2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)

Personalized Synthesis of Intentional and Emotional Non-Verbal Sounds for Social Robots

Hannes Ritschel, Ilhan Aslan, Silvan Mertes, Andreas Seiderer and Elisabeth André
Human-Centered Multimedia, Augsburg University
Augsburg, Germany
{ritschel, aslan, seiderer, andre}@hcm-lab.de

Abstract—**Non-verbal sounds are an essential communication channel for social robots. However, it requires expert knowledge to create and compose synthesizers, develop melodic structures or record samples which express a robot's internal intentions and emotions. This paper presents an approach for adapting a robot's timbre based on non-expert human comparative feedback in order to personalize the sonic interaction design to an individual user's preferences. An evolution strategy learns parameters of real-time sound synthesis for different intentions and emotions. Ultimately, the strategy aims to improve the perceived goodness of how well a specific melody's sound maps to a specific emotion or intention. In order to demonstrate the feasibility of the approach, we report on a user study with a robot, 6 exemplary melodies and 27 participants. Our study results show that the strategy indeed results in improved and preferred sound designs and that many participants are willing to apply such a process to improve their robots' expressivity.**

Index Terms—**sound synthesis, non-verbal sounds, emotions, intents, adaptation, social robots, appearance-constrained robots**

I. Introduction

Social robots mimic human behaviors in order to communicate their intentions and emotions. However, only a handful of robots offer all of natural language, facial expression, gestures and body movement. In fact, colored light, motion, and sound are the most frequently available output modalities of commercially available robots and research prototypes [1]. Many low-cost and consumer robots are appearance-constrained: their embodiment is limited so that they are not able to communicate with gestures or facial expression [2]. Thus, non-verbal sounds [3] are even more important for communicating the robot's affective and internal state effectively. Since non-verbal sounds benefit of being independent of any particular language, they are used in Human-Computer Interaction (HCI) and Human-Robot Interaction (HRI) to express messages in a short time [3]. Moreover, they also shape the perceived robot's personality, which is important in order to make interaction more interesting and desirable, as well as to establish a relationship between user and robot in the long run [4].

Since the expression of intentions and emotions is an important aspect to make robots appear socially intelligent research already investigated how to communicate them with multimodal cues. Besides facial expression, posture, motion and light, sound is of central importance. Research primarily focuses on imitating human, animal or natural sounds, such as imitation of speech, crying or thunder. However, Bethel et al. [2] point out that the sonic interaction design in HRI should use robot-specific social cues, which do not anthropomorphize through human or animal sounds and tones. Thus, some experiments also investigate how to express emotions while relying on established concepts of music theory [5], such as tempo and intonation, to make a robot express its internal state.

Usually, sounds are prefabricated: professional sound designers record and process samples of real-world sounds or compose characteristic melodies with natural or synthetic instruments. Once created, these sounds cannot easily be adapted or modified during the interaction. In contrast, real-time sound synthesis describes and produces sounds based on functions and parameters. In HRI, this approach has only been used in recent years. Luengo et al. [3] point out that a robot's sonic design and synthesis should be adaptable online to make it more expressive, e.g. to synchronize sounds with the robot's motion. In the long run, this will also help to address the initial novelty effect by varying sounds to a certain degree instead of replaying the same sound samples again and again.

Furthermore, generating sounds during runtime enables the personalization of a robot's non-verbal sounds to a user's individual taste. This allows for fine-grained adaptation far beyond a discrete set of voices (e.g., male and female) or prefabricated samples, similarly to spoken language, where Natural Language Generation can be used to adapt the robot's linguistic style accordingly [6]–[12].

Therefore, we propose an approach based on an Evolution Strategy (ES) and human comparative feedback to interactively shape the timbre of exemplary non-verbal emotional and intentional expressions of a robot. We report in detail on a study with 27 users demonstrating for example, that the process results in preferred robot sounds, that users are willing to apply such a process, and what musical characteristics appear in the adapted sounds.

II. Related Work

Different cues are used in the literature to express emotions and intentions. Some are directly related to musical structures and properties (e.g. intonation, tempo) [1], [3], [5], [13], [14], others use human, animal, natural or artificial sounds (e.g. imitation of speech, crying, thunder, sci-fi sounds) [15]–[17]. Technically speaking, there are two basic options for sonic design of a robot's non-verbal sounds: either by using prerecorded or preprocessed audio samples [14]–[16], [18]–[20]

978-1-7281-3888-6/19/$31.00 ©2019 IEEE

or by generating or manipulating samples in real-time, ranging from simple beep sounds to complex synthesizers [3], [17].

