An Evaluation Protocol for the Subjective Assessment of Text-to-Speech in Audiobook Reading Tasks

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

[1]Quality and Usability Lab, Deutsche Telekom Laboratories, TU Berlin, Germany
[2]Digital Signal Processing and System Theory, CAU Kiel, Germany

florian.hinterleitner@telekom.de, georginaneitzel@yahoo.de,
sebastian.moeller@telekom.de, cno@tf.uni-kiel.de

Abstract

This paper presents an evaluation protocol for the subjective assessement of text-to-speech in audiobook reading tasks. We developed a questionaire with 11 scales an tested it on TTS data from 4 different synthetic voices, plus one optimized version.
A MANOVA on the data gathered with the questionnaire showed that the *text type* has a significant influence on 7 of the 11 scales. Moreover, the *level of familiarity* does not have any influence on the ratings.
A subsequent Principal Axis Factor (PAF) analysis with Promax rotation resulted in 2 underlying dimensions. The first factor represents the *listening pleasure* the tested systems achieved. The second dimension comprises scales that evaluate the *prosody* of the synthesized speech signal.
After the analysis of the results we propose to perform slight modifications to the developed questionaire.
Index Terms: speech synthesis, evaluation protocol, audiobooks

1. Introduction

The constant improvements of text-to-speech (TTS) systems over the past decade lead to synthetic voices that no longer remind listeners of the robot-like voices from the eighties, but of real human beings. Even though they can still easily be distinguished from human-produced speech the increase of naturalness made it possible to use TTS for every day applications like email readers, information services, or smart-home assistants. Even more challenging tasks such as audiobooks come into focus.
Since 2005 the Blizzard Challenge (BC) gathers developers of TTS systems to compare techniques in building corpus-based speech synthesizers. The fact that all participants get the same speech corpus to build their systems on assures a comparability between all synthesizers. This year the BC has announced a special task concerning audiobooks read by TTS systems for 2012. With this new application area quality aspects of TTS like listening effort, the ability for emotional speech or the placing of speech pauses in a way that supports the comprehension of the text, get more important.
Depending on which aspect of the system is to be evaluated, different types of listening tests are recommended: articulation and intelligiblity experiments [1] test whether the TTS signal is able to carry information on a segmental or supra-segmental level; comprehension tests [2] show if the listener can discern the content; and overall quality tests [3] capture different global quality aspects and dimensions [4]. However, non of these methods is specialized on measuring quality aspects like the ones mentioned in the previous paragraph. Therefore new ways of evaluating the quality of TTS systems have to be designed.
This paper presents an evaluation protocol developed especially for the subjective assessment of text-to-speech in audiobook reading tasks. In Section 2 we give an overview of the experimental-setup including the used speech material, the TTS stimuli, the design of the experiment and the test procedure. The analysis of the test data and a discussion of the results is presented in Section 3. Finally, in Section 4 suggestions on the use of the proposed test protocol in the upcoming Blizzard Challenge are given.

2. Experimental-setup

2.1. Test preparation

2.1.1. Speech material

We used passages from the German issues of the books in Table 1 as material for the listening test. We tried to cover a wide variety of writing styles and book categories including thrillers, funny books, action-packed passages, books for children, books with very long sentences, and passages containing almost only direct speech.

2.1.2. TTS systems and stimuli

We used the German synthesizers CereProc CereVoice (CV) with the voices Alex (m) and Gudrun (f) and the IVONA (IV) TTS system with the voices Hans (m) and Marlene (f). Each of the systems was used to synthesize the same passage from the books listed in Table 1. TTS systems that had problems synthesizing English names had to be manually adjusted to ensure a normal pronunciation. Additionally, 4 samples were synthesized by a manually optimized version of the IVONA (IVO) voice Marlene. The optimization included an adjustment of wrong articulated words and an improvement of pauses between sentences and paragraphs. The mean length of all stimuli was 54.7s with an average of 138 words.

