FROM VOCAL SKETCHING TO SOUND MODELS BY MEANS OF A SOUND-BASED MUSICAL TRANSCRIPTION SYSTEM

Claudio Panariello, Mattias Sköld†, Emma Frid, Roberto Bresin
Sound and Music Computing
KTH Royal Institute of Technology
†KMH Royal College of Music
{claudiop,maskold,emmafrid,roberto}@kth.se

ABSTRACT

This paper explores how notation developed for the representation of sound-based musical structures could be used for the transcription of vocal sketches representing expressive robot movements. A mime actor initially produced expressive movements which were translated to a humanoid robot. The same actor was then asked to illustrate these movements using vocal sketching. The vocal sketches were transcribed by two composers using sound-based notation. The same composers later synthesized new sonic sketches from the annotated data. Different transcriptions and synthesized versions of these were compared in order to investigate how the audible outcome changes for different transcriptions and synthesis routines. This method provides a palette of sound models suitable for the sonification of expressive body movements.

1. INTRODUCTION

In this paper we present work conducted within the scope of the SONAO project, introduced in [1]. SONAO aims to improve the comprehensibility of robot non-verbal communication (NVC) through an increased clarity of robot expressive gestures and non-verbal sounds. The purpose of the SONAO project is to incorporate movement sonification in Human Robot Interaction (HRI), i.e. to use movement sonification to produce expressive sounds. Up to this point, movement sonification has only been used to a very limited extent in social robotics (see e.g. [2, 3]). Despite the fact that sounds produced by robots can affect the interaction with humans, sound design is often an overlooked aspect in HRI. Although some research has focused on developing sounds for humanoid robots such as NAO [1] (see e.g. [4–6]), sounds used in HRI have traditionally been based on rather simple synthesis methods, or on pre-recorded samples. Design decisions as well as mapping strategies are rarely described and motivated in these contexts. Moreover, those who design the robot sounds often lack musical training.

[1] https://www.softbankrobotics.com/emea/en/nao

In the study presented in this paper, a framework for sound design in HRI is proposed, based on a work-flow starting from recordings of expressive gestures performed by a mime-actor, translated into non-linguistic sounds through vocal sketches, which in turn are annotated using a music annotation system. By incorporating composers in the design process, we hope to gain insight into how vocalizations could be used as a design material in the context of Human Robot Interaction (HRI), through translations into abstract musical representations.

2. BACKGROUND

The current study makes use of vocal sketching as a prototyping tool for exploration of sound design in HRI. Vocal sketching involves the use of the voice and body to demonstrate the relationship between actions and sonic feedback [7] and has successfully been used in a wide range of different projects, for example in SkAT-VG [8].

The notation system used for transcription in this study is part of an ongoing research project at KTH Royal Institute of Technology and KMH Royal College of Music, exploring the possibilities of representing pitch-based and sound-based music for composition [9,10]. By using notation that combines the possibilities of electroacoustic music analysis with traditional music notation, we can describe sound structures with great detail. The notation symbols were adapted from concepts and symbols by Thoresen and Hedman [11], whose notation system for music analysis combines Pierre Schaeffer's ideas on sound classification [12] with Denis Smalley's theories of spectromorphology [13]. Placing symbols, aimed for phenomenological analysis, over a fixed time-frequency grid enables the transcription and re-synthesis of sound structures. The notation system presented in [9,10] had previously been successfully tested with several students at KMH Royal College of Music, where findings suggested that different composers could synthesize very similar sonic results starting from same notation.

Up to this point, there has been relatively little research on how musical transcription could be used in the context of sonification. In particular, few attempts have aimed to merge the fields of electronic music with HRI. In seminal work by choreographer Åsa Unander-Scharin, expressive robot movements have been used for choreographing contemporary versions of classical music compositions by

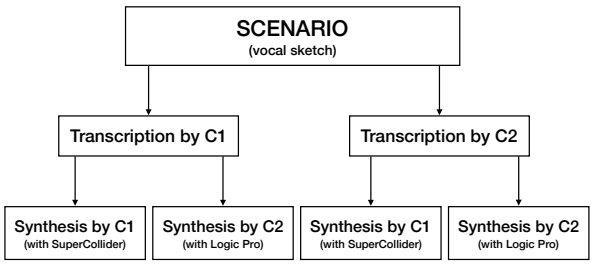

Figure 1. Flow chart of the transcription and synthesis process for composers C1 and C2, for each of the three scenarios (vocal sketches).