A. The sound of R2-D2 and Wall-E

Jee et al. [13] focus on communication via musical themes for socially interactive robots. They analyze in detail the sounds of *R2-D2* and *Wall-E*: intonation, pitch and timbre are identified as dominant musical parameters for expressing intentions and emotions as universally as possible. Based on these findings, the authors produce musical scores for their *Silbot* English teaching robot for two emotions (happiness, sadness) and five different intentions (self-introduction, affirmation, denial, encouragement, question). Results of their experiments show that the sounds are sufficient for communicating intended emotional and intentional expressions.

B. Sound and Multimodal Cues for Communicating Emotions

Sounds are often explored in combination with additional non-verbal social behaviors. Scheeff et al. [15] present a robot which expresses emotions with facial expression, posture, motion and sound. The robot mimics adult Peanuts-like speech for affective speech output depending on its emotional state (neutral, happy, sad, angry, surprised, fearful, inquisitive, nervous, and sleepy). Its vocalizations are described as “muffled speech”, which is generated based on filtered affective human speech samples. The authors observe that subjects did not realize that the speech content was nonsensical by design; they tried to understand, talk to and answer the robot in “pet speech”. According to the authors, the mixture of actual human speech and synthetic beeps and chirps confused people.

Song and Yamada [14] use colors (white, green, blue, red), beep sounds and vibrations with different intensity for expressing four emotions based on the circumplex model of affect [21] with their appearance-constrained *Maru* robot. Different volumes are used for the *relaxed* and *happy* emotion, *sadness* and *anger* are expressed with falling, or respectively raising beep sounds. The authors identify a set of appropriate combinations based on whether participants mapped the presented cues to the intended emotions correctly. While there is no sufficient expression for the happy emotion and the relaxed one only includes color, both sadness and anger also include the respective beep sounds, indicating that they are important and contribute to communicating the robot’s emotion.

Häring et al. [16] explore eight different expressional designs for *anger*, *sadness*, *fear* and *joy* with the *NAO* robot. Multimodal cues include body movement, sounds and colors emitted from the robot’s eye LEDs. In contrast to Song and Yamada, sounds are composed from individual sounds. For example, anger is accentuated with gibberish voice, a “Bleep” sound (as used in television when censoring words), a rolling thunder, or growling like an angry dog. Fear is expressed with a metallic bang or a modulated sound that gets louder. Different types of crying are used for sadness while the robot uses language via the text-to-speech module (“Jippie Yay!”) or a sampled cowboy’s voice (“Yehaa”) to express joy. The evaluation of single cues identifies body movements and half of the sounds as appropriate, but eye colors as unreliable with respect to emotional expression. For example, the crying sound was identified as too extreme, the gibberish voice was interpreted as not serious enough and the metallic bang alone had not much expressivity. In addition, the text-to-speech module was perceived as monotonous.

Similar research has also been conducted recently by Löffler et al. [1] with respect to color, motion and sound of an appearance-constrained robot. They use tone patterns to express emotions with differently pitched beep sounds based on sine waveforms and different tempo. Joy (high pitch, longer inter-beep-interval), sadness (“slow” sounds, falling tone), fear (high pitch, alternating tone, short inter-beep-interval) and anger (low pitch, short inter-beep-interval) were evaluated together with the other modalities. The authors come to the conclusion that joy could not be communicated with sound easily. Apart from this, sound was very important in combination with the other modalities for the other emotions. Moreover, it was the main driver for communicating sadness.

C. Affective Sound Synthesis

In contrast to the formerly outlined research, which primarily relies on pre-recorded samples or beep sounds, this section focuses on real-time synthesis approaches in HRI, which have only been used in recent years and are by no means standard.

Schwenk and Arras [17] present a sound system which is able to produce an “expressive synthetic human voice of a humanoid robot”. It is based on vowels and consonants, which can be combined to words and sentences. Based on the robot’s perception, motion planning, emotional and inner states, it modulates synthesis parameters, such as oscillator frequencies, amplitudes or filter intensities, to generate abstracted robot-specific sounds and tones. The system is used to communicate happiness (high pitch, large pitch variations, fast changing frequencies, high loudness, chirping sounds), sadness (low speech rate, loundess and pitch, minimal pitch variation), fear (short-duration A-vowel phonemes, dog-like whining sound), embarrassment (O-like sound), disappointment (low pitch, decreasing intonation, O-vowels for imitating “no”) and curiosity (high pitch, increasing intonation, randomized phonemes for imitating short utterances). Sounds are synchronized in real-time with posture, motion and LED colors.