2.1.3. Rating scales

In our experiment, we modified selected items which have already been proven useful for the evaluation of TTS systems [3]. Furthermore, we added new items to assess specific quality aspects that should be considered when evaluating TTS audiobooks. The set contained the 11 items listed in Table 2.
The additional items were selected based on the review of current literature. Prosodic elements like communicative, structuring, aesthetic, and emotional aspects can be seen as the

ID	Category	Author	Book
1	Long sentences	Sven Regener	*Der kleine Bruder*[1]
2	Direct speech, incomplete sentences	Douglas Adams	*The Hitchhiker's Guide to the Galaxy*
3	Higher level of lexis, complex sentence structure	Charles Dickens	*The Adventures of Oliver Twist*
4	Poetic, picturesque	Antoine Saint-Exupéry	*Wind, Sand and Stars*
5	Direct speech, basic language	Tommy Jaud	*Resturlaub*[1]
6	Action, short sentences	Thomas Harris	*Hannibal*
7	Children's book	Astrid Lindgren	*Pippi Longstocking*
8	Thriller	Ken Follett	*Code to Zero*

Table 1: Books that were used for the listening test.

most important factors for reading and interpreting books [5], most of the selected scales are focused on prosodic evaluation (ACCT, COPR, SPPA, INT, EMO).

ITU-T Rec. P.85	Additional items
Overall impression (MOS)	Speech pauses (SPPA)
Voice pleasantness (PLT)	Intonation (INT)
Accentuation (ACCT)	Emotion (EMO)
Listening effort (LSTE)	Content (CONT)
Comprehension problems (COPR)	Level of familiarity (LOF)
Acceptance (ACCP)	

Table 2: Items that were selected for the listening test.

In the following, the scales from ITU-T Rec. P.85 [3] as well as the additional items are described. The questionaire as it was used in the listening test can be seen in Figure 2.

Overall impression
This scale evaluates the overall quality of the synthesized signal.

Voice pleasantness
Measures the degree of voice pleasantness from unpleasant to pleasant.

Accentuation
Since unnatural stress' and accentuations are often perceived as very annoying and thus also have a great influence on the text comprehension [5] we used the scale from ITU-T Rec. P.85 with slight modifications.

Listening effort
Describes the effort a listener is required to make when listening to this voice over a longer period of time.

Comprehension problems
This scales captures any comprehension problems that might occur due to badly synthesized speech.

Acceptance
The acceptance-item from the ITU-T Rec. P.85 questionnaire was modified into a continous rating scale in order to make it possible to determine whether this factor falls into one of the main factors.

Speech pauses
Evaluates if punctuation marks (e.g. period, comma, question mark, exclamation mark, colon, etc.) have been converted into appropriate speech pauses between words, sentences, and paragraphs in a way that supports the comprehension of the text [6].

Intonation
This scale captures if the produced pitch curve fits to the type of sentence, e.g. the pitch of interrogative sentences usually increases at the end of a sentence whereas the pitch of declarative sentences decreases [5] [6].

Emotion
Variation of emotion is achieved by variations of sound pressure, intonation, speech pauses and volume [5]. To ensure an authentic reading experience, the voice should reflect the atmosphere of the scene and the moods of the characters [7].

The purpose of the items CONT and LOF was to test, if these scales influence the subject's judgement on the other scales. The items PLT, LSTE, and ACCP were taken from ITU-T Rec. P.85 to assess the subjects ease of listening.

2.2. Test procedure

25 naïve subjects (13 female, 12 male) aged between 19 and 32 years (average 25 years) took part in the test. All of them were native German speakers. Non of them suffered from any hearing problems or dyslexia. All subjects were paid for their participation. The stimuli were presented via headphones (AKG K601) in a sound proof booth. The listening test was designed within subjects, i.e. all participants listened to all stimuli.
We decided to use separate scale and end points as proposed in [8] to reduce the impact of two effects often noticed when subjects use rating scales: first, most subjects avoid to give ratings on the end of scales because they expect even better or worse stimuli to come. Moreover, scales with fixed end points make it hard for subjects to differentiate between stimuli of an either very good or very bad quality.
The subjects were instructed to first rate the overall impression of the stimulus on a continuous rating scale ranging from *bad* to *excellent*. Subsequently, quality estimates for the other 9 scales had to be given via a slider presented on the GUI. While adjusting the slider the selected value (1 to 7) was shown above each slider. In the end, the subjects had to rate if they knew the presented chapter prior to the listening test. To avoid any impact with regard to the order, the sequence of scales (except *overall impression* and *level of familiarity*) was randomized between subjects. To make themselves familiar with the test procedure

[1] no English translation available

all subjects first had to pass a training phase with 2 stimuli that were not included in the main test. The main test consisted of 2 blocks with 18 stimuli and a 5 minute break in-between.
After the test, boxplots with the ratings of every participant on all scales were computed for all stimuli. 2 subjects had more then 5% outlier ratings and were thus excluded from the dataset.

