The sonification would then be able to provide peripheral sonic information about states in a system, providing cues about changes and developments, while the visual information would provide in-depth information for the user if needed. Furthermore, in such a setting the amount of training on the system would be substantially greater than the introduction and training trials used in the present study.

The sonification conditions used in the present pilot study, were designed to be able to function in combinations. Such combinations could either be two or three sonification conditions mapped to the same data, creating an even stronger sonic cue about the conditions in the data, or used simultaneously to sonify different data providing even more information to a user. However, this was not evaluated in the present study.

The stereo panning was experienced as supportive in finding the position (left, center, right) where the data deviated from the normal level. Positions are consequently of help for separating the sonification and position that deviated from the normal. In the present study the sound level of the stereo panned sounds were compensated to be at the same sound level in the stereophonic sound (see Table 3). This compensation could be omitted making the center sound the loudest, thus making the sound in focus for the user more pronounced and perceivable. The stereo width could also be extended, moving the positions not in focus further away from the center position.

The data were sonified differently for the three different positions, by using tones in different pitch. It would be a near impossible task to determine which position at which point in time that deviates from the normal level without using different pitch. However, in the present study there was an overlap in pitch between positions, consequently the range in pitch could be extended

which might further support the distinction between positions. As a possible consequence, the use of different pitch did not seem to be experienced as that useful for the participants.

The sonification used in the present study was not normalized in audibility. The three positions in different pitch were not normalized in relation to each other, while some degree of masking might have occurred. Consequently, this might have made the perception of the changes in the higher pitch somewhat harder to discern as these frequencies to some degree might have been masked by a lower pitch. The perception of the variations in the different musical elements caused by the sonification of the data, were also not normalized, which night have made some of the sonification conditions easier to perceive than others. This is something that must be taken into account in future stages of research, but also when analysing the results in the present study. Despite this, the results found in the present pilot study gives promising suggestions on musical elements to be used in further stages of investigations of sonification for monitoring of dynamic processes.

5. CONCLUSION

The present pilot study investigated the use of different musical elements in sonification of simulated data using an interactive search task. This pilot work has provided a good foundation towards sonification in monitoring of dynamic processes.

The results suggest that all sonification conditions used in the experiment provided enough information for the participants, regardless of musical experience, to solve the experimental task. Even if there were no statistically significant differences between sonification conditions, when studying mean performance and 95% confidence intervals *harmony* and *vibrato* seemed to be less good in providing information to the participants. The changes in sonification conditions were also, in general, rated as easy to perceive and supported in finding both the position (left, center, right) and the time point that deviated most from the normal level.

6. FUTURE WORK

The pilot work done in the present study has provided interesting ideas for future work to further explore the use of sonification for monitoring of dynamic processes. Future research could:

- assess if there is differences in performance between sonification conditions by measuring and analysing response time.
- further explore the use of musical elements and combinations of musical elements in sonification for monitoring.
- evolve and expand the use of stereophonic/spatial sound for sonification in relation to monitoring.
- investigate simultaneous use of different musical elements to sonify different data for different positions, and not only changes in data for one position at a time. This type of inquiry would answer if it is possible to discern changes in different or the same data variable for different positions, the sonification would then both provide an overview of the entire system as well as providing detailed information of the different positions.
- deploy sonification of real data sets in real and streaming monitoring situations with domain experts to further understand the usefulness, the support and benefit, of sonification in a real-life setting/environment, for example process control in industrial manufacturing, air traffic control, or monitoring of steam and gas turbines.

7. REFERENCES

[1] Johannes Zagermann, Ulrike Pfeil, and Harald Reiterer, "Measuring cognitive load using eye tracking technology in visual computing," in *Proceedings of the sixth workshop on beyond time and errors on novel evaluation methods for visualization*. ACM, 2016, pp. 78–85.

[2] Sharon Oviatt, "Human-centered design meets cognitive load theory: designing interfaces that help people think," in *Proceedings of the 14th ACM international conference on Multimedia*. ACM, 2006, pp. 871–880.

[3] Qiuzhen Wang, Sa Yang, Manlu Liu, Zike Cao, and Qingguo Ma, "An eye-tracking study of website complexity from cognitive load perspective," *Decision support systems*, vol. 62, pp. 1–10, 2014.

[4] James W Marcum, "Beyond visual culture: the challenge of visual ecology," *portal: Libraries and the Academy*, vol. 2, no. 2, pp. 189–206, 2002.

[5] Yung-Ching Liu, "Comparative study of the effects of auditory, visual and multimodality displays on drivers' performance in advanced traveller information systems," *Ergonomics*, vol. 44, no. 4, pp. 425–442, 2001.

[6] Jeremy M Wolfe, Todd S Horowitz, and Naomi M Kenner, "Cognitive psychology: rare items often missed in visual searches," *Nature*, vol. 435, no. 7041, pp. 439, 2005.

[7] René Marois, Do-Joon Yi, and Marvin M Chun, "The neural fate of consciously perceived and missed events in the attentional blink," *Neuron*, vol. 41, no. 3, pp. 465–472, 2004.

[8] Colin Ware, *Information Visualization: Perception for Design*, Morgan Kaufmann Publishers Inc., San Francisco, 3nd edition, 2013.

[9] R. Beau Lotto, S. Mark Williams, and Dale Purves, "Mach bands as empirically derived associations," in *Proc. National Academy of Sciences*, Los Alamitos, 1999, vol. 96, pp. 5245–5250, National Academy of Sciences of the United States of America.

[10] Thomas Hermann, Andy Hunt, and John G. Neuhoff, *The Sonification Handbook*, Logos Publishing House, Berlin, Germany, 1nd edition, 2011.

[11] Trevor Pinch and Karin Bijsterveld, *The Oxford Handbook of Sound Studies*, Oxford University Press, 2012.

[12] Karmen Franinovic and Stefania Serafin, *Sonic Interaction Design*, MIT Press, 2013.

[13] Stephen Barrass and Gregory Kramer, "Using sonification," *Multimedia systems*, vol. 7, no. 1, pp. 23–31, 1999.

[14] Quan T. Tran and Elizabeth D. Mynatt, "Music monitor: Ambient musical data for the home," in *Proc. IFIP WG 9.3 International Conference on Home Oriented Informatics and Telematics (HOIT 2000)*, Kluwer, 2000, vol. 173, pp. 85–92, IFIP Conference Proceedings.

[15] Muhammad Hafiz Wan Rosli and Andres Cabrera, "Gestalt principles in multimodal data representation," *IEEE Computer Graphics & Applications*, vol. 32, pp. 80–87, 2015.

[16] Seyed Yaghoub Mousavi, Renae Low, and John Sweller, "Reducing cognitive load by mixing auditory and visual presentation modes.," *Journal of educational psychology*, vol. 87, no. 2, pp. 319, 1995.

[17] Paul Vickers, "Sonification for process monitoring," in *The Sonification Handbook*, Thomas Hermann, Andy Hunt, and John G. Neuhoff, Eds., pp. 455–491. Logos Publishing House, Berlin, Germany, 2011.

[18] William W. Gaver, "Auditory icons: Using sound in computer interfaces," *HumanComputer Interaction*, vol. 2, no. 2, pp. 167–177, 1986.

[19] William W Gaver, Randall B Smith, and Tim OŚhea, "Effective sounds in complex systems: The arkola simulation," in *CHI*, 1991, vol. 91, pp. 85–90.

[20] Thomas Hermann and Andy Hunt, "The discipline of interactive sonification," in *Proc. of the Int. Workshop on Interactive Sonification Workshop (ISON-2004)*, Germany, 2004, pp. 1–9, Bielefeld University.

[21] Andy Hunt and Thomas Hermann, "The importance of interaction in sonification," in *Proc. of the 10th Meeting of the International Conference on Auditory Display (ICAD 2004)*, Sydney, Australia, 2004, pp. ICAD04–1–ICAD04–8.

[22] Jonathan Cohen, "Monitoring background activities," in *Santa Fe Institute Studies in the Sciences of Complexity - Proceedings Volume-*. Addison-Wesley Publishing Co, 1994, vol. 18, pp. 499–499.

[23] Ralf Jung, "Ambience for auditory displays: Embedded musical instruments as peripheral audio cues," in *Proc. International Conference on Auditory Display (ICAD 2008)*. International Community for Auditory Display, 2008.

[24] Anssi Kainulainen, Markku Turunen, and Jaakko Hakulinen, "An architecture for presenting auditory awareness information in pervasive computing environments," in *Proc. International Conference on Auditory Display (ICAD 2012)*. Georgia Institute of Technology, 2006.

[25] Maria Barra, Tania Cillo, Antonio De Santis, Umberto Ferraro Petrillo, Alberto Negro, Vittoro Scarano, Teenie Matlock, and Paul P Maglio, "Personal webmelody: Customized sonification of web servers," in *Proc. International Conference on Auditory Display (ICAD 2001)*. Georgia Institute of Technology, 2001, pp. 1–9.

[26] Maria Barra, Tania Cillo, Antonio De Santis, Umberto Ferraro Petrillo, Alberto Negro, and Vittorio Scarano, "Multimodal monitoring of web servers," *IEEE MultiMedia*, vol. 9, no. 3, pp. 32–41, 2002.

[27] Takahiko Tsuchiya, Jason Freeman, and Lee W. Lerner, "Data-to-music api: Real-time data-agnostic sonification with musical structure models," in *Proc. 21st International Conference on Auditory Display (ICAD 2015)*, Graz, Styria, Austria, 2006, pp. 244–251, Georgia Institute of Technology.

[28] Niklas Rönnberg and Jonas Löwgren, "The sound challenge to visualization design research," in *Proc. EmoVis 2016, ACM IUI 2016 Workshop on Emotion and Visualization*, Sweden, 2016, vol. 103, pp. 31–34, Linköping Electronic Conference Proceedings.

[29] Niklas Rönnberg, "Sonification supports perception of brightness contrast," *Journal on Multimodal User Interfaces*, pp. 1–9, 7 2019.

[30] Niklas Rönnberg, Jonas Lundberg, and Jonas Löwgren, "Sonifying the periphery: Supporting the formation of gestalt in air traffic control," in *Proc. 5th Interactive Sonification Workshop (ISON-2016)*, Germany, 2016, pp. 23–27, CITEC, Bielefeld University.

[31] Thomas Hermann, Tobias Hildebrandt, Patrick Langeslag, and Stefanie Rinderle-Ma, "Optimizing aesthetics and precision in sonification for peripheral process-monitoring," in *Proc. International Conference on Auditory Display (ICAD 2015)*. Georgia Institute of Technology, 2015.

[32] Tobias Hildebrandt, Thomas Hermann, and Stefanie Rinderle-Ma, "Continuous sonification enhances adequacy of interactions in peripheral process monitoring," *International Journal of Human-Computer Studies*, vol. 95, pp. 54–65, 2016.

[33] James McCartney, "Supercollider: A new real-time synthesis language," in *Proc. International Computer Music Conference (ICMC)*, Hong Kong, China, 1996, pp. 257–258, Michigan Publishing.

[34] James McCartney, "Rethinking the computer music language: Supercollider," *IEEE Computer Graphics & Applications*, vol. 26, pp. 61–68, 2002.

[35] Iréne Deliége and John Sloboda, *Perception and Cognition of Music*, Psychology Press Ltd., Hove, East Susse, 1997.

[36] Stefanos A. Iakovides, Vassiliki M. Iliadou, Vassiliki T. H. Bizeli, Stergios Kaprinis, Konstantinos Fountoulakis, and George S. Kaprinis, "Psychophysiology and psychoacoustics of music: Perception of complex sound in normal subjects and psychiatric patients," *Annals of General Hospital Psychiatry*, vol. 3, pp. 1–4, 2004.

[37] Patrik. N. Juslin and Petri Laukka, "Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening," *Journal of New Music Research*, vol. 33, pp. 217–238, 2004.

[38] Ying Liu, Guangyuan Liu, Dongtao Wei, Qiang Li, Guangjie Yuan, Shifu Wu, Gaoyuan Wang, and Xingcong Zhao, "Effects of musical tempo on musicians and non-musicians emotional experience when listening to music," *Frontiers in Psychology*, vol. 9, pp. 2118, 2018.

[39] Alexander J Ellis, "Musical scales of various nations," *RSA Journal*, vol. 33, pp. 485, 1884.

[40] Fritz Winckel, *Music, Sound and Sensation: A Modern Exposition*, Dover Publications, Inc., New York, 1967.

[41] Gareth E. Roberts, *From Music to Mathematics: Exploring the Connections*, Johns Hopkins University Press, Baltimore, 2016.

[42] William A. Sethares, *Tuning, Timbre, Spectrum, Scale*, Springer, London, 2nd edition, 2005.

[43] David Creasey, *Audio Processes: Musical Analysis, Modification, Synthesis, and Control*, Taylor & Francis, New York, USA, 2017.

Proceedings of ISon 2019, 6th Interactive Sonification Workshop, KTH Royal Institute of Technology, Sweden, November 18, 2019

INTERACTIVE SONIFICATION FOR CORRECTION OF POOR SITTING POSTURE WHILE WORKING

Kotaro Okada

Graduate School of Kyoto Sangyo University
Department of Frontier Information Science
Kyoto, Japan
i1888024@gmail.com

Shigeyuki Hirai

Kyoto Sangyo University
Faculty of Information Science and Engineering
Kyoto, Japan
hirai@cse.kyoto-su.ac.jp

ABSTRACT

People who sit while working may unconsciously take a bad posture, such as a stoop. It causes a high physical load or may result in poor work efficiency. Since a bad posture occurs when people are in a state of concentration, they may not notice it by themselves. In order to solve this problem, we propose a system to make the posture state noticeable without disturbing the work significantly. This system indicates a bad posture via sounds in real-time, including the ambient music, by applying interactive sonification. This paper describes our prototype system for interactive sonification of postures while sitting, along with the sound designs. We also discuss results from a preliminary evaluation of our sound designs, with regard to their usefulness in helping users notice and correct bad postures.

1. INTRODUCTION

Changes in the social workforce over the past few decades have forced office workers to spend a long time sitting in the workplace. This, coupled with a lifestyle that tends to make people sit at home, results in health problems such as back and neck injuries[1]. Keeping a bad posture for a long time has been shown to exacerbate health problems[2]. There is evidence that links the use of computers to the risk of developing musculoskeletal pain and disorders. A survey of 512 office workers found that the prevalence of neck pain for 12 months was 45.5%[3]. Reports of lifetime prevalence of neck pain in the general population range from 67-80%[4]. Without proper measures, an increase in the prevalence of neck pain is expected.

Over time, poor posture can cause pain, muscle pain, tension, headaches, and long-term complications such as osteoarthritis[5]. Most upper limb disorders and symptoms (neck, shoulder, elbow, and wrist pain) are associated with the use of computers on poorly-positioned workstations[6].

As one of the factors necessary to maintain a stable sitting position, Shibata cites "attention to sustain in work activities"[7]. A person can detect a poor posture based on physical factors such as stable sitting balance ability and sensory feedback and can maintain a sitting posture in an appropriate posture. However, maintaining an ideal sitting posture during work activities requires the ability to handle dual tasks, that is, the ability to concentrate on the work activity while simultaneously detecting and correcting bad postures. It is very difficult for most people to check their posture while working on a task[8]. Many techniques for solving this problem use visual feedback, but on-screen warnings may not be appropriate as warnings because they can interfere with the tasks being performed on the computer[9].

During this decade, there has been an increase in the recognition of interactive sonification[10] using non-verbal sounds. This is a field of information audibility and auditory display research[11]. Interactive sonification is defined as the use of sound within a tightly closed human computer interface where the auditory signal provides information about the data being analyzed or the interaction itself. By applying this method, it is considered possible to notify posture deterioration without interfering with the user's work. The sound requires design aspects that considers the balance between work concentration and notification of posture deterioration.

This research aims to examine the sound design of the posture correction system that converts posture into sounds interactively and does not disturb the user s work. This paper describes the outline of the prototype system, its sound design, and the result of a preliminary evaluation. However, the current research is at an early stage, and the design policy and content of each sound set have aspects of trial and challenge.

2. RELATED WORK

2.1. Interactive Sonification

Research on the application of interactive sonification includes studies by Matsubara et al.[12], Hirai et al. Bathonify[13], and Cesarini et al.[14]. In a study by Matsubara et al., which was conducted in the field of rehabilitation, they developed and evaluated a system that makes slight angular changes in the ankle according to the frequency pitch of a continuous sine wave source. It was shown that auditory feedback is not inferior to visual feedback. Hirai et al.'s Bathonify converts bathers' movements and biological information into sound effects and music to improve the experience of bathers and manage their health and safety from outside the bathroom. In a study by Cesarini et al., An interactive acoustic representation of hydrodynamic pressure changes caused by a swimmer's hand-water-interaction induced more symmetrically the hand motion in sports swimming. As a result, it was evaluated that the functional sound helps to change the interaction between hands and water.

2.2. Posture Correction by Interactive Sonification

Studies on the audibility of posture include the posture improvement assist systems developed by Enokibori et al. [15] and Itami et

al.[16]. The system developed by Enokibori et al. detects the type and degree of deterioration of posture from small accelerometers that are worn by the user and converts the posture state into various warning sounds having different pitch and tempo. The system developed by Itami et al. measures the tilt of the back with a tilt angle sensor and notifies the user with a warning sound of a different frequency when the tilt exceeds the threshold of two steps.

Other smart chairs that can be used for posture correction by interactive sonification include IntelliChair[17], SenseChair[18], and sensingChair[19]. In particular, Intellechair cites a simple parameter mapping sonification that reflects the pressure values of the eight force sensors in the parameters of the eight parallel audio streams arranged in stereo space as an example of design of interactive sonification for sitting posture.

In these studies, the posture information is notified to the user by sound as in this study, but the sound design in these studies do not consider the degree of concentration on the work.

2.3. Posture Correction without Interactive Sonification

There are some studies on posture correction without the use of sound. Kikukawa et al.'s system[20] detects a forward leaning posture by measuring the distance between the user's head and the display with Kinect, and notifies the degree of leaning by varying the degree of blur on the screen. In Ishimatsu et al.'s system[21], Kinect and pressure sensors detect the user's forward tilt and leg setting and notify them by displaying a pop-up window on the display. The system developed by Kuwabata et al.[22] gives an illusion of the presence of others by using a prototype system that includes the functions of center of gravity measurement, posture detection by a neural network, and gaze presentation of others using HollowFaceIllusion. This system helps the individual desk work workers maintain concentration. An evaluation study showed that visual posture correction is effective in reminiscent of vision loss due to long-term posture deterioration and that it can be used easily and intuitively without training.

3. SYSTEM OVERVIEW

The system developed in this study is used while sitting on a chair and doing desk work using a PC. The users postures are obtained from images of the upper body of the user with an RGBD camera, such as Kinect or RealSense, and the load balance of the sitting surface is measured with a strain gauge installed under the seating surface. These posture data are interactively sonified, and they are fed back to the user through a wearable neck speaker in order to prompt posture correction (see Figure 1). Figure 2 shows an overview of the current sonification system. The depth images from an RGBD camera are processed by a computer to preprocess the spatial coordinates and angle of each joint of the user. The center of gravity of the seating surface is calculated from the signal values of the strain gauges. These data are used for calculating the degree of posture deterioration, which interactively controls some designed sounds or music. In this system, three types of posture deterioration are detected: a drooping head with its head bent forward relative to the trunk, a stoop, and a bias in the center of gravity.

The current system consists of Intel RealSense D435 as an RGBD camera, four minebea strain gauges with LT1167 amplifier, an Arduino for obtaining measurements from the strain gauges,

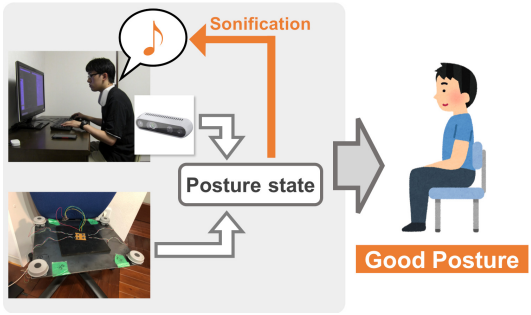


Figure 1: *Overview of System Usage.*

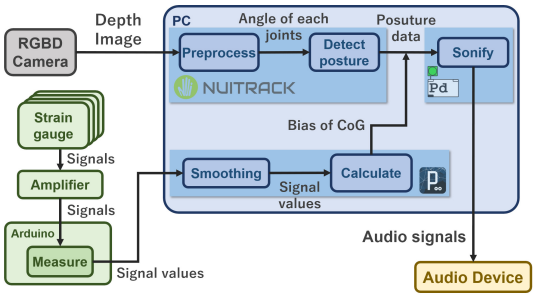


Figure 2: *System Overview.*

and the main computer for calculating the posture states and performing sonification. The strain gauges are sandwiched between two steel plates and fixed at the four corners. The sitting load is sensed by installing the strain gauges under or on the seating surface.

4. POSTURE DETECTION

Figure 3 shows an overview of the posture detection process, in which the type of posture, the degree of deterioration, and the center of gravity on the seat are detected.

As shown in the upper part of Figure 3, six joints, namely the head, neck, right shoulder, left shoulder, chest, and torso, are detected using an RGBD camera and a skeletal tracking library. Currently, we use Nuitrack library for skeletal tracking. First, the normal vector of the surface containing the right shoulder, left shoulder, and torso point is calculated. In addition, two angles, one between the normal vector and the vector from the torso to the chest, and another between the normal vector and the vector from the neck to the head point, are also calculated. Using these two angles, the type of posture and the degree of deterioration are obtained according to the difference between the ideal posture and the current angle. The normal vectors are stored as the ideal posture and used for calculations. To reduce small fluctuations of the vectors, a simple moving average filter for two seconds is adopted.

Meanwhile, as shown in the lower part of Figure 3, the center of gravity on the seat is calculated from the signals obtained from the strain gauges under the seat surface. For center of gravity measurement, when the four signals are FR, FL, RR, and RL (front/rear-right/left), the coordinates of the center of gravity when

the center of the seating surface is the origin are calculated using the following formula: x = (FR + RR) - (FL + RL), y = (FR + FL) - (RR + RL). Then, the bias of the center of gravity is obtained by taking the difference between the ideal position and the current coordinates. In this part, only the deviation of the center of gravity in the lateral direction with respect to the front of the user is used. This bias is utilized separately as a sonification parameter for the degree of deterioration as mentioned above.

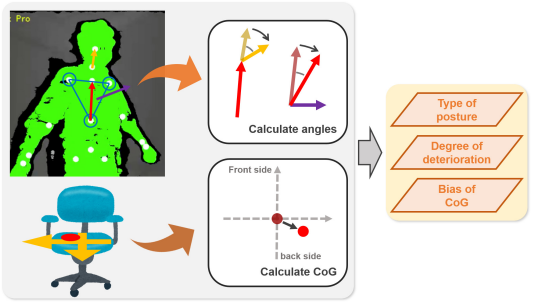


Figure 3: *Posture Detection.*

5. SONIFICATION DESIGN

The sound design described in this chapter can be confirmed on YouTube(https://youtu.be/6MFNBoDEKIQ). The posture information mentioned in the previous section is used as sonification parameters, and they control the sound expression in real-time. This section describes some examples of the sound design for interactive sonification of the posture while working.

5.1. Common Designs of Sonification

Each set of sound design is made to be a different sound expression, but there are also common parts. The fundamental concept of sound design involves the use of ambient music. In our design, the lateral orientation of the posture is used for the left and right localizations of the sound sources. If a user puts more weight on the right, the sound is heard from the left, and if the user puts more weight on the left, the sound is heard from the right. Users can hear the sound from the opposite direction of the laterally imbalanced posture. When the amount of change between the center of gravity and the localization corresponded linearly, it was difficult to sense the change in localization. Therefore, the sigmoid function is utilized for the localization of sounds to emphasize the change specifically close to the center of gravity.

5.2. Use of Environmental Sound

The sonification of this design consists of two streams: ambient music and environmental sounds. When the user is in the ideal position, the user can only hear ambient music from the sound device. It was designed in such a way that additional environmental sounds such as wind and/or rain sampled can be heard when the posture deteriorates. The posture information, such as posture type and deterioration, calculated by the posture detection part is reflected in various sound expressions of environmental sounds.

Figure 4: *Sound Design Using Environmental Sounds.*

5.2.1. Motivation for the Sound Design

The intention of using environmental sounds is to notify the posture deterioration to the user while being in harmony with the ambient music that can be heard at the same time. If they are expressed in a manner where the combination of sounds is less harmonious, there is a high possibility that the user will feel uncomfortable and the use of this system will be suspended. The choice of wind and rain sounds is intended to correlate bad weather with poor posture. In addition, these are designed to make it easier to notice the change in sound density and volume.

5.2.2. Types of Environmental Sounds

The type of environmental sound to be played depends on the type of posture. When the user's head is drooping, a sound of wind blowing will be played, and when the user sits with stooped shoulders, a raining sound will be played.

5.2.3. Severity of Environmental Sound

The intensity of the reproduced environmental sound changes depending on the degree of deterioration. For instance, if the user is in the stooped posture and the degree of deterioration is small, a sound of light rain is heard, and when the degree of deterioration increases, the sound of rain becomes intense. For each type of environmental sound, there are three WAVE files, prepared for different intensities such as light/mid/heavy rain, that are played simultaneously. Each sound volume ratio is controlled by the change in the degree of deterioration. Figure 5 shows a map between the sound volumes and the degree of deterioration. The volume of light rain sound corresponds to 100% to 0% when the degree of deterioration of the back angle is between 0 and 7.5 degrees, and 0% at 7.5 degrees or more. The volume of mid rain corresponds to 0% to 100% at 0 to 7.5 degrees, and 100 to 0% between 7.5 and 15 degrees. The volume of heavy rain sound is 0% at less than 7.5 degrees, and corresponds to 0 to 100% between 7.5 and 15 degrees. Above 15 degrees, the heavy rain sound remains at 100%.