Luengo et al. [3] present the *sonic expression system*, a synthesizer system, which is able to express intentions (approval, rejection, hesitation, greeting), affection (joy, calmness, sadness), human nature sounds (laughter, weeping, coughing, yawning, heartbeat) and narrative communication messages (depending on recent occurences or events). The minimal basis are so-called *quasons*, which represent distinguishable sound units with a set of acoustic features, including amplitude (envelope, volume), frequency (envelope, timbre) and time (duration). *Sonic utterances* can be created by combining different quasons. They also have features for amplitude (volume, dynamics), frequency (melody, harmony, texture) and time (tempo, rhythm, articulation). Results of the evaluation with

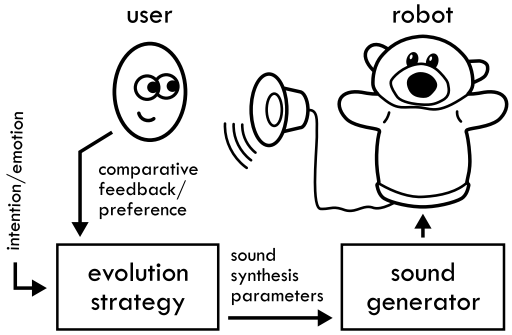

Fig. 1: Overview of the non-verbal sound generation and personalization process.

the *Maggie*, *Mini* and *Mbot* robots show different degrees of recognizability for the generated non-verbal sounds.

So far, previous research has primarily handcrafted robots' non-verbal sounds in advance. To our best knowledge, this paper is the first which explores how to adapt them to the individual user's preferences based on explicit, comparative feedback by non-experts. While we build on the melodies by Jee et al. [13] for expressing intentions and emotions, we are interested in personalizing the robot's timbre with real-time sound synthesis in order to have control over the sound generation process during runtime. We do not use human, animal or natural sounds in order to not anthropomorphize the robot's sounds [2]. Moreover, manipulating this kind of samples could result in odd sounds, as indicated by [15].

III. Personalized Non-Verbal Sound Synthesis

In order to explore the personalization of a robot's non-verbal sounds, we combine an ES and human comparative feedback to interactively shape a robot's timbre. Figure 1 illustrates the general approach: a human listens to an appearance-constrained robot, which plays back generated sounds. The melodies proposed by Jee et al. [13] are used to express exemplary emotions (*happiness* and *sadness*) and intentions (*affirmation*, *denial*, *encouragement*, and *question*). We did not include *introduction* since we are interested especially in more frequent and recurring non-verbal sounds.

The user repeatedly listens to pairs of generated sounds with the same intention or emotion. After she or he selects the preferred sound version, an ES mutates parameters for sound synthesis and thus optimizes the sound for the corresponding intention/emotion. This process is repeated several times for each of the robot's intentions and emotions to learn about the individual user's preferences.

A. Sound Synthesis Parameters

The melodies by Jee et al. [13] with up to four voices (soprano, alto, tenor, bass) serve as tonal basis. When playing back a melody, each voice uses a synthesizer with six continuous parameters. Two parameters control the amplitude of a sine wave (soft timbre) and a saw tooth oscillator (sharp timbre) while another two parameters control the sustain level (volume while holding a note after pressing it) and release time (how fast the note fades out after releasing it). The last two parameters control the *low pass* and *high pass* filter for cutting off high or low frequencies. By mixing these six parameters, a large range of sounds with varying volume can be produced, ranging from very soft to very sharp timbre, short to long-lasting notes, muffled to treble sounds, as well as combinations of them. Additionally, there is one global parameter to adjust the melody's tempo. Table I plots the value ranges of all parameters.

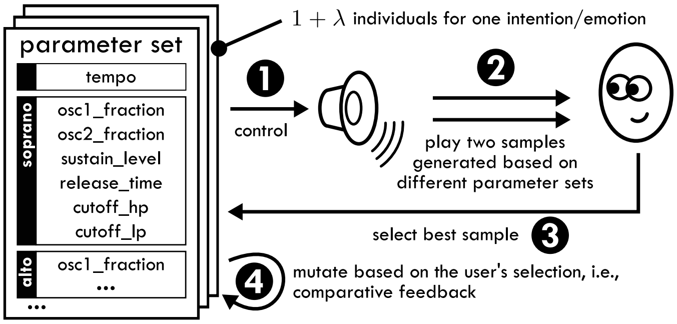

Fig. 2: In each generation of the ES, the population of each of the robot's intentions and emotions contains several *parameter sets*. The parameters of each set define the melody's tempo and the timbre of each voice.