3. Analysis and discussion

3.1. Multivariate analysis of variance (MANOVA)

We run a MANOVA to examine how the dependent variables *text type* and *voice* (CV, IV and IVO with male and female voices) behave simultaneously and to find out more about the relationship between the scales (dependent variables).

3.1.1. Assumptions of MANOVA

The assumptions under which MANOVA is reliable are the following [9]:

- Independence: the observed scores should be statistically independent.
- Random sampling: the data should be randomly sampled from the population of interest and measured at least at an interval level.
- Multivariate normality: the dependent variables collectively have multivariate normality with groups. This assumption cannot be tested in SPSS but it provides tests on univariate normality for each dependent variable in turn. Univariate normality is a necessary condition for multivariate normality, but not a sufficient condition.
  Therefore we used the Shapiro-Wilk test, a powerful statistical test on normal distribution, especially for small sample sizes ($N < 50$) [10]. We examined each combination of *voice* and *text type* on normal distribution. Less distributions, clustering no specific sample, text type or scale, deviate significantly from normality ($p < 0.05$). One reason for that could be the small sample size ($N < 30$) or the fact, that the Shapiro-Wilk test is very sensitive to outliers [11].
- Homogenity of covariance matrices: the variance of each dependent variable and also the correlation between any two dependent variables in the same groups should be homogenous.
  We used the Levene's test to check this assumption, which should ideally be non significant ($p > 0.05$) for each scale. MOS ($p = 0.009$), ACCT ($p = 0.032$), and EMO ($p = 0.030$) were significant, which means that the variance of the underlying groups is significantly different from each other.
  The variance-covariance matrices were checked by the Box's test, which should also be non significant ($p > 0.05$), if the matrices are equal. In our case, the Box's test was highly significant ($p < 0.01$), which means that the variance-covariance matrices are not equal. One reason for that could be the relatively high number of assumed independent variables ($N = 9$) in our experiment.
  If this assumption is violated, MANOVA is still relatively robust as long as the sample sizes are equal [9]. We decided to exclude the optimized version of IVONA Marlene for our analysis, because only 4 of the 8 possible text samples were synthesized and tested with this voice. Thus the sample size for each of the groups is the same.

Finally, the assumption of multivariate normality was broken, so the accuracy of the MANOVA can be seen as compromised.

3.1.2. Results of MANOVA

Multivariate Test Statistics
We computed the multivariate test statistics (Pillai's Trace V, Wilk's Lambda Λ, Hotelling's Trace T, Roy's Largest Root Θ) for assessing the statistical significance of any MANOVA effect.
For each main effect (*voice*, *text type*) the test statistics were highly significant ($p < 0.01$), which indicates that each of the independent variable has a significant influence on the rating scales.
For the test statistics of the *combined effect*, opposed results were found: V ($p = 0.736$), Λ ($p = 0.722$), T ($p = 0.707$) were not significant while Θ ($p < 0.01$) was highly significant. A basic criterion for selecting a valid test statistic is, if the four assumptions of MANOVA were met or not. Because of its robustness under violated assumptions V ($p = 0.736$) is the best indicator of significance in this case, while Θ ($p < 0.01$) is the most powerful test statistic if assumptions are met [12]. As a result, our model does not have any interactions between *voice* and *text type*.

Main effects
The main effect shows the effects of the independent variables *voice* and *text type* on each scale. Table 3 displays the effect of *voice*.

Scales	F	p
PLT	49.791	0.000
SPPA	12.262	0.000
ACCT	43.592	0.000
INT	28.673	0.000
EMO	24.960	0.000
LSTE	58.382	0.000
COPR	37.379	0.000
ACCP	49.293	0.000
CONT	1.092	0.352
MOS	61.288	0.000

Table 3: Effect of *voice* on each scale.

The main effect of voice is shown by the F-ratio for each scale: the most variation of variance through voice is explained by MOS ($F = 61.288$) and LSTE ($F = 58.382$), while for the scale CONT ($p = 0.352$) *voice* accounts for the same variance as the error. Consequently, for each scale (except CONT) the result is highly significant ($p = 0.000$), which means that the factor *voice* has a significant influence on these rating scales.