5.3. Use of the Warning Sound

In this sonification design, the user is notified of a poor posture state by a warning sound with musical adjustment. In the ideal posture, only the background accompaniment is played. When the posture deteriorates, a warning sound is generated at half-note intervals.

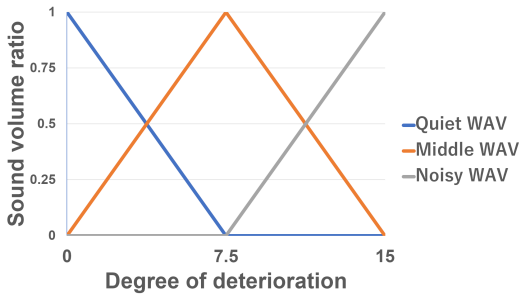


Figure 5: *Volume Control for Each Sound Source.*

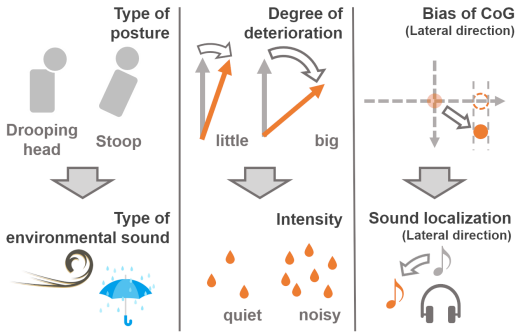


Figure 6: *Combination of Posture Information And Environmental Sound Expression.*

5.3.1. *Motivation for the Sound Design*

In general, a warning sound is used for drawing attention when the user performs an incorrect operation. Although it is considered to be an effective method from the viewpoint of notifying users, it is highly likely that these objectives will be hindered from the viewpoint of relaxing at the work desk or concentrating on work. While adjusting this to harmonize with the ambient music, we try to notify the posture state.

5.3.2. *Timbre of the Warning Sound*

The tone of the warning sound changes depending on the type of posture. When the user sits with stooped shoulders, a warning sound will be heard with a sine wave tone, and when the user sits with a drooping head, a sawtooth wave tone will be heard. These timbers are synthesized in Pure Data with the [phasor] and [osc] objects. In addition, in order to reduce the sharpness of the sound, the release part of the ADSR envelope of the warning sound is set on from 100 to 500 milliseconds and applied the Delay effect with 250 milliseconds.

5.3.3. *Pitch of the Warning Sound*

The pitch of the warning sound is controlled according to the degree of deterioration (see Figure 7). When it is small, the pitch of the warning sound will be lower, and when it is large, the pitch will be higher. Note that the pitch does not change continuously on the frequency axis, but is mapped discretely and musically considering the tonality of the accompaniment so that it does not become dissonant. For example, the pitch of the sine wave warning sound maps the degree of deterioration from 0 to 15 degrees of D Aeolian from D4 to E6 (from 62 to 88 in MIDI note numbers). The FO of the sawtooth sound is set on one octave above the sine wave.

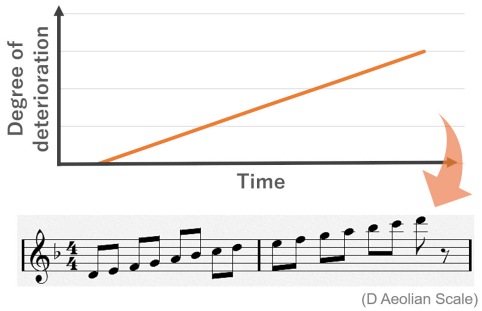


Figure 7: *Pitch Control by Degree of Deterioration.*

5.4. Use of Ornament Melody

In this sonification design, the user is notified of a poor posture state by changing the ornaments of music melody The music in this design consists of a basic note part and an ornamental note part (see Figure 8). The basic part consists of sparse melodies of about one note in two bars and the background accompaniment, and always loops constantly regardless of the degree of deterioration. The ornament part changes dynamically according to the degree of deterioration, and a melody with different timbre and melody density is generated according to the posture state.

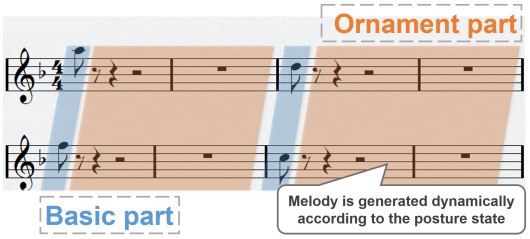


Figure 8: *Composition of Melody.*

5.4.1. *Motivation for the Sound Design*

In general, ambient music is not intended to be focused on listening. This sound design is based on that characteristic, the density of the melody is often sparse to not attract more attention than necessary. And also this design aims to notify the user of the posture state by increasing the density of the melody and emphasizing it only when the posture deteriorated. As mentioned in Section 5.3, the pitch of the melody is controlled and harmonized in the tonality of the background ambient music.

5.4.2. Timbre of the Melody

The timber of the ornament melody part varies depending on the type of posture. When the user is in a stooped posture, the timber of the melody will be a sine wave, and when the user is sitting with a drooping head, it will be a sawtooth wave which is mentioned in Section 5.3.

5.4.3. Density of the Melody

The density of the melody note is controlled by the degree of deterioration. The procedure of increasing the number of notes in the ornamental part is predetermined, and the ornament notes will increase more with the increase in the degree of deterioration. These notes are distributed in a balanced manner within two bars showed in Figure 9). Therefore, the melody notes are designed to have a low density when the deterioration is a small number and a high density when the deterioration is a large number. Also, this melody line is organized in the random pitch of D4 to E6 within the D Aeolian accompanied.

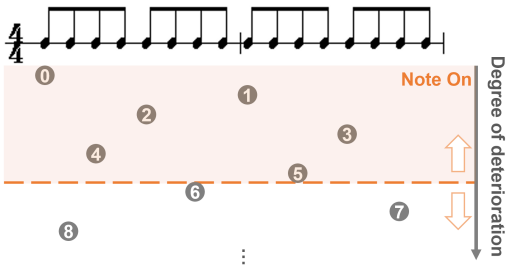


Figure 9: *Increasing pattern of melody density.*

6. DISCUSSION

In this section, we discuss and consider the use of interactive sonification for sitting posture correction.

6.1. Hardware Configuration for Practical Use

The outline of the system described in Section 3 is the current system configuration, and the image of the system used for practical use in the future could be different from that shown here. For instance, with regard to the use of RGBD cameras, it is necessary to capture the image of the users upper body. In this study, we located an RGBD camera at the back of the user. From the point of view of practical use, it should be located on a desk or on the ceiling without being a hindrance for other people. It is hard to be used on a desk or on the ceiling because the user's torso is not visible from the top of a desk or from a ceiling. However, in the future, it may be possible to acquire detailed posture information using machine learning techniques and a small camera that can be located at a convenient location without being a hindrance for other people. On the other hand, with regard to the use of the strain gauge, the pressure sensors will be embedded in a seat with soft cushions, and the chair itself may become an IoT device. The current audio output device is a neck speaker; however, the preference of the device depends on the user. It is possible to use earphones, headphones, or loud speakers, with different audio specifications. It is necessary to consider a sound design that is suitable for such changes in device and posture information. Besides, it is also necessary to consider interference with the other people in such as shared office spaces.

6.2. Sonification Design

6.2.1. Motivation for the Sound Design

As mentioned at the beginning of Section 3, this system is supposed to be used for desk work. Since the aim is to achieve both work efficiency and posture correction, the sound design of this sonification system was designed by considering several points.

The sound of the system should be such that the users can focus on their work without being concerned when they are in the ideal posture and the change in sound can be noticed when the posture deteriorates. However, although notification with an intense sound may be good in terms of emphasizing awareness of the change in posture, we think it should be avoided because it may interfere with mental stability and work. The method of notifying information by controlling the volume level is highly likely to be noticed by the user when applied to a sound source with a low density of notes. If there are many elements that change sounds, it will lead to the user having to listen carefully to catch the change, and this may not fulfill the purpose of this sonification. Therefore, as mentioned in a previous section, this sonification was designed with as little sound changes as possible.

The intention for the use of ambient music as the core component of the sonification design is to help maintain the concentration on the work and relax while preventing excessive concentration on the music. In addition, the intention of making the deviation of the center of gravity correspond to the sound localization is that the load balance can be returned to the direction in which the sound is heard when the posture is incorrect. The sound localization is a guide to move the body. The work of Hammerschmidt et al. indicates the effectiveness of spatial panning for supporting the speed adjustment of car drivers [23]. Therefore, our sound design using sound localization ought to be effective to control good posture.

6.2.2. Preliminary Evaluation of Sonification Design

A preliminary experiment on this system and its sonification designs was conducted. This section describes the experiment and several opinions from participants.

The experiment had held at a house that is built in Kyoto Sangyo University [24]. Participants of this preliminary experiment consisted of 14 people (13 men and one woman) between the ages of 20 and 55. Although none of them had used the system, they use computers every weekday for their work or study. While this experiment, participants wore a neck speaker, Sony SRS-WS1, to hear sounds from the system.

The experiment was held with the following procedure. First, each participant was presented an explanation of the purpose of this study about a posture correction and a concentration of work. They were asked to experience the three sound designs mentioned in the previous section 5 in random order and to evaluate each design with a questionnaire sheet. After explaining the correspondence between the posture state and the sound expression, a participant was instructed to reproduce the deteriorated postures and asked them to hear the sound changes. For each sound design, it took about 10 minutes for evaluation.

As for the design using environmental sounds, most participants (86%) answered that it was intuitive and easy to understand. On the other hand, two people (14% participants) did not associate posture deterioration with the bad impression of the wind and rain as intended. One-third of participants (36%) answered environmental sounds distracted the participants less from their main work task than other musical elements, and among these three designs, there was a tendency to be relatively motivated to use. This indicates the user does not feel uncomfortable when listening to notification sounds and the other sounds at the same time.

Regarding the design using warning sound, most participants (79%) answered that it was easy to understand, and at the same time, four people (29%) responded that they did not like this warning sound. The range of changing pitches was too wide, which could have made participants feel uncomfortable with the high-frequency sound; this may have had an impact on the sound preference. On the other hand, a participant could not immediately know the degree of posture deterioration from the pitch because of the weak sense of sounds.

Six participants (43%) had the opinion that the design using the ornamental melody was basically neither good nor bad. However, a participant was worried about whether it was a musical change or whether they could remember how to change the sound, or that there was no sense of unity in how the sound changed and it was difficult to understand the deterioration of posture. These issues may be solved by fixing the change in pitch of the melody of the ornamental part instead of making it random. However, if the pattern is completely fixed, the user may get bored with a loop that does not change, so we think that a random element that partially prevents boredom is necessary. There was also an opinion that the change might not be noticed when concentrating on work.

Regarding the notification for the deviation of the center of gravity by the change in sound image localization, five participants (36%) answered that they did not find the sound change without aware, and it was difficult to understand. Instead of providing this notification by manipulating the localization of the sound that has already been played, it may be provided by sounding a new sound that has been manipulated to the left or right. Although five people had negative comments for this sound design, the other participant (64%) could understand and find the sound change. Therefore, this sound design can be used.

A participant proposed a new sound design, in which motivation for posture correction is provided by the sounds that cause discomfort. This should be considered carefully because it leads to a decrease in the willingness to use the system; however, it is worth to try. As the reason for hesitating to use the system, two participants stated that they wanted to listen to other music and videos. For this reason, it is necessary to consider a sound design that does not make the user feel uncomfortable, even when listening to other sounds simultaneously. The solution for this matter may be to use sound design with environmental sounds. However, if a similar sound is generated from the other system than this sonification system that is used for posture correction, the notification sounds may not be noticed; hence, it is necessary to consider the selection of the timbre.

7. CONCLUSION

We built a system that facilitates sitting posture correction by interactive sonification. This paper provided an overview of this system and discussed the sound designs for interactive sonification. We also designed and developed three sound sets based on environmental sound, warning sound, and ornamental melody. In this paper, the preliminary user evaluation of this sonification system and sound designs. Our future work is to further devise and implement sound designs and to verify the usefulness of the sound design and system through user evaluations.

8. REFERENCES

[1] Jan Hartvigsen, Charlotte Leboeuf-Yde, Svend Lings, and Elisabeth H Corder, "Is sitting-while-at-work associated with low back pain? a systematic, critical literature review," *Scandinavian journal of public health*, vol. 28, no. 3, pp. 230–239, 2000.

[2] Owen Evans and Kim Patterson, "Predictors of neck and shoulder pain in non-secretarial computer users," *International Journal of Industrial Ergonomics*, vol. 26, no. 3, pp. 357–365, 2000.

[3] Barbara Cagnie, Lieven Danneels, Damien Van Tiggelen, Veerle De Loose, and Dirk Cambier, "Individual and work related risk factors for neck pain among office workers: a cross sectional study," *European Spine Journal*, vol. 16, no. 5, pp. 679–686, 2007.

[4] Bart N Green, "A literature review of neck pain associated with computer use: public health implications," *The Journal of the Canadian Chiropractic Association*, vol. 52, no. 3, pp. 161, 2008.

[5] Lance T. Twomey PhD and James R. Taylor MD PhD FAFRM(Sci), *Physical Therapy of the Low Back*, Churchill Livingstone, 3 edition, 4 2000.

[6] Amy J Haufler, Michael Feuerstein, and Grant D Huang, "Job stress, upper extremity pain and functional limitations in symptomatic computer users," *American journal of industrial medicine*, vol. 38, no. 5, pp. 507–515, 2000.

[7] Jin Asai and Isao Nara, Eds., *Shisei Seigyo To Rigaku Ryouhou No Jissai (Posture control and physical therapy)*, Bunkodo, 5 2016.

[8] S Phillips, "The continuing problem of oos in the office," *Ergonomics Australia*, vol. 14, no. 2, 1999.

[9] Roel Vertegaal et al., "Attentive user interfaces," *Communications of the ACM*, vol. 46, no. 3, pp. 30–33, 2003.

[10] "Interactive Sonification Webpage," `http://interactive-sonification.org`.

[11] Hermann Thomas and Hunt Andy, Eds., *The Sonification Handbook*, Logos Verlag Berlin, 2 2014.

[12] Masaki Matsubara, Hideki Kadone, Masaki Iguchi, Hiroko Terasawa, and Kenji Suzuki, "The effectiveness of auditory biofeedback on a tracking task for ankle joint movements in rehabilitation," in *Proceedings of the 4th interactive sonification, workshop (ISon2013)*, 2013, pp. 1–6.

[13] Shigeyuki Hirai, "Embedded smarthouse bathroom entertainment systems for improving quality of life," in *Entertaining the Whole World*, pp. 85–114. Springer, 2014.

[14] Daniel Cesarini, Thomas Hermann, and Bodo Ungerechts, "An interactive sonification system for swimming evaluated by users," in *Sonification of Health and Environmental Data-York 2014. Conference Proceedings*, 2014.

[15] Yu Enokibori, Yuma Mori, Kenji Mase, et al., "Performance evaluation for long-term and repeated use of voice notification type posture maintenance assist system assuming daily use (japanese)," *IPSJ SIG technical reports*, vol. 2015, no. 11, pp. 1–6, 2015.

[16] Kimiwa Itami and Mikiko Kurushima, "Development and evaluation of a body mechanics learning system equipped with a function to generate "sound" at a dangerous angle for improving nursing posture (japanese)," *Journal of Japanese Society of Nursing Research*, vol. 33, no. 2, pp. 2_95–2_102, 2010.

[17] Thomas Hermann and Risto Koiva, "tactiles for ambient intelligence and interactive sonification," in *International Workshop on Haptic and Audio Interaction Design*. Springer, 2008, pp. 91–101.

[18] Jodi Forlizzi, Carl DiSalvo, John Zimmerman, Bilge Mutlu, and Amy Hurst, "The sensechair: The lounge chair as an intelligent assistive device for elders," in *Proceedings of the 2005 conference on Designing for User eXperience*. AIGA: American Institute of Graphic Arts, 2005, p. 31.

[19] Hong Z Tan, Lynne A Slivovsky, and Alex Pentland, "A sensing chair using pressure distribution sensors," *IEEE/ASME Transactions On Mechatronics*, vol. 6, no. 3, pp. 261–268, 2001.

[20] Mariko Kikukawa and Hideaki Kanai, "Examination of correction effect of wrinkles by long-term behavioral results (japanese)," *Interaction*, vol. 2012, pp. 696–700, 2012.

[21] Haruna Ishimatsu and Ryoko Ueoka, "Bitaika: development of self posture adjustment system," in *Proceedings of the 5th Augmented Human International Conference*. ACM, 2014, p. 30.

[22] Kengo Kuwahata, Yuichi Ito, and Ryoko Ueoka, "Construction of a system to support maintenance of concentration when working alone (japanese)," in *Proceedings of the 23rd Virtual Reality Society of Japan annual conference*, 2018, pp. 22E–5.

[23] Jan Hammerschmidt and Thomas Hermann, "Slowification: An in-vehicle auditory display providing speed guidance through spatial panning," 2016.

[24] Shigeyuki Hirai and Hirotada Ueda, "Towards a user-experience research in a living laboratory? home (ksu-ihome) (japanese)," *Proceedings of SI2011*, 2011.

Proceedings of ISon 2019, 6th Interactive Sonification Workshop, KTH Royal Institute of Technology, Sweden, November 18, 2019

COMMUNICATING GAIT PERFORMANCE THROUGH MUSICAL ENERGY: TOWARDS AN INTUITIVE BIOFEEDBACK SYSTEM FOR NEUROREHABILITATION

Prithvi Kantan

Aalborg University
Copenhagen, Denmark
pkanta18@student.aau.dk

Sofia Dahl

Aalborg University
Copenhagen, Denmark
sof@create.aau.dk

ABSTRACT

The use of rhythmic auditory cues in gait rehabilitation has been shown to improve walking performance across numerous neurological conditions. The utility of interactive sonification in such settings has also been increasingly researched, and the use of musical stimuli has been of considerable interest due to their emotional appeal and movement-inducing capabilities. This paper presents the design and implementation of musical gait sonification system capable of real-time temporal gait parameter measurement and organic synthesis of layered rhythmic music. We introduce the use of *musical energy* as an intuitive and interesting sonic feedback dimension, and outline the design of our feedback model based on timbre embodiment research. The feedback model was evaluated by means of a listening test with 14 cognitively unimpaired participants. The results showed that musical energy changes were perceived easily and as intended by the majority of participants with no prior training, although the perceived changes were generally modest in magnitude. Future work primarily includes the exploration of suitable gait mapping strategies to the musical energy dimension, and the design of additional feedback strategies to enhance feelings of musical agency and engagement in the rehabilitation setting. Future studies must also include user-centric system tests involving real patients and clinicians.

1. INTRODUCTION

The use of interactive sonification in physical rehabilitation possesses considerable potential as a therapeutic feedback tool. Diverse sonification paradigms have been conceived and tested for both upper and lower limb rehabilitation of patients afflicted by *strokes, Parkinson's Disease (PD), Acquired Brain Injury* and other acute or chronic neurological conditions (full review in [1]). In the context of gait (walking) rehabilitation, a previously adopted approach is that of measuring spatiotemporal gait features using instrumented footwear [2, 3], and mapping these quantities to audio synthesizer or processor parameters, effectively representing movement qualities in the form of sonic manipulations. Recent studies (see [4, 5]) advocate the use of musical feedback signals in such applications, owing to the universal ability of musical stimuli to elicit emotion, as well as motivate, monitor and modify bodily movement [4]. Furthermore, music-based interventions such as *Rhythmic Auditory Stimulation (RAS)*, *Patterned Sensory Enhancement (PSE)* and others have been repeatedly shown [1, 6] to dramatically improve gait performance in multiple neurological conditions. Recent research investigating human-music interaction in exercise has also found that feelings of musical agency during strenuous physical performance reduce perceived exertion [7], pain [8] and improve mood [9].

A number of musical sonification systems have been developed for exercise in general [7, 10] and specifically gait [11, 12]. The core sonic interaction varies widely on a case-by-case basis, from matching gait cadence to music tempo/music choice in the D-Jogger [11] to rewarding the compliant user with richer musical instrumentation in the MoBeat system [10] or direct modulation between movement and spectral bandwidth of the music in the Jymmin system [7]. Such systems are examples of *mediation technology*, wherein technology mediates human perception and action - giving the human mind an extension in the digital musical domain. Such interactions conceivably have a vast design space, although in the domain of healthcare this is constrained by the need to make the auditory display *perceivable, intuitive and pleasant* for a largely non-musician user base with a potentially wide range of cognitive and physical impairments. This in turn requires designers to cope with huge variability among people's abilities and demands [13]. The ideal system would afford clear and unambiguous inference of movement performance from the auditory display, while still providing a clearly discernible sense of musical agency, causality and control to the lay user.

Our primary objective is to lay the foundation for a musical gait sonification system based on low-level audio synthesis, with novel movement-sound metaphors and deepened embodied music interaction. The purpose is to enhance enjoyment, motivation and subsequent adherence to therapy among patients. The main focus of this paper is the introduction of *musical energy* as an intuitive and pleasant auditory dimension for sonification in gait rehabilitation (based on recent embodiment theories of timbre perception), as well as a user-baseline-specific dynamic mapping architecture. The ensuing sections contain a more in-depth treatment of relevant past literature and the current system design. The sonification model was evaluated by means of a listening test with 14 cognitively unimpaired individuals. Ethical approval for testing the system on real patients could not be procured during the current study.

2. RELATED RESEARCH

Rhythmic Auditory Stimulation, (RAS) is a rehabilitation technique of rhythmic motor cuing to facilitate movements that are intrinsically and biologically rhythmical, such as walking. RAS has been used in the rehabilitation of several neurological diseases (reviewed in [1]). In essence, it is the application of a rhythmic pulse (or beat) to organize periodic bodily movements in a process that occurs below conscious perception and functions to improve

movement efficiency. The temporal structure of the auditory stimulus serves as a physiological template to improve positional and muscular control [14], thereby improving gait performance parameters [15]. The use of rhythmic music may confer several additional advantages apart from the salient cuing pulse it provides. Maes et. al. [4] advocate music-based feedback systems, as they leverage the strong motivational qualities inherent to people's interactions with music. This is explained by the neuropsychological mechanisms of arousal and motor resonance, reliant on prediction processes in the brain. Music not only affects timing, but also the *vigor* of movements. In a fixed-tempo synchronized walking experiment, Leman et. al. [16] found a significant effect of music type on walking velocity, and documented the characteristics of physically 'activating' and 'relaxing' music. Additionally, Barrass et. al. [17] explored sonification designs in the realm of fitness and sports, and found a user preference for algorithmic music.

Factoring in the ability of the auditory system to effortlessly deduce rhythm [6] and concurrently monitor multiple auditory data sets [18], the use of multilayered rhythmic music in an interactive gait sonification system emerges as an attractive possibility. The rhythmic component of the music can provide temporal cues for gait alignment, and independent acoustic changes in several music layers can simultaneously supply real-time feedback on multiple dimensions of gait performance. The use of real-time music synthesis instead of pre-recorded music would make it feasible to manipulate both global musical parameters such as tempo, and individual instrument *tracks* without engendering audio artifacts. Even though this opens up a large design space of data mappings to representational acoustic variables, it is important to consider how much of the intended message is received by the listener, and how closely the perceived information matches this intended message [19]. The design of feedback delivery mechanisms must therefore be cautiously approached. Target individuals exhibit wide ranges of perceptual ability, cognitive impairment and display comprehension skill. If multiple gait parameters are arbitrarily mapped to audio dimensions, the resulting ensemble will likely exhibit a large number of continuously and simultaneously varying sounds, whose individual variations may or may not be directly relatable to distinct aspects of gait, even if salient enough to be perceptible. Principles of auditory streaming [20] may certainly be useful here, although it may still not be inherently clear to users without extensive training what the sonic variations mean or how they are to be interpreted.

In a systematic review of sonification mappings conducted by Dubus et. al. [21] based on 179 scientific publications, 58.6% of 495 total mappings were found to use auditory dimensions related to pitch, loudness, duration and spatialization. The reason for this predominant use of these dimensions is unclear, in particular considering the documented lack of evaluation in sonification studies [22]. On the other hand, brightness, timbre and instrumentation accounted for only 12.3% of all mappings, and we suspect that their potential may be under-exploited, particularly from the perspective of embodied music interaction. Schedel et. al. [23] tested rhythmic and timbral distortion to indicate gait dysfunction, finding them to be well perceived by PD patients. In a prototype by Kantan et. al. [24], deviations from typical gait were 'punished' by detrimental modifications to the musical stimulus such as white noise, ring modulation and melody suppression. However, their expert interview revealed another perspective, namely that the use of unpleasant sonic manipulations might be excessively harsh towards ailing patients in possibly fragile mental and physical states [24]. This motivated us to explore more empathetic and novel sonification strategies that promote real feelings of musical agency, a more challenging proposition from the perspective of universal comprehension. The effectiveness of mapping gait to musical structure would depend greatly on individual music background and ability. In this work, we argue that one visceral aspect of music that affords less variability in its interpretation is musical energy related to instrumentation, dynamics or articulation. Most individuals are exposed to music that exhibits both short and long-term musical energy evolution. Even untrained music listeners are sensitive to these variations, which contribute greatly to the overall emotional appeal of music by arousal and predictive neural mechanisms described by Maes et. al. [4].

'Musical energy' in this sense is a high-level attribute, conceivably correlated to signal properties such as dynamics, timbre and regularity of occurrence. Timbre is perhaps the most significant attribute here, and certain specifics pertaining to human timbre perception are worthy of remark. As per Wallmark [25], there is good evidence that timbre perception is embodied in a motor-mimetic sense. There is certainly survival value attached to the ability to identify sources and their states from timbre, as well as to derive meaning from vocal timbre [26]. In timbre perception, we deduce mimetic similarities to vocal expression [27], which is the reverse of emotional state affecting the acoustic output of the produced voice - 'emotion is connected to motion'. This is carried out by the brain using *inverse modelling processes* [28]. Brain scans also show evidence that suggests the presence of subvocalization in timbre perception [29], which is essentially related to motor resonance of the voice in response to certain 'objects'. Wallmark argues that physical exertion and arousal are thereby linked with acoustic characteristics such as brightness (high spectral centroids), noise and roughness. This can be understood in terms of the modifications to human vocal timbre in stressful situations, and Wallmark hypothesizes that it may also be generalized to instrument timbres owing to shared perceptual mechanisms for vocal and instrument timbre. The presence of upper partials increases stimulus content in sensitive frequency regions, improving intelligibility and indicating proximity [30]. In the current study, we argue that these principles can guide the design of sonic textures to provide easily perceptible, unambiguous performance feedback during gait rehabilitation.

