Synthesising expressive speech – which synthesiser for VOCAs?

Jan-Oliver Wülfing, Chi Tai Dang, Elisabeth André

Angaben zur Veröffentlichung / Publication details:

Wülfing, Jan-Oliver, Chi Tai Dang, and Elisabeth André. 2020. "Synthesising expressive speech – which synthesiser for VOCAs?" *Lecture Notes in Computer Science* 12284: 397–408. https://doi.org/10.1007/978-3-030-58323-1_43.

Synthesising Expressive Speech
Which Synthesiser for VOCAs?

Jan-Oliver Wülfing, Chi Tai Dang, and Elisabeth André

Human-Centred Multimedia, University of Augsburg,
Universitätsstrasse 6a, 86159 Augsburg, Germany http:\\www.hcm-lab.de
{wuelfing,dang,andre}@hcm-lab.de

Abstract. In the context of people with complex communication needs who depend on Voice Output Communication Aids, the ability of speech synthesisers to convey not only sentences, but also emotions would be a great enrichment. The latter is essential and very natural in interpersonal speech communication. Hence, we are interested in the expressiveness of speech synthesisers and their perception. We present the results of a study in which 82 participants listened to different synthesised sentences with different emotional contours from three synthesisers. We found that participants' ratings on expressiveness and naturalness indicate that the synthesiser CereVoice performs better than the other synthesisers.

Keywords: Complex Communication Needs, Voice Output Communication Aid, Expressive Speech Synthesis, Online Survey

1 Introduction

How often do we vocally speaking people use our tone of voice to communicate our intentions, wishes, or desires to a communication partner throughout the day? Depending on the emotions to be conveyed, the tone of voice is portrayed by a variation of prosodic features (rhythm, speed and pitch, etc.) and voice quality [5]. For instance, a sad person has different tone of voice than a happy one. The first one typically speaks slower and lower pitched than the latter one.

As Hoffmann and Wülfing pointed out in a survey [11] with 129 participants, people who cannot or almost not articulate themselves vocally would like to do the same with the help of their VOCA (*Voice Output Communication Aid*). These VOCAs fall into the group of technologies in the domain of AAC (*Alternative and Augmentative Communication*). VOCAs take text input and synthesise the input as auditory output. Yet, the possibilities and potentials of synthesisers have not been used extensively. In the past, industry has mainly focused on naturalness neglecting variability in expressive style. However, as text-to-speech synthesisers continue to improve, the question arises of whether synthesisers may help people use VOCAs to express their feelings, wishes, and intentions as well.

As a first step to answer this question, we investigate the expressiveness of three freely available synthesisers in terms of recognised emotions and their naturalness in terms of perceived pronunciation quality. For this purpose, we conducted a survey with 82 participants in order to investigate which of these synthesisers shows the highest expressiveness and naturalness for sentences generated for the German language. We decided to evaluate MaryTTS v5.2[1] developed collaboratively by the German Research Center for Artificial Intelligence and Saarland University and eSpeak[2] v1.48.04 developed by Jonathan Duddigton and maintained by Reece Dunn. Both are open source synthesisers. eSpeak provides voices created by using formant synthesis. MaryTTS provides both unit selection and voices based on Hidden-Markov Models (HMM) [13]. As a third synthesiser, we chose the commercial CereVoice unit selection speech system[3] v4.0.6 developed by CereProc's Ltd. CereVoice is a commercial-grade real-time ESS (*Expressive Speech Synthesis*) system [2]. For our study, we used an academic licence provided by CereProc Ltd.

All three synthesisers have capabilities to manipulate prosodic features and make use of a markup language that more or less follows the industry standard SSML (*Speech Synthesis Markup Language*) v1.1[4]. eSpeak uses SSML, however, with fewer options to manipulate. For MaryTTS, MaryXML[5] serves as its own data representation format which facilitates the synthesis of prosodic utterances - the syntax is similar to SSML. In addition to SSML support, CereVoice offers CereVoice XML extensions[6] for emotional synthesis control. In our previous work [16], we evaluated how a VOCA that enables the specification of certain emotional states via Emojis would be perceived by users with CCN. To this end, we presented them with a first prototype VOCA 'EmotionTalker' (ET) in their daily environment. Here, we focus on which speech synthesiser to use for enhancing a VOCA with expressive speech. To this end, we compared three publicly available speech synthesisers (eSpeak, MaryTTS, CereVoice) in a perception study with 82 participants. Our long-term objective is to pave the way towards a new generation of VOCAs that convey emotions and personality.

