How Should Pepper Sound - Preliminary Investigations on Robot Vocalizations

Felix Burkhardt[1], *Milenko Saponja*[1], *Julian Sessner*[2] *and Benjamin Weiss*[1]

[1]*audEERING GmbH,* [2]*Lehrstuhl für Fertigungsautomatisierung und Produktionssystematik, Friedrich-Alexander-Universität Erlangen-Nürnberg.*
fburkhardt@audeering.com

Abstract: We describe a perception experiment as a preliminary investigation to identify an appropriate vocal expression for a robot interacting with autistic children. 18 audio tracks with different voice characteristics but the same text were generated using speech synthesizers and acoustic modification and then used as over dubs for a short video of a robot. When participants in a perception experiment were asked how far the audio fits with the robot and how agreeable it sounds, most selected less artificial sounding samples.

1 Introduction

Interaction with robots becomes more and more part of our daily life, irrespective of form and size or if they are physically present robots or just virtual. If they can speak, the voice of robots is usually not generated by mechanical vocal organs but by digital speech synthesis, so the match between voice and appearance does not come natural. If the match is bad, users might not even realize that the voice originates from the robot. Another aspect of robotic vocal expression is the speaking style which would in an ideal case adapt to the communication situation, but at least should be appropriate to the task the robot is designed to do. If for example the robot is designed as a toy to interact with children, an anchor-man style of voice is probably not a good choice.

As pointed out in [1] it is not self-evident that robot voices should be as human-like as possible. Designing them more artificial might even be a solution to avoid the uncanny valley effect [2] and could be suited to lower the expectation on the robot's general world knowledge.

In [3], robot voices from movies and games were investigated and described according to acoustic parameters, which in generally feature a voluntary high degree of artificiality.

In this paper we describe a perception experiment as a preliminary investigation to identify an appropriate vocal expression for a robot interacting with autistic children. Within the German national BMBF project ERIK[1] interaction of a robot with autistic children is planned with the aim to tutor them with respect to interpretation and expression of emotional arousal. The robot will most probably be the Pepper[2] robot from Softbank Robotics.

The aim of the project is to help the children to enhance their emotional communication and part of the training will be that the robot expresses emotions adequately for a given context. Therefore one part of the project is to implement an appropriate voice for the robot that is capable to express emotional arousal. The dialog will probably be scripted, so the use of human recordings is a viable alternative to text-to-speech synthesizers.

A short sequence of Pepper performing a greeting arm-gesture was recorded with video and the test audio files were replaced with the original audio track. These videos were then shown

[1]https://www.technik-zum-menschen-bringen.de/projekte/erik
[2]https://www.softbankrobotics.com/emea/en/pepper

to test subjects who rated firstly how good the fit between robot and voice is for this video and second how agreeable the voice sounds.

This paper is structured as follows: after this introduction we describe the selected audio alternatives in Section 2. In Section 3 we discuss how the stimuli were prepared and in Section 4 how the perception experiment was done. This is followed by a description and discussion of the results in Section 5. The paper concludes with Section 6.

2 Selecting voice alternatives

As stated above, the dialog between Pepper and the children will probably be scripted, so the use of human recordings is a viable alternative to text-to-speech technology. We grouped several alternatives to generate vocal expression into four categories, which can be combined.

- Human voice
- Modified voice
- TTS (Text to speech)
- Sounds (extralinguistic)

As a carrier sentence for the robot we used the German utterance "*Hallo, ich bin Pepper! Ich freue mich dich zu sehen. Lass uns ein Spiel zusammen machen.*" (*Hello, I'm Pepper, I'm happy to see you. Let's do a game together.*), as it seemed appropriate for the target scenario, which was not finally specified at the time of this experiment.

For each of the above mentioned categories, we manually crafted several examples of possible emotional expressive speech, using the same textual content. These are described in the next sections.

2.1 Human voice

Two humans, a female and a male, recorded the target sentences in a quiet surrounding with a ZOOM H5 audio recorder. Both speakers are native Germans and in the early 40s. From five versions, stylistically directed to children, the most natural one was selected by the authors.

2.2 Modified voice

According to [3], robot voices "... *can be achieved by a small increase in pitch, followed by adding back the original (c.f. 'harmony') and introducing some echo*". We added these manipulations on the natural male recording to our test set with the Audacity wave editor[3]. The pitch was raised by a perfect third, two duplicate versions raised and lowered by octaves added and echo added (delay time: 0.1 seconds, decay factor: 0.4).