Another consideration is the mapping function from gait to audio parameters. One approach is to compare ongoing performance to typical unimpaired performance and sonify the difference between these [31]. Torres et. al. [32] developed a system that allowed multiple strategies for this type of error sonification, with fixed and adaptive deviation thresholds for their sonification triggering. Kantan et. al. [24] used a fixed threshold performance error sonification approach, but their expert interview revealed problems with this approach due to diverse principal gait problems among patients. A middle ground between fixed and adaptive systems is a paradigm where the system rewards improvement over individual baseline performance. This could, for example, be realized with a mapping function that maximizes musical energy when a *target improvement* relative to the baseline is attained - resulting in clear and automatic positive reinforcement for the patient. System design should ideally make it possible to tailor the action-sound coupling on the fly, to cater to individual needs.

We finally discuss gait measurement and the choice of sonifiable gait parameters. The model proposed by Lord et. al. [33] suggests that both spatial and temporal measures vary between PD-

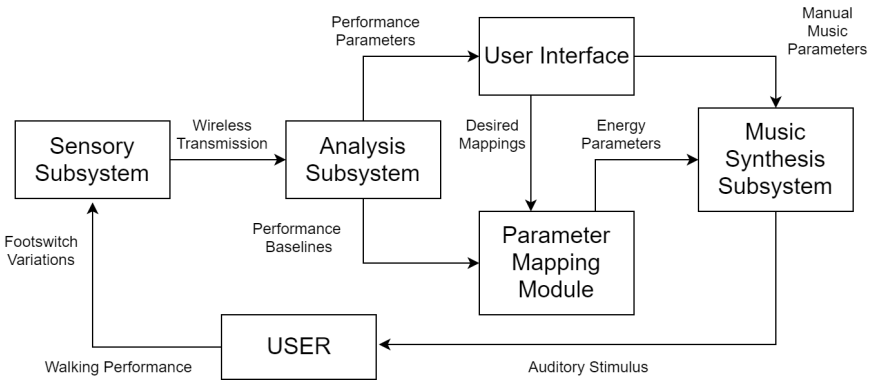


Figure 1: Overall System Schematic

impaired individuals and controls. Post-stroke hemiplegic patients also exhibit increased spatiotemporal asymmetry [34]. These works provide a good starting set of target gait parameters to measure and sonify, which would essentially provide 'feedback of result' [35]. Our measurement system captures exclusively temporal parameters, and a footswitch approach inspired by Blanc et. al. [36] provides good accuracy for both high-level measures like step duration and the finer roll-over characteristics.

3. SYSTEM DESIGN

The current implementation was designed to fulfill the following requirements:

- Real-time generation of expressive rhythmic music.
- Parametric control of musical energy characteristics.
- Light, non-invasive and durable measurement hardware.
- Real-time calculation and display of temporal gait parameters, with storage possibilities.
- Performance baseline calculation for patient-specific sonification scaling.
- User-defined parameter mapping capabilities for sonification customization.

The system may be seen as the combination of multiple functionally distinct subsystems, as illustrated in Figure 1. In this section we briefly discuss each in turn.

Sensory Subsystem: This subsystem is responsible for sensor-based gait performance detection, digitization and wireless transmission. A footswitch approach as in [36] is adopted with switches placed at locations corresponding to the heel, metatarsal and toes, whose digitized data is transmitted at 200 Hz by an Arduino ESP32 microcontroller.

Analysis Subsystem: The raw switch variations are processed into meaningful gait-related information by the analysis subsystem. Foot contact states are determined in real time, and temporal gait performance parameters (mean duration, variability coefficients and asymmetry indices) are calculated and displayed on the user interface.

Parameter Mapping Module: Gait-audio parameter mapping is individualized, and baselines for each gait parameter are calculated from the first 20 steps of a session. The music synthesis subsystem provides five parameters for JUCE control that select one of three musical energy levels for each track. The mapping module simultaneously maps the target-baseline differences for multiple gait parameters to these musical energy levels. The topology is user-definable in real-time through a 5×5 mapping matrix. The majority of chosen gait parameters are among those identified by Lord et. al. in [33]: *cadence, step-time CoV, swing time asymmetry, flat-foot asymmetry and step-time asymmetry*.

Music Synthesis Subsystem: This component is responsible for the generation of the multitrack music ensemble, and real-time modification of musical energy in response to gait pattern changes. It is implemented in FAUST, an audio domain-specific programming language and exported as a JUCE-compatible C++ class using the Faust2Api library, enabling direct parameter control from the mapping module in JUCE. Its main functions are the spontaneous generation of suitable musical structures and their synthesis. The mechanisms and methods used here are treated in greater detail in the next section.

4. SOUND MODEL

The auditory feedback is synthesized and manipulated in real time by a single engine that is responsible for temporal organization, musical timekeeping, musical structure generation and audio synthesis.

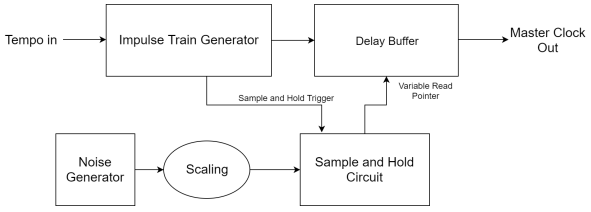


Figure 2: Master Clock Generation

Clocking and Musical Timekeeping: A central master clock serves as the global time reference for musical structure creation, trigger for music synthesis and parameter determinant for time-based effects. It is simply an isochronous impulse train obtained from a FAUST library function, whose frequency is 4 times the externally configured musical tempo (beat rate). This allows 16th

Musical Energy Level	**Track 1 Bass Drum**	**Track 2 Snare Drum**	**Track 3 Hi Hat**	**Track 4 Melody Synth**	**Track 5 High Melody**
L1	Djembe Model Tuned to tonic	Djembe Model Tuned to fifth	Marimba Model Tuned to 800 Hz Cutoff freq - 5000 Hz	Sine Wave	Sine Wave 8th Dotted Delay
L2	Sine + Noise S: Env-modulated freq fPeak = 412 Hz N: Bandlimited: 1500 Hz - 5200 Hz	Filtered WGN 9 parallel modal BPFs Mode Freq Range: 130 Hz - 1960 Hz	Filtered WGN Bandlimited b/w 5000 Hz - 10000 Hz	FM - 3 Mod ModFreq Ratios - 6, 11, 19 Mod Indices - 280, 140, 35	Triangle Wave LPF @ 1 KHz 8th Dotted Delay
L3	Sine + Noise S: fPeak = 206 Hz N : Same as Level 2 Cubic soft clipper LPF Fc = 8 Khz	Filtered WGN 9 parallel modal BPFs Mode Freq Range: 165 Hz - 7360 Hz	Filtered WGN Bandlimited b/w 10000 Hz - 20000 Hz	FM - 3 Mod ModFreq Ratios - 4, 7, 19 Mod Indices - 520, 260, 65	Triangle Wave LPF @ 5 KHz Leslie Simulation Effect : 50% wet 8th Dotted Delay

Table 1: Track-wise synthesis methods used to generate the three musical energy levels. WGN : White Gaussian Noise, LPF/BPF : Low/Band Pass Filter. Djembe/Marimba models are from the FAUST physical model library.

note subdivisions in the music ensemble. To avoid a machine-like quality, small timing variations in the performance are achieved by delaying each of the originally regular-spaced clock pulses by a different random amount. The impulse train passes through a non-interpolated circular buffer, whose read pointer is modulated by a white noise generator scaled to yield delay times ranging from 0 ms to 10 ms (Figure 2). To reproduce evolving musical structures, the sequencer keeps track of musical time with one set of counters that store the musical 'present' at the sixteenth note, beat and bar level, as well as another that keeps track of elapsed bars.

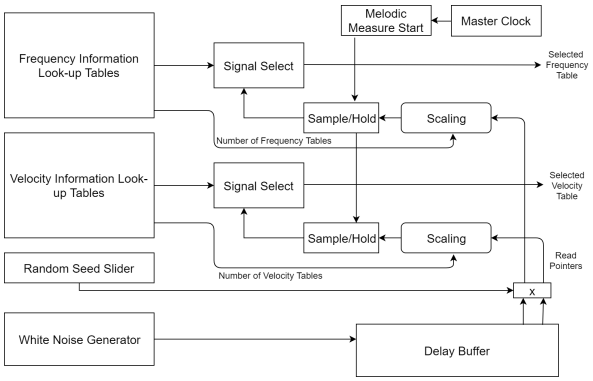


Figure 3: Melodic Pattern Randomization

Musical Structure Generation: The instrument-wise arrangement of the music consists of five tracks, each playing a specific role within the overall ensemble. The musical roles within the ensemble are the *bass drum, snare drum, hi-hat, main melody synth and high melody synth*. A collection of pre-defined patterns for each track is stored in a series of lookup tables, containing elements corresponding to the 16 temporal subdivisions of a musical bar. For percussion tracks, these are simply gain multipliers (velocity information) ranging from 0 to 1. For melody tracks, separate look-up tables store pitch and velocity information, and these tables are randomly 'cross-bred' to create a large number of motif possibilities. Pitch information is stored and processed in the form of scale degrees, allowing flexible pattern reharmonization and transposition in real time by modifying scale and tonic choices respectively. A within-bar sixteenth note counter acts as the read-index pointer for all look-up tables of a track simultaneously, yielding the contents of each one in the form of a time-domain signal dependent on the master clock. A parallel array of signals is thus generated from all tables, one of which is randomly selected every four bars for each track and this causes the music to evolve in a random fashion (see Figure 3).

Pattern Randomization: Multiple levels of randomization are used in the musical pattern generation process, to create both local and global musical variations. White Gaussian noise is the method used to achieve this in all cases, applied in different ways. As far as pattern selection is concerned, white noise is first made unipolar and scaled to fit the number of look-up tables, and its instantaneous values are delayed, sampled and held at the beginning of each percussive/melodic measure, subsequently fed to the signal selector for each look-up table array as seen in Figure 3. An equally important aspect is the presence of incidental and improvisatory local variations, along with some degree of performance humanization. Random variations from the temporal grid ('swing') are obtained by randomly delaying master clock pulses. Apart from this, note velocity and articulation are also suitable candidates for performance humanization. Pseudo-random addition of hi-hat pulses toward the end of measures, as well as simulation of hi-hat openness variations through random release time modulation are done here. Portamento-like variations in the melody are achieved by passing the frequency information signal of the motif through a FAUST smoothing filter with integration time 25 ms. This contour smoothing replaces the sharp and discontinuous inter-note frequency transitions by a smoother articulation, adding yet another form of variety and flavour.

Audio Synthesis: The musical structure generation in terms of temporally organized frequency and amplitude information is then used as control data for the synthesis of each track at the externally determined musical energy level. First, the musical patterns themselves are stripped down at the lowest musical energy level by cancelling select subdivisions in a random fashion (we call this pattern complexity reduction). Rather than continuously varying one or more synthesis parameters to modulate musical energy, a more coherent result is obtained by designing three sonic textures per track, and selecting one of them at any instant. The synthesis

techniques and effect chains for each level (L1, L2, L3 from low to high musical energy) are appropriately pre-defined and tuned by an analysis-by-synthesis approach, and the techniques used are simple in principle but considerably variegated.

The methods range from bandlimited noise and custom-built multiple modulator FM synthesis to simple physical models (using FAUST libraries). A track-by-track signal chain description is provided in Table 1, and the general principles followed are those enumerated by Wallmark [25].

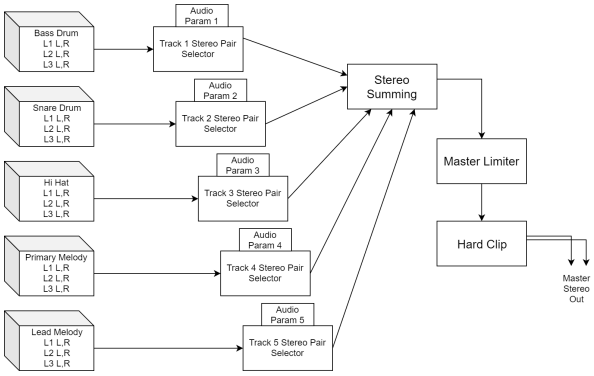


Figure 4: Signal selection and master processing.

Musical Energy Level Selection and Master Summing: Based on control inputs from the mapping module, only one musical energy level per track is selected to play at a given time. With the time-varying nature of the control inputs, tracks must be made to seamlessly switch between levels of musical energy in an aesthetic fashion. A custom signal selection function is implemented to handle this task, with gain factors for musical energy levels smoothed using FAUST one-pole smoothing filters having a 0.5 sec integration time. This fades musical energy levels smoothly in and out during transitions, albeit adding some auditory feedback latency.

Five stereo pairs corresponding to each track are obtained as a result, which are linearly summed to a single stereo buss (master buss), where final processing takes place. In order to maximize distortion-free loudness, a stereo limiter is implemented with fast attack and slow release to minimize low frequency waveform distortion effects. The compression ratio is set relatively high (10:1), but the threshold is set at 0dBFS so as to only act on intermittent stray signal peaks. The output of the limiter is then hard-clipped at -1 and 1 to ensure FAUST stability in all situations, although this leads to no audible effects since the signal gains are staged to be well within the available headroom. The signal selection and master summing signal flow is depicted in Figure 4. Examples and source code have been made available as well [1] [2].

5. EVALUATION

The musical energy levels were designed and implemented such that L3 would be perceived to be the most energetic, and L1 the least. As the efficacy of the auditory feedback in this form depends upon individual ability to perceive and interpret these changes in musical energy as intended, it was important to evaluate listeners' impressions of these energy changes, specifically a) Whether the changes were consistently and judged in the correct direction (increase or decrease), b) Whether the changes were perceived with sufficient magnitude and c) Whether musical energy changes were only perceived when an actual transition occurred. To gauge this, a brief listening test was performed with a set of participants, described in this section.

5.1. Participants

A convenience sample of 14 participants (4 women, age 20-25) were recruited among students of Aalborg University. Participants had no documented hearing or cognitive impairments. Before giving their informed consent, participants were briefed about the aim of the experiment, and that they could withdraw at any time without losing their remuneration, which was an array of breakfast items.

5.2. Experimental Setup

The test was conducted in a small room on campus. The FAUST engine was set up to run on a test computer, in the FAUST web browser in real-time. Pattern complexity reduction was not applied, so as to test the effectiveness of purely timbral information. The engine was specially modified such that a 3 kHz sine beep would be played with every change in musical energy level. A button was also added to trigger the same beep with no change in musical energy. The audio output was simultaneously monitored on headphones by both the experimenter and the participant using the headphone outputs of a Focusrite 18i8 audio interface. Participants were asked to bring their own headphones, to test perceptibility of musical energy changes on different reproduction systems.

5.3. Procedure

Participants were asked to listen to the presented music on their headphones and pay close attention to the musical energy. They were informed that when they heard beeps, the level of musical energy would increase, decrease or remain the same, and they were asked to quantify this musical energy change as a subjective *percentage* relative to before the beep. For e.g. 'heard no change' or 'decreased by 20% after the beep'. Participants were seated across the experimenter such that they could not see the changes that were being made to the engine parameters. The energy changes were made to the music from time to time in exactly the same order, starting at L3, and down to L2, L1 and back up to L2, L3. This was done in three configurations - a) All tracks, b) Only percussion tracks (melody at L3 throughout), c) Only melody tracks (vice-versa). Loudness was maintained constant between energy levels through appropriate gain staging of the synthesized waveforms. This was done to exclusively study the effect of timbre and instrumentation changes on perceived musical energy, independent of loudness cues. In between real transitions, a total of four false beeps unaccompanied by changes in musical energy were randomly triggered in each trial to prevent order habituation. In all cases, participants communicated the perceived change in percentage verbally to the experimenter.

6. ANALYSIS AND RESULTS

We first carried out a directional analysis for reported transitions in musical energy in each track configuration. Observed frequen-

[1] Audio examples and interface demonstration found here [37].
[2] Source code available here [38]

Transition Type	**Direction Judged Incorrect (%)**	$\chi^2(2)$, ***p***	**Judged Change (%) [↑ Level Shifts]**	**Judged Change (%) [↓ Level Shifts]**	**t(13), *p***	**Effect Size *d***
All Tracks	5.36	53.33, <.001				
L1 ⇔ L2			20 (12.55)	-23.21 (15.64)	6.368, <.001	3.047
L2 ⇔ L3			13.57 (9.69)	-14.28 (6.46)	8.434, <.001	3.382
Percussion Only	23.21	40.5, <.001				
L1 ⇔ L2			13.57 (12.47)	-7,87 (7.77)	4.449, .001	2.062
L2 ⇔ L3			7.85 (8.25)	-6.78 (8.22)	4.691, <.001	1.777
Melody Only	17.85	38.43, <.001				
L1 ⇔ L2			8.57 (5.69)	-9.64 (7.95)	5.824, <.001	2.634
L2 ⇔ L3			8.21 (7.23)	-4.28 (9.16)	4.301, .001	1.513
False Beeps	28.57					

Table 2: Comparison of perceived musical energy changes in upward and downward level transitions between adjacent levels in all track configurations. Chi-Squared Test results are first shown for directional judgments in each configuration. Next, judged change percentages are depicted as Mean across participants with Standard Deviation in parenthesis. The final columns present statistical significance and effect size (Cohen's *d*) between perceived musical energy changes during upward and downward transitions, obtained by paired-samples t-tests.

cies of correct identified direction (increase, decrease, same) were compared to random expectancies (evenly distributed as 1/3) using Pearson's Chi-Squared tests. As shown in Table 2, significant associations were found in each track configuration, with the percentage of incorrect judgments minimum for All Tracks and maximum for the Percussion Only subset. 28.57% of the false beeps (20/56) were wrongly reported to be accompanied by energy transitions. Thereafter, we analyzed the perceived magnitudes of the musical energy changes across participants for each adjacent pair of levels for all track configurations using 2-tailed paired-samples t-tests. In all track configurations, statistically significant differences were found between perceived change percentages for *true* upward and downward musical energy transitions between all adjacent pairs of levels (see Table 2). Absolute mean magnitudes were greater in the 'All Tracks' configuration, with greater effect sizes in comparison to Percussion Only and Melody Only subset transitions. Mean perceived changes were also greater for L1-L2 transitions than L2-L3 transitions. All statistical analysis was carried out in SPSS 25.0.

7. DISCUSSION

During this study, we developed a prototype device for real-time sonification of temporal gait parameters through changes in musical energy. We tested the effectiveness of this auditory dimension through a listening test with cognitively unimpaired participants. From the results, it is evident that the participants' impressions of what constituted an increase or decrease showed good agreement with our designed musical energy level hierarchy. Transitions were correctly judged with particularly good regularity for All Tracks. Reported changes for upward and downward transitions exhibited statistically significant differences. The fact that participants were able to achieve this level of agreement and accuracy with no prior training would indicate that motor-mimetic mechanisms sensitive to brightness, noise and roughness [25] were indeed invoked by them to gauge energy on a uniform basis. However, participants also reported musical energy change in false no-change conditions which indicates that the communication also involves some error in detection, at least in the absence of training.

As far as multiple streams are concerned, the increased incidence of errors with Percussion Only and Melody Only suggests that energy shifts in track subsets were less reliably perceptible on the whole, also corroborated by the smaller perceived change magnitudes and effect sizes. This could be the combination of results of the timbral shifts of individual instruments not being large enough by themselves, and auditory masking of 'narrower-band' low energy levels of an instrument by relatively broadband high energy levels of others. Misjudged musical energy changes during false beeps could be attributed to short-term local dynamic variations over the course of the music. An interesting finding was the greater judged change magnitude for L1-L2 transitions than L2-L3 in general. This could be ascribed to the synthesis methods used; L1 used physical models and sine waves, while L2 and L3 used similar filter-based methods with different synthesis parameters related to noise and brightness. This may suggest that complete changes in instrumentation may be more effective musical energy cues than simple changes to synthesis parameters, although the magnitude of these parameter changes certainly matters. Reducing musical pattern complexity at lower musical energy levels could considerably exaggerate inter-level differences. It is important to note that the perceived changes were small-to-moderate although the participants were cognitively unimpaired, and future studies must evaluate the perceived magnitude of these musical energy changes on individuals from the target group.

A larger question is how musical energy-related mappings can be optimally used in a gait sonification system. As identified by Dubus et. al. [21], the most popular mappings follow the logic of ecological perception, corresponding to natural perceptual associations. This is also relevant for embodied music interaction in motor re-learning, with continuity and contingency being important factors for associative learning processes [28]. Musical energy is a high-level sonic feature, and it follows that the dimension may not be equally effective at representing every aspect of gait performance. For example it may be intuitive to map gait *vigor* or walking speed to musical energy [16], but not step time asymmetry. There are certainly potential use-case situations, for instance musical energy variations could also augment RAS-based systems to give clear feedback when the patient matches the target cadence. But in general, spatiotemporal gait parameters mapped to musical energy changes may not be optimal in contexts where feedback must be provided on finer aspects of gait technique, because simply put, the feedback only indicates that there is a problem and not

how to fix it. Therefore, while the musical energy dimension can certainly deepen the embodied experience with the correct mapping choices, it must be supplemented with other novel musical feedback channels to make individual gait technique more explicit.

Our study also has some limitations. We are yet to gauge the perceived pleasantness of the musical energy dimension in comparison to more conventional parameters such as pitch, and study the effect of individual musical background and ability in its perception. Our listening test design also makes it difficult to separate the contributions of the melodic and percussive content to perceived changes in musical energy, and regression models could reveal more about these relationships. The system needs further testing with actual stakeholders in the clinical domain, and we will proceed to do this by involving patients and practitioners in future iterations through a participatory design process. As far as the developed system is concerned, we have already identified several areas for improvement. User familiarity with the presented music has been shown to have a significant influence on gait [39], and we are developing a framework to replace the random music generation of the current prototype by a protocol that can be used to encode and resynthesize patient-chosen musical pieces. The addition of chord harmony, evolving musical structures, more discrete levels of musical energy and general improvements in sound quality and expressiveness are also important priorities. We also aim to alter our sonification philosophy from a largely 'feedback of result' system to one that provides 'feedback of performance' [35], so as to enhance the real-time embodied experience and self-awareness of gait technique. This is attainable by the addition of IMU systems, as in [32] or flex sensors or a combination. Individual baselining, measurement and storage of performance results is still very useful from a clinician's perspective [24] and will still be part of future systems.

8. CONCLUSION

The work done during this study was a step towards the realization of a musical gait sonification system with compelling movement-music metaphors, promoting deeper feelings of agency and embodiment during gait rehabilitation. In this paper, we have proposed an approach for the communication of gait performance through musical energy changes based on timbre embodiment principles. A complete prototype with sensory, analysis and audio synthesis functionality was implemented. Evaluation of the sonic model by means of a listening test indicated the promise of musical energy as an intuitive auditory dimension for real-time auditory feedback. Future work includes exploration of mapping strategies for the musical energy dimension, upgrades to the sensory and synthesis engines, and user-centered studies with members of the real target group and clinical stakeholders. We believe that the potential of embodied music interaction technologies in rehabilitation is prodigious, and when fully exploited will transform the lengthy and arduous recovery process into a more engaging and rewarding experience.

9. AUTHOR CONTRIBUTIONS

Author Kantan designed and implemented the system, designed the listening test and analyzed the results as part of a MSc. study in Sound and Music Computing with author Dahl as supervisor. Dahl's contribution was partially funded by NordForsks Nordic UniversityHub Nordic Sound and Music Computing Network NordicSMC, project number 86892.

10. ACKNOWLEDGEMENTS

We would like to thank the 14 participants who participated in the experiment, as well as Jesper Greve, Peter Williams and Daniel Overholt for their advice and assistance with building the hardware prototype.

11. REFERENCES

[1] Nina Schaffert, Thenille Braun Janzen, Klaus Mattes, and Michael H. Thaut, "A Review on the Relationship Between Sound and Movement in Sports and Rehabilitation," *Frontiers in Psychology*, vol. 10, pp. 244, 2019.

[2] Anna-Maria Raberger, Lena Schön, Ronald Dlapka, Jakob Doppler, Michael Iber, Christian Gradl, Anita Kiselka, Tarique Siragy, and Brian Horsak, "Short-Term Effects of Real-Time Auditory Display (Sonification) on Gait Parameters in People with Parkinson's Disease-A Pilot Study," *Biosystems and Biorobotics*, vol. 15, pp. 855–859, 10 2016.

[3] Matthew Rodger, William Young, and Cathy Craig, "Synthesis of Walking Sounds for Alleviating Gait Disturbances in Parkinson's Disease," *IEEE Transactions on Neural Systems and Rehabilitation Engineering : A Publication of the IEEE Engineering in Medicine and Biology Society*, vol. 22, 10 2013.

[4] Pieter-Jan Maes, Jeska Buhmann, and Marc Leman, "3mo: A Model for Music-Based Biofeedback," *Frontiers in Neuroscience*, vol. 10, pp. 548, 2016.

[5] Pieter-Jan Maes, Luc Nijs, and Marc Leman, "A conceptual framework for music-based interaction systems," in *Springer Handbook in Systematic Musicology*, pp. 793–804. Springer, 2018.

[6] Michael Thaut, *Rhythm, Music, and the Brain: Scientific Foundations and Clinical Applications*, Routledge, 2013.