Monteverdi [2] and Tchaikovsky [14].

3. METHOD

3.1 Procedure

The current paper emanates from material presented in the dataset described in [15]. This dataset consists of videos, motion capture data and audio recordings of a mime-actor portraying five inner states and emotions. A subset of these videos was used in a workshop with the same mime-actor, in which he vocalized sounds associated with respective emotion (and corresponding expressive gesture). Videos of the mime actor performing the three gestures used in this study are available online [3] . An example of the mime-actor performing one of the gestures is displayed in Fig. 3. We also interviewed the mime actor about which parts of the body that were essential in the communication of the emotions through respective gestures. In the current study, a selection of recordings from this vocal sketching session was used as basis for a composition task. Vocalizations expressing the following emotions were opted for: frustrated, relaxed and sad.

Two composers, author 1 (C1) and author 2 (C2), listened to the vocal sketches and transcribed them using the notation system described in section 2. Each composer worked on the transcription independently, resulting in a total of two transcriptions per scenario. Then, all the transcriptions where used by both composers as a starting point for synthesis of new sonic sketches. Every composer produced two different sonic sketches per scenario, one for each transcription. This methodology was used to ensure that the composers did not only re-synthesize their *own* transcriptions. In the end, the number of sonic sketches was four for respective scenario, giving us a total of 12 sonic sketches. This process in outlined in Fig. 1. The final synthesized sketches were then compared in order to investigate how they were affected by the transcription and the different synthesis routine adopted by the two composers.

3.2 Material

Three of the vocalizations performed in the vocal sketching experiment described above were used in the current study: one vocalization of a frustrated gesture (called ”Scenario 1”), one vocalization of a relaxed gesture (called ”Scenario 2”), and one vocalization of a gesture going from sad to reassuring (called ”Scenario 3”). The three scenarios included the following dialogues:

Scenario 1 Actor: *”Everyone can you please line up to the left.”*

Scenario 2 Actor: *”Everyone can you please line up to the left.”*

Scenario 3 Actor: *”Sorry I broke this glass.”*
Interlocutor: *”No problem, I’ll fix it.”*

The same phrase was used for Scenario 1 and 2, however, the level of emotional expression was different for the two. Sound files are available online [4] . Spectrograms of respective vocal sketch are shown in Fig. 2.

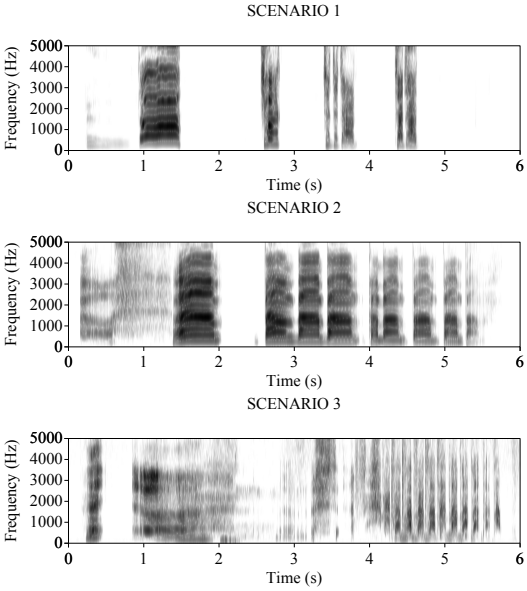

Figure 2. Spectrograms of the vocalizations for scenario 1-3.

Figure 3. Mime-actor performing a *”frustrated”* gesture.