[4] describe a system called DAVID that was used to "emotionalize" the human voice samples by altering global pitch, adding inflections, vibrato and applying spectral filters. To simulate a happy, agreeable voice, we used the program's default to simulate happy arousal and added 30 Hz to the F0, used the default increase for inflection and applied a 8kHz high shelf filter. The result was quite subtle.

[3] https://www.audacityteam.org/

2.3 TTS (Text-to-Speech) voice

Speech synthesizers can be distinguished into roughly five categories with varying support to add emotional expression:

- DNN Synthesis: Quite the newest addition to speech synthesis algorithms are artificial neural networks or deep neural nets. TTS with neural nets has been done since many decades but now have strong attraction because of the advancements in hardware power and data collections. They replace currently the HMM approach to predict the best acoustic parameters for a given sequence of symbols representing text. [5] describe an approach to learn emotional prosody styles within an ANN framework.

- HMM Synthesis: Synthesis based on Hidden Markov Models is a statistical approach to model the transition probabilities of the acoustic parameters based on the speech to be generated. The approaches are trained on a relatively large data corpus, but have a small footprint for synthesis because they don't operate on the wave data directly but on some parameter representation. However, this is also the reason they tend to produce artifacts, especially when generating emotional speech.

- Non-uniform unit-selection: Best fitting chunks of speech from large databases get concatenated, minimizing a double cost-function: best fit to neighbor unit and best fit to target prosody. Sounds very natural (similar to original speaker), but is inflexible with respect to simulate out-of-database styles.

- Diphone-synthesis: Speech concatenated from diphone-units (two-phone combinations), prosody-fitting is done by signal-manipulation which depends on unit-coding. It has a relatively small footprint but does not sound very natural due to the small database and high degree of added signal manipulation. Of course this is an advantage if you want to simulate emotional arousal and in fact many systems operate on top of diphone synthesis, e.g. [6].

- Formant-synthesis: Speech synthesized by physical models (formants are resonance frequencies in vocal-tract). This is obviously very flexible and has the smallest footprint, but sounds rather unnatural if not manually crafted. Early attempts to simulate emotions were done with formant synthesis, e.g. [7].

The original Pepper voice is licensed from Nuance Vocalizer and can be prosodically modified with markup[4] comparable to SSML (Speech Synthesis Markup Language)[5]. This was used to simulate emotional arousal. We modified the middle sentence ("*I'm happy to see you*") with

```
rspd=50, vct=150, emph=2
```

which slows the speech rate, raises the pitch and adds additional emphasis, according to the documentation[6].

We used the Mary [8] framework with the (male) emotional Pavoque speech database to generate the test sentences as a neutral version and once with the emotion "happy" added to the middle target sentence.

More versions were generated with the Emofilt [6] toolkit which is based on the Mbrola diphone speech synthesizer [9]. The voices de6 (male) and de7 (female) were used to generate

[4]http://doc.aldebaran.com/2-4/naoqi/audio/altexttospeech-tuto.html
[5]https://www.w3.org/TR/speech-synthesis11/
[6]http://doc.aldebaran.com/2-4/naoqi/audio/altexttospeech-tuto.html

neutral and "happy" versions respectively. The happy version was generated using the so-called "wave model" where the main stressed syllables are raised and the syllables, that lie equally distanced in between, are lowered. Its parameters are the maximum amount of raising (150 %) and lowering (100 %) and connected with a smoothing of the pitch contour, because all F0 values are linearly interpolated.

As a last synthesizer, the open source formant synthesizer eSpeak[7] was used to generate a default (male) version and one that sounds a bit more female by shifting the pitch 14 semitones upwards.

With these synthesizers we had samples from non-uniform unit-selection approach (Vocalizer and Mary) as well as diphone synthesis (Mbrola) and formant synthesis (eSpeak).

2.4 Sounds (extralinguistic)

Because a focus of this investigation was emotional expression, we added two sound generating systems to the test field that don't produce understandable language but emotional sounds.

The Sony pet voice samples from Oudeyer [10] were used to give Pepper a meaningless cartoon style voice. Oudeyer used an algorithm based on Mbrola [9] combined with emotion expression rules derived from literature.

As a second example, we used demonstration samples[8] from the MIT Kismet voice [11]. The system assembles *strings of phonemes with pitch accents on the fly to produce a style of speech that is reminiscent of a tonal dialect* and with a Klatt-style [12] formant synthesizer modified by emotion-expression rules inspired by Janet Cahn's work [7].