[7] Thomas Fritz, Samyogita Hardikar, Matthias Demoucron, Margot Niessen, Michiel Demey, Olivier Giot, Yongming Li, John-Dylan Haynes, Arno Villringer, and Marc Leman, "Musical Agency Reduces Perceived Exertion During Strenuous Physical Performance," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 110, 10 2013.

[8] Thomas Fritz, Daniel L Bowling, Oliver Contier, Joshua Grant, Lydia Schneider, Annette Lederer, Felicia Höer, Eric H. Busch, and Arno Villringer, "Musical Agency During Physical Exercise Decreases Pain," in *Frontiers in Psychology*, 2017.

[9] Thomas Fritz, Johanna Halfpaap, Sophia Grahl, Ambika Kirkland, and Arno Villringer, "Musical Feedback During Exercise Machine Workout Enhances Mood," *Frontiers in Psychology*, vol. 4, pp. 921, 12 2013.

[10] Bram Vlist, Christoph Bartneck, and Sebastian Mueler, "Mobeat: Using Interactive Music to Guide and Motivate Users During Aerobic Exercising," *Applied Psychophysiology and Biofeedback*, vol. 36, pp. 135–45, 03 2011.

[11] Bart Moens, Chris Muller, Leon van Noorden, Marek Frank, Bert Celie, Jan Boone, Jan Bourgois, and Marc Leman, "Encouraging Spontaneous Synchronisation with D-Jogger, an Adaptive Music Player that Aligns Movement and Music," *PLOS ONE*, vol. 9, 12 2014.

[12] Roberto Bresin, Anna Dewitt, Stefano Papetti, Marco Civolani, and Federico Fontana, "Expressive Sonification of Footstep sounds," *Proceedings of the Interaction Sonification Workshop (ISon) 2010*, 05 2010.

[13] Micheline Lesaffre, *Investigating Embodied Music Cognition for Health and Well-Being*, pp. 779–791, 01 2018.

[14] Nina Schaffert, Michael Thaut, and Klaus Mattes, "Rhythm-based Regulation/Modification of Movements in High-Performance Rowing and Neurologic Rehabilitation," in *Proceedings of the ISon 2013, 4th Interactive Sonification Workshop*, 12 2013.

[15] Michael H. Thaut and Mutsumi Abiru, "Rhythmic Auditory Stimulation in Rehabilitation of Movement Disorders: A Review Of Current Research," *Music Perception: An Interdisciplinary Journal*, vol. 27, no. 4, pp. 263–269, 2010.

[16] Marc Leman, Dirk Moelants, Matthias Varewyck, Frederik Styns, Leon van Noorden, and Jean-Pierre Martens, "Activating and Relaxing Music Entrains the Speed of Beat Synchronized Walking," *PLoS ONE*, vol. 8, no. 7, July 2013.

[17] Stephen Barrass, Nina Schaffert, and Tim Barrass, "Probing Preferences Between Six Designs of Interactive Sonifications for Recreational Sports, Health and Fitness," *Proceedings of the Interaction Sonification Workshop (ISon) 2010*, 05 2010.

[18] W. T. Fitch and G Kramer, *Auditory Display: Sonification, Audification, and Auditory Interfaces*, chapter 12, pp. 307–326, 1994.

[19] Thomas Hermann, Andy Hunt, and John G Neuhoff, *The Sonification Handbook*, Logos Verlag Berlin, 2011.

[20] Albert Bregman, "Auditory Scene Analysis: The Perceptual Organization of Sound," *Journal of The Acoustical Society of America - J ACOUST SOC AMER*, vol. 95, 01 1990.

[21] Gaël Dubus and Roberto Bresin, "A Systematic Review of Mapping Strategies for the Sonification of Physical Quantities," *PLOS ONE*, vol. 8, 12 2013.

[22] Katharina Vogt, "A Quantitative Evaluation Approach To Sonifications," *Proceedings of the 17th International Conference on Auditory Display, Budapest, Hungary*, 2011.

[23] Margaret Schedel, Daniel Weymouth, Tzvia Pinkhasov, Jay Loomis, Ilene Berger Morris, Erin Vasudevan, and Lisa Muratori, "Interactive Sonification of Gait: Realtime BioFeedback for People with Parkinson's Disease," in *Proceedings of the ISon 2016, 5th Interactive Sonification Workshop*, 2016.

[24] Prithvi Kantan and Sofia Dahl, "An Interactive Music Synthesizer for Gait Training in Neurorehabilitation," *Proceedings of the 16th Sound and Music Computing Conference, SMC 2019*, 5 2019.

[25] Zachary Thomas Wallmark, *Appraising Timbre: Embodiment and Affect at the Threshold of Music and Noise*, Ph.D. thesis, UCLA, 2014.

[26] Stephen Handel, "Listening: An Introduction to the Perception of Auditory Events," *Trends in Neurosciences*, vol. 13, no. 6, 1990.

[27] Arnie Cox, "Embodying Music: Principles of the Mimetic Hypothesis," *Music Theory Online*, vol. 17, 07 2011.

[28] Pieter-Jan Maes, Marc Leman, Caroline Palmer, and Marcelo Wanderley, "Action-based Effects on Music Perception," *Frontiers in Psychology*, vol. 4, pp. 1008, 01 2014.

[29] Andrea Halpern, Robert Zatorre, Marc Bouffard, and Jennifer A Johnson, "Behavioral and Neural Correlates of Perceived and Imagined Musical Timbre," *Neuropsychologia*, vol. 42, pp. 1281–92, 02 2004.

[30] John Middlebrooks and David M. Green, "Sound Localization by Human Listeners," *Annual Review of Psychology*, vol. 42, pp. 135–59, 02 1991.

[31] Frédéric Bevilacqua, Eric Boyer, Jules Françoise, Olivier Houix, P Susini, Agnes Roby-Brami, and Sylvain Hanneton, "Sensori-motor Learning With Movement Sonification: A Perspective From Recent Interdisciplinary Studies," *Frontiers in Neuroscience*, vol. 10, 08 2016.

[32] Andrés Villa Torres, Viktoria Kluckner, and Karmen Franinovic, "Development of a Sonification Method to Enhance Gait Rehabilitation," in *Proceedings of the ISon 2013, 4th Interactive Sonification Workshop*, 2013, pp. 37–43.

[33] Sue Lord, Brook Galna, and Lynn Rochester, "Moving Forward on Gait Measurement: Toward a more Refined Approach," *Movement Disorders*, vol. 28, no. 11, pp. 1534–1543, 2013.

[34] Sheng Li, Gerard Francisco, and Ping Zhou, "Post-Stroke Hemiplegic Gait: New Perspective and Insights," *Frontiers in Physiology*, vol. 9, pp. 1021, 08 2018.

[35] Martin Eriksson and Roberto Bresin, "Improving Running Mechanics by Use of Interactive Sonification," *Proceedings of the Interaction Sonification Workshop (ISon) 2010*, 05 2010.

[36] Yves Blanc, Claude Balmer, Theodor Landis, and François Vingerhoets, "Temporal Parameters and Patterns of the Foot Roll Over during Walking: Normative Data for Healthy Adults," *Gait & Posture*, vol. 10, pp. 97–108, 11 1999.

[37] Prithvi Kantan, "ISon 2019 Sound Examples and Demonstration," `https://drive.google.com/drive/folders/1lBIxFApRSHzW1ufAywraGIeOEH_wQpta?usp=sharing`, 2019.

[38] Prithvi Kantan, "Musical Energy Sonification Source Code," `https://github.com/prithviKantanAAU/iSON_SourceCode`, 2019.

[39] Kyoung Shin Park, Chris Hass, Bradley Fawver, Hyokeun Lee, and Christopher Janelle, "Emotional States Influence Forward Gait During Music Listening Based on Familiarity with Music Selections," *Human Movement Science*, vol. 66, pp. 53–62, 03 2019.

Proceedings of ISon 2019, 6th Interactive Sonification Workshop, KTH Royal Institute of Technology, Sweden, November 18, 2019

CARDIOSCOPE: ECG SONIFICATION AND AUDITORY AUGMENTATION OF HEART SOUNDS TO SUPPORT CARDIAC DIAGNOSTIC AND MONITORING

Andrea Lorena Aldana Blanco[1], Marian Weger[2], Steffen Grautoff[3] , Robert Höldrich[2], Thomas Hermann[1]

[1]Ambient Intelligence Group, CITEC, Bielefeld University, Bielefeld, Germany
[2]Institute for Electronic Music and Acoustics (IEM),
University of Music and Performing Arts, Graz, Austria
[3]Klinikum Herford, Emergency Department, Herford, Germany
aaldanablanco@techfak.uni-bielefeld.de

ABSTRACT

CardioScope is a sonification/auditory augmentation tool intended to support cardiac diagnosis and monitoring. It allows users to record and visualize synchronized Electrocardiogram (ECG) and Phonocardiogram (PCG) signals, to sonify the electrical activity of the heart or to augment the sound produced by its mechanical behaviour. As first step towards a realtime-interactive auditory augmentation, we here propose an auditory augmentation method using amplitude modulation that allows users to accentuate specific segments of the heart sound in order to make pathological signals from the heart sound more salient. We present a set of sound examples illustrating the proposed method, and discuss results of a preliminary qualitative test with two physicians who are doing their residency in cardiology.

1. INTRODUCTION

The idea of using our listening capabilities to support medical diagnosis has been present in humanity for a long time. In 1816 there was a giant leap in this domain when the French doctor René Laënnec rolled a piece of paper to make a tube that he could place between his ear and the chest of one of his patients in order to better listen to the internal sounds of the body. This moment marks the invention of the stethoscope. It was also Laënnec who introduced the term auscultation, which means listening to the sounds of the body for diagnostic purposes. In particular, the heart, the lungs and the bowel movements are first assessed by clinicians using auscultation.

Over the years, the stethoscope has become a primary tool for initial medical assessment. Auscultation is part of the basic physical exam that physicians carry out whenever they examine a patient. If they detect an abnormality, they can order further tests to obtain a deeper insight into the problem. Being able to detect such abnormalities through auscultation requires training, as the listening skills need to be developed with practice [1]. Besides the need to acquire expertise, noise conditions in medical environments can also increase the challenge of the task.

As medical technology advances, there are new tools that provide a deeper look to pathologies that are first noticed through auscultation. Nevertheless, the availability of such tools within a clinical setting vary greatly, as not all medical centers can afford the latest equipment, or the location and transportation conditions can also make it difficult to acquire specific equipment in remote places. Thus, even with new medical technology developments, the use of the stethoscope as a primary assessment tool maintains. First, because its portability and low cost make it easy to acquire even in remote places, but second, because auscultation makes it possible to get immediate feedback about the state of the patient, and thus to proceed accordingly in a proper time frame.

Given our listening abilities and our capabilities to discern patterns and changes in the signal [2], other forms of auditory feedback have become an integral part of medical monitoring, an example of this is sonification. The term sonification was officially introduced to the research community during the first half of the nineties by Scaletti and Craig [3] and it was later defined by Kramer et al. as transforming data into sound with the purpose of conveying information [4].

In the medical field, sonification is already used as a monitoring tool. A well-known example is the pulse oximeter, a device that measures the oxygen saturation level in the blood and produces a short duration sound in synchronization with the pulse rate. The pitch of the sound decreases if the oxygen saturation level diminishes, thus calling for immediate attention of clinicians. Current technology developments provide opportunities to integrate auditory tools such as auscultation and sonification to support medical diagnosis and monitoring.

As already mentioned, the field of auscultation focuses on several body systems (respiratory, cardiovascular and gastrointestinal). In this research project, we focus on the cardiovascular system [1], which is of particular relevance since according to the World Health Organization cardiovascular diseases are the first cause of death in the world [5].

In terms of sonification of cardiac signals, there are already research efforts to sonify features of electrocardiograms (ECG) with medical purposes[2] [6, 7, 8]. ECG sonification aims to support diagnosis and monitoring in situations where the visual sense is already focused on a primary task (e.g., performing a cardiac procedure) or in situations where the capabilities of the auditory system could help discern patterns that are not evident in the visual representation.

Besides sonification and auscultation, there is another approach that could enhance medical auditory based tools, in particular auscultation. Auditory augmentation is defined by Bovermann et al. [9] as "a paradigm to vary the objects sonic characteristics such that

[1]Cardiovascular means that is related to the functioning of the heart, blood and blood vessels

[2]Electrocardiograms are visual representations of the electrical activity of the heart.

their original sonic response appears as augmented by an artificial sound that encodes information about external data. All this manipulation does not affect the sounds original purpose". In digital stethoscopes, there are already approaches to enhance the characteristics of the sounds mainly through noise and interference reduction and amplification of the signal [10]. However, taking into account the nature of heart sounds, it is possible to propose other types of sound enhancement so that the features of interest in the signal are made more noticeable. Inspiration can be drawn from auditory contrast enhancement, a method to make intrinsic features of a sound (or differences between multiple sounds) more salient so that the underlying information can be better perceived [11].

In this research work, we propose the *CardioScope* system that combines ECG and heart sound signals in order to provide a broader overview about the state of the heart. First we record synchronized ECG and PCG signals and use the ECG signal as a reference point to find important segments of the heart sound signal. In the following section we explain the main characteristics of ECG signals and heart sounds. The signal acquisition process is described in Sec. 3. Furthermore, the method for the ECG-synchronized auditory augmentation of heart sounds is explained in Sec. 4. In Sec. 5 we introduce the *CardioScope* User Interface that allows users to acquire ECG and stethoscope signals simultaneously and interactively control basic features of the auditory augmentation in real-time. Finally, a discussion as well as conclusions are presented in Sec. 7 and Sec. 8.

2. ECG AND HEART SOUNDS

The stethoscope (1816) and the ECG (1903) are two of the earliest developments for cardiac assessment, they are also two of the most used methods. The sounds heard through the stethoscope give important cues about the state of the heart; however in order to give a precise diagnostic further tests need to be carried out, for example, electrocardiograms.

2.1. ECG and PCG signals

For the purposes of this research project, the ECG and the phonocardiogram and its timely connection play a major role. The cardiac cycle and the relationship between the electrical impulses displayed by the ECG and the heart sounds visualized as the phonocardiogram are shown in the Wiggers diagram, Fig. 1, alongside different pressures and ventricular volume of the heart.

In healthy persons the heart rhythmically pumps blood through the vascular system of the body to maintain oxygen supply. As part of the electromechanical coupling electrical impulses initiate the contraction of the heart muscle and therefore maintain a steady blood flow. The initiator of electrical impulses is called the sine node. From the sine node the electrical impulse is transferred via the atria to the atrioventricular node and further to the ventricles of the heart. The electrical activity can be visualized by recording an electrocardiogram (ECG).

Oxygen depleted blood is transported by the venous system to the heart during the diastole which is the relaxation phase. The blood enters the right atrium and is forwarded in the tension phase to the right ventricle. Right atrium and right ventricle are separated by the tricuspid valve. The right ventricle pumps the blood during the systole through the pulmonary valve in the pulmonary circulation. The blood enters the lungs and becomes oxygenated before returning to the left atrium. During the tension phase blood flows

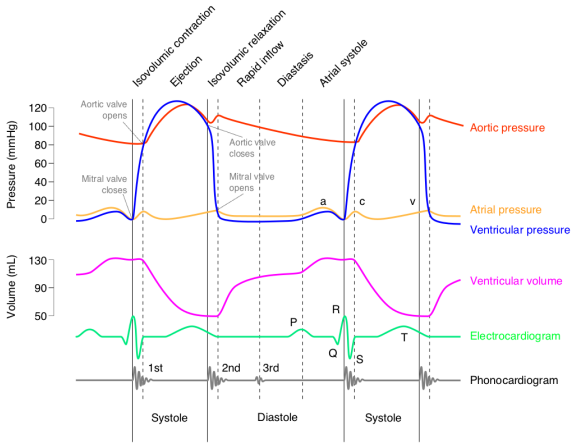


Figure 1: Wiggers Diagram. Image by adh30 revised work by DanielChangMD who revised original work of DestinyQx from Wikimedia Commons. Licensed under CC BY-SA 4.0.

from the left atrium to the left ventricle passing the mitral valve before finally entering the systemic circulation through the aortic valve during the systole. The blood becomes distributed via arteries and arterioles to the organs of the body and in the periphery. In the different parts of the venous and arterial system, pressures can be measured during different phases of the heart cycle. See Fig. 1. A stethoscope can be used by medical providers to auscultate heart sounds. Heart sounds are physiologically generated by the closing of the heart valves. However, heart sounds are not generated directly by the valves itself. The sound waves are due to turbulence caused by their closing.

The first heart sound (S1 or 1st) is produced during the time of the closing of the mitral and tricuspid valves. See Fig. 1. Since the closing of the mitral valve precedes the tricuspid valve by a fraction of a second the first heart sound is physiologically split. This splitting is hard to discriminate for a human ear when auscultating. The second heart sound (S2 or 2nd) is audible during the closing of the aortic and pulmonary valve. Normally the aortic valve closes just before the closing of the pulmonary valve which can also lead to a splitting of varying duration.

S1 and S2 can be auscultated in healthy adult humans. There might be additional sounds in adults in cases of pathological conditions or physiologically in children. For example, the third sound (S3 or 3rd) is considered normal in children or athletes but pathological in adults. S3 is a very low frequency sound produced during ventricular filling, a period also known as diastole (See Fig. 1). The cause of the sound is not yet very clear, it has been said to be related to volume-overload in the ventricles [12], but latest research has also linked it to the diameter of the mitral valve [13]. Furthermore, there might be heart murmurs which are always referred to as pathological.

2.2. The ECG signal

The ECG is a visual representation of the electrical activity of the heart. A standard ECG has twelve leads or channels that depict the activity of the heart from different angles.

In each cardiac cycle, the resulting signal depicts a set of intervals and reference points that physicians evaluate in order to detect anomalies in specific regions of the heart. Figure 2 shows the standard reference points found in every cycle. The beginning of the cycle is given by the P wave that represents atrial depolarization, then there is the QRS complex that represents ventricular depolarization. This is followed by the isoelectric ST segment, which depicts the time between depolarization and repolarization of the ventricles. Finally, there is the T wave as a result of the ventricular repolarization.

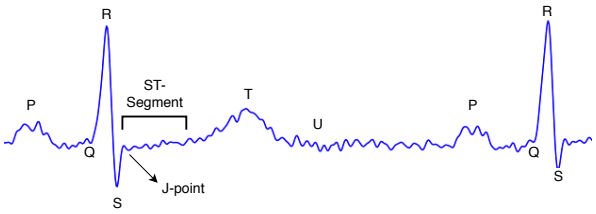


Figure 2: ECG standard reference points (P, Q, R, S, J point, and T), and ST-segment

2.3. Heart sounds

Recordings of heart sounds are known as Phonocardiograms (PCG). As explained in section 2.1 during the cardiac cycle the heart makes a set of sounds that result from the vibrations created by the closing of the valves. Figure 3 shows S1 and S2 during a cardiac cycle.

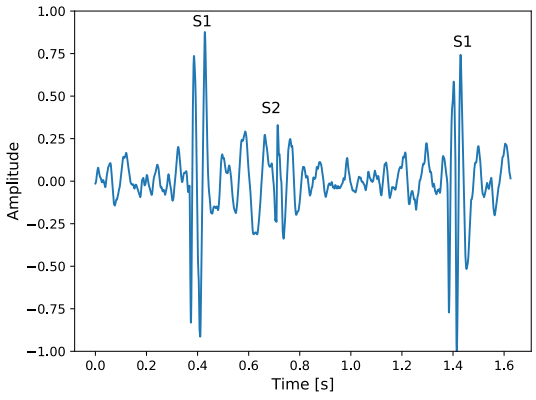


Figure 3: Heart sound depicting S1 and S2

When an abnormality in the heart sound occurs, it is called a murmur. Nonetheless some murmurs are considered harmless [12]. The most common murmurs involve turbulence due to improper closing of the valves after ejection of the blood or due to obstruction of the valves. Such murmurs cause changes in the frequency and intensity of the heart sounds. A broader overview about heart sound murmurs can be found in [14]. Figure 4 depicts changes in the envelopes of heart sounds when different types of murmurs are present. For example, case B illustrates a pathology known as aortic stenosis. This murmur is caused by an obstruction in the aortic valve (See Sec. 2.1), as a result, the second sound (S2) is not a clean short duration valve-closing sound as in case A, but instead it produces a noisy bell-shaped waveform that starts before the actual closing of the valves.

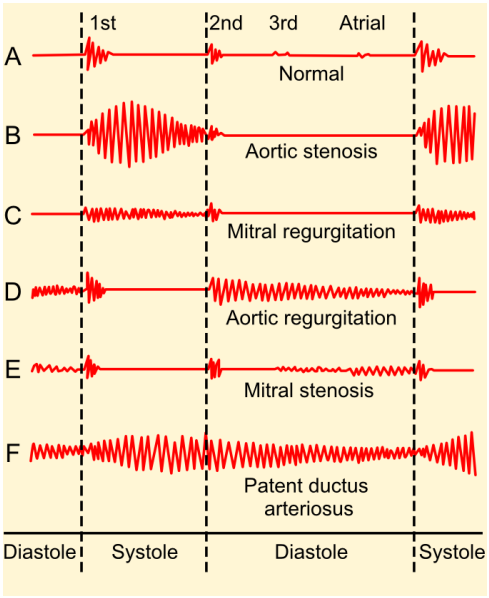


Figure 4: Heart murmurs. Image by Madhero88 from Wikimedia Commons. Licensed under CC BY-SA 3.0.

In order to better hear specific murmurs that are related to a particular heart valve, physicians place the stethoscope in specific areas of the chest. Common heart auscultation places are shown in Fig. 5 and are marked with the letters P, A, T and M, indicating each of the valves (Pulmonary, Atrial, Tricuspid and Mitral) .

3. SYNCHRONIZED SIGNAL ACQUISITION

ECG signals are usually recorded through a medical data acquisition system, while the stethoscope signal can be easily recorded by a microphone connected to an audio interface. Both are analog electrical signals which can be captured by stereo Analog-to-Digital Converters (ADC).

For the first prototype, we use the ECG Sensor from the BITalino (r)evolution plugged kit[3] [15]. It requires an input voltage V_{cc} between 2.0 and 3.5 V (see [16]) which we supply through a 3.7 V rechargeable battery. Additional $V_{cc}/2$ is provided through a simple voltage divider with two 10 k Ω resistors. The ECG sensor then outputs a value between 0 and V_{cc}. An additional voltage regulator

[3] BITalino: `https://bitalino.com`

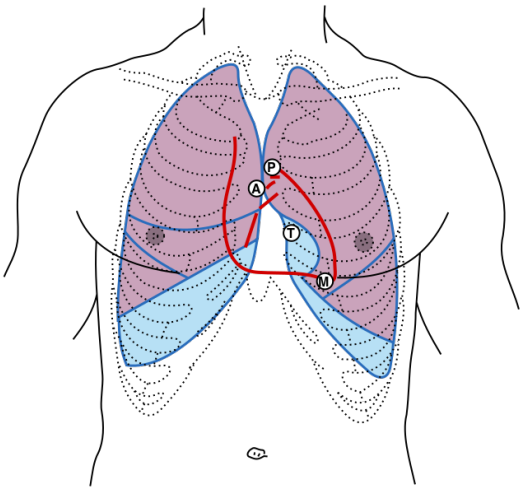


Figure 5: Common heart auscultation locations. Image by Henry Vandyke Carter via Wikimedia Commons. Derivative work by Huckfinne. Licensed under CC0 1.0

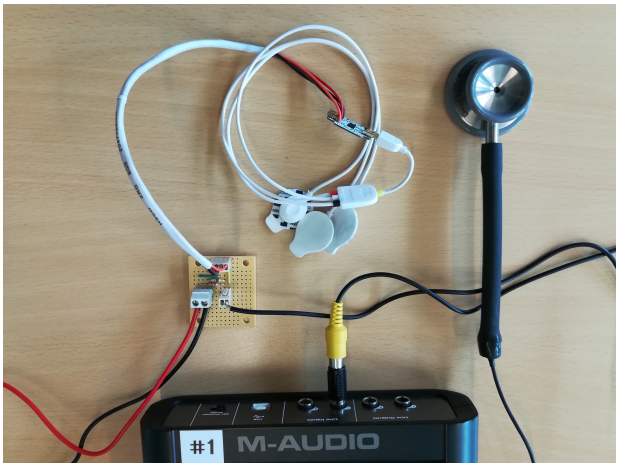


Figure 6: ECG and PCG recording system

can produce a stable reference voltage of 3.3 V and would thus theoretically allow reconstruction of the absolute voltage values from the digital signal. However, we omitted this step for the first prototype. To create a centered/bipolar audio signal, we rely on the analog DC-removal high-pass filter that is already built into the audio interface. The centered signal between $-V_{cc}/2$ and $+V_{cc}/2$ then sufficiently matches the specification of a +4 dBu line level audio signal which goes from $-V_{peak}$ to $+V_{peak}$, with $V_{peak} = 1.736$ V. A photo of the ECG and stethoscope hardware is shown in Fig. 6.

For stethoscope recording, we use a DocCheck Advance II dual head stethoscope chest piece attached to a short rubber tubing. An AKG C417 PP miniature microphone is inserted into the open end of the tubing and connected to the microphone preamplifier of the audio interface.

Both ECG and stethoscope signal are recorded by an M-Audio Mobile Pre USB audio interface. The downside of this procedure is that the ECG sensor and thus electrodes are electrically connected to the audio interface and thus computer which could introduce additional noise. The electrical connection between the human body and the audio interface through the ECG electrodes is no greater risk than a microphone held in the hand.

In the case of medical equipment products, there is a safety standard that needs to be met when devices are in direct contact with the patients. If there is an electrical connection to the heart of the patient the CF type standard needs to be taken into consideration.

4. ECG-SYNCHRONIZED AUDITORY AUGMENTATION OF HEART SOUNDS

Considering the nature of heart murmurs explained in section 2.3, we propose *ECG-synchronized auditory augmentation* as a method to accentuate segments of interest in the heart sound cycle. The augmentation is achieved using amplitude modulation. The idea is to accentuate sound segments that correspond to heart murmurs while attenuating the amplitude of segments that are outside of the region of interest.