4. RESULTS

4.1 Transcriptions

The two composers transcribed all three vocal sketches independently, resulting in a total of six scores. Comparing the two transcriptions for each scenario, we could observe that the transcriptions were similar in terms of the overall gestures, rhythm and pitch. Fig. 4 shows the two

[2] http://www.operamecatronica.com/node/1171
[3] https://kth.box.com/v/robotsonification
[4] https://kth.box.com/v/robotsonification

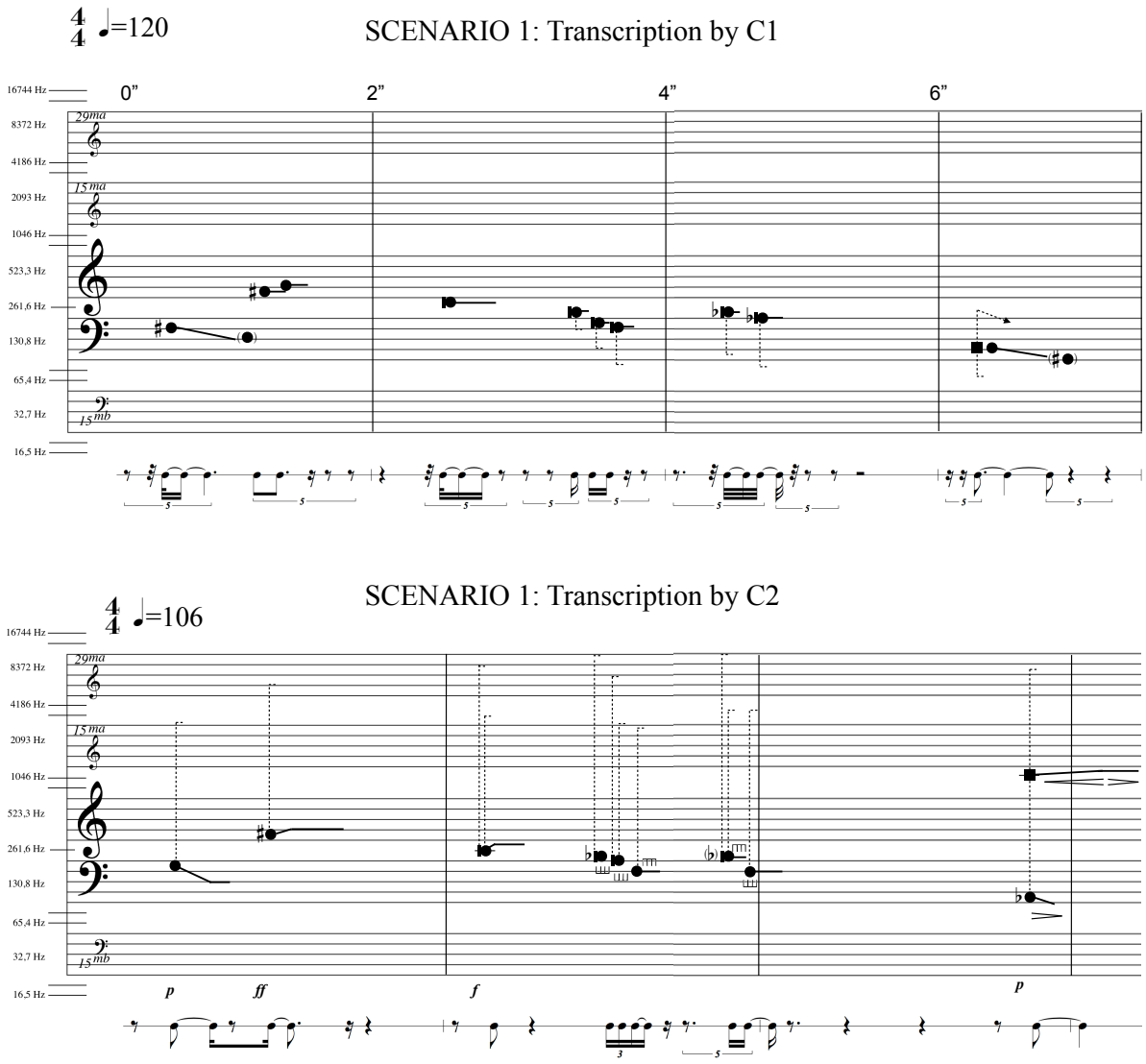

Figure 4. Transcriptions of Scenario 1 by the two composers.