The effect of the factor *text type* is displayed in Table 4. In comparison to the factor *voice*, the results for factor *text type* show a minor influence on the scales. The highest values for the F-ratio are reached on the scales CONT ($F = 5.768$), EMO ($F = 3.747$), and SPPA ($F = 3.365$). The fact, that the value for ACCP ($F = 0.620$) is less than 1 indicates more unexplained

Scales	F	p
PLT	2.209	0.032
SPPA	3.365	0.002
ACCT	2.496	0.015
INT	2.155	0.036
EMO	3.747	0.001
LSTE	1.978	0.056
COPR	3.288	0.002
ACCP	0.620	0.740
CONT	5.768	0.000
MOS	1.440	0.186

Table 4: Effect of *text type* on each scale.

variance than the variance that could be explained by *text type*. Accordingly, PLT, SPPA, ACCT, INT, EMO, COPR, and CONT are significantly influenced by the *text type* ($p < 0.005$), where as LSTE, ACCP, and MOS are not significantly affected by the type of text.

3.1.3. Effect size of MANOVA

To get an objective measure of the importance of the experimental effects, standardized effect sizes (ω) are shown in Table 5.

	voice	*text type*
ω_{PLT}	0.824	0.223
ω_{SPPA}	0.573	0.305
ω_{ACCT}	0.806	0.247
ω_{INT}	0.739	0.210
ω_{EMO}	0.714	0.340
ω_{LSTE}	0.845	0.202
ω_{COPR}	0.783	0.301
ω_{ACCP}	0.823	* [2]
ω_{CONT}	0.063	0.414
ω_{MOS}	0.851	0.137

Table 5: Standardized effect sizes for *voice* and *text type*.

The table shows, that *voice* has a large effect ($\omega > 0.5$) on each scale except CONT ($\omega < 0.1$). However, *text type* has not more than a medium effect ($\omega > 0.3$) on SPPA, EMO, COPR, and CONT respectively no effect on ACCP and a small influence ($0.3 > \omega > 0.1$) on the remaining scales [13].

3.2. Influence of familiarity of texts

We assumend that subjects could be biased if they knew some of the books prior to the listening test. Therefore, we split the gathered data in one set of ratings that were given by subjects that did not know the read passages and one set with ratings by subjects that knew the passages. We compared the mean values of each stimuli from each dataset and found no significant differences. Therefore, the *level of familiarity* of the text does not introduce any bias on the ratings of other scales. For future listening tests we propose to omit this item from the questionaire.

[2] because the error explained more variance than the experimental effect, the squared effect size has a negative value, thus ω cannot be computed in this case.

3.3. Factor analysis

To get an impression of the perceptual dimensions that were captured in the listening test, a Principal Axis Factor analysis (PAF) was carried out. We used the data from all scales except *level of familiarity* which only had nominal scale level and *overall impression* since it comprises the information from the other scales.
2 factors were extracted which account for 61.53% of the total variance. Residuals between the observed and reproduced correlations were computed: 6 (16%) of them were nonredundant with absolute values greater than 0.05. Subsequently, we opted for an oblique rotation method (Promax rotation with $\kappa = 4$) since we assumed highly correlated dimensions but wanted to obtain interpretable scales as an output. The resulting factor pattern matrix can be seen in Table 6. To ensure readability values below 0.2 are suppressed.
Since we used an oblique rotation method the factors are no longer orthogonal. Factors 1 and 2 reach a correlation of 0.70.

	Factor loadings	
	1	2
Voice pleasantness	0.852	
Listening effort	0.800	
Acceptance	0.789	
Intonation		0.935
Speech pauses		0.593
Emotion	0.223	0.537
Accentuation	0.285	0.473
Comprehension problems	0.403	0.414
Content		

Table 6: Factor pattern matrix.

To ensure a meaningful interpretation of the perceptual space, scales with high cross-loadings (| loading on factor A - loading on factor B | < 0.2) will not be taken into account. In our case this applies to *accentuation*, *comprehension problems* and *content*.
Factor 1 includes the scales voice *voice pleasantness*, *listening effort*, and *acceptance* and thus covers the **listening pleasure** the TTS systems achieve. With the high loading of the scale *intonation*, the second dimension seems to reflect the **prosody** of the signal. Moreover all other scales that account for factor 2 express natural rhythm or stress.
It is also noticeable that the highest loading of the scale *content* is only 0.16. This means that this scale did not provide much to any of the dimensions. Moreover, the communality of this item is only 0.05 and thus lies far behind the communalities of the other items (mean value: 0.59). This also implies that the suggested factor structure does not represent the scale *content*.