4.1. R-peak detection and heart sounds segmentation

Signals recorded through a stethoscope can be quite noisy. For example, moving the stethoscope around the skin or ambient noises in the medical environment can mask the signal of interest, which on the one hand can make the auscultation labor more difficult and on the other hand can make the segments detection in the heart signal more challenging when using automatic systems.

There are several approaches to perform segment detection in heart sounds [17]. One of them is to use the ECG signal as a reference for the segmentation [18, 19] and, since we already have synchronized recordings of the two signals (ECG and PCG), we opt to use the ECG signal for the segmentation process.

Taking into consideration the temporal relation between the electrical and mechanical signals shown in Fig. 1, first, we detect the R-peaks in each cardiac cycle. In order to perform the R-peak detection, initially we apply a low-pass filter with a cutoff frequency of 70 Hz to eliminate frequencies that are outside of the range for ECG diagnostics [20]. Then we implement the method proposed by Worrall et al. [21] in a time window of 200 ms.[4].

Figure 7 depicts the R-peaks detected in one cardiac cycle and their temporal relation to S1 and S2.

4.2. The amplitude modulation signal

The amplitude modulation signal is calculated for each cardiac cycle. The reference to determine the duration of each cycle are the R-peaks found in the ECG signal. Once a cardiac cycle in the ECG signal is detected, we create the amplitude modulation signal for the current heartbeat with period T_i. In the off-line implementation the period is calculated based on the time difference between consecutive R-peaks Δt_{RR}. In the real-time implementation, the interval between subsequent heartbeats can be predicted by applying a linear regression to a series of past *M* inter-beat intervals.

[4]The typical QRS duration for a healthy adult with a heart rate of 60 beats per minute (bpm) is 100 ms

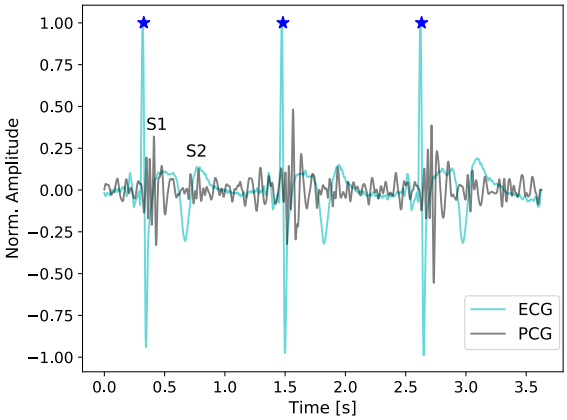


Figure 7: R-peak detection

We define the amplitude modulation signal in the ith heartbeat as

$$\text{mod}_i(t) = (1 - g_a) + g_a \cdot w\left(\frac{t - \beta \cdot T_i}{\alpha \cdot T_i}\right) \quad (1)$$

for $0 \leqslant t \leqslant T_i$, where g_a is the accentuation parameter, β is the relative time lag before the window starts and α is the relative window length.

The window w is defined as

$$w(t) = 0.5 \cdot (1 - \cos(2\pi t)) \quad (2)$$

for $0 \leqslant t \leqslant 1$, otherwise 0.

Sound H1 [5] corresponds to a heart sound of a healthy subject recorded using the method described in Sec. 3. The healthy heart sound is presented in order to provide an auditory reference of a healthy signal. Next, we present the auditory augmentation method using pathological data.

In order to test how the *ECG-synchronized auditory augmentation* method performs in pathological heart sound signals, we use a selection of heart sounds obtained from the *Classifying Heart Sounds Challenge* database [22]. This database contains a series of recordings including normal sounds, murmurs, extra heart sounds or artifacts. We selected a group from the murmur category to create the sound examples.

Sound P1.1 corresponds to a pathological heart sound from the previously described database. *Sound P1.2* is the auditory augmentation of *Sound P1.1* using parameters $g_a = 0.9, \beta = 0.4$ and $\alpha = 0.3$ (See equation 1). Figure 8 depicts the *Sound P1.2* heart sound modulation. This auditory augmentation focuses on S2, thus making more noticeable the murmur related to these heart valves (aortic and pulmonary).

A different pathology can be heard in *Sound P2.1*. In this case, the murmur is longer than in the previous example, it starts around S2 and it prolongs until S1. An auditory augmentation of *Sound P2.1*, is presented in *Sound P2.2*. The parameters used for the augmentation are: $g_a = 0.9, \beta = 0.5$ and $\alpha = 0.4$. Figure 9 illustrates the previously described augmentation.

[5] See section 9 for supplementary material.

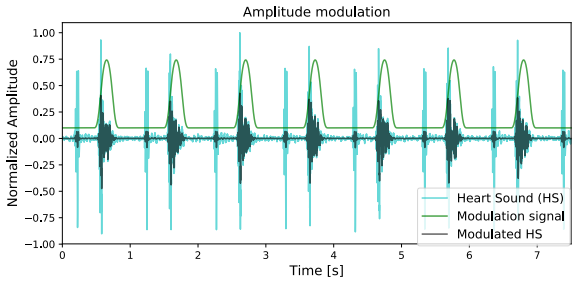


Figure 8: Amplitude modulation of *Sound P1.1*. The original waveform is shown in cyan color, the modulation signal is presented in green and the modulated heart sound is shown in black color.

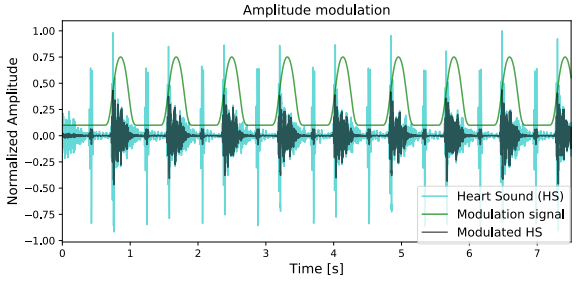


Figure 9: Amplitude modulation of *Sound P2.1*. The heart sound is shown in cyan color, the modulation signal is presented in green and the modulated heart sound is shown in black color.

5. THE CARDIOSCOPE INTERACTIVE GUI

Auscultation itself is a highly interactive process. Physicians place the stethoscope at different locations to better hear specific sounds (See Fig. 5). Furthermore, they can choose between the two sides of the stethoscope (diaphragm/bell) to accentuate a given frequency range.

The *CardioScope* interactive GUI (See Fig. 10) allows users to acquire and visualize synchronized ECG and PCG signals in real time. Moreover it provides basic controls to enhance the signal using filtering and gain controls.

At present, the amplitude modulation method is not yet implemented in real time, however, it is planned to be included in the next development of the *CardioScope* application.

5.1. Listening modes and ECG sonification methods

CardioScope has a listening mode module that allows users to select between three listening options (1) ECG sonification: to listen to the sonified ECG signal, (2) Stethoscope: to listen to heart sounds and (3) ECG and stethoscope: which plays the ECG sonification on the left channel and the heart sound on the right channel.

The implemented ECG sonification methods are the result of our previous work, in which we proposed a set of sonification designs focused on arrhythmias[6] [8, 23] and the elevation of the ST-

[6] Rhythm disturbances in the heart

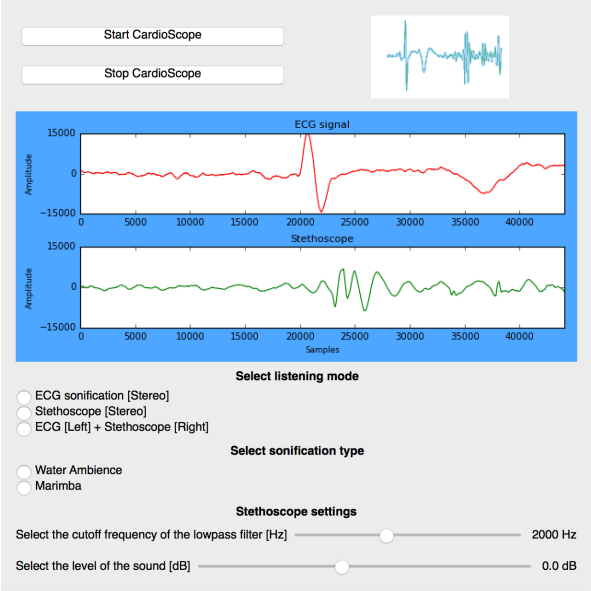


Figure 10: CardioScope GUI

segment [24] (see Fig. 2), which can be an indicator of myocardial infarction (MI)[7].

6. PRELIMINARY QUALITATIVE TEST

We consulted two physicians who are doing their residency in cardiology in order to have their opinion about the use of the *ECG-synchronized auditory augmentation* method in cardiac auscultation tasks. The physicians were asked to listen to a set of auditory augmented heart murmurs and to provide their feedback. Considering that this is a preliminary step towards the design of a qualitative user study, they were not asked to evaluate the interactive tool. One of the physicians regarded the proposed augmentation method as helpful, as it was possible to focus more on the pathological heart murmur by attenuating/filtering unwanted sounds in the heart signal. However, the physician also stated that it is important to not fully attenuate other sounds within the heart cycle as then it would be harder to identify which part of the cycle is being listened to. The other consulted physician had a more doubtful opinion about the benefit of the proposed method. First, because there is a traditional way of doing auscultation that hasn't been changed in a long time, and adjusting to the new method would require time from doctors and students. The second argument was that if a physician listens to what might be a heart murmur, then there would be a follow-up check using ultrasound or other methods. However, the proposed augmentation method could then be helpful in remote areas or countries where they don't have such possibilities about the availability of several medical devices.

[7]Lack of oxygen supply in the heart due to a blocking of the coronaries

7. DISCUSSION

The *CardioScope* system is a tool to support cardiac diagnostic and monitoring based on auditory displays. *CardioScope* aims to provide a broader overview about the state of the heart by combining ECG and PCG signals.

Current medical technology developments have made it possible to create multiple tools for cardiac diagnosis. Most of them rely on visual feedback, technologies such as echocardiography, Heart CT scan and Cardiac Magnetic Resonance Imaging (MRI) are among the most common ones. At present, there is an ongoing discussion about how these new tools could replace the role of the stethoscope [25, 1]. Nevertheless, it is important to keep into account that the availability of such tools depend on how well equipped medical centers are. For example, in remote places access to modern medical equipment is rather limited, which raises the question on how to improve health care conditions in these communities [26]. In addition, there is a part of the medical community that believes the stethoscope shouldn't be replaced but instead used in conjunction with tools of increasing demand such as portable ultrasound devices for echocardiography [25].

It is known that the human auditory system is a very powerful tool to detect patterns and changes in the signals, and auscultation has proved that these abilities can be used in cardiac assessment. Moreover, the use of ECG sonification as a medical supporting tool had also shown promising results [7, 24, 23]. Thus *CardioScope* serves as a tool that focuses on the powerful abilities of our listening system and uses different auditory display techniques to enhance cardiac features. The idea is that physicians can overcome current challenges in the interpretation of ECG signals that are a limitation of visual displays and limitations in the auscultation process due to internal and external noises and poor audio quality. State of the art digital stethoscopes already implement noise reduction to provide better signal quality, nevertheless there is still room to improve the quality and salience of signal features acquired through a stethoscope.

The proposed method of *ECG-synchronized auditory augmentation* provides new possibilities to the auscultation process, not only from the diagnostics perspective, but also for educational purposes, since it makes features of interest in the heart sound that are the cues to the detection of cardiac pathologies more salient. For example, the proposed augmentation method allows physicians to emphasize murmurs that derive from specific regions, such as the mitral and tricuspid valves (S1) or the aortic and pulmonary valves (S2), thus making such murmurs more noticeable during the auscultation process.

The feedback that we received from the consulted physicians falls in line with the current medical discussion previously described, about the role of the stethoscope and its use in conjunction with other cardiac medical technologies. On the one hand, the proposed auditory augmentation is regarded as useful as it can attenuate unwanted sounds and thus make pathological patterns more salient, however on the other hand, there are physicians that would rather rely on the traditional auscultation method or use other technologies such as echocardiography in order to make a diagnostic. A reason for that, however, could lie in the fact that the cardiologists listened to the sonifications outside a closed-interaction loop: if they would have controlled parameters such as α interactively, e.g. by adjusting a slider, then it would have been directly clear when in time the heard sound occurs relative to S1 and the R-peak. We believe that establishing a closed-loop interaction is crucial for

acceptance and profitable use of CardioScope.

If the *ECG-synchronized auditory augmentation* method is introduced as an auscultation tool, a training phase would have to be made so that medical doctors learn to recognise the patterns using the augmented method, however, such training is already the base to learn how to do traditional auscultation. In order to have a better idea about the scope and limitations of the proposed method, it is necessary to carry out a user study. We plan to conduct a qualitative user study involving experts from the medical field in order to determine the significance of the auditory augmentation of heart sounds in the auscultation task.

In the current system development version the auditory augmentation is not yet implemented in real-time mode. However, this is plan to be included in the next development. Moreover, gathering a database of pathological ECG and PCG recordings it is also one of the future tasks of this project.

8. CONCLUSIONS

We introduced the CardioScope system that allows users to acquire synchronized ECG and PCG signals and listen to the signals in three different modes: (1) ECG sonification, (2) auditory augmentation of PCG (3) ECG and PCG. Additionally it offers a set of controls to filter and amplify/attenuate the heart sound signal in real-time. Furthermore, we presented and auditory augmentation method for heart sounds based on amplitude modulation to accentuate specific segments of the signal in order to better detect pathological sounds.

We regard the combined use of ECG and PCG in CardioScope as offering a versatile tool, of relative low cost, which could be applied for cardiac diagnostic in remote areas, and could support physicians in situations when a visual display is not an ideal option or when the noise conditions difficult the auscultation task.

A further quantitative user study still needs to be conducted in order to determine how *CardioScope* can support physicians in their everyday activities involving interpretation of ECG signals and auscultation of heart sounds.

9. RESOURCES

Examples of the segmented amplitude modulation are provided in: http://dx.doi.org/10.4119/unibi/2938001

10. ACKNOWLEDGMENT

This work has been supported by the German Academic Research Service (DAAD) and the Cluster of Excellence Cognitive Interaction Technology "CITEC" (EXC 277) at Bielefeld University, which is funded by the German Research Foundation (DFG).

11. REFERENCES

[1] Tom Rice, "Learning to listen: auscultation and the transmission of auditory knowledge," *The Journal of the Royal Anthropological Institute*, vol. 16, pp. S41–S61, 2010.

[2] Brian Moore, *Psychoacoustics*, pp. 459–501, Springer New York, New York, NY, 2007.

[3] Carla Scaletti and Alan B. Craig, "Using sound to extract meaning from complex data," 1991, vol. 1459.

[4] G. Kramer, B. Walker, T. Bonebright, P. Cook, J. Flowers, N. Miner, and J. Neuhoff, Eds., *Sonification Report: Status of the field and research agenda*, Palo Alto, 1997.

[5] "Cardiovascular diseases (cvds)," `https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)`, Accessed: 05.08.2019.

[6] Hiroko Terasawa, Yota Morimoto, Masaki Matsubara, Akira Sato, Makoto Ohara, and Masatoshi Kawarasaki, "Guiding auditory attention toward the subtle components in electrocardiography sonification," Georgia Institute of Technology, 2015.

[7] Jakob Nikolas Kather, Thomas Hermann, Yannick Bukschat, Tilmann Kramer, Lothar R. Schad, and Frank Gerrit Zllner, "Polyphonic sonification of electrocardiography signals for diagnosis of cardiac pathologies," *Scientific Reports*, vol. 7, pp. 44549, 2017.

[8] Andrea Lorena Aldana Blanco, Steffen Grautoff, and Thomas Hermann, "Cardiosounds: Real-time auditory assistance for supporting cardiac diagnostic and monitoring," in *Proceedings of the 12th International Audio Mostly Conference on Augmented and Participatory Sound and Music Experiences*, New York, NY, USA, 2017, AM '17, pp. 45:1–45:4, ACM.

[9] Till Bovermann, René Tünnermann, and Thomas Hermann, "Auditory augmentation," *IJACI*, vol. 2, pp. 27–41, 2010.

[10] Shuang Leng, Ru San Tan, Kevin Tshun Chuan Chai, Chao Wang, Dhanjoo Ghista, and Liang Zhong, "The electronic stethoscope," *BioMedical Engineering OnLine*, vol. 14, no. 1, pp. 66, Jul 2015.

[11] Marian Weger, Thomas Hermann, and Robert Höldrich, "Real-time auditory contrast enhancement," in *ICAD*, 2019.

[12] Turnquest AE Dornbush S, "Physiology, heart sounds," `https://www.ncbi.nlm.nih.gov/books/NBK541010`, Accessed: 20.08.2019.

[13] Hesham R Omar and Maya E Guglin, "Mitral annulus diameter is the main echocardiographic correlate of s3 gallop in acute heart failure.," *International journal of cardiology*, vol. 228, pp. 834–836, 2017.

[14] Bernard M. Karnath and William Norman Thornton, "Auscultation of the heart," 2002.

[15] Hugo Plácido Da Silva, José Guerreiro, André Lourenço, Ana LN Fred, and Raúl Martins, "Bitalino: A novel hardware framework for physiological computing.," in *PhyCS*, 2014, pp. 246–253.

[16] PLUX – Wireless Biosignals, S.A., *Electrocardiography (ECG) Sensor Data Sheet*, 2016, Rev. A.

[17] Babatunde Emmanuel, "A review of signal processing techniques for heart sound analysis in clinical diagnosis," *Journal of medical engineering & technology*, vol. 36, pp. 303–7, 07 2012.

[18] M. B. Malarvili, I. Kamarulafizam, S. Hussain, and D. Helmi, "Heart sound segmentation algorithm based on instantaneous energy of electrocardiogram," in *Computers in Cardiology, 2003*, Sep. 2003, pp. 327–330.

[19] Noemi Giordano and Marco Knaflitz, "A novel method for measuring the timing of heart sound components through digital phonocardiography," *Sensors*, vol. 19, pp. 1868, 04 2019.

[20] Gari D Clifford, Francisco Azuaje, and Patrick Mcsharry, "Ecg statistics, noise, artifacts, and missing data," *Advanced Methods and Tools for ECG Data Analysis*, vol. 6, pp. 18, 2006.

[21] David Worrall, Balaji Thoshkahna, and Norberto Degara, "Detecting components of an ecg signal for sonification," Georgia Institute of Technology, 2014.

[22] P. Bentley, G. Nordehn, M. Coimbra, and S. Mannor, "The PASCAL Classifying Heart Sounds Challenge 2011 (CHSC2011) Results," http://www.peterjbentley.com/heartchallenge/index.html.

[23] Andrea Lorena Aldana Blanco, Steffen Grautoff, and Thomas Hermann, "Cardiosounds: A portable system to sonify ECG rhythm disturbances in real-time," in *Proceedings of the 24th International Conference on Auditory Display (ICAD)*, Houghton, MI, USA, 2018.

[24] Andrea Lorena Aldana Blanco, Steffen Grautoff, and Thomas Hermann, "Heart Alert: ECG Sonification for supporting the detection and diagnosis of ST segment deviations," 2016.

[25] Francis Fakoya, Maira du Plessis, and Ikechi B Gbenimacho, "Ultrasound and stethoscope as tools in medical education and practice: considerations for the archives," *Advances in Medical Education and Practice*, vol. Volume 7, pp. 381–387, 07 2016.

[26] Donna Goodridge and Darcy Marciniuk, "Rural and remote care: Overcoming the challenges of distance," *Chronic Respiratory Disease*, vol. 13, 02 2016.

Proceedings of ISon 2019, 6th Interactive Sonification Workshop, KTH Royal Institute of Technology, Sweden, November 18, 2019

DEVELOPING MOVEMENT SONIFICATION FOR SPORTS PERFORMANCE: A SURVEY OF STUDIES DEVELOPED AT THE INSTITUTE OF MOVEMENT SCIENCE

Benjamin O'Brien, Adrien Vidal, Lionel Bringoux, Christophe Bourdin

Aix Marseille Univ, CNRS, ISM, Marseille, France
{benjamin.o-brien, adrien.vidal, lionel.bringoux, christophe.bourdin}@univ-amu.fr

ABSTRACT

This paper offers a survey of movement sonification studies conducted over the last four years at the Institute of Movement Science. Our research focuses on studying the effects of online sonification on sporting performance and movement in golf and cycling. Given the different goals and motor control skills required to be successful, our experiences have provided us with significant insight when considering experimental design and analysis. Skill level and the complexity and ease of repeating motor tasks are major factors when developing sonification strategies and studying its effects. Decisions regarding which movement parameters and the presentation of sonification are equally important depending on study goals. The following outlines our perspectives and methodologies when developing and studying sonification and its effect on sports performance and movement.

1. INTRODUCTION

Improving sport performance is a major focus in the field of sciences. It traditionally involves multi-disciplinary knowledge, from sociology to psychology, biomechanics, and also neurosciences. A popular area of research over the last couple of decades is studying the effects of augmented reality and multi-sensory feedback on human movement and performance [18]. While vision is significant when performing motor control tasks, humans are multi-sensory, and thus the influence of other stimuli has to be more accurately studied. Our research interest focuses on studying the effects of online sonification on sporting performance and movement.

In general, *sonification* is the use of sound to represent data [13, 19]. Natural sonification happens all the time in our daily lives [10], which is typically understood as acoustic feedback generated by the contact of two surfaces, such as dropping a glass marble onto a marble table. However, more recently, artificial sonification is the process of synthesising sound from data abstracted from a source, such as human movement [18]. Artificial sonification of human movement can be presented as a history of performance (offline) or in real-time, concurrent with motor control tasks (online). An important focus then is to design (and present) sounds that relate to motor control tasks required to be successful.

Previous research has suggested that the repetition of auditory-motor activities promotes neural coupling [20], which can enhance motor learning, performance, and rehabilitation. An example of the influence of training these actions can be found in studies on professional pianists, such as [3], which showed auditory feedback enhances the learning of coordinated motor-related actions. Like musicians, athletes also require precise movements and timings, and studies by [9] and [2] have shown the effectiveness of training with sound.

Over the last four years, we have developed numerous sonification studies at the Institute of Movement Science based on golf and cycling. Although each sport appears quite different, they both require demands on vision - eyes on the ball and the road, respectfully. Because of this there is an opportunity to develop and use sound to augment and convey information regarding performance and movement.

Golf putting is a discrete complex motor skill sport, which requires considerable concentration and precision to move the club at a speed in which impact is sufficient enough for the ball to reach a target [7]. Although the putting gesture can be partitioned into backswing and downswing sub-movements or phases, there are numerous ways to swing the putter, such as increasing or decreasing wrist or elbow movement. Research has shown, however, that despite the innumerable ways of applying forces during the swing phases, most successful putts have comparable velocity profiles [11]. This observation supports findings by [8], which found *club head velocity* at impact strongly correlates to ball distance, which of course relates to performance.

Alternatively cycling requires continuous and coordinated movements to be performed across distances over time. A complex process, the pedal stroke consists of pushing and pulling phases and high and low transitions [17]. The most difficult is the pulling phase, as well as transitioning in to and out of it, and research has been dedicated to evaluate efficient pedalling techniques [4].

This paper details our studies on the effects of online sonification on sporting performance and movement in golf and cycling. The first part outlines early studies and pilot tests we developed to determine which movement parameters and sonification strategies affected performance with novices. Because of the obvious differences in skill levels with novices, the following section addresses research conducted with expert participants. As sound is perceived differently among humans, these sections seek to address some of the different ways sonification affects performance differently depending on sonification strategies and skill levels. The following section offers methodologies and results to golf putting and cycling sonification studies that focussed on learning and performance enhancement when vision was either limited or there was a greater demand on it. We conclude with a recent study for error-based sonification of golf putting and future work.

2. SELECTING SONIFICATION PARAMETERS

As there are many ways to map data to sound [12], our first goal was to select factors that play important roles when performing the motor control task. Moreover, it was equally important that those features map to sound in ways that participants are able to perceive and associate with their movements, which, in turn, they can use

to enhance performance.

2.1. Golf putting parameters & sonification strategies

To study the effects of sonification on putting performance, we first devised several pretests to examine which swing parameters and sound characteristics best conveyed swing information to novices. Using the CodaMotion CX1 scanner, we placed infrared markers near the hand grip and club head of the putter and recorded the kinematic data of an expert golfer performing putts at 3 m, 6 m, and 9 m (sampling rate: 200 Hz). With this data, we synthesised sound in MATLAB (offline) by using different combinations of swing parameters (club head velocity, 'time to arrival' [8, 16]) and their sound mappings (frequency, psychometric) and asked 15 novices to identify which sonifications best conveyed swing information. Based on these results, we developed a second pretest to observe any behavioural effects of 20 participants performing the golf putting task while listening to sounds synthesised by different combinations of psychometric ranges (3), mapping functions (2), and displays (2). These sounds were based on the club head speed of an expert golfer performing putts at a similar distance. Based on the RMSE method, we averaged all participant trials and compared the means to the expert movement in four dimensions: maximum velocity during the (1) backswing and (2) downswing, and the standard deviation from velocity across the (3) backswing and (4) downswing. The results of both pretests are available in [16], all of which helped us develop the sonification methodology used for our golf putting learning study (see: **Section 4.1.1**). This first step was immeasurable, as it provided us with an opportunity to observe (some of) the limitations of novices and their ability to identify and associate golf putting swing features in sounds they heard.