2 Related Work

Recently, the naturalness of synthesised speech has significantly improved. In some cases, it has become hard to distinguish artificially created voices from human voices. This is in particular true for commercial speech synthesisers, such as CereVoice. In the area of speech synthesis, basically two approaches have been used: unit selection approaches and statistical parametric synthesis approaches (see [3] for a recent survey). Unit selection approaches make use of a large inventory of human speech units that are subsequently selected and combined based on the sentence to be synthesised. Statistical parametric synthesis approaches create acoustic models from recorded speech (for example, using

[1] http://mary.dfki.de (accessed 02/06/20)

[2] http://espeak.sourceforge.net (accessed 02/06/20)

[3] https://www.cereproc.com/en/products/academic (accessed 11/06/20)

[4] https://www.w3.org/TR/speech-synthesis11/ (accessed 02/06/20)

[5] http://mary.dfki.de/documentation/maryxml/ (accessed 12/06/20)

[6] https://www.cereproc.com/de/products/sdk (accessed 12/06/20)

Hidden Markov Models or Deep Neural Networks) that are used to reconstruct synthesised speech from the generated parameters. Usually, more natural synthesis results are obtained by unit selection approaches. However, unit selection approaches offer little flexibility to manipulate speech parameters in a way that different emotional styles are conveyed. To give users more control over the synthesised speech, specific extensions for the industry standard SSML have been developed, such as CereProc XML extensions or MaryXML, that enable users to create different styles of expressive speech.

To evaluate the quality of the produced speech, a variety of perceptual quality dimensions of synthetic speech, such as intelligibility and naturalness, have been defined (see [10] for an overview) that are also employed in the annual 'Blizzard' challenge[7] on advancing speech synthesis. Also, the emotional atmosphere of a scene and the moods of the characters have been included as a quality dimension in audiobook synthesis tasks. Wagner et al. [15] point out that the evaluation of TTS is still using criteria from the early days of speech synthesis research and argue for a user-centered approach that considers a larger diversity of users including gender and age. A user-centered approach is in particular recommended for AAC users who would like to communicate with expressive voices as shown in our previous research [16]. When developing VOCAs with expressive speech, the question arises of how to enable AAC users to control the quality of speech in an easy manner. Recent work on expressive VOCAs (see [16] and [8]) makes use of expressive keyboards that include Emojis to specify the emotions to be conveyed. While such interfaces enable an easy specification of the emotional content, they provide the AAC user only with a limited amount of control over the synthesised speech. However, when being engaged in a conversation, the fine-grained control of a large number of parameters that would ensure a high quality of expressive speech is no option. For this reason, we decided to focus in our study on a few set of parameters that can be easily mapped on emotions to be conveyed without requiring extensive fine-tuning.

3 Study

In order to evaluate the expressive capabilities of the three synthesisers to be considered for integration into a VOCA, we performed an online survey. Participants were acquired through a mailing list at the first author's home university, the news-site of the department to which the authors are affiliated, and a forum entry especially for AAC users and their personal assistants.

3.1 Online Survey

The online survey consisted of 27 WAV-files (3 sentences * 3 emotions * 3 synthesisers) which were prepared in advance. For the study, we relied on German voices. In particular, we used the following voices: eSpeak (Formant, male, de),

[7] http://www.festvox.org/blizzard (accessed 02/06/20)

MaryTTS (HMM, female, bits1-hsmm), CereVoice (Unit Selection, female, Gudrun). Following Murray et al. [12], we selected three emotionally neutral sentences. 'Emotionally neutral' means that the semantics of a sentence does not provide any clue on the speaker's emotion. For example, one of the sentences was "Ich kann da drüben Leute sehen“ (engl. "I can see people over there“). In order to convey the emotions (happy, sad, angry), we used SSML-markups to manipulate pitch, volume, rate and contour. In light of later integration into an easy-to-use VOCA GUI, we did not exploit the full potential of XML extensions to enable more sophisticated emotional control. The online survey and sentences were reviewed by several researchers in terms of wording and conveyed emotions.

