PERCEPTUAL QUALITY DIMENSIONS OF TEXT-TO-SPEECH SYSTEMS IN AUDIOBOOK READING TASKS

Florian Hinterleitner[1], *Christoph Norrenbrock*[2], *Sebastian Möller*[1]

[1]*Quality and Usability Lab, TU Berlin, Germany*
[2]*Digital Signal Processing and System Theory, CAU Kiel, Germany*
florian.hinterleitner@tu-berlin.de

Abstract: In this paper we present research on perceptual quality dimensions of text-to-speech systems in audiobook reading tasks. Therefore, we proposed a newly developed evaluation protocol for the assessment of synthetic speech in audiobook reading tasks for the Blizzard Challenge 2012. We illustrate the experimental setup of the special audiobook reading task of the Blizzard Challenge 2012 and analyze and interpret the results of the subjective listening test. Via a factor analysis, two quality dimensions could be extracted. Through the correlation between the values of the rating scales and the factor values, the dimensions could be assigned to prosody & rhythm and to the listening pleasure of the user. This confirms the results of the previous study in which the current evaluation protocol was created. Also, a comparison with the perceptual quality dimensions of text-to-speech systems in different use cases led to significant similarities.

1 Introduction

From the beginnings of text-to-speech (TTS) to the high quality voices of modern unit-selection or HMM synthesizers, the synthetic generation of speech has come a long way. Especially, the continuous improvement of TTS quality over the past years lead to synthetic voices that no longer sound like entirely emotionless computers. Even though, TTS can still easily be distinguished from human speech, it is now possible to deploy synthetic voices in more and more different use cases. Every day applications like email readers, information systems and smart-home assistants can already be considered as "classic" domains of TTS. Today, even more challenging tasks such as audiobooks come into focus.

With this new application area, additional quality aspects of TTS like listening effort and the ability for emotional speech get more important. Therefore, standard listening tests like articulation and intelligibility tests [1], comprehension tests [2], and overall quality tests [3] are not specialized on measuring quality aspects that became significant through the recent extension of the use cases for TTS systems.

Within the scope of a bachelor thesis [4] a set of 11 scales was developed particularly for the quality assessment of TTS in audiobook reading tasks. These scales were partially based on the ITU-T Rec. P.85 [3]. Furthermore, some scales were added to cover specific quality aspects that should be considered when evaluating TTS audiobooks like speech pauses and the ability for emotional speech. Additionally, passages from 8 different books were collected, synthesized by 2 different TTS systems and judged by 23 naïve subjects in a listening test. A Principle Axis Factor (PAF) analysis led to 2 perceptual quality dimensions which could be assigned to listening pleasure and prosody & rhythm. Furthermore, the analysis of the results led to changes in the evaluation protocol.

The Blizzard Challenge (BC) is an annual competition for developers of speech synthesis systems. In 2012 the scope of the BC was extended and thus for the first time it also featured an audiobook reading task. The online listening test consisted of 7 rating scales that were taken from the previous modified evaluation protocol. In this paper we analyze the results of this task. Furthermore, we investigate if the aforementioned perceptual quality dimensions can be confirmed with the modified evaluation protocol in a large-scale study like the BC 2012.
In Section 2 we give an overview of the experimental setup including a short introduction to the BC, the test database that was used in the experiment as well as the design of the rating scales and the test procedure. The analysis of the test data consisting of a factor analysis and the interpretation of the resulting quality dimensions is shown in Section 3. Moreover, since TTS audiobook readers are likely to feature perceptual quality dimensions that differ or extend the set of quality dimensions that were found in previous studies [5][6], we compare the results from this study with the perceptual quality dimensions of previous work in Section 4. Finally, in Section 5 we conclude the outcome of this study.

2 Experimental setup

This section gives a brief description of the annual BC. Moreover, it shows an overview of the books that were synthesized and presented in the listening test. We define the scales that were used to evaluate the stimuli and we describe the test procedure of the online listening test.

2.1 Blizzard Challenge

Since 2005 the BC gathers developers of TTS systems to compare techniques in building corpus-based synthesizers. The fact that all participants get the same speech corpus to build their systems on assures a comparability between all synthesizers. In 2012 for the first time the BC ran a special task concerning audiobooks read by TTS systems.

Table 1: Selected books for the audiobook task of the BC.