3 Stimulus preparation

We took as short (about eight seconds) video of Pepper uttering the carrier utterances with its default voice, see Figure 1 for a screen shot. At the end Pepper performs a wave gesture with its right arm. We used the kdenlive video editing software[9] to overdub the movie with the respective audio files as described in the previous section. In order to normalize the volume we normalized all samples to -0.9 DB with the software. We also added reverb (room size 20) to all tracks with kdenlive, because the original audio track contained strong room characteristics.

The final number of stimuli was 18, the videos were coded as Ultra-high Definition, MP4-H265 which is available in kdenlive version 17.12.3.

4 Perception experiment

The aim was to obtain a first impression on consistency and preference of the fit and agreeableness of this voice-robot combination, before presenting a smaller set to autistic children, preferable in an interactive scenario.

For the perception experiment, all stimuli were randomly presented to 28 adult participants which we collected by an appeal to colleagues at audEERING and the ERIK team. We used the IHEARu-PLAY platform [13] which is a web platform to annotate auditory data with a gaming component that also supports the display of videos. The participants did the experiment at their homes or offices with their own headphones, it took about 10-15 minutes. Regrettably we don't know who exactly participated and what were their age, gender and nationality.

[7] http://espeak.sourceforge.net/

[8] http://www.ai.mit.edu/projects/sociable/expressive-speech.html

[9] https://kdenlive.org/

Figure 1 – Screenshot of the video

Each video was rated on the two scales, operated with sliders on a 5-level scales: 1. *How does this voice fit to the Pepper robot?* and 2. *How agreeable appears Pepper with this voice?* It appears from feedback we had from our participants that some were not very sure how the question about the "agreeableness" is meant, but we hope that most of them understood the distinction.

Although most of the participants were native Germans, some were not and all of them speak English at a high level.

5 Results and Discussion

The Fit and agreeableness correlate only moderately (Spearman's $\rho = .54$, $p < .0001$), but given that the two questions were quite different, we are surprised that they correlate at all. Both scales were analyzed similarly. The inter-rater consistency is good for the fit ($ICC(3,28) \to .926$), and sufficient for agreeableness ($ICC(3,28) \to .724$).

Concerning the fit between robot and voice, a Friedman test shows a significant difference between the stimuli ($\chi^2_{(17)} = 191.21$; $p < .0001$). See Figure 2 for a boxplot of ratings separated for condition. Nemenyi post-hoc tests confirm that the original pepper voice is considered as better fit than 'Espeak+14', 'M-robotic', and the 'Oudayer' non-speech sound, while it is not significantly less fitting than any other voice.

A similar test for agreeableness of the audio-visual stimulus also results in significant differences between the condition ($\chi^2_{(17)} = 76.407; p < .0001$). The corresponding box plot is depicted in Figure 3. Apart from a generally smaller variability, the ranking of stimuli is roughly similar, with differences for 'EmofiltDe6Happy', 'F', and 'PepperMOD', which are ranked lower in agreement than in fit, and 'M-MODhappy' and 'Pepper', which are both ranked higher according to the median.

From the results we might conclude that generally non-speech seems not to be a good fit which indicates that Pepper appears convincingly humanoid. This is supported by the fact that the natural versions were perceived as fitting to the robot and in general more natural version preferred over the more robotic ones.