analyses of Scenario 1: both composers agreed in notating the initial glissando pitched sound followed by another short glissando in the middle register and then in notating five pitched sounds with a complex onset; the two analysis both end with a short glissando in the low register overlapped with a complex sound. However, there are some differences in the notation of the spectral width. These discrepancies with regard to timbre were to be expected since notating the spectral content of a sound over a fixed time-frequency grid is not a standardized method of analysis for composers, even in the field of electroacoustic music. In so saying, the notation method leads to some approximations in the graphical representation that affects the synthesis. As a matter of fact, the two transcription of Scenario 3 are the ones that show the most significant differences, as can be seen in Fig. 5: the original vocal sketch was indeed composed by a high number of non-pitched throat sounds with a complex timbre, which are hard to notate in an univocal way. As can be noticed, there were different notation solutions for the vocal sketches' more growling sounds, where C1 choose to notate them as inharmonic sounds with diamond noteheads, while C2 used the comb-like symbol that signifies a granular energy articulation. Nevertheless, the two composers agreed on the rhythmic transcription and on the general trend of the sonic structure of Scenario 3 (starting in the middle register, going to the low, then raising to the high register and ending with an iterated sound in the low register again).

4.2 Sound Synthesis

Both composers realized sound synthesis from all six scores, resulting in a total of 12 synthesized sound files. Composer C1 realised them using only the SuperCollider programming environment [5] ; composer C2 used only Logic Pro [6] .