Structure The survey had three parts. First of all, participants had to agree to a DPA (Data Processing Agreement) in order to continue. Then, they had to provide demographic data including age, sex, and cultural background (in order to exclude any disposition). Next, participants had to listen to the 27 sentences. After each audio clip, they were asked in the online survey to type the sentence heard, to indicate the emotion perceived, and how satisfied they were with their choice of the selected emotion. After evaluating all 27 sentences, the participants were presented again with three sentences explicitly indicated as happy, sad, or angry. This time, participants had to mark how satisfied they were with the naturalness and expressivity of the corresponding speech synthesiser. This third part was designed as a double check of the second part.

Questions In the second part of the survey, we presented the participants with a forced response choice. Following the approach of Murray et al. [12], we included two additional emotions (fear and disgust) and a neutral state as distractors. That is, we disguised the number and the category of the emotions actually being tested. The participants had to listen to short sentences played back through WAV-files in a randomised order of the speech synthesisers. The first question in part two "Please, write the heard sentence into the box“ (transl.) was asked in order to identify any acoustic issues. The second question "Which emotion do you link to the sentence“ provided us with the perceived emotion. The last question "How satisfied are you with the choice of the perceived emotion“ served as a confidence measure for the previous answer. The third part of the online survey served to get information on the participants' subjective impression of the speech synthesiser. Participants were asked "Please, evaluate the synthesiser XX in respect ...“, "... to its articulation“, and "... to its expressivity“.

4 Results

The online survey was conducted between February and May 2019 with 82 German-speaking participants, who filled in the survey completely. We had 32 male/50 female participants aged between 18 and 65 years ($M = 28.78, SD = 10.61$). In addition, the participants had the opportunity to state their origin.

The large majority came from Germany. In addition, Austria, Poland, Russia, Asia, Latin, and Turkey were stated. Participants needed on average 798 seconds ($SD = 151.41$) to complete the survey.

In total, participants had to evaluate three sentences for three synthesisers for each of the three emotions, i.e., for each synthesiser, they had to correctly assign emotions to nine sentences. The highest number of correctly assessed emotions were: Seven correct for CereVoice by one participant, six correct for MaryTTS by one participant, and four correct for eSpeak by nine participants. Regarding the emotions, the highest number of correctly assessed emotions was achieved for Sadness (eight hits by 16 participants), followed by Angry (five hits by six participants) and Happiness (four hits by three participants).

4.1 Average Number of Recognised Emotions

Which of the synthesisers expresses which of the emotions best? In order to answer this question, we look at the correctly assessed emotions. Table 1 gives an overview of the mean values and the corresponding standard deviations for the number of correctly assessed emotions from the synthesised sentences across all three synthesisers and separately for each of the three emotion classes. If we consider only the correctly assessed emotions independent of the synthesisers (aggregated over all synthesisers), then the class *Sad* was recognised best with 3.68 sentences ($SD = 1.92$), followed by the class *Angry* with $M = 2.22, SD = 1.60$. The class *Happy* was expressed the worst of all ($M = 1.28, SD = 1.11$). When all emotion classes are considered together for each of the synthesisers, CereVoice scores best with an average of 2.78 ($SD = 1.42$) correctly assessed emotions, closely followed by MaryTTS with 2.62 ($SD = 1.54$) correctly assessed emotions. The synthesiser eSpeak has the worst average score of 1.78 ($SD = 1.31$) correctly assessed emotions.