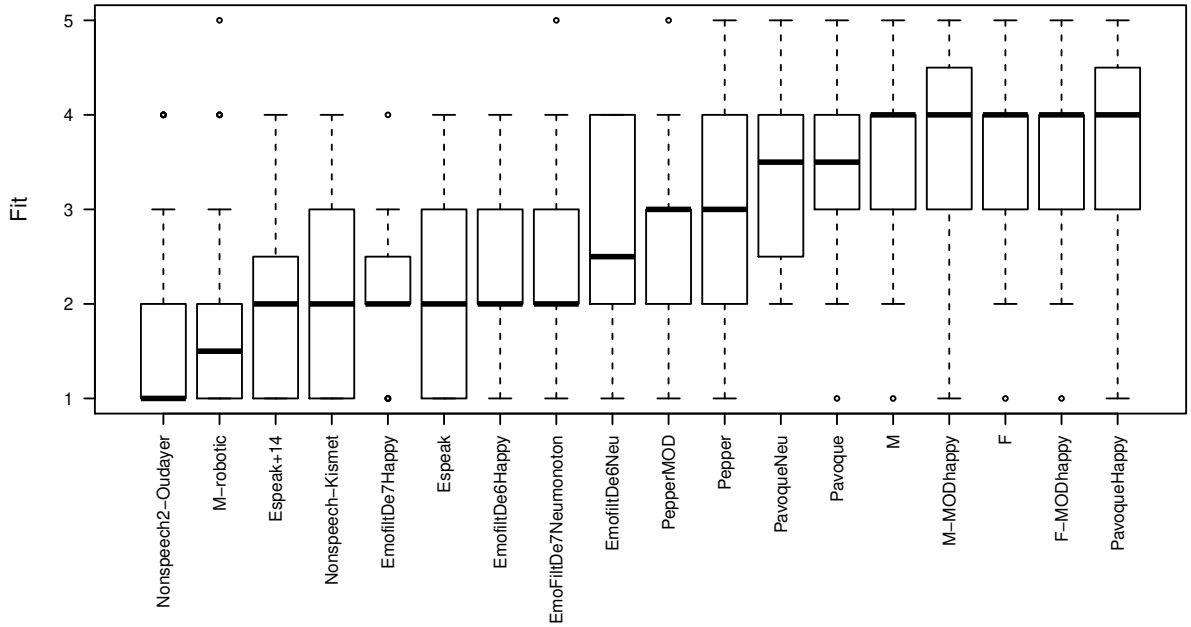

Figure 2 – The fit between voice and the robot. Ordered by median and mean.

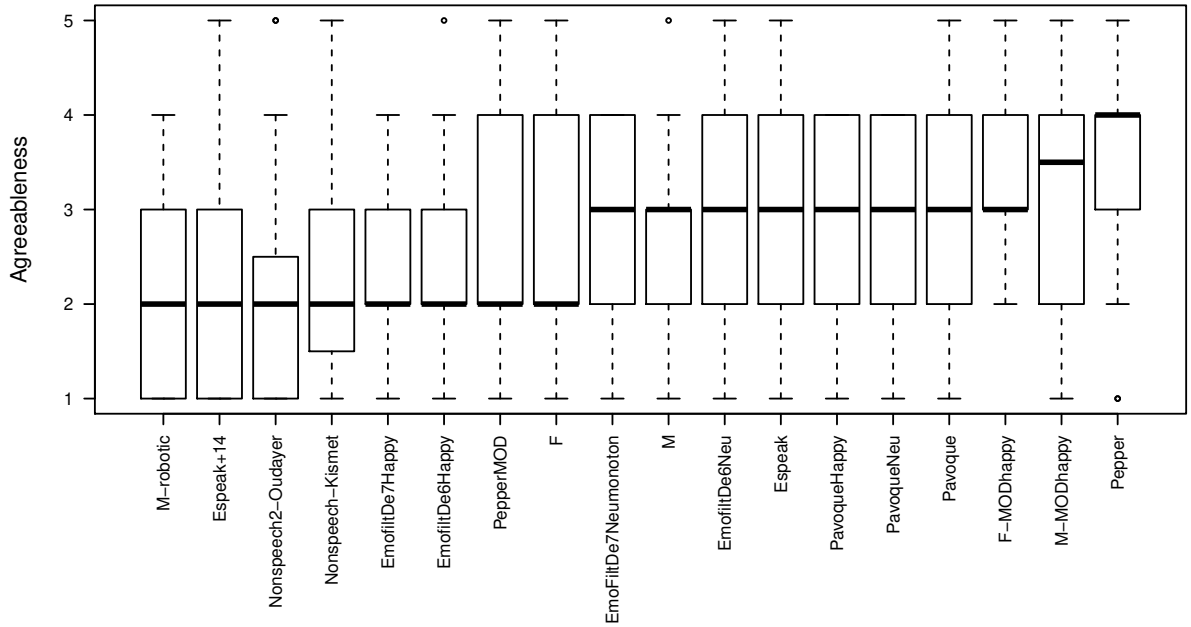

Figure 3 – Agreeableness of the whole video. Ordered by median and mean.

The agreeableness shows similar results compared to fit, but here, the two most agreeable videos have manipulated or clearly synthesized voices: 'Pepper' and 'M-MODhappy' (MARY), it seems that the engineers of Softbank made a good job with the standard voice.