The mapping of the stimuli in the perceptual space is displayed in Figure 1, in which the subscripted characters (f/m) represent the speaker gender and the subscripted numbers (1-8) the text ID from Table 1. As it can be seen the stimuli form one cluster for each of the synthesis systems. The stimulus IV_{f3} is the only outlier. Its value for the dimension *listening pleasure* as well as for *prosody* is far below the mean value of the IV system. This impression is confirmed by the ratings this stimulus achieved in the listening test: it scored lowest on intonation and voice pleasantness while getting rated with

the most comprehension problems and highest listening effort of all female IV stimuli. However, on average the stimuli produced with the IVONA TTS system achieve higher values in both dimensions.
It is suprising that 2 stimuli of the manually optimized version of the female voice of IVONA (IVO_{f4}, IVO_{f7}) are inferior in both dimensions to the stimuli synthesized by the original TTS system. However, IVO_{f8} performed better than IV_{f8} and even achieved the best prosody value of all stimuli. This shows that simply by adjusting wrong articulated words and improving the pause lenghts between sentences and paragraphs even the quality of good synthesizers can be improved.
In addition, it is noticeable that stimuli of the same text type and the same synthesizer reach similiar prosody values. Whereas for a given system the listening pleasure of the male stimuli is always superior to the female voices. This accounts for the data from the CV as well as the IV synthesizer (except for text 8). Furthermore, it can be stated that both synthesizers have difficulties with different kinds of texts, e.g. the stimuli with text type 4 are one of the best rated stimuli of synthesizer CV whereas IV_{m4} is one of the worst rated of IV.
Hence, the quality of the synthesized speech does mostly depend on the synthesizer and not on the type of text, even though the quality of the tested synthesizers varies highly between text types.

4. Conclusions and suggestions for the Blizzard Challenge

We developed a questionaire with 11 scales for the subjective assessment of TTS in audiobook reading tasks. A set of 8 texts comprising different writing styles and book catagories was collected. The texts were synthesized by 2 TTS systems each with male and female voice. Additionally, 4 passages were synthesized by a manually optimized version of one of the TTS systems. The questionaire was then used to evaluate the TTS database.
A MANOVA revealed that the *voice* (combination of TTS system and male/female voice) has a significant influence on all scales (except CONT). Moreover, *text type* has also a significant influence on the scales PLT, SPPA, INT, EMO, COPR, and CONT. However, since not all assumptions under which a MANOVA is reliable were fulfilled the accuracy of the MANOVA might be compromised.
A subsequent factor analysis revealed 2 factors that were labeled *listening pleasure* and *prosody*. The mapping of the stimuli in the perceptual space (see Figure 1) made clear that the IVONA system scored higher values in average in both dimensions. Most of the stimuli synthesized with male voices achieved better values on the dimension listening pleasure than stimuli synthesized by the same system but with female voice.
For the audiobook reading task of the Blizzard Challenge 2012 we suggest to use the presented evaluation protocol with the following modifications:

- the *text type* has a significant but minor influence than the *voice* on the ratings on most of the scales, thus we propose to use a large variety of texts of different categories and with different writing styles.
- the *level of familiarity* does not have a significant influence on the ratings on other scales. Therefore, this item can be dropped in further tests.
- the scale *comprehension problems* has high crossloadings on both dimensions from the factor analysis. Thus, we propose to drop this scale since it does not help to discern between the 2 dimensions.
- the scale *content* does not account much for any of the 2 dimensions. Hence, we propose to conduct further tests without evaluating the *content*.

5. Acknowledgements

The present study was carried out at Deutsche Telekom Laboratories, Berlin. It was supported by the Deutsche Forschungsgemeinschaft (DFG), grants MO 1038/11-1 and HE 4465/4-1.

6. References

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

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

[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] F. Hinterleitner, S. Möller, C. Norrenbrock, and U. Heute, "Perceptual Quality Dimensions of Text-to-Speech Systems," *Proceedings of the 12th Annual Conference of the ISCA (Interspeech 2011). International Speech Communication Association (ISCA)*, 2011.

[5] U. Rautenberg and T. Schnickmann, *"Das Hörbuch - Stimme und Inszenierung"*. Harrassowitz Verlag, Wiesbaden, 2007, ch. "Die Stimme im Hörbuch: Literaturverlust oder Sinnlichkeitsgewinn?", pp. 21–54.