TABLE I: Value ranges of the sound synthesis parameters

Parameter	Description	Value range
osc1_fraction	saw tooth oscillator amplitude	[0.0001, 2]
osc2_fraction	sine wave oscillator amplitude	[0.0001, 2]
sustain_level	volume while holding a note	[0.0001, 1]
release_time	note fade out after releasing it	[0.0001, 1] sec
cutoff_hp	high pass filter cutoff	[1, 500] Hz
cutoff_lp	low pass filter cutoff	[600, 9000] Hz
tempo	tempo for playing back the melodies	[30, 240] bpm

While the melody of each intention and emotion is fixed, the timbre and tempo of the resulting sound is defined by a *parameter set*. The parameter set of each melody contains the tempo parameter and the six sound synthesis parameters for all voices (see Figure 2), resulting in $1 + c_v \cdot 6$ parameters where c_v is the melody's voice count. For example, a parameter set for sadness has 13 parameters since its melody contains only two voices (see Figure 3).

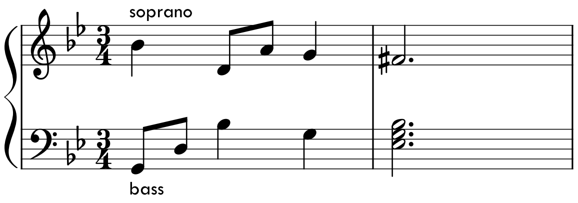

Fig. 3: The musical score for sadness [13] with two voices.

B. Evolution Strategy

The adaptation of the robot's timbre is realized with an $(1 + \lambda)$-ES [22], where λ denotes the offspring size. For each intention/emotion, a complete population consists of $1 + \lambda$ individuals (parameter sets) that are evaluated to find the parent of the new generation, which again is expanded with λ new slightly mutated versions of itself. A similar approach has formerly been used by Mitchell and Pipe [23], who could achieve sufficient results in optimizing a set of real-valued Frequency Modulation (FM) parameters for sound generation with a $(\mu/\rho, \lambda)$-ES that uses analogical principles as the $(1 + \lambda)$-ES. We did not use the $(\mu/\rho, \lambda)$-ES because the selection of multiple parents in order to allow crossover-operations would desire much more user feedback.

At the beginning, all parameter sets of all intentions and emotions are initialized randomly, respecting the corresponding value ranges (see Table I). The user replaces the need for an explicit fitness function by selecting the preferred sound samples. Therefore, the fitness function is the result of the human comparative feedback: the preferred parameter set is the fittest. After every comparison, the preferred parameter set is kept in the population, the other one is removed. For the experiment, we set $\lambda = 4$: after four comparisons, only the fittest one is left. In each iteration, the fittest parameter set of one population is mutated. All parameters in the set are mutated by means of a Gaussian distribution, which uses the old parameter value as mean. The mutation is constrained by the parameters' upper and lower bounds.

Since the human acts as fitness function and needs to compare many samples by hand, a tradeoff between training time (to not overwhelm or tire the user), a sufficiently big population (to explore the search space of the parameter sets) and generations (many iterations are necessary to achieve best results) had to be found. In order to meet these requirements, we chose an iteration size of 8, i.e. 8 consecutive generations per intention/emotion were generated. By assuming that a single comparison of two different sounds lasts approximately 10 seconds, this choice allowed us to accomplish the training process in roughly 30 minutes per user.

Common methods for adapting the mutation strength (i.e. the variance of the Gaussian distribution) rely on much more iterations [24]. Due to the small number of 8 generations we decided to choose a constant value for each intention and emotion which we found by experimental tests. Thus the variance for every parameter p is calculated as follows, where σ_p^2 denotes the variance of p, l_p the lower bound of p and u_p the upper bound of p: $\sigma_p^2 = \frac{|u_p - l_p|}{5}$. The calculated values are an adequate compromise to equip the algorithm with the ability to explore the search space properly without resulting in extensive overshooting of the target values.

C. Hardware and Software Setup

The setup (see Figure 4) consists of (1) BärBot, the appearance-constrained robot, (2) a loudspeaker[1] with a USB powered PAM8403 amplifier, placed inside the robot, and (3) a computer for audio synthesis and processing the user's feedback. Similar to [17], sound synthesis is implemented with SuperCollider[2] and the Open Sound Control (OSC)[3] protocol for communicating with the adaptation process.

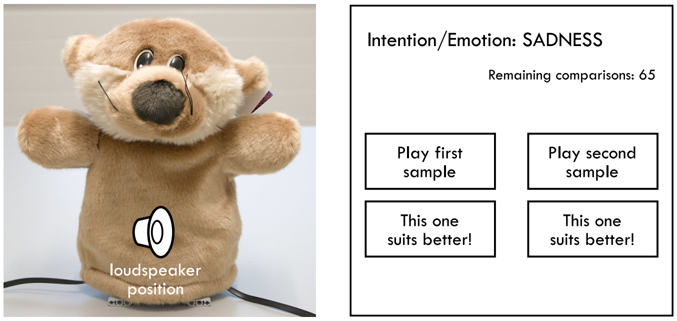

Fig. 4: Robot and graphical user interface schematic for playing back generated sounds and giving feedback.