AUTHOR	BOOK
Jane Austen	Emma
Charles Dickens	Oliver Twist
Arthur Conan Doyle	The Hound of the Baskervilles
Alexandre Dumas	The Three Musketeers
Jerome K. Jerome	Three Men in a Boat
Franz Kafka	The Trial
Edgar Allan Poe	The Fall of the House of Usher
Mary Shelley	Frankenstein or the Modern Prometheus
Mark Twain	Alonzo Fitz
Mark Twain	Those Extraordinary Twins
Jules Verne	Twenty Thousand Leagues Under the Sea
H.G. Wells	Time Machine
P.G. Wodehouse	My Man Jeeves

2.2 Test database

Since the online listening test was open to the public, the texts that were synthesized were limited to copyright-free and out-of-copyright sources [7] [8]. We chose passages out of 13 different books and authors with a mean stimulus duration of 44.5s. We selected passages with the

idea to cover a wide variety of writing styles and book categories. The full list of authors and books can be seen in Table 1.

2.3 Scales and test procedure

The online listening test consisted of 9 sections of which 2 were dedicated for the evaluation of audiobook stimuli. One of the audiobook sections also included a natural reference voice while the other one did not. Both sections consisted of 10 synthetic stimuli from 10 different systems. All in all 230 audiobook stimuli were evaluated during the listening test. Each stimulus was rated by at least 20 participants.
Taking into account the results of [4] [9], as well as the ITU-T Rec. P.85 [3], the following 7 scales were chosen for the evaluation: overall impression, voice pleasantness, speech pauses, word stress, intonation, emotion, and listening effort. The scores were given on a continuous slider. Each participant was giving the score as the distance of the handle from the left end of the slider, but could not see the actual number. We decided to use scales with separate scale and end points as proposed in [10] to ease the effect that often occurs on rating scales: most users try to avoid ratings on both ends of the scales because they expect to rate even better/worse stimuli. Moreover, standard rating scales make it hard for subjects to differentiate between either very good or very bad stimuli. A screenshot of the GUI from the online listening test can be seen in Figure 1.

3 Data analysis

3.1 Factor analysis

A Principal Axis Factor (PAF) analysis was conducted on the 7 items (scales). The Kaiser-Meyer-Olkin (KMO) measure verified the sampling adequacy for the analysis, KMO = .90, and all KMO values for the individual items were > .86, which is well above the acceptable limit of .50. Bartletts's test of sphericity, $\chi^2(21) = 21459.16$, indicated that correlations between the items were sufficiently large for a PAF. An initial analysis was run to obtain the eigenvalue for each component in the data. Only one item had an eigenvalue over Kaiser's criterion of 1. The scree plot also indicated to retain one component. However, given the very large sample size (4809), the easy interpretability after an extraction of 2 factors, and the fact that a 2-factor model allows a deeper insight into the human perception of TTS signals, we chose to retain 2 components in the final analysis. These 2 factors together explained 75.61% of the total variance.
Since we assumed correlated quality dimensions we opted for an oblique rotation method (Promax rotation with $\kappa = 4$). The resulting factor pattern matrix can be seen in Table 2. For clarity, values below .40 are suppressed. Due to the oblique rotation the factors are not orthogonal, rather they correlate with .78.

3.2 Resulting quality dimensions

To achieve a meaningful interpretation of the resulting quality dimensions, we omitted all scales with high cross-loadings, meaning high loadings on both factors, before the interpretation. In our case this applies to the item *listening effort*.
Given the high loadings of the scales *intonation* and *stress* and the medium loading of *speech pauses* on factor 1, this dimension is clearly linked to **prosody & rhythm**. The scale *emotion* which assesses the ability of the TTS systems to synthesize appropriate emotions, i.e., through prosodic modulations, also contributes to this interpretation.

Table 2: Factor pattern matrix.

	FACTOR LOADINGS	
SCALE	1	2
intonation	.86	
stress	.76	
emotion	.60	
speech pauses	.54	
voice pleasantness		.90
overall impression		.87
listening effort	.44	.47

Note: For better readability values below .40 are suppressed.

Dimension 2, however, correlates highly with the scales *voice pleasantness* and *overall impression*. This indicates that dimension 2 is related to the voice of the speaker as well as the naturalness of the signal and the overall experience. Thus, this dimension can be associated with the **listening pleasure**. The affiliation of the scale *overall impression* indicates that this dimension is in fact more important for the quality impression of the listener than dimension 1.