[6] J. Häusermann, K. Janz-Peschke, and S. Rühr, *"Das Hörbuch - Medium, Geschichte, Formen"*. UVK Verlags-Gesellschaft, Konstanz, 2010.

[7] M. Burkey, *"Sounds Good to Me: Listening to Audiobooks with Critical Ear"*. Booklist, 2007.

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

[9] A. Field, *Discovering Statistics Using SPSS*. SAGE Publications Ltd, London, 2005.

[10] S. S. Shapiro and M. B. Wilk, "An Analysis of Variance Test for Normality (Complete Samples)," *Biometrica, Vol. 52, No. 3/4*, pp. 591 – 611, 1965.

[11] E. Seier, "Comparison of Tests for Univariate Normality," *Proc. Interstat 2002*, 2002.

[12] R. F. Haase and M. V. Ellis, "Multivariate Analysis of Variance," *Journal of Counseling Psychology, Vol. 34, No. 4*, p. 404 Ű 413, 1987.

[13] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*. Routledge Academic, 2 edition, 1988.

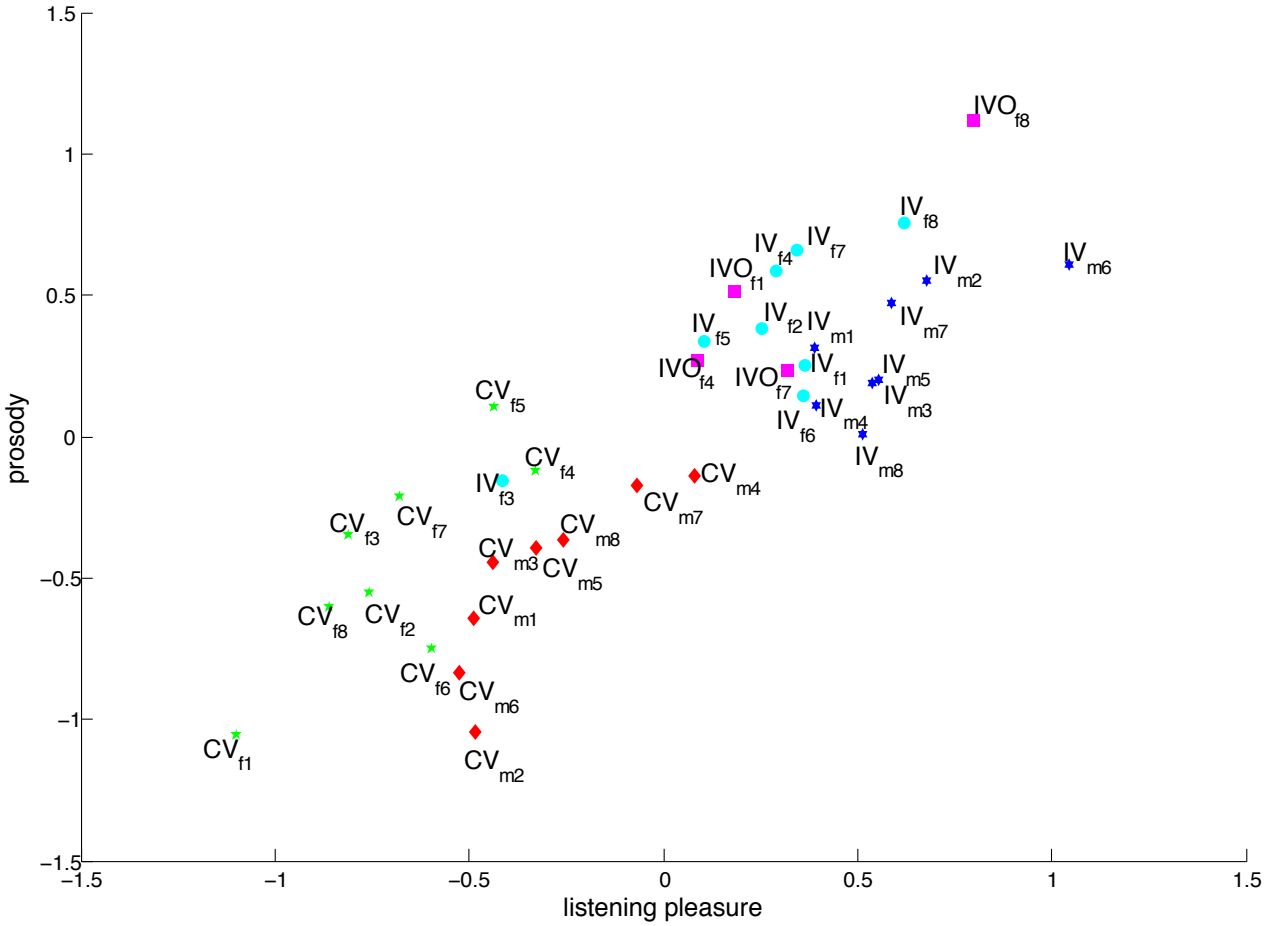
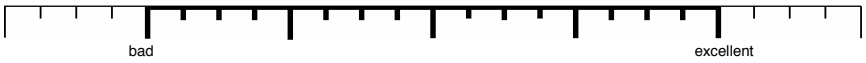
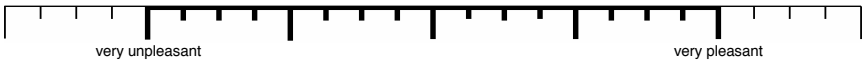