Fig. 5: Study setup: a participant listens to the robot's generated sounds and gives feedback to the adaptation process.

IV. Study

A. Participants, Apparatus and Procedure

27 participants (12 female, 15 male), aged from 19 to 83 ($M = 30.5$, $SD = 14.4$), were recruited for the evaluation. We were specifically interested in whether the proposed adaptation approach would succeed in helping subjects to adapt the robot's timbre to their own preferences and whether the interactions would result in improvements. The participants were informed that they should repeatedly listen to two sound versions of a melody and choose the preferred one considering the intention/emotion the melody was designed for.

After collecting demographic data, including preferred music genres and previous knowledge or experience regarding robots and synthesizers, the interaction with the robot started. The robot was placed in front of them together with the graphical user interface (see Figure 5).

Each participant started with randomly initialized parameter sets for each of the robot's intentions and emotions. In total, each subject had to listen to 192 generated sound pairs:

[1] http://www.visaton.de/en/products/fullrange-systems/fr-7-4-ohm

[2] https://supercollider.github.io/

[3] http://opensoundcontrol.org/

the evolution strategy required 8 iterations à 4 comparisons for each of the 6 intentions/emotions, which were presented consecutively, but in random order. Participants could play back the samples as often as they wanted.

The final parameter sets were saved after collecting the user's preferences. Subjects had to rate on a five-point Likert scale how well the timbre, as well as whether the melody maps to the corresponding intention/emotion generated with the final parameter set. Moreover, the sound of each intention/emotion was presented again both generated with the initial (random) parameter set, as well as with the final, adapted parameter set. Both versions were played back in random order and subjects had to select the preferred one. We did this to measure if the approach resulted indeed in preferred sounds.

B. Results

1) User preferences and comments: Figure 7a depicts how well the final adapted sounds mapped to target intentions/emotions. Considering Figure 7a a paired t-test revealed that the adaption process resulted in a significant difference between how well the adapted sound of the melody and the melody itself maps to the target emotion/intend for encouragement ($t = 2.8$, $p = 0.008$, $r = 0.40$) and affirmation ($t = 2.7$, $p = 0.012$, $r = 0.48$). The difference for the other four melodies was not statistically significant. Interestingly the result for denial seems inverse, which could mean that the process works better for melodies that sound positive, such as an affirmation or an encouragement melody. An intriguing implication could be that users may prefer (implicitly, knowing that their goal is to find a mapping sound) sounds that sound good, and thus, such adaptation processes will inevitably result in designs that map less good to negative target intentions/emotions, but will still be preferred by users. Figure 7b shows that participants have over proportionally (i.e., 86 %) preferred the sound of each melody with the adapted parameters. All but two participants out of 27 preferred (out of the 6 melodies) the adapted versions in majority (i.e., more than 3). Both other participants stated that they did to not enjoy the process itself. Consequently, for 25 of 27 participants the adaptation process indeed resulted in improvements. They also reported on a five point Likert scale their willingness to apply the process in order to adapt their robots' sounds ($M = 3.7$, 95 % CI of 0.8), and how much they enjoyed the process ($M = 3.3$, 95 % CI of 0.8). Both values indicate that participants would want to apply such a process.

Indeed, only one person mentioned that a robot's timbre is not important, three participants explicitly stated that personalization of robots is important and that robots should sound according to an individual's preferences. The main shortcoming was the duration of the interaction: five subjects reported that it took too much time (car or train journeys were suggested at good times to apply the process) and that over time there could be difficulties to keep comparing the sounds. Two participants with background knowledge in music production mentioned that manual control over individual parameters would also be beneficial, but that the automatism allowed to explore new timbres without knowing the corresponding parameter values. One of them stated that the preferred timbre could be achieved based on the feedback if one had an idea of the desired final timbre already in the beginning, which suggests that expert participants felt control and is in line with our own experience using the approach. Five participants mentioned that the melodies were pleasant, that comparisons were simple and efficient, that it was interesting to hear the differences, and that finding the preferred timbre was easy. One person