4 Comparison with perceptual quality dimensions from previous work

Comparing the results from this study with previous work on perceptual quality dimensions in audiobook reading tasks [9] reveals major similarities. The assignment of scales to quality dimensions is nearly the same as in the previous study. Thus, both dimensions represent about the same perceptual quality impression of the user as before. Nevertheless, there are some differences: first of all, the scale *listening effort* which highly correlated with the dimension *listening pleasure* before has very high cross-loadings in the current study and thus did not help to discern between the factors.

Interestingly, the importance of both dimensions changed in the current study. While in [9] the dimension *listening pleasure* explained most of the variance in the data and was thus labeled as *dimension 1*, the order of importance is reversed in the current study. Nonetheless, since the scale *overall impression* correlates highly with factor 2 this dimension is clearly the most relevant when it comes to the overall quality impression of the user.

Even in comparison to the results of multidimensional studies on TTS used in domains such as email and short message readers as well as smart-home systems more similarities can be found. The broad dimension *naturalness* from [5] which correlates highly with scales like naturalness, voice pleasantness, accentuation and rhythm in fact covers both dimensions from this study (*prosody & rhythm* and *listening pleasure*). With regard to the dimensions found in [6] we assigned the dimension *listening pleasure* to the dimension *naturalness of voice* which correlates highly with scales like voice pleasantness, naturalness, and intelligibility. Whereas the dimension temporal distortions from [6], with its high correlation with the scale rhythm, includes the prosodic and rhythmic part of dimension 1.

5 Conclusions

In the audiobook task of the BC 2012 we tested the evaluation protocol that was designed especially for the subjective quality assessment of TTS in audiobook reading tasks. The set of pre-

sented scales was adjusted according to the recommendations in [9]. The gathered data was analyzed via a PAF with Promax rotation. 2 factors were extracted that represent *prosody & rhythm* and *listening pleasure* of the user. We found great similarities with the dimensions extracted previously [9]. They were even found to be similar to the perceptual quality dimensions of TTS used in different use cases. This shows that prosody and rhythm as well as the naturalness of the voice play a crucial role across different domains of synthetic speech.

6 Acknowledgements

The present study was carried out at Quality and Usability Lab, TU Berlin. It was supported by the Deutsche Forschungsgemeinschaft (DFG), grants MO 1038/11-1 and HE 4465/4-1. The authors would like to thank Simon King and Vasilis Karaiskos from the Blizzard Challenge team for their support.

References

[1] R. Van Bezooijen and V.J. van Heuven. Assessment of Speech Output Systems. In D. Gibbon, R. Moore, and R. Winski, editors, *Handbook of Standards and Resources for Spoken Language Systems*, pages 481–563, Berlin, 1997. Mouton de Gruyter.

[2] C. Delogu, S. Conte, and C. Sementina. Cognitive Factors in the Evaluation of Synthetic Speech. In *Speech Communication*, pages 153–168, 1998.

[3] ITU-T Rec. P.85. *A Method for Subjective Performance Assessment of the Quality of Speech Voice Output Devices*. International Telecommunication Union, Geneva, 1994.

[4] G. Neitzel. *Entwicklung eines Evaluationsverfahrens zur Bestimmung der Qualität synthetisch erzeugter Hörbücher*. Bachelor Thesis, Quality and Usability Lab, TU Berlin, 2011.

[5] F. Hinterleitner, S. Möller, C. Norrenbrock, and U. Heute. Perceptual Quality Dimensions of Text-to-Speech Systems. *Proc. of the 12th Annual Conference of the International Speech Communication Association (Interspeech 2011)*, pages 2177–2180, 2011.

[6] F. Hinterleitner, C. Norrenbrock, S. Möller, and U. Heute. What makes this voice sound so bad? A multidimensional analysis of state-of-the-art text-to-speech systems. *Proc. of the 2012 IEEE Workshop on Spoken Language Technology (SLT 2012)*, pages 240–245, 2012.

[7] Project Gutenberg. http://www.gutenberg.org.

[8] Many Books. http://manybooks.net/.