Figure 1: Factor loadings for the dimensions *listening pleasure* and *prosody*.

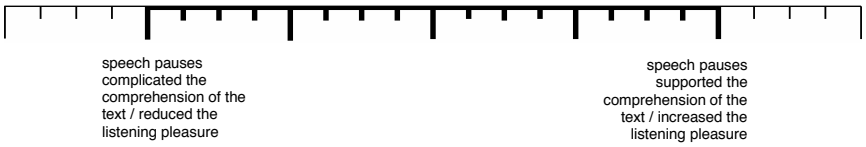
Overall impression
How do you rate the overall quality of the sound considering all aspects?



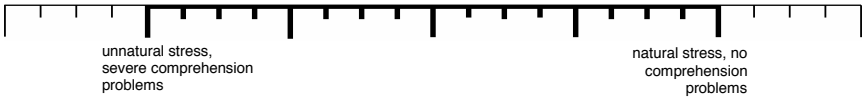
Voice pleasantness
How pleasant was it listening to the voice?



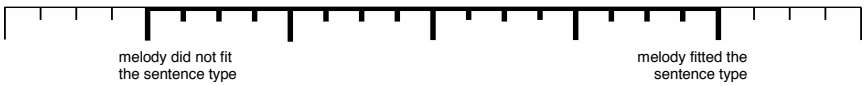
Speech pauses
Did the placement of speech pauses support the comprehension of the text or increase the listening pleasure?



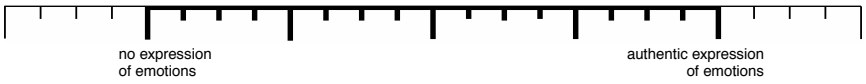
Accentuation
Did you notice any word stress anomalies which affected your text comprehension?



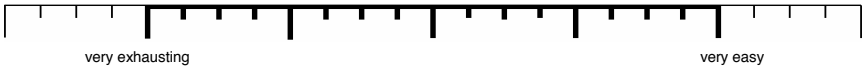
Intonation
Did the pitch of the speaker fit to the type of sentence?



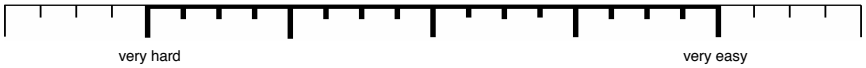
Emotion
Did the accent carry emotions that fitted the situation?



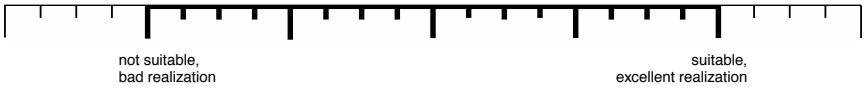
Listening effort
How would you describe the effort to listen to this voice over a longer period of time?



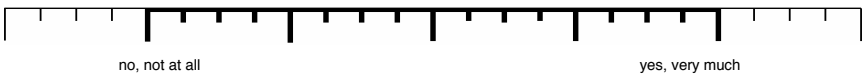
Comprehension problems
How hard was it to comprehend the text?



Acceptance
Do you think that this voice could be used for synthesizing this audiobook?



Content
Did you like the content of this section?



Level of familiarity
Did you know the section of the presented audiobook prior to the listening test?

☐ yes
☐ no

Figure 2: Questionnaire for the assessment of TTS audiobooks.