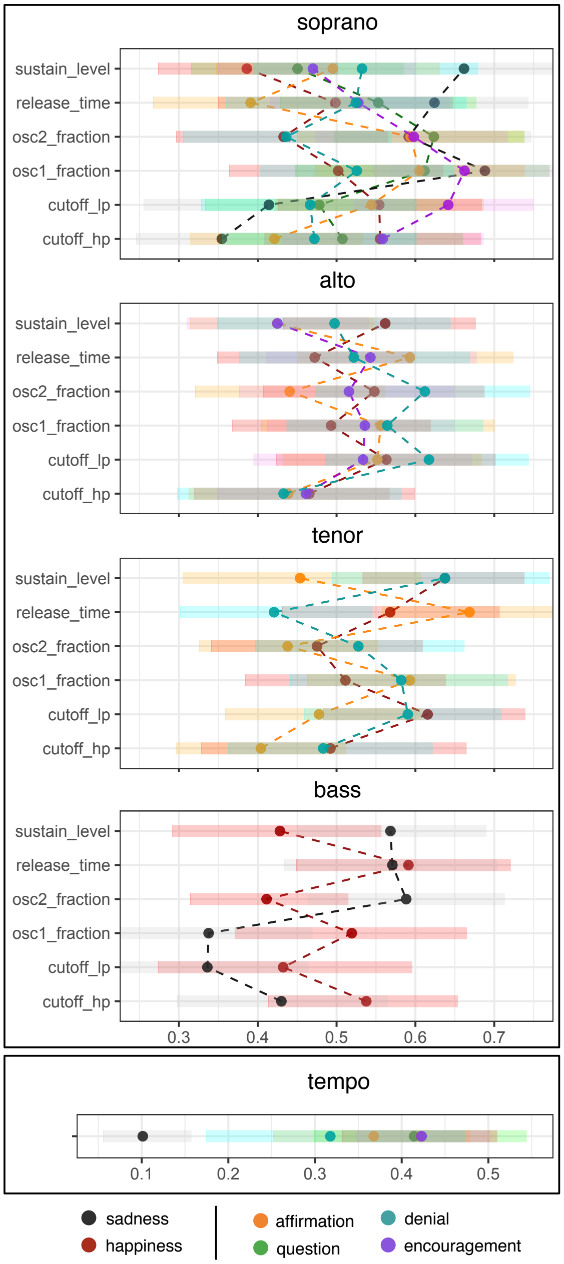

Fig. 6: Final parameter sets (averaged over all users) for each voice of the robot's intentions and emotions. Parameter values are normalized according to their value ranges (see Table I). Error bars denote 95 % CI.

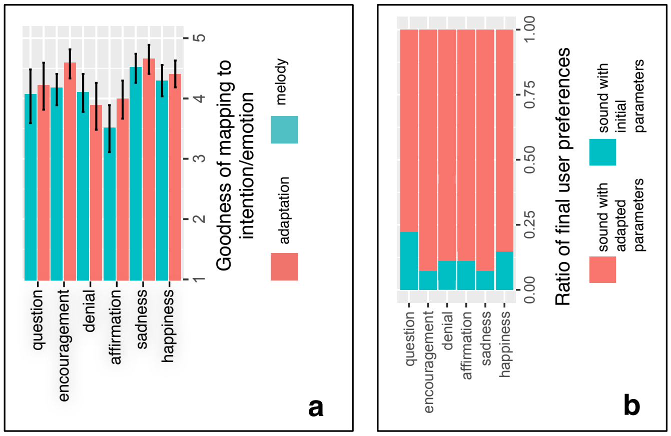

Fig. 7: a) Goodness of how well each melody and associated final adapted sound map to emotion/intention, and b) Ratio of final user preferences over all participants and melodies.

also mentioned enjoying the adaptation process itself with the different emerging sounds. Reasons for using this approach also included the fact that "it worked", was playful, sounded well, that a learning effect was noticeable, and that the final timbre matched the intended intentions and emotions better.

2) Synthesizer parameters: In Figure 6 we report the adapted, final parameter sets for all melodies and voices, excluding the data (14 %) of participants who did not prefer the adapted sounds. Overall, we observed variation in participants' preferences, which was also visible in the distribution of the final parameters. One possible reason could be that the participants in our sample had divers music genre preferences, which we collected as part of the demographic data.

However, a general trend for certain aspects can be observed from the averaged data. For example, it demonstrates that a distinctive characteristic for expressing sadness is slow tempo, which is a common stylistic instrument in music composition [5]. Furthermore, the average timbre of happiness and sadness shows very different characteristics. When looking at the bass voice, the amplitude of the saw tooth oscillator is much smaller than the sine wave, resulting in a softer timbre for sadness. This is reinforced by the lower cutoff of the high pass and low pass filter, which also applies to the soprano. These two aspects were also noticed by one subject with previous knowledge in music production, who stated that tempo and sharpness of the sound were the most prominent and noticeable features. Happiness has a smaller sustain level, resulting in shorter notes, which is particularly noticeable in the soprano. All in all, these average tendencies reflect typical stylistic instruments in music composition, which seem to be reflected in both experts' and non-experts' subjective feedback for the expression of emotion, too.