#A. emotions per synthesiser (0..9)		**#A. emotions overall (0..9)**	
E - eSpeak	$M = 1.78, SD = 1.31$	a - Angry	$M = 2.22, SD = 1.60$
M - MaryTTS	$M = 2.62, SD = 1.54$	h - Happy	$M = 1.28, SD = 1.11$
C - CereVoice	$M = 2.78, SD = 1.42$	s - Sad	$M = 3.68, SD = 1.92$

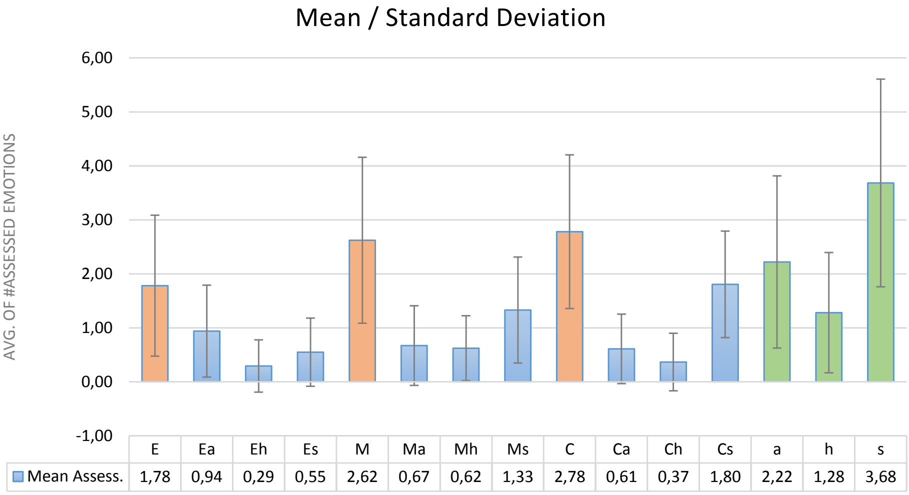
#A. emotions per synthesiser and emotion class (0..3)			
Emotion	E - eSpeak	M - MaryTTS	C - CereVoice
a - Angry	$M = .94, SD = .85$	$M = .67, SD = .74$	$M = .61, SD = .64$
h - Happy	$M = .29, SD = .48$	$M = .62, SD = .60$	$M = .37, SD = .53$
s - Sad	$M = .55, SD = .63$	$M = 1.33, SD = .99$	$M = 1.8, SD = .99$

Table 1. Means and standard deviations for the number of correctly assessed emotions. ”Assessed“ is abbreviated with A.

4.2 Performance between Synthesisers

Which of the synthesisers has the best/worst numbers of correctly assessed instances across all emotion classes? The answer to this question is provided by a comparison of the mean values for assessed emotion instances of all three synthesisers (see Table 1, "#Assessed emotions per synthesiser").

Here, a repeated measurement ANOVA [9] showed highly significant differences between the synthesisers ($F(2, 162) = 18.05, p < .001$). The post-hoc pairwise comparisons (with Bonferroni corrections) for each measured synthesiser revealed that there are significant differences between the synthesiser eSpeak (denoted by E) and MaryTTS (denoted by M) (E-M: $p < .001$, $-.84, 95\% - CI[-1.27, -.42]$) as well as eSpeak and CereVoice (denoted by C) (E-C: $p < .001$, $-1.00, 95\% - CI[-1.40, -.60]$), meaning that emotions were in general recognised significantly better with MaryTTS and CereVoice than with eSpeak. Figure 1 shows all mean values and standard deviations of Table 1, whereby the blue bars indicate an unequal distribution of the correctly assessed emotion classes within the synthesisers, which we discuss further in the following.



	E	Ea	Eh	Es	M	Ma	Mh	Ms	C	Ca	Ch	Cs	a	h	s
■ Mean Assess.	1,78	0,94	0,29	0,55	2,62	0,67	0,62	1,33	2,78	0,61	0,37	1,80	2,22	1,28	3,68

Fig. 1. Means and standard deviations of correctly assessed emotions from sentences.

Performance within Individual Synthesisers. *If we consider the mean values within individual synthesisers, which of the emotion classes is expressed better/worse than the other classes?* To answer this question, we have a more detailed look at the individual emotion classes across the three synthesisers.

Repeated measurement ANOVAs for each of the emotion classes and synthesisers revealed further differences of the correctly assessed emotions for each of the synthesisers as indicated by the blue bars in Figure 1. We found highly significant differences in the recognition of synthesised emotions within the synthesiser eSpeak ($F(2, 162) = 22.00, p < .001$). Post-hoc pairwise comparisons showed that the recognition of the emotions *Angry, Sad, Happy* (in that order) differ significantly from well to badly recognisable. With the synthesiser CereVoice, the highly significant differences between all three emotion classes ($F(2, 162) = 97.59, p < .001$), i.e., post-hoc comparisons, showed that *Sad* was recognised best and *Happy* worst. Smaller amount of differences were found with MaryTTS ($F(2, 162) = 23.63, p < .001$). The significant differences with MaryTTS from post-hoc pairwise comparisons showed that *Sad* could be better distinguished from the other emotion classes.