[5] https://supercollider.github.io/
[6] https://www.apple.com/logic-pro/

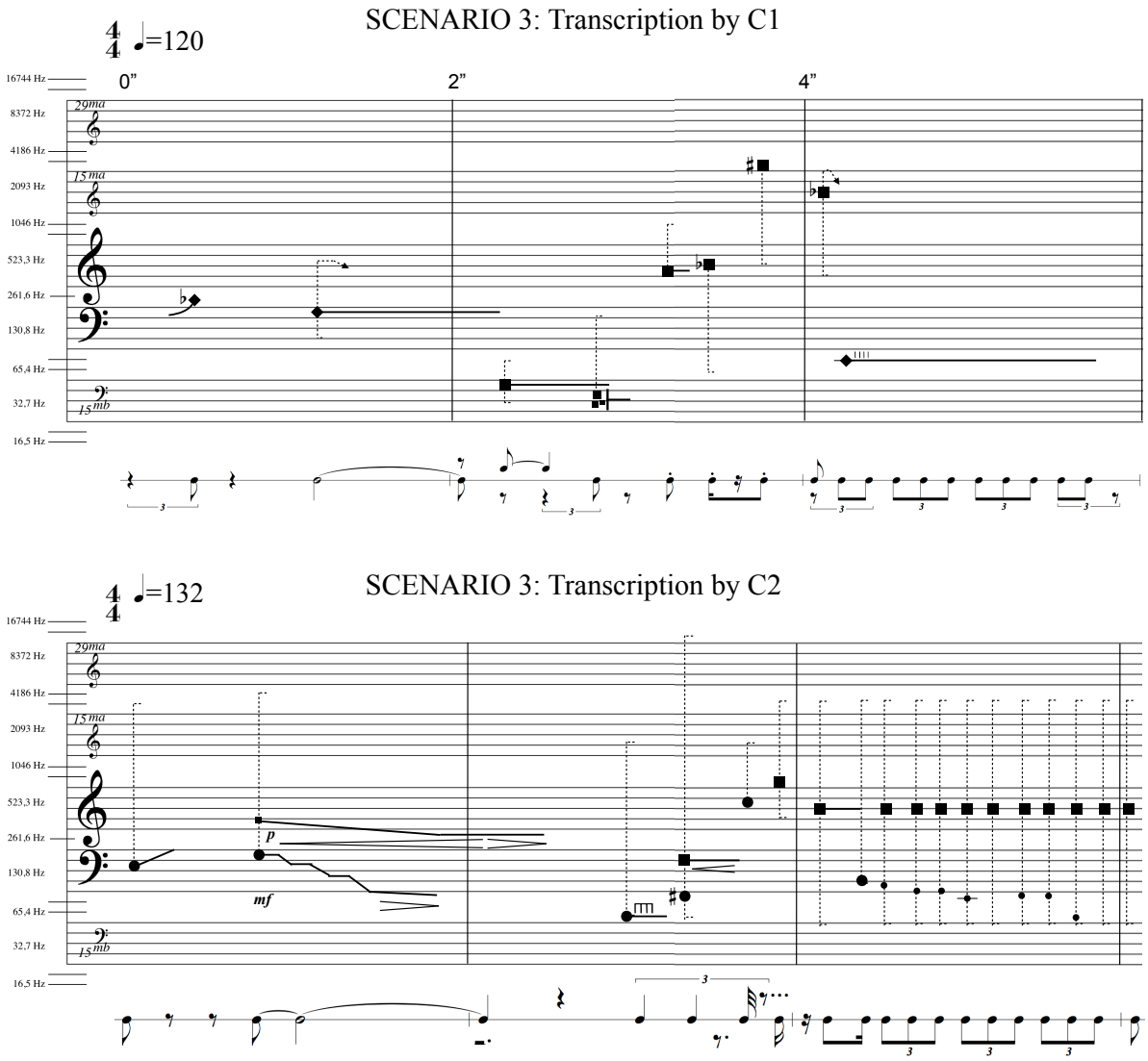

Figure 5. Transcriptions of Scenario 3 by the two composers.

4.2.1 *Synthesis with SuperCollider*

For the synthesis in SuperCollider, three main `Synths` were built in order to recreate the three main categories used in the analysis. The noise-based sound was created using different instances of subtractive synthesis; the pitched sound was realized using a filtered sawtooth; the pitched sound with inharmonic spectrum was designed using an inharmonic additive synthesis of filtered noise generators.

All the `Synths` had the possibility to be shaped with a parametric envelope and to be granularized using the `GrainIn` unit generator. The output of each `Synth` was sent into a reverberation module.

The score was then created on the client side using the `Task`, that is a pauseable process. Two arrays were initialized with the durations and the main pitches found in the analysis step, and they were used to schedule all the sound events. For each of them, one or more Synths were initialized with all the appropriate parameters. This process is summarized in the code presented in Listing 1.

The SuperCollider patches are available online [7] .

Listing 1. SuperCollider patch structure.

```
//Definition of Synths
SynthDef(\noise, {... }).send(s);
SynthDef(\pitch, {...}).send(s);
SynthDef(\dystonic, {...}).send(s);
SynthDef(\rev, {...}).send(s);
//
(
//Score
~durations = [...];
~pitch = [...];
t = Task({
        ... //Sound events//...
}).start;
)
```

[7] `https://kth.box.com/v/robotsonification`

4.2.2 Synthesis with Logic Pro

For the synthesis in Logic Pro, three instances of the ES2 virtual analog synthesizer plugin were used. The layout of the ES3 is similar to that of the Minimoog, but with some digital advantages such as 100 single-cycle waveforms for the oscillators. The sound objects of the notation were synthesized using combinations of filtered sawtooth oscillators and noise generators. For articulation and dynamics, the ES2 volume envelope was used for short durations and Logic's track volume automation was used for longer durations. For more flexible control and also automation of spectral width, separate channel EQs with low-pass and high-pass filters were added, mainly for instances playing the non-pitched noise-based sounds. Iteration and granularity were generated using LFOs controlling amplitude modulation in the ES2 modulation matrix.

4.2.3 Results

Despite the differences in choices of sound synthesis software, the produced sound files showed great similarities. Many of the vocal sketch sounds were either pitched or complex (non-pitched) sounds, which for both sets of the synthesized scores translated into filtered saw-tooth waves and filtered noise. Fig. 6 shows the original vocal sketch compared to the two sound synthesis of Scenario 1 made by C1 and C2. Moreover, there is also great compatibility between these sound synthesis and the ones the composers realized from the transcription of the other: C1's synthesis of C2's transcription is really similar to C1's synthesis of his own transcription, and vice-versa. This shows that, starting from the same transcription, the different synthesized versions sound the same, proving the effectiveness of the notation system.

Similar results could be observed for Scenario 2: the transcriptions were similar and there were no doubts in identifying the sound events as complex or pitched. The sound synthesis results were very similar as well.

Interestingly, the case of Scenario 3 was a bit different from the prior scenarios. The original vocal sketch was harder to notate in regards to the spectral content. The two transcriptions lead to synthesized sounds that barely resemble the original vocal sketch. Despite this, when the composers synthesized over the other's transcription, the results are again compatible with the previous synthesis, as expected.