C. Participants' Feedback

Final open questions addressed what participants liked and disliked about the robot's non-verbal sounds and adaptation. Seven participants disliked the robot's/synthesizer's timbre (e.g. described as "synthetic", "artifical" or "metallic"), one of them could not associate them with the robot. Disliked aspects also included the duration/monotony of the interaction (mentioned by three subjects), the lack of different melodies (noted by two subjects) and one participant disliked the melody for expressing happiness. Two participants mentioned that it was sometimes difficult to compare samples which sounded nearly identical. One participant with background knowledge in music production would have liked more control over tempo and sharpness of the sound, another two pointed out that sometimes there was also a change for the worse.

The general idea was received very positively by six subjects, including the adaptation of non-verbal sounds based on subjective likings, the expression of emotions with melodies and sounds. Another six participants stated that they liked the melodies; the melody for encouragement was highlighted by two subjects. One also mentioned that listening to this melody while looking at the robot induced a positive feeling, one described it as "stimulating", and another one associated denial with animal cheeping sounds. Six participants pointed out that the robot's timbre adapted to their own preferences, which was noticed e.g. with regard to tempo, the fact that sounds became more euphonious gradually towards the end, and that the robot was "aware of what you like and dislike". Four subjects mentioned that the system was easy to use. Two participants described the robot's appearance as "cute".

V. Conclusion

In order to effectively communicate a robot's affective and internal state we have outlined an interactive approach for personalizing non-verbal sounds in real-time. It is based on an evolution strategy and human comparative feedback, which allows even non-experts to participate in the sound design process of a robot's sonic expressions. Tempo and sound synthesis parameters are modified iteratively to adapt a robot's timbre towards an individual's preferences. We reported results of a user study with 27 participants, which also verified the approach's effectiveness (i.e., 25 of 27 participants preferred the personalized outcome over randomized baseline versions). We believe that with the increasing amount of robotic products on the consumer market, it will become more and more important to study easy to use and computational ways for personalizing non-verbal robot sounds. We also believe that there is a possibility to further optimize the sonic interaction design considering implicit human reactions to a robot's expressions in real-time, such as with reinforcement learning and human social signals [6], [7], [25]–[28]. Therefore, we aim to study in a next step possible correlations between users' affective states and situated sound preferences.

Acknowledgments

This research was funded by the Bavarian State Ministry for Education, Science and the Arts (StMWFK) as part of the ForGenderCare research association.

References

[1] D. Löffler, N. Schmidt, and R. Tscharn, "Multimodal expression of artificial emotion in social robots using color, motion and sound," in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction, HRI 2018, Chicago, IL, USA, March 05-08, 2018*, 2018, pp. 334–343.

[2] C. L. Bethel and R. R. Murphy, "Survey of non-facial/non-verbal affective expressions for appearance-constrained robots," *IEEE Trans. Systems, Man, and Cybernetics, Part C*, vol. 38, no. 1, pp. 83–92, 2008.

[3] F. J. F. de Gorostiza Luengo, F. Alonso-Martín, Á. C. González, and M. A. Salichs, "Sound synthesis for communicating nonverbal expressive cues," *IEEE Access*, vol. 5, pp. 1941–1957, 2017.

[4] C. Breazeal, *Designing sociable robots*. MIT press, 2004.

[5] P. N. Juslin and J. A. Sloboda, *Music and emotion: Theory and research.* Oxford University Press, 2001.

[6] H. Ritschel, T. Baur, and E. André, "Adapting a robot's linguistic style based on socially-aware reinforcement learning," in *26th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN 2017, Lisbon, Portugal, August 28 - Sept. 1, 2017*, 2017, pp. 378–384.

[7] H. Ritschel, "Socially-aware reinforcement learning for personalized human-robot interaction," in *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS 2018, Stockholm, Sweden, July 10-15, 2018*, 2018, pp. 1775–1777.

[8] H. Ritschel, A. Seiderer, K. Janowski, S. Wagner, and E. André, "Adaptive linguistic style for an assistive robotic health companion based on explicit human feedback," in *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments, PETRA 2019, Island of Rhodes, Greece, June 5-7, 2019*, 2019, pp. 247–255.

[9] H. Ritschel and E. André, "Shaping a social robot's humor with natural language generation and socially-aware reinforcement learning," in *Workshop on Natural Language Generation for Human–Robot Interaction at INLG 2018*, Nov 2018.

[10] H. Ritschel, K. Janowski, A. Seiderer, S. Wagner, and E. André, "Insights on usability and user feedback for an assistive robotic health companion with adaptive linguistic style," in *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments, PETRA 2019, Island of Rhodes, Greece, June 5-7, 2019*, 2019, pp. 319–320.