4.3 Performance between Emotion Classes

Considering the recognition of emotion classes aggregated over all synthesisers, which emotion class is recognised best? And which synthesiser performs best on which emotion class? For the answer to these questions, we combine the assessed rates of the different emotion classes across all synthesisers (c.f., Table 1). For the analysis within an emotion class we take a more detailed view on the individual synthesisers (c.f., green bars in Figure 1).

A repeated measurement ANOVA comparing the correctly assessed emotions for each emotion class showed a highly significant difference ($F(2, 162) = 59.80, p < .001$) in correctly assessed emotions between the emotion classes. Post-hoc pairwise comparisons revealed that the class *Sad* was most frequently and thus significantly more often correctly identified by participants (a - s: $p < .001$, $-1.46, 95\% - CI[-2.04, -.88]$; h - s: $p < .001$, $-2.40, 95\% - CI[-2.97, -1.84]$) than for the other classes. Furthermore, the class *Angry* was significantly more often identified than the class *Happy* (a - h: $p < .001$, $.94, 95\% - CI[.47, 1.41]$), meaning that the class *Happy* was the worst recognisable.

Performance within Individual Emotion Classes. *If we consider the mean values within individual emotion classes, which of the synthesisers expresses the emotions better/worse than the other synthesisers?* To answer this question, we have a more detailed look at the means across the three synthesisers for each of the emotion classes (c.f., blue bars in Figure 1).

For the emotion class *Angry*, we found significant differences ($F(2, 162) = 6.03$, $p < .005$) between the synthesisers, where post-hoc analysis identified eSpeak as significantly better recognisable than the other synthesisers. For the emotion class *Happy*, there was a significant difference ($F(2, 162) = 10.52, p < .005$) between the synthesisers in favour of MaryTTS revealed by post-hoc pairwise comparisons. Finally, for the emotion class *Sad*, the statistics showed highly significant differences ($F(2, 162) = 59.78, p < .001$), where the post-hoc analysis identified that each of the synthesisers significantly differed from each other in the order *CereVoice, MaryTTS, and eSpeak* from best to worst.

4.4 Satisfaction with the Choice of Assessed Emotions

For each of the assessed emotions, participants were asked to rate on a Likert scale (*”not satisfied at all - 1“*, *”undecided - 3“*, *”very satisfied - 5“*), how satisfied they were with the choice of the assessed emotion class. Table 2 contains all mean values and standard deviations. Participants seemed to have different degrees of satisfaction with their choice between the synthesisers. While participants rated on average with less than *”undecided-3“* for eSpeak, the ratings for MaryTTS and CereVoice tended to be higher towards *”satisfied - 4“*. However, the mean values for satisfaction hardly differed between the emotion classes (a, h, s), with mean values slightly above *”undecided - 3“*.

For chosen emotions per synthesiser		**For chosen emotions**	
eSpeak	$M = 2.81, SD = 0.94$	a - Angry	$M = 3.12, SD = 0.67$
MaryTTS	$M = 3.26, SD = 0.63$	h - Happy	$M = 3.12, SD = 0.71$
CereVoice	$M = 3.27, SD = 0.63$	s - Sad	$M = 3.11, SD = 0.74$

Satisfaction ratings for chosen emotions per synthesiser and emotion			
Emotion	E - eSpeak	M - MaryTTS	C - CereVoice
a - Angry	$M = 2.91, SD = 1.10$	$M = 3.27, SD = 0.70$	$M = 3.12, SD = 0.68$
h - Happy	$M = 2.76, SD = 1.03$	$M = 3.34, SD = 0.75$	$M = 3.26, SD = 0.79$
s - Sad	$M = 2.76, SD = 1.00$	$M = 3.17, SD = 0.64$	$M = 3.38, SD = 0.78$

Table 2. Overview of the means and standard deviations for the satisfaction ratings (on a scale of 1 ... 5) for a chosen emotion (and emotion class).