2.2. Pedal stroke parameters & sonification strategies

Selecting which parameters to sonify and their methods was fundamental to our first sonification of the pedal stroke study. Our goal was to examine the effects of different sounds on right pedal performance. Both novice (12) and experienced (16) cyclists participated in our study, which consisted of five 2-minute sessions on a stationary bike. During all sessions, forces applied to the pedal, or *torque*, and pedal angle were measured with the Rotor crank and application (sampling rate: 50 Hz), which uses ANT+ transmission. Participants were presented no sound during sessions 1, 5, but were randomly represented three auditory conditions during sessions 2-4:

- *Squeak*: When the torque applied on the pedal was negative, a 'squeak' sound was produced from a custom Max/MSP synthesiser using wide-band noise ($f_c = 300$ Hz, $Q = 3$).
- *Dynamic*: The centre frequency of band-pass noise was correlated to pedal phase, such that frequency rose when the pedal ascended (and vice-versa).
- *Music*: Instead of sonification, the song "Gimme all your Lovin?" by ZZ Top was played, which emphasises the tempo (120 BPM) with strong attacks on snare and bass drums.

A major finding was that both novice and experienced cyclists were able to use sonification to improve average *torque effectiveness*[1] (TE). RM ANOVA revealed a main effect on mean TE of $F_{4,104} = 8.23$, $p < 0.001$, and post-hoc Bonferroni-adjusted t-tests showed that, with the exception of Dynamic, the Squeak condition had a higher average TE than the Silence and Music conditions. Another takeaway of this study was that an error-based sonification work best. Our results suggest listening to sonification while cycling was not attention demanding, which was an important finding moving forward with future studies on the effects of sonification on the pedal stroke.

3. SKILL LEVEL

Participant skill level is also an important factor when considering the effects of sonification on sports performance. When studying novices, our goals are to examine whether they are able to use sonification to enhance performance or aid their learning of a new motor control task. When studying experts, we seek to examine whether they can use sound to sustain or improve upon their already high-level of performance.

3.1. Expert golfers

To date we have not yet conducted a study on the effects of sonification on expert golfers performing putts. A study by [14] found expert golfers were able to identify their own swings, associated with 65 m, targets by sound associated with them. This was an important reference for our development of a study on golf swing sonification, which is described in **Section 6**. As it too is a difficult aspect of the game, we might imagine developing a study that has expert participants performing putts at multiple distances, where personalised sonification is presented in ways that might help them associate putting club head speeds with distance.

3.2. Expert cyclists

Based on the results reported in **Section 2.2**, we developed a study for expert cyclists to examine the effects of bilateral sonification. As our previous study only presented sound relative to right pedal performance, we wanted to examine whether unilateral sonification might increase pedalling asymmetries, which, as reported in [6], would make sonification counter-productive by decreasing overall performance. Using the same sonification parameterisation and strategy, 24 expert cyclists performed five 4-minute pedalling sessions, each with different auditory conditions: sessions 1, 5 were silent, whereas sessions 2-4 were randomly selected from three sonification display conditions (left, right, stereo). RM ANOVA revealed a significant effect of auditory condition $F_{4,88} = 19.23$, $p < 0.001$. Post-hoc Bonferroni-adjusted t-tests showed the left foot sonification was significantly higher for all conditions, except for the right foot sonification where performances were similar. The main takeaway from our findings show participants greatly improved their average torque effectiveness (TE) when presented bilateral sonification, as compared to unilateral sonification. Additionally, while unilateral sonification did improve the concerned foot, there was only a slight improvement for the opposite foot.

This work led to our current study examining the effects of sonification on torque effectiveness, kinematics, and muscular activity in experienced cyclists. While our previous research found

[1] Torque effectiveness $= \frac{\tau^+ + \tau^-}{\tau^+}$, where τ^+ and τ^- are the total positive and negative torque values over the cycle, respectfully.

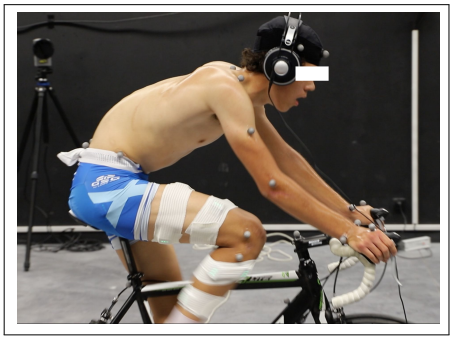


Figure 1: Experimental setup for current sonification of pedal stroke study. Qualysis (66) passive-markers were placed on the head, body, and limbs to measure kinematics. EMGs (10) were placed on the right leg to measure muscular activity. Headphones were used to deliver sonification.

that both novices and experienced cyclists were able to use sonification to become more efficient at pedalling, the goal of this study is to observe whether these changes are physiological or biomechanically costly. By focusing on only experts, we want to examine whether, given any significant effects of sonification, there are also any significant performance correlations or 'moments' when is sound used to enhance (already high-levels of) torque effectiveness and kinematic or muscular performance. Participants performed four 6-minute cycling sessions, where each session randomly presents sonification differently (silent, right, left, stereo). During each session, sonification was only presented for 20 seconds at the start of 1, 2, 3, 4 minutes, which allowed us to measure time it takes for participants to associate sound with their pedal performance. To measure the kinematic effects of sonification, 66 Qualysis markers (size: 19 mm, weight: 2.5 g) were placed on participants and bike, while 10 EMGs were placed on their right legs to measure muscular activity. **Figure 1** illustrates our setup. **Figure 2** illustrates one participant's average torque per angle for each auditory condition. In this instance, we observe that, in comparison to the silent condition, the participant reduced her negative torque for both feet when presented any sonification. Testing is on-going.

4. EFFECTS OF SONIFICATION

The previously discussed studies address some of the ways we look at studying the effects of sonification on sports performance. Each of them, in some way, position sonification as a tool to enhance performance in real-time. However, a different study goal is to examine whether sonification can be used by novices to learn complex motor skills or for experts to improve and sustain performance. Additionally, because of the visual demands of each sport, another perspective is to study how sonification can enhance motor skill performance when vision is either limited or there is a greater demand on it.

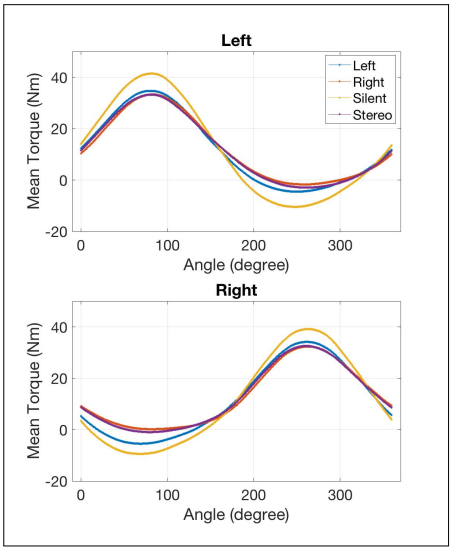


Figure 2: Participant's average torque per angle for each pedal when presented each auditory condition.

4.1. Learning studies

4.1.1. Golf putting learning study

The aim of our first comprehensive sonification of golf putting study was to examine whether novices can use sensory cues developed from expert swing performance to enhance motor learning of the golf putting gesture [5]. Thirty participants were divided into control, audio, and visual groups ($n = 10$) and participated in putting sessions over eight weeks. All participants putted balls across three distances (3 m, 6 m, 9 m) on an artificial terrain installed in our lab. During their putts, participants' swing movements and putting distances were measured with CodaMotion and analysed in MATLAB. Participants in the auditory and visual groups were presented offline sensory stimuli based on the previously collected expert kinematic data and were instructed to syncopate their movements. Following our pretesting (see: **Section 2.1**), we decided to linearly map the club head speed of an expert performing putts at similar distances to the centre frequency of a BPF with white noise input, and used different frequency ranges for each distance [16]. As reported in [5] all groups improved target distance over the course of the experiment, but the improvement in auditory and visual feedback groups was more pronounced. RM ANOVA found, among other things, a main effect on target distance $F_{1.18,30.7} = 38.18$, $p < 0.01$, $\eta_p^2 = 0.59$. Although post-hoc Bonferroni-adjusted t-tests reported no group effects, both audio and visual feedback groups showed more pronounced improvements when compare to the control group. However, both types of sensory guidance led to display dependences, as performance dropped when participants were no longer exposed to sensory stimuli. Based on these results, we believed that developing sonification concurrent with putting movements might diminish any 'guiding effects' [1] and improve motor learning, which was reported in a study by [9]. This point served as the basis of our putting study discussed in **Section 4.2.1**.

4.1.2. Cycling learning study

We expanded our study described in **Section 3.2** by examining whether a small group of experts were able to use sonification to improve and stabilise their torque effectiveness over successful training sessions. Using the same experimental setup and protocol, participants (n =3) repeated one session per week over seven weeks, as well as a final session to examine retention one month later. Because of the small population of participants, statistical analysis was not conducted. However, we found participants improved their average torque effectiveness (TE) by the fourth session and then stabilised. The retention test also showed they were able to maintain this performance.

4.2. Limits of vision

4.2.1. Golf putting with limited vision

Despite swing idiosyncrasies and immeasurable strategies that separate successful golfers, they are all required to focus their vision on the ball in order to make precise contact with the ball. Because of these visual demands, we developed a study to examine whether online sonification had a direct behavioural or perceptual effect on golf putting with limited visual feedback. 20 novices performed a random sequence of 25 3.5 m putts. During each putt, they were exposed to a different online sonification of their club head velocity, which was synthesised from a combination of mapping (3), synthesisers (2), timbral modulations (2) and scale (2) types, whose constructions are described in [16]. Participants performed this 25-putt sequence five times (125 putts). At impact with the ball, shutters worn by participants were closed, whereupon they were asked to estimate the location of their ball. **Figure 3** illustrates the experimental setup. For target distance error standard deviation, we found, among other things, a main effect for types of synthesiser $F_{2,38}$ = 41.2, $p < 0.001$, η_p^2 = 0.68, and our post-hoc Bonferroni-adjusted t-tests found both synthesisers had lower means when compared to the pink noise trials (8.02 $\pm$ 1.7; 6.91 $\pm$ 1.73), $p < 0.001$. For zone estimation error standard deviation there was, among other things, a main effect for types of synthesiser $F_{2,38}$ = 31.89, $p < 0.001$, η_p^2 = 0.63, and post-hoc Bonferroni-adjusted t-tests showed that one synthesiser had a significantly lower mean when compared to both pink noise (0.26 $\pm$ 0.1) and the other synthesiser (0.13 $\pm$ 0.05) trials, $p < 0.05$. Our analysis showed that, despite vision being limited at impact, participants significantly reduced variability in their distance from the target and ball location estimation when presented auditory feedback that was concurrent with their movements. We found the effect of online sonification with one type of synthesiser on these two performance features yielded a significant correlation (R^2 = 0.51, $p < 0.001$). Thus our findings illustrate that novices were able to use a particular (synthesiser) type of sonification to reduce variability in putting distance and estimations based on their performance despite vision limitations to their vision.

4.2.2. Cycling and visual demands

Because of the visual demands required to cycle in the real-world, we developed a study to examine the effects of auditory and visual stimuli on torque effectiveness and reaction times when identifying (virtual) obstacles while pedalling. 24 novices participated in six 2-minute sessions. The first three training sessions randomly presented subjects three conditions when the torque applied to the pedal was negative:

Figure 3: Experimental setup for sonification of golf putting gesture with limited visual feedback (shutter glasses).

- Auditory feedback: the 'squeak' sound delivered by headphones
- Visual feedback: red circles generated in Jitter via a small digital screen positioned near the handle bars
- Control: no auditory or visual feedback

During all of these sessions, based on their pedal performance participants were presented a real-time animation depicting a typical road-cycling experience, which was developed in UNITY and presented on a monitor in front of them. For the remaining three sessions, participants were similarly presented the conditions, but were now asked to verbally identify when they noticed an obstacle ('a large white dome') displayed win the animation. **Figure 4** illustrates the experimental setup. In comparison to the Control condition, both visual and auditory feedbacks significantly enhanced participants' pedalling techniques $F_{2,46}$ = 8.265, $p < 0.001$. During the visual condition, the gaze behaviour was partially oriented towards the small screen on the handlebars, which was where the visual feedback presentation was located. Thus, participants were less attentive to the 'road' - the real-time animation of road-cycling experience. Our comprehensive results are reported in [21]. Moving forward, our findings suggest that participants benefit from artificial multi-sensory feedback, but what remains unclear is how do we determine which type - auditory, visual, haptic, or multi-modal - suits the individual best.

5. ERROR-BASED SONIFICATION FOR GOLF PUTTING

As evidenced we have studied the effects of both offline and online sonification on sports performance. As we found both novice and expert cyclists were able to use error-based sonification to improve torque effectiveness, we wanted to examine whether novices could use a similar strategy to improve putting performance. But given our previous experiences observing the immense swing variability between and within novices performing the golf putting gesture, we first required personalised swing models to calculate errors for which to develop sonification.

Figure 4: Experimental setup for cycling study with visual and auditory feedback with real-time animation of road-cycling experience (screen), visual feedback (handlebars), and sonifciation (headphones).

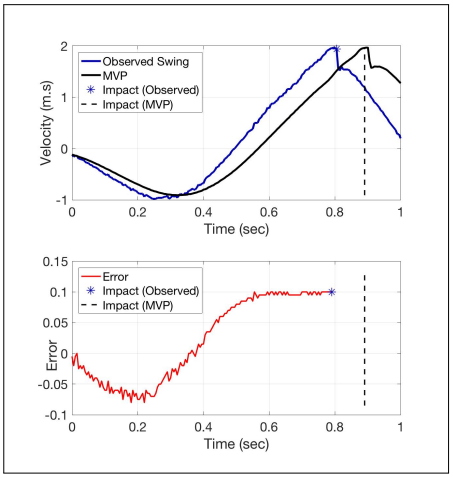


Figure 5: Top: Comparison between observed (blue) and MVP (black) swings. Bottom: Error (red) calculated from the difference between observed and MVP swings.

Described in [16], we developed a method that synchronised any number of putting trials at ball impact, shifted their swing velocities, and calculated the mean velocity profile (MVP). Our method then estimated the time to impact with the ball by using club head marker values to calculate its acceleration and distance from the ball. This estimated time to impact was then compared in real-time to the MVP, which, in turn, gave us a real-time difference, or *error*, between a participant's observed and MVP swings. **Figure 5** illustrates the real time difference (error) between a participants observed and MVP swings for a 2 m putt.

Forty participants first performed 20 2 m, 4 m putts, which were used to calculate their MVPs. Next participants were randomly assigned to a different group ($n = 10$), where they then performed 20 2 m, 4 m (total: 80 putts). During these trials, participants were presented different auditory conditions depending on their group: static pink noise ('Control'); MVP velocities mapped psychometrically to a sinusoidal oscillator ('MVP'); errors modulated the stereo display of the MVP auditory signal ('Directivity'); and errors modulated the 'roughness' of the MVP auditory signal ('Roughness').

Among other results, we found a main effect on group for percentage of improvement for average swing velocity deviation $F_{3,36}$ = 3.17, $p < 0.05$, η_p^2 = 0.21, and post-hoc Bonferroni-adjusted t-tests showed MVP participants significantly lowered improved in comparison to the Control group, $p < 0.05$. In addition, for temporal ratio standard deviation we found an interaction on trial type * group $F_{3,36}$ = 3.02, $p < 0.05$, η_p^2 = 0.2 and post-hoc tests showed Directivity participants significantly lowered their means when presented sonification, $p < 0.01$. These results provide further evidence of the benefits of sonification for novices learning new motor skills and suggest the use of personalised templates for sonification reduces variability in the execution and timing of complex movements. Our findings also suggest that sonification of real-time errors (auditory feedback) can be more influential on novice performance than personalised sonification (auditory guidance).

6. CONCLUSION

The studies presented in this paper demonstrate how sonification can be used as a tool to aid novices and experts alike in golf putting and pedalling tasks. Because of the complexity and speciality of motor skills required to improve performance, the scientific questions are numerous and quite challenging. Although learning, training, and improving this difficult motor task is de facto complex, it also presents an exciting opportunity for study, especially when considering the development of augmented reality tools for expert athletes, who wish to maintain or improve their performance and mechanics.

Our work has lead to recent research on the effects of online error-based sonification on a very rapid and complex motor control task: the golf swing. By adapting the Nesbit and McGinnis optimisation model of the golf swing [15], which identifies three swing parameters as possible for optimising swing velocity at impact, we developed a protocol that calculates and sonifies the real-time difference, or error, between observed and optimised swing paths for each swing parameter. The general idea of this study follows closely to our error-based sonification of golf putting study with novices (see: **Section 5**), which demonstrated how novices benefitted from error-based sonification. A major advantage of our model is that it adjusts to the kinematic capacities of the individual, which may prove useful in both healthy and rehabilitation research.

Due to the inter-individual performance differences between experts, the choice in media may relate to participant characteristics. Multi-sensory integration theories have shown that different modalities can be integrated and used differently among humans. Relative to sound, psychoacoustics tells us that sound is perceived differently among humans due to psychological and physiological differences, which may offer a general explanation as to why some participants used sounds developed from one sonification strategy, whereas others found another to be more easily useable. Nevertheless, audition appears to be important for performance improvement, but remains one - and very new - tool among others.

Acknowledgements

This research was funded by the Carnot STAR institute under the PEPS project, and by the French National Research Agency (ANR) under the SoniMove project (ANR-14-CE24-0018).

7. REFERENCES

[1] Adams, J. (1971) "A closed-loop theory of motor learning," Journal of Motor Behaviour, Vol. 3(2): 111-149.

[2] Agostini, T., Righi, G., Galmonte, A., & Bruno, P. (2004). The Relevance of Auditory Information in Optimizing Hammer Throwers Performance, *Biomechanics and Sports*. Vienna: Springer, 67–74.

[3] Baumann, S., Koeneke, S., Schmidt, C., Meyer, M., Lutz, K., & Jancke, L. (2007) A network for audio-motor coordination in skilled pianists and non-musicians, *Brain Research*, *1161*, 65–78. doi:10.1016/j.brainres.2007.05.045

[4] Bibbo, D., Conforto, S., Bernabucci, I., Carli, M. Schmid, M., D'Alessio, T. (2012) Analysis of different image-based biofeedback models for improving cycling performances., Image Processing: Algorithms and Systems X; Parallel Processing for Imaging Applications II. doi:10.1117/12.910605.

[5] Bieńkiewicz, M., Bourdin, C., Bringoux, C., Buloup, F., Craig, C., Prouvost, L., Rodger, M. (2019) The Limitations of Being a Copycat: Learning Golf Putting Through Auditory and Visual Guidance, *Frontiers 10*, 92. doi:10.3389/fpsyg.2019.00092

[6] Bini, R., & Hume, P. (2014) Assessment of bilateral asymmetry in cycling using a commercial instrumented crank system and instrumented pedals, *International Journal of Sports Physiology and Performance 9*(5): 876–881. doi:10.1123/ijspp.2013-0494.

[7] Burchfield, R. & S. Venkatesan (2010) A Framework for Golf Training Using Low-Cost Inertial Sensors, *Proceedings of the 2010 International Conference on Body Sensor Networks*. doi:10.1109/BSN.2010.46

[8] Craig, C. M., Delay, D., Grealy, M. A., & Lee, D. N. (2000) Guiding the swing in golf putting, *Nature*, 295–6. doi:10.1038/35012690

[9] Effenberg, A., Ursula, F., Schmitz, G., Krueger, B., & Mechling, H. (2016) Movement Sonification: Effects on Motor Learning beyond Rhythmic Adjustments, *Frontiers in Neuroscience*. doi:10.3389/fnins.2016.00219

[10] Gaver, W. (1993) "What in the World Do We Hear?: An Ecological Approach to Auditory Event Perception," Ecological Psychology, Vol. 5: 1-29.

[11] Grober, R. (2009) Resonance in putting, *arXiv*, doi:0903.1762.

[12] Grond, F. & Berger, J. (2011) Parameter mapping sonification. In Hermann, T., Hunt, A., Neuhoff, J. G., editors, *The Sonification Handbook*: 363-397. Logos Publishing House: Berlin.

[13] Hermann, T. (2008) Taxonomy and definitions for sonification and auditory display, *Proceedings of the 14th International Conference on Auditory Display*. Paris, France.

[14] Murgia, M., Prpic, V., O, J., McCullagh, P., Santoro, I., Galmonte, A., & Agostini, T. (2017). Modality and Perceptual-Motor Experience Influence the Detection of Temporal Deviations in Tap Dance Sequences, *Frontiers in Psychology*, *8*: 1340. doi:10.3389/fpsyg.2017.01340

[15] Nesbit, S.M. and McGinnis, R. (2014) Kinetic Constrained Optimization of the Golf Swing Hub Path, *Journal of Sports Science and Medicine 13*, 859-873. doi:10.1080/10671315.1979.10615598

[16] O'Brien, B., Juhas, B., Bienkiewicz, M., Pruvost, L., Buloup, F., Bringnoux, L., and Bourdin, C. (2018) Considerations for Developing Sound in Golf Putting Experiments. *Post-proceedings of CMMR 2017 - Music Technology with Swing*, Lecture Notes in Computer Science, Springer-Verlag Heidelberg. doi:10.1007/978-3-030-01692-0

[17] Patterson, R. & M. Moreno. (1990) Bicycle pedalling forces as a function of pedalling rate and power output, *Medicine and science in sports and exercise 22*(4), 512–516.

[18] Sigrist, R., Rauter, G., Riener, R., Wolf, P. (2013) Augmented visual, auditory, haptic, and multimodal feedback in motor learning: A review, *Psychonomic Bulletin & Review*, *20*(1), 21–53. doi:10.3758/s13423-012-0333-8

[19] Scaletti, C. (1994) Sound Synthesis algorithms for auditory data representation. G. Kramer (ed.), *Auditory Display (XVIII) of Santa Fe Institute, Studies in the Science of Complexity Proceedings*, 223-252. Addison-Wesley, Reading, MA, 1994.

[20] Schaffert, N., Janzen, T., Mattes, K., & Thaut, M. (2019) A Review on the Relationship Between Sound and Movement in Sports and Rehabilitation, *Front Psycho* 10: 244. doi:10.3389/fpsyg.2019.00244

[21] Vidal, A., Bertin, D., Kronland-Martinet, R., & Bourdin, C. (2019) Pedalling technique enhancement: a comparison between auditive and visual feedbacks, *CMMR*, Oct 2019, Marseille, France. hal-02264340

SPEED SONIFICATION IN A UNIMANUAL TIMING AND A BIMANUAL COORDINATION TAPPING TASK

Magdalena Gippert

Biopsychology & Cognitive Neuroscience, Faculty of Psychology & Sports Science

Bielefeld University, Bielefeld, Germany
magdalena.gippert@uni-bielefeld.de

*Tobias Heed**

Biopsychology & Cognitive Neuroscience, Faculty of Psychology & Sports Science and CITEC Center of Excellence

Bielefeld University, Bielefeld, Germany
tobias.heed@uni-bielefeld.de

*Thomas Hermann**

Ambient Intelligence Group Faculty of Technology and CITEC Center of Excellence

Bielefeld University, Bielefeld, Germany
thermann@techfak.uni-bielefeld.loc

ABSTRACT

This paper introduces systematic research on how uni- and bimanual tapping tasks can profit from interactive sonification of hand-surface interactions to support coordination. To that end, we developed a new experimental platform featuring a web app that allows multi-touch tablet-based interaction for sonification experiments. We present and discuss two experiments to test tap-interval-to-tone-frequency-mappings, to assess their ability to support rhythmic, resp. polyrhythmic tapping. The results, although negative with regards to our initial hypothesis, provide guidance for subsequent experiments to better understand how sonification can best be used to support coordination tasks, particularly in motor control contexts that involve discrete, goal-directed movements.

1. INTRODUCTION

Motor learning is a central challenge in many contexts. For instance, rehabilitation and sports training often focus on enhancing specific movements, such as learning how to walk after stroke, or learning how to do a backflip in gymnastics.

A commonly used technique in motor learning is augmented feedback, such as video feedback of sports performance, pacing or timing movement by clapping, and touching a body part during an exercise to indicate that its posture should be corrected. Although these practical examples illustrate that all sensory systems can be addressed to convey potentially relevant information, basic research has primarily focused on investigating the principles involved in processing visual feedback. While feedback through other sensory systems has received less attention, the recent years have seen an increasing interest in sonification as a means of providing performance-relevant information auditorily (Alfred O. Effenberg, Fehse, Schmitz, Krueger, & Mechling, 2016; Sigrist, Fox, Riener, & Wolf, 2016; Sigrist, Rauter, Marchal-Crespo, Riener, & Wolf, 2015; Vinken et al., 2013).

There are some apparent advantages of using audition over vision as the modality provides movement-relevant information: First, temporal discrimination of events is best perceived in the auditory modality (Shea, Wulf, Park, & Gaunt, 2001). Second, there is no obvious disruption of intrinsic information processes, as is the case, for instance, when participants have to fixate a screen during movement to obtain visual feedback. Third, some experimental results have suggested that retention may be superior after auditory as compared to visual learning, possibly because the auditory feedback more strongly supports creation of an internal model of the concurrently perceived proprioceptive movement consequences (Dyer, Stapleton, & Rodger, 2017; Ronsse et al., 2011).

Whereas the investigation of sonification principles in practical contexts may be advantageous for transferring experimental knowledge into real-world applications, real life tasks often do not lend themselves to systematic investigation of the many factors that are potentially involved in the underlying cognitive processes, such as motor learning and motor control. Therefore, we present here first results of a line of studies that aim at providing an experimental paradigm that allows finetuned experimental manipulations for multiple sonification implementations, and in addition is suited for implementing visual control conditions that are analogous to the tested sonifications. To this end, we developed a bimanual coordination paradigm with sonification. All conditions can be ported to visual feedback conditions; we note, however, that the present paper focuses only on auditory conditions.