[11] H. Ritschel, A. Seiderer, K. Janowski, I. Aslan, and E. André, "Drink-o-mender: An adaptive robotic drink adviser," in *Proceedings of the 3rd International Workshop on Multisensory Approaches to Human-Food Interaction*, ser. MHFI'18. ACM, 2018, pp. 3:1–3:8.

[12] H. Ritschel, I. Aslan, D. Sedlbauer, and E. André, "Irony man: Augmenting a social robot with the ability to use irony in multimodal communication with humans," in *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, ser. AAMAS '19. International Foundation for Autonomous Agents and Multiagent Systems, 2019, pp. 86–94.

[13] E. Jee, Y. Jeong, C. H. Kim, and H. Kobayashi, "Sound design for emotion and intention expression of socially interactive robots," *Intelligent Service Robotics*, vol. 3, no. 3, pp. 199–206, 2010.

[14] S. Song and S. Yamada, "Expressing emotions through color, sound, and vibration with an appearance-constrained social robot," in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction, HRI 2017, Vienna, Austria, March 6-9, 2017*, 2017, pp. 2–11.

[15] M. Scheeff, J. Pinto, K. Rahardja, S. Snibbe, and R. Tow, *Experiences with Sparky, a Social Robot*. Springer US, 2002, pp. 173–180.

[16] M. Häring, N. Bee, and E. André, "Creation and evaluation of emotion expression with body movement, sound and eye color for humanoid robots," in *20th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN 2011, Atlanta, Georgia, USA, July 31 - August 3, 2011*, 2011, pp. 204–209.

[17] M. S. Schwenk and K. O. Arras, "R2-D2 reloaded: A flexible sound synthesis system for sonic human-robot interaction design," in *The 23rd IEEE International Symposium on Robot and Human Interactive Communication, IEEE RO-MAN 2014, Edinburgh, UK, August 25-29, 2014*, 2014, pp. 161–167.

[18] M. Seif El-Nasr and M. Skubic, "A fuzzy emotional agent for decision-making in a mobile robot," in *1998 IEEE International Conference on Fuzzy Systems Proceedings. IEEE World Congress on Computational Intelligence (Cat. No.98CH36228)*, vol. 1, May 1998, pp. 135–140 vol.1.

[19] C. L. Bethel and R. R. Murphy, "Non-facial/non-verbal methods of affective expression as applied to robot-assisted victim assessment," in *Proceedings of the Second ACM SIGCHI/SIGART Conference on Human-Robot Interaction, HRI 2007, Arlington, Virginia, USA, March 10-12, 2007*, 2007, pp. 287–294.

[20] R. Read and T. Belpaeme, "People interpret robotic non-linguistic utterances categorically," *I. J. Social Robotics*, vol. 8, no. 1, pp. 31–50, 2016.

[21] J. POSNER, J. A. RUSSELL, and B. S. PETERSON, "The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology," *Development and Psychopathology*, vol. 17, no. 3, p. 715–734, 2005.

[22] H. Beyer and H. Schwefel, "Evolution strategies - A comprehensive introduction," *Natural Computing*, vol. 1, no. 1, pp. 3–52, 2002.

[23] T. J. Mitchell and A. G. Pipe, "Convergence synthesis of dynamic frequency modulation tones using an evolution strategy," in *Applications of Evolutionary Computing, EvoWorkshops 2005: EvoBIO, EvoCOMNET, EvoHOT, EvoIASP, EvoMUSART, and EvoSTOC, Lausanne, Switzerland, March 30 - April 1, 2005, Proceedings*, 2005, pp. 533–538.

[24] K. Liang, X. Yao, and C. S. Newton, "Adapting self-adaptive parameters in evolutionary algorithms," *Appl. Intell.*, vol. 15, no. 3, pp. 171–180, 2001.

[25] T. M. Moerland, J. Broekens, and C. M. Jonker, "Emotion in reinforcement learning agents and robots: a survey," *Machine Learning*, vol. 107, no. 2, pp. 443–480, 2018.

[26] J. Broekens and M. Chetouani, "Towards transparent robot learning through tdrl-based emotional expressions," *IEEE Transactions on Affective Computing*, pp. 1–1, 2019.

[27] H. Ritschel and E. André, "Real-time robot personality adaptation based on reinforcement learning and social signals," in *Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction, HRI 2017, Vienna, Austria, March 6-9, 2017*, 2017, pp. 265–266.

[28] K. Weber, H. Ritschel, I. Aslan, F. Lingenfelser, and E. André, "How to shape the humor of a robot - social behavior adaptation based on reinforcement learning," in *Proceedings of the 2018 on International Conference on Multimodal Interaction, ICMI 2018, Boulder, CO, USA, October 16-20, 2018*, 2018, pp. 154–162.