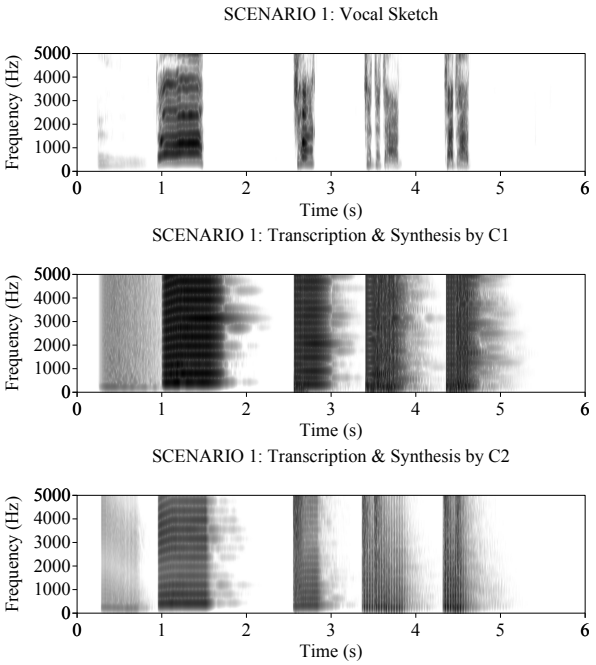

Figure 6. Spectrograms of the vocalization for Scenario 1, and synthesized versions by the two composers from their own transcriptions.

5. DISCUSSION

The challenge in transcribing non-musical sounds is similar to that of analysing electroacoustic music. One must decide what parameters to account for and with what level of detail. Clearly audible onsets of purely pitched or noisy sounds are easier to describe (as shown in the cases of Scenario 1 and Scenario 2) than intricate combinations of sound where elements of pitch and noise are intertwined and transformed over time (case of Scenario 3). Still the *"musical identity"* of the vocal sketches remained intact as they were translated into scores and back into synthesized sound. This was also noted when the same notation system was used for the interpretation of musical structures [10]. Indeed the synthesized versions of each scenario made by the two composers were judged to be perceptually very similar in informal listening tests made by expert listeners at KTH. Still, there are discrepancies between the vocal sketches and the synthesized versions. There are two ways of dealing with them: one is to aim for transcriptions with much greater detail in an attempt to capture the voice more fully, the other is to think of the notation's function as the preserver of a sound structure's basic identity and consider some features of the vocal sketch the interpretations of the sound structure itself.

The new method presented in this paper, based on the sonic rendering of transcriptions of vocal sketches of body movements, provides a palette of sound models suitable for the sonification of expressive body movements. In particular, in the framework of the SONAO project [1] we are interested in identifying a set of sound models which can be used as a starting point for the design of real-time sonic representation of humanoid-robots expressive movements.

6. CONCLUSIONS

We have showed how notation developed for sound-based musical structures can be used for representing vocal sketches depicting robot movements. Traditional music notation will typically capture the fundamental sound structure of the music, leaving interpretation and emotional expression to the performer. Similarly, what constitutes a sad vocal sound structure will not necessarily translate into a sad synthesized version of its score. This depends on what vocal features that were used to convey the feeling and with what level of detail the sound passage

was notated. However, using notated sound structures as blueprints for sonified movements is conceptually different from other forms of sonification in that the movements are not directly sonified, but connected to notated structures in the form of a scores to be interpreted. This way of working opens a space for the sonic interpretation of the movements where certain structural relations between specific movements and their sounding counterparts remain the same while other features are interpreted depending on the context.

7. FUTURE WORK

The study presented in this paper will be followed by formal and extensive listening experiments focusing on the perceptual distance between vocal sketches and their synthesis. Some possible applications and future work are described below.

7.1 Sonification of Robot Gestures

During the interview with the mime-actor, he emphasized that the following parts of the body were important in the communication of the sad gesture in Scenario 3 were the hands, and possibly also the shoulders. For the frustrated and relaxed gestures in Scenario 1 and 2, he also mentioned that the hands should be emphasized. This connection between the body movement and the vocal sketch will be used in a future stage of the project: having all the MoCap data of the mime gestures, it will be possible to use them to control the sound synthesis, i.e. sonification, focusing on the parts of the body that was indicated by the mime actor himself as being the most important ones.