Moving two limbs independently of each other is a key requirement of motor tasks in everyday life. Asymmetrical bimanual movement tasks provide the opportunity to study how the tendency to move in symmetry and with equal speed can be suppressed giving insight into fundamental motor mechanisms. Accordingly, bimanual coordination is an area of motor control that has been addressed with a multitude of motor tasks, such as rhythmically bending the two wrists (Amazeen, DaSilva, & Amazeen, 2008), wiggling the index fingers (Heed & Röder, 2014), making circular hand movements (Mechsner, Kerzel, Knoblich, & Prinz, 2001), and finger tapping (Heed & Röder, 2014; Mechsner et al., 2001). Accordingly, a rich literature has addressed the principles according to which visual feedback interacts with behavioral performance (for extensive reviews, see Hommel, Müsseler, Aschersleben, & Prinz, 2001; Shea, Buchanan, & Kennedy, 2016; Swinnen, 2002). The principles underlying the use of auditory feedback, in contrast, have received much less attention in the literature (Dyer, Stapleton, & Rodger, 2015; Sigrist, Rauter, Riener, & Wolf, 2013). By linking sonification directly to this already established area of

* T. Heed and T. Hermann contributed equally to this work

psychology, in particular to visuomotor processing (Mechsner et al., 2001; Shea et al., 2016; Sisti et al., 2011), we aim to systematically derive processing principles that generalize across motor tasks and help the design of specific mapping designs in practical contexts.

Here, we focus on a rhythmic, bimanual tapping paradigm. Tapping tasks have been used to investigate the ability to perform asymmetrical motor patterns (Mechsner et al., 2001). In addition, unimanual tapping has long been used to examine rhythmic coordination of perception and action (Repp, 2005). Our paradigm requires participants to tap, with each hand, on four targets at a constant tapping rate in such a way that one hand makes two, and the other hand three rounds in the same amount of time.

There are many movement parameters that could lend themselves to improving motor control through sonification. There are examples in the literature for sonifying movement speed (Schaffert, Mattes, & Effenberg, 2011; Schmitz et al., 2013), acceleration (Chiari et al., 2005), force (Coull, Tremblay, & Elliott, 2001; A.O. Effenberg, 2005), object size (Säfström & Edin, 2006), specific movement reversal points (Ronsse et al., 2011) and body segmental alignment (Baudry, Leroy, Thouvarecq, & Chollet, 2006). For our tapping paradigm, we first focused on speed sonification. Sonification of velocity has been successfully used to support swimming and rowing (Schaffert et al., 2011; Schmitz et al., 2013). In motor tasks where movement velocity or rhythm are of significant importance for successful execution, sonification of speed of motion seems promising since sound is especially effective in displaying temporal aspects (Sigrist et al., 2013). In our paradigm, the hands must move at different speeds to fulfil the task requirement of executing a different number of rounds with each hand during a given time interval.

Similarly, in our sonification we could manipulate many different parameters in the present task context. In our experience pitch is the most salient quality of pitched sounds (compared to qualities such as loudness, roughness, modulation, and attack time), and we therefore chose it as the first sonification parameter to test in the present setting. Moreover, it has been demonstrated that presenting the sonification of a reference movement to one ear, and the sonification of actual task performance to the other ear, is an effective way to improve performance (Sigrist, 2015), likely because this sonification highlights deviations from an optimal movement pattern. Because participants had to tap with both hands in our task, there was no optimal movement velocity in the present study. Instead, participants had to achieve a particular relative timing of taps with the two hands. We reasoned that this scenario can, however, be conceptualized similarly to presenting a correct reference: effectively, participants had to match the sonification relating to one hand with that of the other. Correct performance resulted in identical pitch for the two hands.

2. EXPERIMENT 1: BIMANUAL 2:3 TAPPING

2.1. Methods

We recruited 46 participants. Data from 6 participants were excluded due to medical reasons or technical difficulties during the study. The sample used in all reported analyses comprised 40 healthy, normal-hearing participants with a mean age of 23.13 ($SD = 4.46$); 3 participants were left-handed, and 1 ambidextrous. Given that both hands were involved in the experiment, we did not control handedness for group assignment. Participants gave written, informed consent and received course credit and/or snacks for participation. The experiment was approved by Bielefeld University's local ethics committee.

Technical setup. Participants tapped on an iPad Pro 12.9-inch Retina display. The display showed two sets of four targets each, one set for the left and one set for the right hand.

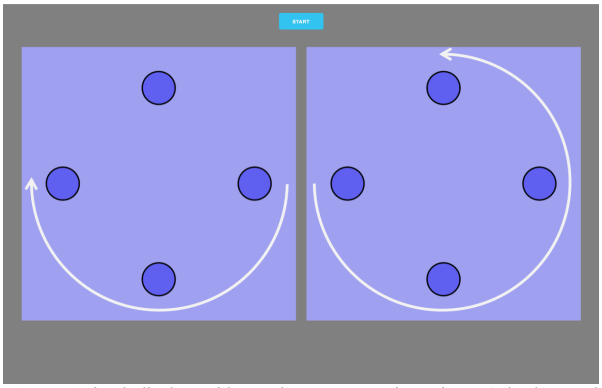


Figure 1. iPad display with tapping targets. Dimensions: 2 background squares – 12x12cm each; 8 targets – 1.5cm diameter each; distance between two adjacent targets within one square – 4.5cm. Arrows indicate the direction of tapping and the correct relative movement speed when the right hand was the instructed faster hand. Arrows are for illustration purposes only and were not presented to participants.

Task. Participants were instructed to perform a 2:3 tapping task. They had to tap on the four targets on the left side, one after another, with the index finger of the left hand progressing clockwise, and the four targets on the right side with the index finger of the right hand progressing counterclockwise (see Figure 1). Participants were instructed to move each hand with a constant velocity but complete two rounds with one hand while the other completed three rounds. In each trial, participants started with tapping the two targets closest to each other at the same time and were instructed to continue tapping in the 2:3 relative velocity until a trial was over. Each trial lasted 30 s. Some trials included sonification of the taps (see below). When no sonification was provided, participants heard white noise through headphones to prevent auditory feedback related to tapping the iPad screen.

Experimental design. Participants were assigned to the experimental or the control group. Both groups underwent identical experiments, with the exception that the experimental group received a sonification of their tapping speed, whereas the control group, instead, always heard with each tap an auditory signal that was independent of speed. In each group, half of the participants were asked to move the right hand faster than the left hand (3:2) and the other half vice versa (2:3). Regardless of their group membership, all participants underwent the same general procedure and task structure. Each participant underwent assessment of baseline performance without sonification (5 trials), testing and familiarization with sonification according to group without performing the target task (2 trials), practice of the 2:3 coordination task with sonification (15 trials), and retention testing without sonification (5 trials). Experiment duration was ~45-60 minutes including instructions and breaks.

Sonification. During practice and familiarization trials, the experimental group received sonification of tapping speed with sounds of 200ms duration. The sonification was designed to support the instructed task of moving the hands at different speeds in a 2:3 ratio. Tapping speed was mapped to the frequency of the tones for each hand, resulting in a speed-depending pitch.

To support the correct relative speed for both hands, the inter-tap interval of the faster hand had to be 2/3 of the length of the slower hand to produce the same frequency. Perfect coordination was achieved when all taps produced tones of the same pitch. For each hand, faster tapping entails a shorter inter-tap interval and was, thus, transformed with increased pitch.

Our custom-written web application renders the GUI as an HTML5 website in a Browser window, using a touchable.js library to process multitouch interactions (a.k.a. tapping), using the WebAudio Javascript API for synthesis and playback. We implemented various alternative solutions (involving synthesis on PC with SuperCollider3 via OSC, and native app development on Android devices using Pure Data for synthesis), yet in the end, the web application on Apple iPad provided the lowest overall latency, i.e. a time delay from touch to sound onset, of 35ms. We used different timbres for each hand, chosen to be easily discriminable and perceived as equally loud. We chose periodic waveforms, i.e. fundamental plus harmonics, to maximally support the perception of pitch. The left hand tone was created by additive synthesis of f_0, $3f_0$ and $5f_0$ sine waves with amplitudes 1, 0.5 and 0.1; the right hand tone was a sawtooth signal. Level was relatively increased for the left hand to counteract the relative spectral paucity of the additive sound pattern (left : right – 0.5 : 0.3). The first tap of each hand elicited a 440 Hz tone. For all subsequent taps, tone frequency depended on the last inter-tap interval of each hand according to the following mapping: The frequency f_s for the instructed slower turning hand was calculated as

$$f_s = 220\text{Hz} / (t - t_{pco,s}), \quad (1)$$

where t is the actual tap onset time and $t_{pco,s}$ the previous tap onset on the slower hand panel. To match frequencies when fulfilling the 2:3 relative velocities, the frequency for the faster hand was correspondingly calculated as

$$f_f = (2/3) \cdot 220\text{Hz} / (t - t_{pco,f}). \quad (2)$$

A tap was considered as "on target" when the position registered by the iPad was maximally 1.1cm away from the target's midpoint; given that the target radius was 0.75cm, a touch was thus considered as on-target even if slightly outside the displayed target. A sonification was presented only if a tap was registered on or near the visually shown tapping targets. Otherwise, taps were not followed by any sonification. Allowing a spatial margin around targets avoided an exceeding number of taps without sonification that would have been due to imperfect target location recording due to finger size or slight tapping imprecision by participants. Timbre assignment to the left and right hand was constant across the entire experiment. Note, however, that due to randomization of which hand had to tap fast or slow, each timbre was paired with each speed across participants.

Contrary to the experimental group, the control group did not receive speed-dependent sonification. Instead, control group auditory feedback was a 440Hz tone for each tap, independent of tapping speed. Nonetheless, timbre was used to disambiguate tones belonging to the right or the left panel (i.e., hand), just like in the experimental group. Thus, the only difference between the two groups was the dependence of the tone's pitch on hand movement speed.

Dependent variable. The iPad yielded location and time of first contact with the tablet for each tap. The raw data was pre-analyzed with Python 3 in a Jupyter Notebook (5.6.0). The Notebook and videos are made available as supplementary material on https://doi.org/10.4119/unibi/2937988. The dependent variable of the experiment was the inter-tap ratio of the two hands. Each inter-tap interval of the faster hand was divided by the most recent inter-tap interval of the slower hand. If a participant performed the task correctly, this ratio should be 2/3, with the hand making more rounds being faster than the hand making fewer rounds.

We coded the ratio of a given tap of the faster hand (with "fast" referring to the instruction, not to the participant's true performance) as a missing value in the following cases:

1) One of the hands did not follow the correct order, so that the tap occurred at an incorrect target in the sequence of that hand. Note, that this procedure allows that the sequences of the two hands are offset relative to each other and relates only to the correct order within one hand. This criterion eliminates data points when participants do not move the hand between taps, or when they adjust their sequence, for instance when they try to "start over" during a trial.
2) Participants paused one or both hands, leading to very long inter-tap intervals. We identified inter-tap intervals of the fast and slow hand which deviated more than two standard deviations from the respective mean of inter-tap intervals of a given 30s trial and excluded these data points.
3) Time intervals of the two hands were too far apart; when participants paused one hand, then new taps of the still moving hand would be continually related to the last tap interval available for the pausing hand. Therefore, we excluded data points when the time span between the end of each interval of the fast hand and the end of the corresponding most recent interval of the slow hand exceeded three times the mean of the inter-tap intervals of the fast hand; for the calculation of this value, inter-tap intervals that deviated more than two standard deviations from the mean were excluded.

To obtain an error measure that weighs relative deviations from 2:3 equally independent of direction (i.e., the fast hand being too slow vs. too fast), we calculated the absolute difference of the natural logarithm of the ratio to the natural logarithm of 2/3. We averaged over all individual data points obtained in a given trial. Baseline and retention performance were assessed as the mean of all five baseline and retention trials, respectively. For performance during practice, we calculated the mean of the last 5 of the 15 trials in this condition. Because of transmission errors during the experiment, 7 trials in total could not be saved. No participant had more than 2 trials missing.

Statistical analysis. We conducted a mixed 3×2 ANOVA with repeated measurement factor Test Phase (baseline vs. practice vs. retention) and between participants factor Sonification Group (experimental vs control group) as a between groups factor. This analysis was conducted using R (3.6.0; R Core Team, 2019), using the packages *tidyr* (Wickham & Henry, 2019), *dplyr* (Wickham, François, Henry & Müller, 2019), *ez* (Lawrence, 2016), and *ggplot2* (Wickham, 2016).

Other potential moderator variables. Following the experiment, participants filled out a paper-and-pencil questionnaire about their experience with sports and music and gave feedback and subjective impressions regarding our sonification.

Hypotheses. We hypothesized that sonification of tapping speed should aid participants of the experimental group, as opposed to those of the control group, in performing the 2:3 rhythm by adjusting their tapping patterns to produce similar frequency tones with the two hands.

2.2. Results

Performance of participants in the experimental and control groups, measured as the ratio of tap intervals of the fast vs. slow hand, is displayed in Fig. 2. A mixed ANOVA with factors Test Phase (baseline vs. end of practice vs. retention) and Sonification Group (experimental vs. control group) as a between groups factor, did not reveal any main effects and interaction. Thus, mean performance was similar across the two groups and all phases of the experiment.

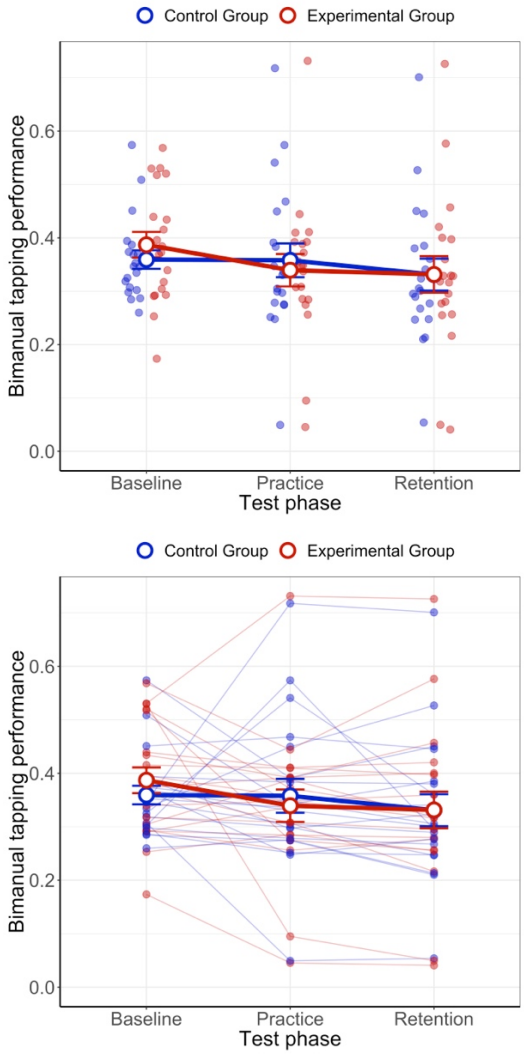


Figure 2. Results of Experiment 1. Bimanual tapping performance was measured as the difference between the ratio of the inter-tap interval of the fast and slow hand and perfect 2:3 tapping. Perfect performance would be indicated by zero values. Large, connected dots indicate group means; small, colored dots indicate individual participants. Error bars depict the standard error of the means.

To put the error measure into context, it is helpful to know how some other tapping ratios that are often erroneously performed by participants correspond to this measure. Tapping in the correct 2:3 ratio would result in an error measure of 0; tapping in a 1:2 ratio (that is, one hand taps twice as fast as the other) would result in an error value of 0.288; and tapping in a 1:1 ratio (that is, both hands at same speed) would result in an error measure of 0.405. Looking at individual performance (see small, filled dots in Fig. 2) one could argue that only three of forty participants were able to execute the required coordination task correctly, because they achieved a mean closer to 0 than to 0.288 in at least one phase of the experiment.

2.3. Discussion

The results of our Experiment 1 suggest that participants could not use the sonification of tapping speed to improve the coordination involved in performing taps at different speeds with the two hands. Thus, the experiment did not confirm our hypothesis that speed sonification can aid a rhythmic tapping task.

There may be two reasons for the lack of a sonification effect. First, participants may be unable to use the information contained in the two auditory streams to adjust their relative hand tapping speeds; in this case, the reason for the failure of Experiment 1 would be related to the difficulty and complexity of the 2:3 coordination task. Second, however, participants may be unable to adjust tapping speed according to our speed feedback more generally, even for a single hand. If one hand alone cannot even maintain sufficiently constant intervals between taps to produce tones that sound alike, then it would be unlikely that our tested speed sonification can help coordinating a second hand to match the (variable) frequencies of the first hand.

Previous studies have suggested that participants can benefit from simple auditory feedback about tapping, that is, a standard tone displayed with every tap (Chen, Repp, & Patel, 2002; Furuya & Soechting, 2010). However, we are not aware of any studies that have tested whether information about tapping speed, as used in Experiment 1 for both hands, can aid in coordinating the precision of rhythmic tapping. We therefore ran a second experiment to address these two potential reasons for the results of Experiment 1.

3. EXPERIMENT 2: UNIMANUAL TAPPING

3.1. Methods

Experiment 2 was conducted after study 1 with 20 healthy, normal-hearing participants (age: $M = 23.55$, $SD = 4.17$) who had not taken part in Experiment 1. Four participants were left-handed.

Technical setup. For the task of this experiment, the right side of the tablet was covered, so that participants only saw the four targets on the left side.

Task. Participants were instructed to tap the four targets one after another with one hand. They were asked to maintain an even pace and focus on producing inter-tap intervals of equal length throughout each trial. Half of the participants were asked to use their right hand, turning counterclockwise, and half the left hand, turning clockwise. Group assignment was randomized regardless of handedness.

Experimental design. Trials were 30s long. All participants were tested in three conditions: baseline (5 trials), speed sonification (10 trials) and control sonification (10 trials). Each sonification condition was preceded by one familiarization trial in which participants were encouraged to familiarize themselves with the respective sonification. All participants started with baseline trials. During this phase, white noise was played through headphones during the task. Participants then performed speed

sonification and control sonification; the order of these latter two phases was balanced across participants. To encourage an approximate tapping pace, we presented an audio template five times during the experiment: before the baseline trials, after each familiarization trial, and after the first five trials in each sonification condition. At the end of the experiment, participants filled out a questionnaire about their sport and music experience as well as their opinion about the usefulness of the two sonifications.

Sonification. For the unimanual study we used Experiment 1's timbre of the left hand (see 3.1) Control sonification was a 261Hz tone. For speed sonification, frequency was calculated according to equation (1), as detailed in section 3.1. The template was approx. 8s long and displayed a 261Hz tone of 200ms duration every 5/6s. Note, that tones of equal pitch (i.e., frequency) would result if participants performed the task well; depending on their individual speed, pitch would be higher with higher tapping speed. Thus, during speed sonification, participants could use consecutive sonifications as feedback about their tapping speed.

Dependent variable. The dependent variable of Experiment 2 was the coefficient of variation (CV) of the inter-tap intervals. First, the inter-tap interval was calculated for each tap. In analogy to Experiment 1, inter-tap intervals were excluded if the tapping order was incorrect, and when they deviated more than two standard deviations from the mean inter-tap interval length. Within each trial, the standard deviation of the remaining inter-tap intervals was divided by the respective mean to obtain the CV as a measure of variability (Zelaznik et al., 2005). Division by the mean tapping interval standardizes data of all participants by removing differences in individual tapping speed.

Statistical analysis. Performance in each condition was assessed as the mean CV of the last five trials in each condition. To compare the three conditions, we conducted a repeated measures ANOVA (baseline vs. speed sonification vs. control sonification).

3.2. Results

A repeated-measures ANOVA over the three sonification phases (baseline vs speed sonification vs control sonification) was significant ($F(2,38) = 4.04$, $p = .026$). Post-hoc, pairwise t-tests between all three conditions, Bonferroni-corrected for multiple comparisons, revealed a difference between baseline and control sonification $t(19) = 3.01$ $p = .022$. The mean CV was 0.059 ($SD = 0.013$) in Baseline, and 0.050 ($SD = 0.013$) during Control Sonification, indicating lower variability in this condition. However, neither the comparison between Baseline and Speed Sonification ($t(19) = -1.22$, $p = .711$), nor between Speed and Control Sonification reached significance ($t(19) = -1.53$, $p = .431$).

3.3. Discussion

The results of Experiment 2 suggest that feedback about the timepoint of unimanual taps – as opposed to a situation in which auditory feedback is prevented through noise – reduces movement variability. This result is consistent with previous reports about the guidance of rhythmic finger tapping by auditory feedback (Chen et al., 2002; Furuya & Soechting, 2010). Note, that auditory feedback of finger taps implicitly conveys information about tapping duration when a participant estimates the time that has elapsed since the previous auditory feedback. Critically, additional sonification of this time interval in the form of a mapping to oscillator frequency (audible as pitch) of the present auditory feedback did not improve tapping variability on average. In fact, variability of performance in the speed sonification phase statistically differed neither from the baseline nor from the control sonification phase, suggesting that performance was in between these two conditions. It is possible that paying attention to the additional information content of the auditory feedback bound additional resources, in effect reducing rather than aiding rhythmic performance.

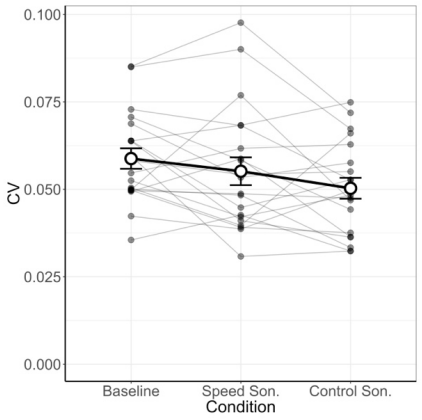


Figure 3. Results of Experiment 2. Variability of the time interval between taps during unimanual tapping was measured as the variance over a 30s period, standardized against the average tapping interval. Error bars depict the standard error of the means; CV = coefficient of variation; Son. = Sonification

However, it is noteworthy that variability *across participants* (as opposed to variability within a trial) was largest in the speed sonification; in fact, three participants performed best in this, rather than in the control sonification phase. Thus, there may be individual differences in the ability to extract sonified information.

Finally, another reason for the lack of improvement by speed sonification may be the specific mapping we employed. We used a linear mapping of interval length to tone frequency (audible as pitch); we chose this strategy because we were not aware of any studies that have systematically tested whether non-linear mappings aid levelling into a particular instructed tapping speed. Furthermore, a linear mapping is ecologically valid and connects to the pitch dependency for instance when moving at variable speed over a ruled grating. A consequence of our linear mapping is that even small changes in the length of successive inter-tap intervals were coupled with easily perceivable changes in pitch. This may have led participants to overcompensate perceived differences, or it may have demotivated them to continually try to achieve a constant pitch.

4. GENERAL DISCUSSION

We tested whether sonification of tapping intervals between individual taps directed to varying spatial targets aids the performance of bimanual and unimanual motor control. We term this type of auditory feedback *speed sonification*. A bimanual tapping task that required moving the two hands at different speeds did not benefit from speed sonification. Further experimentation with a unimanual tapping task suggests that this lack of benefit is likely due to the ineffectiveness of speed sonification in conveying effective information about inter-tap

intervals. More specifically, speed sonification did not improve tapping variability over and above the gain obtained from adding a stereotyped, invariable auditory cue to the tapping task. This result from unimanual tapping suggests that the speed information sonified in the bimanual task of Experiment 1 could not be extracted and used by participants.

In Experiment 1, some participants expressed their frustration during debriefing because they felt that they were unable to master the task. However, even though speed sonification did not improve performance in either experiment, about half of the participants reported during debriefing that they had found speed sonification helpful for task execution. This could be an indication that at least some participants were able to perceive the additional information without being able to take advantage of it. However, when asked about their strategies, several participants mentioned counting and/or focusing on the rhythm of the tapping rather than on pitch. Thus, future work could test whether better use can be made of speed sonification when participants are instructed in more detail. It is also possible that at least some participants would have learned to use our sonification efficiently if they had received more training. The present study focused on whether speed sonification can be effective in a context, in which task success would have required almost immediate use of the sonification for the movement task. It remains possible that a speed mapping can be acquired through more extensive training.

The present results contrast with previous results that have reported that sonification of movement speed can be effective, for instance to improve movement velocity during rowing (Schaffert et al., 2011) and perception of movement velocity during swimming (Schmitz et al., 2013). It is possible that the success of these studies hinges on their more extensive training of the mapping between sonification and movement. However, there are also other differences between these previous and our current study. The present experiment differs from these previous reports in that our tapping tasks involved discrete, goal-directed movements to varying spatial target locations. In contrast, movements performed during rowing and swimming are continuous. Thus, although the task in our present study was continuous and repetitive, it differs from the requirements tested previously in that specific, visual target locations must be reached. We suspect that it is this task characteristic that may make speed sonification ineffective for the type of tasks tested here.

In addition, speed sonification may put salience on a parameter which is not sufficient on its own to guide the movement of the two tested tasks. Visual feedback during bimanual coordination tasks has been found to improve performance in different tasks when it facilitated detection of deviation from optimal movement execution (Mechsner, 2001; Mechsner, 2004; Sisti, 2011).

It may strike as surprising that participants apparently did not extract rhythmic information from our sonification: although our sonification stressed inter-tap speed via pitch, sonifying taps additionally and inherently contains information about inter-tap intervals simply because the sonification occurs at the time at which the tap occurs. Thus, our sonification provided rhythmic information, and participants could have used this information to improve their tapping performance. In other words, our pitch-related sonification presented an aspect of movement in an indirect manner that could have been extracted directly from tap timing. It is possible that our sonification distracted participants from the more direct way of extracting interval length. This conclusion fits with some of our participants reporting that they used a counting or rhythm strategy. Thus, in tasks where successful performances are characterized by a steady rhythm or even a polyrhythm, it may be more advantageous to put salience on deviations from the correct rhythm. For example, playing a polyrhythm template before a bimanual coordination task that produces theses exact sounds when performed correctly was effective in a previous study (Dyer, 2017).

5. FUTURE WORK

Future work in our labs will explore the specific conditions under which sonification of movement speed can aid motor coordination. Ongoing work explores whether speed sonification is helpful for performing continuous, rather than discrete, movements. Moreover, it we are currently investigating which movement parameters other than tapping speed can aid discrete, goal-directed movements, such as those in the tapping task presented here. Finally, it will be important to directly compare the effects of visual vs. auditory feedback of the parameters identified as relevant in these different motor coordination scenarios.