The extra-linguistic sounds and the formant synthesizer voice weren't percieved as neither fitting nor agreeable, the diphone synthesis voices are about in the middle.

With respect to our Project, it is encouraging to see that synthesized voices are a possible alternative to acted voices, which have the disadvantage that the content to be spoken can not be dynamic. The Mary synthesized voice as well as the original Pepper voice from Nuance, like the human voices, were both perceived as both agreeable and fitting to the robot.

6 Conclusions

As a pre-investigation to a project concerned with robot interaction with autistic children, we investigated possibilities to give the Pepper robot a voice. We compiled a set of 18 audio tracks for a short video featuring Pepper and had this videos judged by 28 participants as to how the voice fits to Pepper and how "agreeable" the voice sounds. We found that a significant number of people preferred a "less robotic" voice, meaning less artificial. The both questions correlate moderately which already is surprising as they asked for quite different things. But perhaps Pepper visually already appears to look agreeable, so non-agreeable voices do not seem to fit.

When the project progressed we will measure the adequacy of Pepper's voice in the context of the target environment, i.e. the interaction with autistic children in a game scenario.

7 Acknowledgements

The authors acknowledge the financial support by the Federal Ministry of Education and Research of Germany in the framework of Mensch-Technik-Interaktion (project ERIK: train emotional competence with robots). We also thank our colleagues who participated in the listening test.

References

[1] MOORE, R. K.: *Appropriate voices for artefacts: some key insights*. In *1st Int. Workshop on Vocal Interactivity in-and-between Humans, Animals and Robots (VIHAR-2017)*. 2017.

[2] MORI, M.: *The Uncanny Valley. Energy*, 7(4), pp. 33–35, 1970. doi:10.1109/MRA.2012.2192811.

[3] WILSON, S. and R. K. MOORE: *Robot, alien and cartoon voices: implications for speech-enabled systems*. In *1st Int. Workshop on Vocal Interactivity in-and-between Humans, Animals and Robots (VIHAR-2017)*. 2017.

[4] RACHMAN, L., M. LIUNI, P. ARIAS, A. LIND, P. JOHANSSON, L. HALL, D. RICHARDSON, K. WATANABE, S. DUBAL, and J. J. AUCOUTURIER: *DAVID: An open-source platform for real-time transformation of infra-segmental emotional cues in running speech. Behavior Research Methods*, 50(1), pp. 323–343, 2018. doi:10.3758/s13428-017-0873-y.

[5] WANG, Y., D. STANTON, Y. ZHANG, R. SKERRY-RYAN, E. BATTENBERG, J. SHOR, Y. XIAO, F. REN, Y. JIA, and R. A. SAUROUS: *Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis*. 2018. doi:10.1227/01.NEU.0000297116.62323.15. URL `http://arxiv.org/abs/1803.09017`. `1803.09017`.

[6] BURKHARDT, F.: *Emofilt: The simulation of emotional speech by prosody-transformation*. In *9th European Conference on Speech Communication and Technology*. 2005.

[7] CAHN, J. E.: *The Affect Editor. Journal of the American Voice I/O Society*, 8, pp. 1–19, 1989.

[8] CHARFUELAN, M. and I. STEINER: *Expressive speech synthesis in MARY TTS using audiobook data and EmotionML. Proc Interspeech*, 2013.

[9] DUTOIT, T., V. PAGEL, N. PIERRET, F. BATAILLE, and O. DER VREKEN: *The MBROLA Project: Towards a Set of High-Quality Speech Synthesizers Free of Use for Non-Commercial Purposes. Proc. ICSLP'96, Philadelphia*, 3, pp. 1393–1396, 1996.

[10] OUDEYER, P.-Y.: *The synthesis of cartoon emotional speech. Speech Prosody 2002, International Conference*, pp. 13–16, 2002.

[11] BREAZEAL, C.: *Emotion and sociable humanoid robots. International Journal of Human Computer Studies*, 59(1-2), pp. 119–155, 2003. doi:10.1016/S1071-5819(03)00018-1.

[12] KLATT, D. H.: *Software for a Cascade/Parallel Formant Synthesizer. JASA*, 67(3), pp. 959–971, 1980.

[13] HANTKE, S., T. APPEL, F. EYBEN, and B. SCHULLER: *iHEARu-PLAY: Introducing a game for crowdsourced data collection for affective computing*. In *Proc. 1st International Workshop on Automatic Sentiment Analysis in the Wild (WASA 2015)*, pp. 891–897. IEEE, 2015.