4.5 Satisfaction between Synthesisers

Which synthesiser showed the highest satisfaction with the choice on average when all emotion classes were included? To address this question, we compared the given satisfaction ratings between each of the synthesisers. As already indicated by Table 2, the repeated measurement ANOVA showed highly significant differences between the synthesisers ($F(2, 162) = 27.33, p < .001$). Overall, the participants were significantly more satisfied with their choice of an emotion class while listening to sentences synthesised by MaryTTS (E-M: $p < .001, -.45, 95\% - CI[-.65, -.26]$) and CereVoice (E-C: $p < .001, -.47, 95\% - CI[-.67, -.26]$) than by eSpeak.

Satisfaction within Individual Synthesisers. *With which of the conveyed emotion classes were the participants most satisfied measured by the mean values within the synthesisers?* For answering this question, we compared the values within the individual emotion classes across the three synthesisers.

Only CereVoice showed a measurable significant effect ($F(2, 162) = 3.43, p < .05$), meaning that participants were more satisfied with the choice of the class *Sad* than with the class *Angry* (a-s: $p = .037$, $-.20, 95\% - CI[-.39, -.01]$).

4.6 Satisfaction between Emotion Classes

We also analysed the aggregated satisfaction ratings (all synthesisers together) for the emotions to investigate whether satisfaction with the choice for one of the emotions was rated distinctly better. However, no significant effects were found.

Satisfaction within Individual Emotion Classes. *Which synthesiser elicits the highest satisfaction ratings for individual emotion classes?* To answer this question, we conducted ANOVAs for each of the emotion classes and synthesisers.

For all emotion classes, we found significant differences (*Angry*: $F(2, 162) = 7.56, p < .01$; *Happy*: $F(2, 162) = 22.199, p < .001$; *Sad*: $F(2, 162) = 24.459, p < .001$) between the synthesisers, where post-hoc analysis identified eSpeak as significantly less satisfactory when choosing the emotion class than both of the other synthesisers. In addition, for the emotion class *Sad*, the post-hoc pairwise comparisons also revealed that satisfaction with emotions generated by CereVoice resulted in significantly higher ratings than with MaryTTS.

4.7 Pronunciation / Emotion

In the final part of the online survey, participants had to rate both the pronunciation and the synthesised emotions on a Likert scale from very poor (1) to very good (5). The synthesisers were presented one after the other, and for each synthesiser sentences with all three emotion classes were generated, which could be listened to by the participant as often as desired before both ratings were given. Table 3 contains the mean values and standard deviations for both ratings. The mean values indicate that eSpeak was rated worst and CereVoice was rated best for pronunciation as well as synthesised emotions.

Rating of Pronunciation / Emotion (1 .. 5)			
Emotion	E - eSpeak	M - MaryTTS	C - CereVoice
Pronunciation	$M = 1.74, SD = .93$	$M = 2.77, SD = .99$	$M = 3.27, SD = .89$
Emotion	$M = 1.60, SD = .65$	$M = 3.88, SD = .95$	$M = 4.06, SD = .78$

Table 3. Means and standard deviations for the ratings of pronunciation and synthesised emotions on a scale of 1 ... 5.

Rating of Pronunciation / Synthesised Emotion. A repeated measurement ANOVA on the ratings for pronunciation showed significant differences between the synthesisers ($F(2, 162) = 88.83, p < .001$). Post-hoc pairwise comparisons revealed that pronunciation of the generated sentences were rated from best to worse in the order CereVoice, MaryTTS, and eSpeak (E-M: $p < .001$, $-1.02, 95\% - CI[-1.32, -.72]$; E-C: $p < .001$, $-1.52, 95\% - CI[-1.81, -1.24]$; M-C: $p < .001$, $-.50, 95\% - CI[-.72, -.23]$).