7.2 The Notation of Movement and Sound

Expanding on the possibilities of notation with regard to a robot's expressive movements and sounds, there is the possibility of also notating the movements, placing both gesture and sound on the same conceptual level. There is a rich tradition of notating both movement and sound in dance, and notation systems like Labanotation [16], often used for notating dance movements, have already been used in the design of movement-based interaction [17] and in interactive dance performances (see for example recent works by Daniel Zea [8]).

Acknowledgments

The authors would like to thank Simon Alexanderson and Alejandro Bonnet for their valuable contributions to the SONAO project. We also thank the three anonymous reviewers for very helpful comments that contributed to improve the quality of our paper. This project was funded by Grant 2017-03979 from the Swedish Research Council and by NordForsk's Nordic University Hub "Nordic Sound and Music Computing Network - NordicSMC", project number 86892.

[8] http://danielzea.org/works/

8. REFERENCES

[1] E. Frid, R. Bresin, and S. Alexanderson, "Perception of Mechanical Sounds Inherent to Expressive Gestures of a NAO Robot-Implications for Movement Sonification of Humanoids," in *Proceedings of the Sound and Music Computing Conference*, 2018.

[2] R. Zhang, M. Jeon, C. H. Park, and A. Howard, "Robotic sonification for promoting emotional and social interactions of children with ASD," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*. ACM, 2015, pp. 111–112.

[3] J. Bellona, L. Bai, L. Dahl, and A. LaViers, "Empirically Informed Sound Synthesis Application for Enhancing the Perception of Expressive Robotic Movement," in *Proceedings of the International Conference on Auditory Display*. Georgia Institute of Technology, 2017.

[4] J. Monceaux, J. Becker, C. Boudier, and A. Mazel, "First Steps in Emotional Expression of the Humanoid Robot NAO," in *Proceedings of the 2009 International Conference on Multimodal Interfaces*. ACM, 2009, pp. 235–236.

[5] M. Häring, N. Bee, and E. André, "Creation and Evaluation of Emotion Expression with Body Movement, Sound and Eye Color for Humanoid Robots," in *RoMan, 2011 Ieee*. IEEE, 2011, pp. 204–209.

[6] A. Kappas, D. Küster, P. Dente, and C. Basedow, "Simply the BEST! Creation and Validation of the Bremen Emotional Sounds Toolkit," in *International Convention of Psychological Science*, 2015.

[7] D. Rocchesso, S. Serafin, and M. Rinott, "Pedagogical approaches and methods," *Sonic Interaction Design*, pp. 125–150, 2013.

[8] D. Rocchesso, G. Lemaitre, P. Susini, S. Ternström, and P. Boussard, "Sketching sound with voice and gesture," *interactions*, vol. 22, no. 1, pp. 38–41, 2015.

[9] 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.

[10] Sköld, M., "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.

[11] 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.

[12] P. Schaeffer, *Treatise on Musical Objects: An Essay Across Disciplines*. Univ of California Press, 2017, vol. 20.

[13] D. Smalley, "Spectromorphology: Explaining Sound-Shapes," *Organised Sound*, vol. 2, no. 2, pp. 107–126, 1997.

[14] Å. Unander-Scharin, "Activity: La Robot-Cygne : Choreographic Reflections on Dancing Through a Mechatronical Double," 2009, startdatum: 24/09/2009; Slutdatum: 24/09/2009; Roll: Föreläsare; Typ: Föreläsning / muntligt bidrag.

[15] S. Alexanderson, C. O'sullivan, M. Neff, and J. Beskow, "Mimebot — Investigating the Expressibility of Non-Verbal Communication Across Agent Embodiments," *ACM Transactions on Applied Perception (TAP)*, vol. 14, no. 4, p. 24, 2017.

[16] A. H. Guest, *Labanotation: the System of Analyzing and Recording Movement*. Routledge, 2013.

[17] L. Loke, A. T. Larssen, and T. Robertson, "Labanotation for Design of Movement-Based Interaction," in *Proceedings of the second Australasian conference on Interactive entertainment*. Creativity & Cognition Studios Press, 2005, pp. 113–120.