6. CONCLUSIONS

We have tested whether sonification of the time between taps can aid the regularity of unimanual tapping, and the coordination of individual speeds during bimanual tapping. We found that speed sonification was not helpful in our experimental test scenarios. Thus, we conclude that pitch-mapping sonification of the time taken for discrete movements is not an effective parameter for sonification in motor control contexts that involve discrete, goal-directed movements.

7. ACKNOWLEDGMENTS

We thank Julia Thomas for help with data acquisition and Carsten Schwede for technical and programming help with the iPad implementation. MG was partly supported by a stipend of Bielefeld University. This work was further supported by the German Research Foundation (DFG) through an Emmy Noether grant to ToH (He 6368/1-3) and the Excellence Cluster Cognitive Interaction Technology (CITEC, EXC 277) at Bielefeld University.

8. REFERENCES

[1] Amazeen, E. L., DaSilva, F., & Amazeen, P. G. (2008). Visual–spatial and anatomical constraints interact in a bimanual coordination task with transformed visual feedback. Experimental Brain Research, 191(1), 13–24. https://doi.org/10.1007/s00221-008-1490-x

[2] Baudry, L., Leroy, D., Thouvarecq, R., & Chollet, D. (2006). Auditory concurrent feedback benefits on the circle performed in gymnastics. Journal of Sports Sciences, 24(2), 149–156. https://doi.org/10.1080/02640410500130979

[3] Chen, Y., Repp, B. H., & Patel, A. D. (2002). Spectral decomposition of variability in synchronization and continuation tapping: Comparisons between auditory and visual pacing and feedback conditions. Human Movement Science, 21(4), 515–532. https://doi.org/10.1016/S0167-9457(02)00138-0

[4] Chiari, L., Dozza, M., Cappello, A., Horak, F. B., Macellari, V., & Giansanti, D. (2005). Audio-Biofeedback for Balance Improvement: An Accelerometry-Based System. IEEE

Transactions on Biomedical Engineering, 52(12), 2108–2111. https://doi.org/10.1109/TBME.2005.857673

[5] Coull, J., Tremblay, L., & Elliott, D. (2001). Examining the specificity of practice hypothesis: Is learning modality specific? Research Quarterly for Exercise and Sport, 72(4), 345–354. https://doi.org/10.1080/02701367.2001.10608971

[6] Dyer, J. F., Stapleton, P., & Rodger, M. (2017). Transposing musical skill: Sonification of movement as concurrent augmented feedback enhances learning in a bimanual task. Psychological Research, 81(4), 850–862. https://doi.org/10.1007/s00426-016-0775-0

[7] Dyer, J. F., Stapleton, P., & Rodger, M. W. M. (2015). Sonification as concurrent augmented feedback for motor skill learning and the importance of mapping design. The Open Psychology Journal, 8(1), 192–202. https://doi.org/10.2174/1874350101508010192

[8] Effenberg, Alfred O., Fehse, U., Schmitz, G., Krueger, B., & Mechling, H. (2016). Movement sonification: Effects on motor learning beyond rhythmic adjustments. Frontiers in Neuroscience, 10, 1–18. https://doi.org/10.3389/fnins.2016.00219

[9] Effenberg, A.O. (2005). Movement sonification: Effects on perception and action. IEEE Multimedia, 12(2), 53–59. https://doi.org/10.1109/MMUL.2005.31

[10] Furuya, S., & Soechting, J. F. (2010). Role of auditory feedback in the control of successive keystrokes during piano playing. Experimental Brain Research, 204(2), 223–237. https://doi.org/10.1007/s00221-010-2307-2

[11] Heed, T., & Röder, B. (2014). Motor coordination uses external spatial coordinates independent of developmental vision. Cognition, 132(1), 1–15. https://doi.org/10.1016/j.cognition.2014.03.005

[12] Hommel, B., Müsseler, J., Aschersleben, G., & Prinz, W. (2001). The theory of event coding (TEC): A framework for perception and action planning. Behavioral and Brain Sciences, 24, 849–878. https://doi.org/10.1017/S0140525X01000103

[13] Lawrence, M. A. (2016). ez: Easy Analysis and Visualization of Factorial Experiments. R package version 4.4-0. https://CRAN.R-project.org/package=ez

[14] Mechsner, F., Kerzel, D., Knoblich, G., & Prinz, W. (2001). Perceptual basis of bimanual coordination. Nature, 414(6859), 69–72. https://doi.org/10.1038/35102060

[15] R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

[16] Repp, B. H. (2005). Sensorimotor synchronization: A review of the tapping literature. Psychonomic Bulletin & Review, 12(6), 969–992. https://doi.org/10.3758/BF03206433

[17] Ronsse, R., Puttemans, V., Coxon, J. P., Goble, D. J., Wagemans, J., Wenderoth, N., & Swinnen, S. P. (2011). Motor learning with augmented feedback: Modality-dependent behavioral and neural consequences. Cerebral Cortex, 21(6), 1283–1294. https://doi.org/10.1093/cercor/bhq209

[18] Säfström, D., & Edin, B. B. (2006). Acquiring and adapting a novel audiomotor map in human grasping. Experimental Brain Research, 173(3), 487–497. https://doi.org/10.1007/s00221-006-0394-x

[19] Schaffert, N., Mattes, K., & Effenberg, A. O. (2011). An investigation of online acoustic information for elite rowers in on-water training conditions. Journal of Human Sport and Exercise, 6(2), 392–405. https://doi.org/10.4100/jhse.2011.62.20

[20] Schmitz, G., Mohammadi, B., Hammer, A., Heldmann, M., Samii, A., Münte, T. F., & Effenberg, A. O. (2013). Observation of sonified movements engages a basal ganglia frontocortical network. BMC Neuroscience, 14(1), 32. https://doi.org/10.1186/1471-2202-14-32

[21] Shea, C. H., Buchanan, J. J., & Kennedy, D. M. (2016). Perception and action influences on discrete and reciprocal bimanual coordination. Psychonomic Bulletin & Review, 23(2), 361–386. https://doi.org/10.3758/s13423-015-0915-3

[22] Shea, C. H., Wulf, G., Park, J.-H., & Gaunt, B. (2001). Effects of an auditory model on the learning of relative and absolute timing. Journal of Motor Behavior, 33(2), 127–138. https://doi.org/10.1080/00222890109603145

[23] Sigrist, R., Fox, S., Riener, R., & Wolf, P. (2016). Benefits of Crank Moment Sonification in Cycling. Procedia Engineering, 147, 513–518. https://doi.org/10.1016/j.proeng.2016.06.230

[24] Sigrist, R., Rauter, G., Marchal-Crespo, L., Riener, R., & Wolf, P. (2015). Sonification and haptic feedback in addition to visual feedback enhances complex motor task learning. Experimental Brain Research, 233(3), 909–925. https://doi.org/10.1007/s00221-014-4167-7

[25] Sigrist, R., Rauter, G., Riener, R., & Wolf, P. (2013). Augmented visual, auditory, haptic, and multimodal feedback in motor learning: A review. Psychonomic Bulletin & Review, 20(1), 21–53. https://doi.org/10.3758/s13423-012-0333-8

[26] Sisti, H. M., Geurts, M., Clerckx, R., Gooijers, J., Coxon, J. P., Heitger, M. H., … Swinnen, S. P. (2011). Testing Multiple Coordination Constraints with a Novel Bimanual Visuomotor Task. PLoS ONE, 6(8), e23619. https://doi.org/10.1371/journal.pone.0023619

[27] Swinnen, S. P. (2002). Intermanual coordination: From behavioural principles to neural-network interactions. Nature Reviews Neuroscience, 3(5), 348–359. https://doi.org/10.1038/nrn807

[28] Vinken, P. M., Kröger, D., Fehse, U., Schmitz, G., Brock, H., & Effenberg, A. O. (2013). Auditory Coding of Human Movement Kinematics. Multisensory Research, 26(6), 533–552. https://doi.org/10.1163/22134808-00002435

[29] Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York

[30] Wickham, H., François, R., Henry, L., and Müller, K. (2019). dplyr: A Grammar of Data Manipulation. R package version 0.8.1. https://CRAN.R-project.org/package=dplyr

[31] Wickham, H., & Henry, L. (2019). tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions. R package version 0.8.3. https://CRAN.R-project.org/package=tidyr

[32] Zelaznik, H. N., Spencer, R. M. C., Ivry, R. B., Baria, A., Bloom, M., Dolansky, L., … Whetter, E. (2005). Timing Variability in Circle Drawing and Tapping: Probing the Relationship Between Event and Emergent Timing. Journal of Motor Behavior, 37(5), 395–403. https://doi.org/10.3200/JMBR.37.5.395-403

Proceedings of ISon 2019, 6th Interactive Sonification Workshop, KTH Royal Institute of Technology, Sweden, November 18, 2019

SONIC FEEDBACK OF PERFORMANCE ERROR WHILE CONTROLLING A LAPTOP TOUCHPAD AS LAPTOP ORCHESTRA CHAMBER MUSIC

Michael V. Blandino

School of Music and Center for Computation and Technology
Louisiana State University
Baton Rouge, Louisiana, USA
mblandi@lsu.edu

ABSTRACT

The sonification model presented here provides an auditory representation of error when following a continuous series of targets with a trackpad. This paper includes a description of a wavetable synthesis model and consideration of competing motivations in a unified musical performance and human computer interaction experiment. The information feedback resulting from the model serves as the basis for a laptop orchestra concert performance but also supports performance improvement in an experiment that investigates human control of continuous or analog control sensors by applying information theory concepts. The extended practice and rehearsal of an ensemble performance bolsters the experimental result through increased learning and developed capability in controlling the two dimensions of the trackpad interface.

1. INTRODUCTION

Musical performance often involves goal oriented movement to achieve desired acoustic outcomes. Such goals may be those set by a composer, with some room left widely or narrowly to engage with a performer's intuition, choice, or limits of capability. Interpretation and thoughtful deviation from the prescriptive encoding of musical intentions is widely considered to be welcome and essential to the vibrancy and richness of expression in music. Nonetheless, precision and agility in control imbue performance with elements of vibrancy and richness alike. Precision in control is gained through mastery of the instrument, and high levels of information would be needed to encode a representation of the precise movements of an expert musician. In this still-early phase of composition for and design of digital musical instruments (DMI) we can engage performers in the pursuit and attainment of high precision and agility of musical control using sensing devices. Understanding the limits of control of such devices will assist in bringing the precision and agility of traditional instrument performance to the realization of sounds through DMIs alike, affording greater opportunity for expression through these media.

Continuous control sensors are commonly used in digital DMI or New Interfaces for Musical Expression (NIME) design to afford performers with gestural control of computed values for music making. Subsets of human computer interface (HCI) research involve experimentation with human subjects to identify information throughput during certain performance tasks using sensing equipment. Pointing tasks have been extensively researched, developing a literature that has established Fitts' law and extensions [1]. In contrast, fewer studies have been completed investigating continuous control tasks and pursuit tracking modes of movement [2, 3], leaving the practical — and musical — application of sensors that afford such control relatively uninformed.

When investigating the ability to perform time-series, goal gestures with continuous control sensors, quantities of training time and repetition of performance movements can emerge as determining factors in the results of accuracy measurements in cases where experimental subject participant time and incentives are significantly constrained. Given the well-established, practical design goal to tailor instrument design to novice performers [4], there could be much value in determining the performance capability of such a class of users. However, in service to a design goal that aspires to the long-term viability of a DMI[5], acquiring an understanding of well-practiced performer accuracy is desirable.

Securing additional practice time and opportunities to repeatedly perform goal gestures should reasonably be expected to improve performance results and a better understanding of human capacities to control sensors for DMI design. Further, real-time feedback beyond visual tracking during performance of the pursuit-tracking task could also improve results in accuracy measurements. Sonification of performance error as a musical performance in an ensemble setting could provide better results through corrective measures and motivations inherent to the musical performance dynamic. To support better performance and to situate this experiment in a concert performance and rehearsal context, a sonification of the experimental data of the human subject performance in real time was prepared. Research into the sonification of error for motor performance improvement has shown mixed results[6, 7], but in a musical setting and in performance as a musical task the sonic feedback should be considered more directly relevant to those trained in a chamber ensemble.

Laptop orchestras have been established in several research university music programs since their inception in 2005 with the Princeton Laptop Orchestra [8]. Experimentation in new performance structures and interactions has been a feature of this movement. However, the current study appears to be the first instance of research using a laptop orchestra piece as a human subjects experiment for better understanding human factors/HCI.

In laptop orchestra contexts, there is an established practice of composition for the laptop to be performed as a DMI[9]. The laptop offers reliable, standardized display interface components and input sensing components, including a keyboard, some form of pointing sensor apparatus, and an integrated microphone and camera. Sometimes provided with a nub-style joystick but more often recently with a touchpad, the ability to follow an intended path in two dimensions with the operating system cursor is a core element of a laptop system. Laptops in a musical performance may

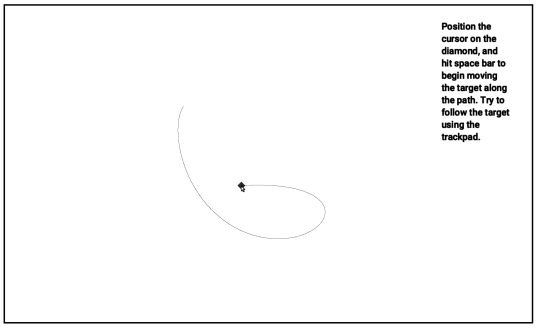


Figure 1: The presentation of a target with one second of preview in the *Pursuit Variations* performance interface. Image color inverted for visibility and converted to black and white.

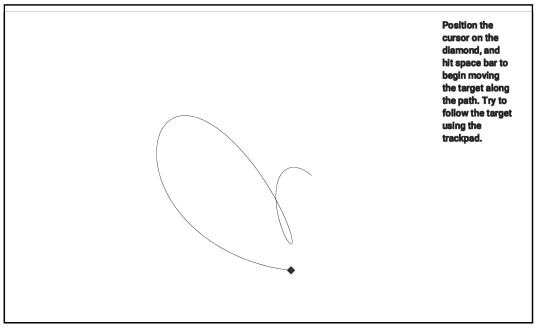


Figure 2: The presentation of a more difficult target with one second of preview in the *Pursuit Variations* performance interface. Image color inverted for visibility and converted to black and white.

be used with an audio interface and external speakers or used to control additional systems, depending on the performance setting.

Given the utility of the standard laptop interface for musical performance, an empirical understanding of the capacity for expression through the laptop input components should be gained. The trackpad in particular is a continuous control interface in two dimensions, affording two degrees of freedom of movement.

2. EXPERIMENTAL CONTEXT

To support investigation of continuous control sensors using an information theory approach, target signals of Gaussian band-limited noise are presented visually (see figure 1) to performers as a diamond shape representing the current coordinate. A two-dimensional curve showing one second of preview is displayed to show the path the diamond target shape will follow, in order to prepare the performer's pursuit tracking movement. The performer is to follow the target shape using the laptop cursor as best as they can while the system records their performance and excites the sonification model based on their error. Later analysis of the recorded performance data using the Shannon-Hartley theorem within a human-computer system model developed in prior research[10] will establish an upper bound of information communication capacity in two dimensions using the touchpad.

The experimental design and subject pool criteria were reviewed by the appropriate governing institutional review board and received approval before experimentation start.

3. COMPOSITION

Driven by the experimental design, a work *Pursuit Variations* proceeds through a succession of sixteen 20 second pursuit tracking gestures for a group of 4-8 performers. These time-series goal targets were prepared in an exponential spacing of bandwidth limits of Gaussian noise from 0.4 Hz to 5.0 Hz, progressing from least to most difficult. This signal type supports the aforementioned information theory analysis model. As a formal consideration, the pauses inherent to completing one 20 sec. gesture and moving on to another one as a group are jarring, so the performers are instructed to proceed at their own pace in continuing from gesture to gesture. This adjustment allows the work to feature continuous sound and blurs the transition from less to more difficult gestures in performance. This procedural and creative adjustment is expected to do little harm to the experimental procedure or resulting data. Rehearsal of the work is currently underway over a period of three months in weekly meetings and independent practice, with a premiere of the work scheduled for a university laptop orchestra concert in November 2019. Data collection throughout this process will be analyzed following this performance.

As an ensemble performance, the sonification of each performer's variance from the target movement is heard alongside that of the other performers. This presentation is made within the context of other performers and their respective variance. At times, the emergence of sections of higher volume resulting from performer movements away from the target of the score can resemble phrases and interactions between performer "phrases." As the piece progresses and the goal gestures require higher rates of movement, the uniformity of sound levels and characteristics increases, creating a more cohesive unity among the ensemble and its participants.

4. SONIFICATION SYNTHESIS AND AESTHETIC CONSIDERATIONS

Error as a digital media aesthetic has roots in both the digital art and computer music traditions[11]. Consistent with the research goal of this experiment and engaging with performer perceptions, the synthesis within this sonification design is intended to convey a sense of erroneous or glitch results of audio signals.

Cycling74's Max software was used to realize the sonification system. The experimental interface is drawn with Jitter from target signals generated in the *numpy* Python scripting library. The position of the cursor as controlled by the touchpad is polled for comparison to the simultaneous target coordinate.

Preliminary analysis of correlation between the horizontal and vertical performance error has shown that these values are independent of one another yet are comparable in magnitude. Using these error values independently for sonification rather than combining them into a single vector magnitude is of little difference in application and provides two values with which to drive the sonification parameters. Accordingly, two distances from the target in these dimensions are calculated, considering the target as an origin in motion.

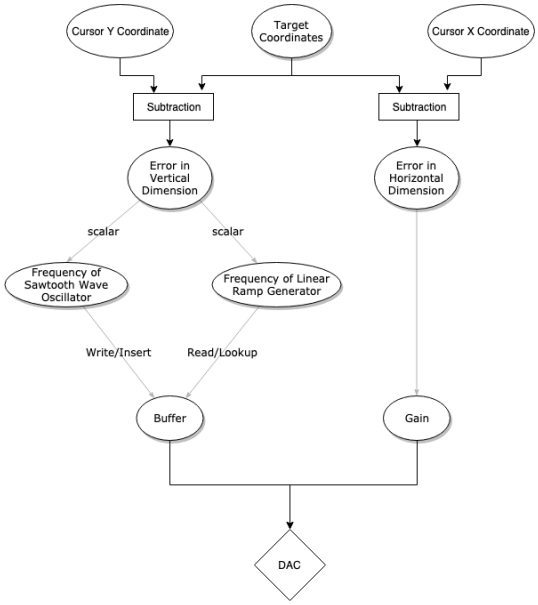


Figure 3: The sonification model, mapping error values in performance to buffer insertion and wavetable readout with gain control.

To provide a direct emergent value of error, the horizontal distance value is directly applied as a control of the amplitude of a synthesis wavetable model. A higher distance of error results in an attendant increase in amplitude level, therefore sonifying the error as a direct and notable increase in sound pressure. One can discern mistakes and corrective movements that performers enact as a result of this mapping.

The vertical distance value is scaled and applied to the frequency parameter control of a sawtooth wavetable signal generator for insertion into a blank wavetable buffer in memory. The vertical error distance value is also separately scaled and applied as a frequency parameter control of a phasor wavetable oscillator that reads through the aforementioned buffer for playback. The resulting pitch and texture of the content of the work is therefore derived directly from the performer error in this dimension. The pitch and texture are thus more complex than a primary oscillator value, adding some refinement of the discernment of movement within the amplitude value from the horizontal error.

Some fixed elements of the synthesis design are independent of the error in performance, affecting the character of the synthesis and of the piece in general. The lookup system that plays back from the written buffer according to the phasor control indexing includes a moving offset value that effectively limits the buffer size dynamically in a repeating small to large pattern. This design introduces a repeating structural pattern and also some discontinuities resulting from looped buffer playback immediately from end values to beginning values. These signal discontinuities sound like audio errors, contributing to the aesthetic of error. The sawtooth signal generator mentioned above also contributes some sonic patterning that, while affected by the performer error, is not directly attributable in its texture and characteristics to their movements. Aside from these exceptions, the sonic content of the piece is derived directly from the relationship between the performed gesture and the target signal. Indeed, if the performer perfectly matched the target signals (an impossible task no doubt), there would be no sound issued by the synthesis system.

The effectiveness of the sonification design is best exemplified by the responsiveness of the sonic interface to the error values. Sounds of similar timbral characteristics are introduced by error, but are not identical, creating a comparable but not overly repetitive or identical result. Very small values of error are noticeable and correctable through adjustment.

5. INTERACTIVITY

With such a responsive system, the participants are able to identify their own performance error in real time. The immediate feedback allows for some corrective actions to be taken. In the rehearsal context, some performers experimented with deliberate error, exciting more sound energy as a result. An inverse motivation to perform poorly could thus be identified, although the performance as an ensemble is somewhat dependent on norms of realizing composer intent. Performer/participants are also motivated by supporting a strong experimental result.

Further, the immediacy of the error sonification feedback loop provides an additional level of interest for the concert performance aspect of this project. As performers engage with the experimental target prompts in their performance, their perception of error is shared with the audience. Hearing these interactions is an important component of the compositional design.

6. AGENCY

In Pursuit Variations, the performer does not determine the intended path of their movements. In many cases, musical scores may very precisely fix certain musical parameters to realize a composition and, by extension, determine the movements of a performer to accomplish this intention. Here, the movement itself is specified without any description of or direct connection to the musical parameters other than that related to the matching of the movement and avoidance of error.

Secondary elements of motivation and attentiveness to their performance are matters of will and capacity as performers. The performer holds agency in engagement with the performance task and with the experimental outcome. Their creative agency, however, is limited by the experimental design.

If one performs less well than the other performers, there is a sense of standing out amongst the group, with possible attendant emotions of embarrassment, guilt, or anxiety. Avoiding these negative feelings and wishing to fulfill the goals of the performance are motivations for better pursuit tracking of the target. Inevitably, the more difficult targets will generate a significant volume level and texture resulting from the presence of error in the measurements.

7. INVESTIGATIVE/CREATIVE ENDEAVORS

The motivations involved in experimentation with human subjects and in musical composition and performance may differ significantly, complicating the conjoining of these activities in one project. In the case of the effort described here, several aspects of the composition design were restricted in order to preserve the integrity

of experimental research findings. The score as presented to the performers consisted of generated paths that conformed to and approach of analysis using information theory to analyze the channel capacity of a system. This limitation does not necessarily pose a conflict because such a design is consistent with the traditions of composition utilizing chance or other randomized generation processes.

Uniformity of score paths supports consistent comparison across subject performances, but prevents definition of multiple, characteristic voices and diversification across the frequency spectrum or across other parameter spaces. As mentioned above, a progression as an ensemble through 16 segments of 20 seconds each with a pause between each would be too disruptive a formal design, sounding more like an experiment than a composition. The participants are allowed to start successive segments at their own rate. Randomizing the difficulty of the segments was also explored, but the formal design and progression of the piece is better supported by a successive increase in difficulty from segment to segment.

8. CONCLUSIONS

Sonification of performance error can be designed in an aesthetic way to engage creative ends alongside scientific observation goals. The results of the experimental aspect of this study should inform understanding of practiced instrumental training and learning progress across the rehearsal and performance phases of a chamber orchestra's use of a digital musical instrument. The interactive sonification within this design is a key component of motivating performers to contribute to the forthcoming results and to engage musical practice and performance dynamics. To wit, laptop orchestras may provide a setting where performance using digital musical instruments can be investigated systematically.

9. REFERENCES

[1] R. William Soukoreff and I. Scott MacKenzie, "Towards a Standard for Pointing Device Evaluation, Perspectives on 27 Years of Fitts' Law Research in HCI," *Int. J. Hum.-Comput. Stud.*, vol. 61, no. 6, pp. 751–789, Dec. 2004.

[2] Johnny Accot and Shumin Zhai, "Beyond fitts' law: Models for trajectory-based hci tasks," in *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 1997, CHI '97, pp. 295–302, ACM.

[3] E. R. F. W. Crossman, "The information-capacity of the human motor-system in pursuit tracking," *Quarterly Journal of Experimental Psychology*, vol. 12, no. 1, pp. 01–16, 1960.

[4] Perry Cook, *2001: Principles for Designing Computer Music Controllers*, pp. 1–13, Springer International Publishing, Cham, 2017.

[5] Fabio Morreale and Andrew P. McPherson, "Design for longevity: Ongoing use of instruments from nime 2010-14," in *17th International Conference on New Interfaces for Musical Expression, NIME 2017, Aalborg University, Copenhagen, Denmark, May 15-18, 2017.*, 2017, pp. 192–197.

[6] Robert Riener, Georg Rauter, Roland Sigrist, and Peter Wolf, "Error sonification of a complex motor task.," *BIO Web of Conferences*, p. 00098, 2011.

[7] Effenberg Alfred Oliver, eFehse Ursula, eSchmitz Gerd, eKrueger Bjoern, and eMechling Heinz, "Movement sonification: Effects on motor learning beyond rhythmic adjustments.," *Frontiers in Neuroscience*, 2016.

[8] Daniel Trueman, Perry Raymond Cook, Scott Smallwood, and Ge Wang, "Plork: The princeton laptop orchestra–year 1.," 2006.

[9] Rebecca Fiebrink, Ge Wang, and Perry R. Cook, "Don't forget the laptop: Using native input capabilities for expressive musical control," in *Proceedings of the 7th International Conference on New Interfaces for Musical Expression*, New York, NY, USA, 2007, NIME '07, pp. 164–167, ACM.

[10] Michael Blandino, Edgar Berdahl, and R. William Soukoreff, "An estimation and comparison of human abilities to communicate information through pursuit tracking vs. pointing on a single axis," in *Advances in Human Error, Reliability, Resilience, and Performance*, Ronald L. Boring, Ed., Cham, 2020, pp. 247–257, Springer International Publishing.

[11] Janne Vanhanen, "Virtual sound: Examining glitch and production.," *Contemporary music review*, vol. 22, no. 4, pp. 45 – 52, 2003.