A similar picture could be found for the ratings of the synthesised emotions. A repeated measurement ANOVA on the ratings for synthesised emotions showed significant differences between the synthesisers ($F(2, 162) = 345.47, p < .001$). The post-hoc pairwise comparisons identified the synthesiser eSpeak as worse than MaryTTS and CereVoice (E-M: $p < .001$, $-2.28, 95\% - CI[-2.56, -2.0]$; E-C: $p < .001$, $-2.46, 95\% - CI[-2.70, -2.23]$) in terms of the synthesised emotions.

5 Discussion

As Aylett et al. [1] mentioned, the time to only mimicry the naturalness of human voice is over. People especially those with CCN (*Complex Communication Needs*) have a great need for speech synthesisers that are able to convey a variety of expressive styles in a natural manner. This aspect is also important in light of the rapidly increasing speech interaction and its acceptance in smarthomes [6], to respond appropriately to the emotions of residents [7].

Researchers spent decades in developing natural sounding TTS (*Text-to-Speech*) incorporating prosodic elements with different approaches. As shown in Table 1, there are differences in correctly assessing emotions per synthesiser (in decreasing order: CereVoice [M = 2.78, SD = 1.42], MaryTTS [M = 2.62, SD = 1.54], eSpeak [M = 1.78, SD = 1.31]). These results are confirmed in the final part of the online survey (see Table 3). It comes as no surprise that the quality of the single synthesisers provided different, but consistent subjective assessments as they are based on different underlying techniques: formant synthesis (eSpeak), HMM-based synthesis (MaryTTS) and unit selection (CereVoice).

Our results are in line with previous studies investigating the quality of different types of speech synthesisers (see, for example, the chapter on Perceptual Quality Dimension by [10]). Formant synthesis tends to sound mechanical and artificial while the greatest amount of naturalness is typically achieved with unit selection. Even though we did not exploit the full potential of MaryXML and CereProc XML to control the quality of the expressive speech, MaryTTS and CereVoice performed better in terms of expressivity than eSpeak. CereProc showed the best results both in terms of satisfaction with the pronunciation, i.e. naturalness, and ability to convey emotional states as a whole, i.e. expressiveness. While it can be argued that we only used simple markups, we have to take into account that CCN users need to be able to control their voices in an easy and quick manner. The next step would be to integrate capabilities for expressive speech into EmotionTalker by enabling AAC users to specify emotions at a higher level of abstraction, but still communicate the intended expressive style in a believable manner.

The current research complements our previous research on the evaluation of EmotionTalker, a first prototype of a VOCA interface that included Emojis to enable people specify the intended emotion. For this experiment, we relied on a small number CCN users who tested EmotionTalker in their daily environment. Even though we aimed to include AAC users in our current evaluation by contacting an AAC forum, the current evaluation was not specifically addressed to

AAC users. This was due to our focus on a perceptive study with a large number of users. For the online survey, it could be objected that we could not control the participants' surroundings and their equipment for listening to the sentences. However, to complete the survey, participants had to listen to all sentences with all synthesisers. So, they had a direct comparison.

6 Conclusion

Our objective was to identify a natural speech synthesiser with variability in expressive style for integration into a VOCA. To this end, we evaluated the ability of three synthesisers (eSpeak / MaryTTS / CereVoice) to convey emotionally neutral utterances in a happy, sad, or angry manner. Our assumption that CereVoice has the best capabilities was confirmed. In our online survey most of the 82 participants rated CereVoice better than MaryTTS - eSpeak was rated worst. As outlined by [4], people with CCN may have deficits in building emotional competencies during childhood. In order to improve their capabilities, it would have potential to equip VOCAs with ESS and usable input methods. CereVoice seems to be an adequate candidate, as our findings show.

The next step will be to extend our tests with EmotionTalker. We plan to have people with CCN test EmotionTalker in their own environment in specified situations over one week. It has to be shown if they can socialise more easily with a VOCA capable of ESS. Furthermore, novel synthesis paradigms should be taken into account, see the recent developments on the MaryTTS architecture to enable synthesis based on Deep Neural Networks [14] or the recently announced neural speech synthesis system CereWave AI by CereProc Ltd.[8]

Acknowledgements

The work presented here is partially supported by 'PROMI - Promotion inklusive' and the employment centre. We thank the students, Lena Tikovsky and Ewald Heinz, for their contribution to this work.