[9] F. Hinterleitner, G. Neitzel, S Möller, and C. Norrenbrock. An Evaluation Protocol for the Subjective Assessment of Text-to-Speech in Audiobook Reading Tasks. In *Proceedings of the Blizzard Challenge Workshop. International Speech Communication Association (ISCA)*, Turin, Italy, 2011.

[10] S Möller. *Assessment and Prediction of Speech Quality in Telecommunication*. Kluwer Academic Publishers, Boston, 2000.

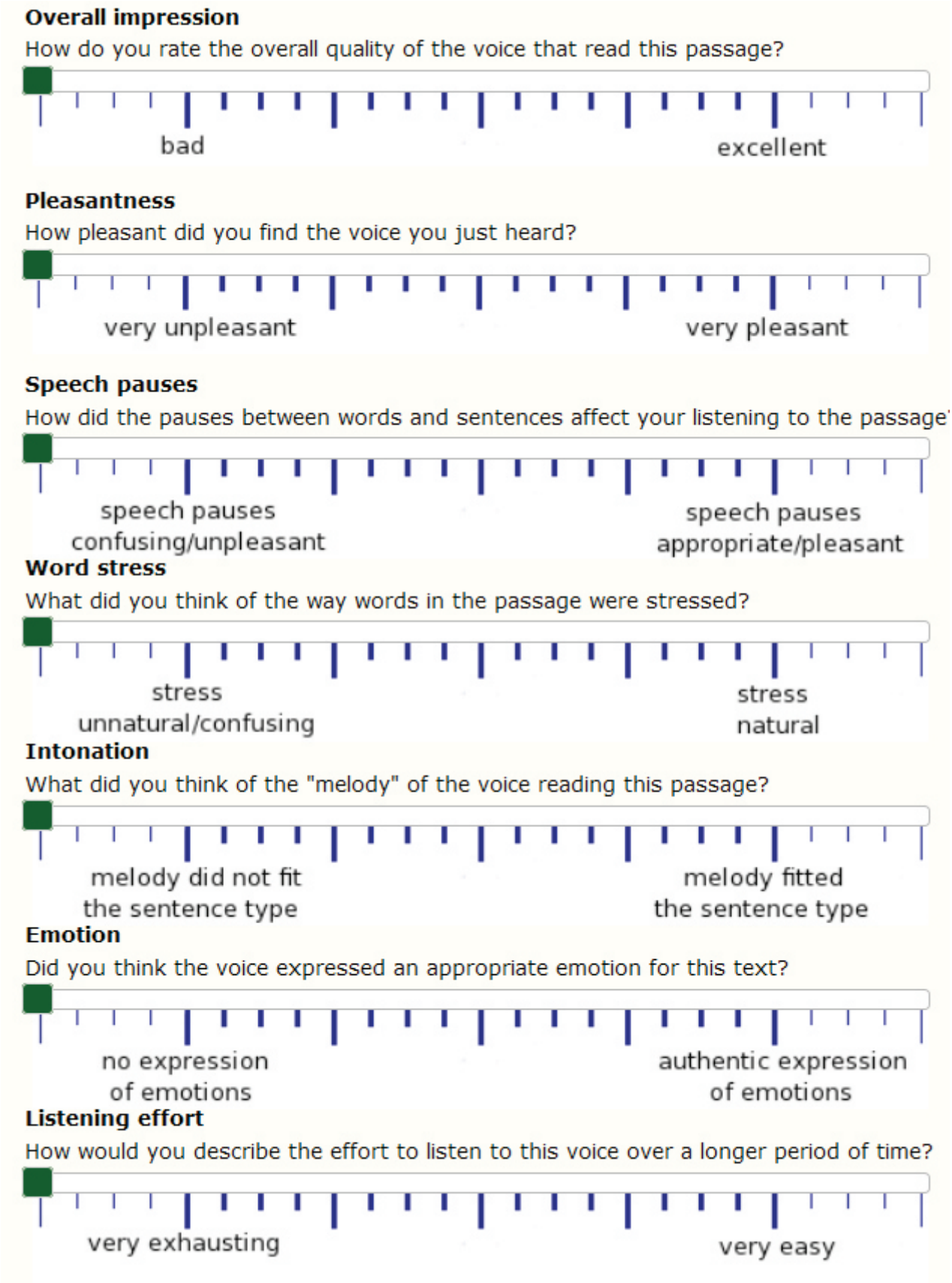

Figure 1: Screenshot of the scales presented in the online listening test.