References

1. Aylett, M.P., Cowan, B.R., Clark, L.: Siri, echo and performance: You have to suffer darling. In: Conference on Human Factors in Computing Systems, Extended Abstracts. ACM, New York, NY, USA, Glasgow, Scotland, UK (2019), https://doi.org/10.1145/3290607.3310422
2. Aylett, M.P., Pidcock, C.J.: Adding and controlling emotion in synthesised speech. Tech. Rep. UK patent GB2447263A (2008)
3. Aylett, M.P., Vinciarelli, A., Wester, M.: Speech synthesis for the generation of artificial personality. IEEE Trans. Affect. Comput. **11**(2), 361–372 (2020). https://doi.org/10.1109/TAFFC.2017.2763134

[8] https://www.cereproc.com/en/v6 (accessed 11/06/2020)

4. Blackstone, S.W., Wilkins, D.P.: Exploring the importance of emotional competence in children with complex communication needs. Perspectives on Augmentative and Alternative Communication **18**(3), 78–87 (2009). https://doi.org/10.1044/aac18.3.78
5. Chafe, W.: Prosody: The music of language. In: Genetti, C., Adelman, A. (eds.) How languages work - an introduction to language and linguistics, pp. 236–256. Cambridge University Press, Cambridge, UK, 2nd edn. (2019)
6. Dang, C.T., Andre, E.: Acceptance of autonomy and cloud in the smart home and concerns. In: Dachselt, R., Weber, G. (eds.) Mensch und Computer 2018 (MuC 2018) - Tagungsband (2018)
7. Dang, C.T., Aslan, I., Lingenfelser, F., Baur, T., André, E.: Towards somaesthetic smarthome designs: Exploring potentials and limitations of an affective mirror. In: Proceedings of the 9th International Conference on the Internet of Things. IoT 2019, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3365871.3365893
8. Fiannaca, A.J., Paradiso, A., Campbell, J., Morris, M.R.: Voicesetting: Voice authoring uis for improved expressivity in augmentative communication. In: Mandryk, R.L., Hancock, M., Perry, M., Cox, A.L. (eds.) Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI 2018, Montreal, QC, Canada, April 21-26, 2018. p. 283. ACM (2018). https://doi.org/10.1145/3173574.3173857
9. Girden, E.R.: ANOVA: Repeated measures. Sage (1992)
10. Hinterleitner, F. (ed.): Quality of Synthetic Speech. T-Labs Series in Telecommunication Services, Springer (2017). https://doi.org/10.1007/978-981-10-3734-4
11. Hoffmann, L., Wülfing, J.–.O.: Usability of electronic communication aids in the light of daily use. In: Proceedings 14th Biennial Conference of the International Society for Augmentative and Alternative Communication. p. 259 (2010)
12. Murray, I.R., Arnott, J.L.: Applying an analysis of acted vocal emotions to improve the simulation of synthetic speech. Computer Speech and Language **22**(2), 107–129 (2008). https://doi.org/10.1016/j.csl.2007.06.001
13. Schröder, M., Charfuelan, M., Pammi, S., Steiner, I.: Open source voice creation toolkit for the MARY TTS platform. In: INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011. pp. 3253–3256. ISCA (2011)
14. Steiner, I., Maguer, S.L.: Creating new language and voice components for the updated marytts text-to-speech synthesis platform. In: Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., Tokunaga, T. (eds.) Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018. European Language Resources Association (ELRA) (2018)
15. Wagner, P., Beskow, J., Betz, S., Edlund, J., Gustafson, J., EjeHenter, G., LeMaguer, S., Malisz, Z., Székely, E., Tåannander, C., Voße, J.: Speech synthesis evaluation - state-of-the-art assessment and suggestion for a novel research program. In: Proceedings of the 10th ISCA Speech Synthesis Workshop. pp. 105–110 (2019). https://doi.org/10.21437/SSW.2019-19
16. Wülfing, J.–.O., André, E.: Progress to a voca with prosodic synthesised speech. In: Miesenberger, K., Kouroupetroglou, G. (eds.) Computers Helping People with Special Needs, vol. LNCS 10896, pp. 539–546. Springer, Cham, Swiss (2018). https://doi.org/10.1007/978-3-319-94277-3_84