WHAT MAKES THIS VOICE SOUND SO BAD?

A MULTIDIMENSIONAL ANALYSIS OF STATE-OF-THE-ART TEXT-TO-SPEECH SYSTEMS

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

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

florian.hinterleitner@tu-berlin.de, cno@tf.uni-kiel.de,
sebastian.moeller@telekom.de, uh@tf.uni-kiel.de

ABSTRACT

This paper presents research on perceptual quality dimensions of synthetic speech. We generated 57 stimuli from 16/19 female/male German text-to-speech systems (TTS) and asked listeners to judge the perceptual distances between them in a sorting task. Through a subsequent multidimensional scaling algorithm, we extracted three dimensions. Via expert listening and a comparison to ratings gathered on 16 attribute scales, the three dimensions can be assigned to *naturalness of voice*, *temporal distortions* and *calmness*. These dimensions are discussed in detail and compared to the perceptual quality dimensions from previous multidimensional analyses. Moreover, the results are analyzed depending on the type of TTS system. The identified dimensions will be used in the future to build a dimension-based quality predictor for synthetic speech.

Index Terms— speech synthesis, perceptual quality dimensions, multidimensional scaling

1. INTRODUCTION

Even though the quality of modern TTS systems has reached a level that makes it possible to use synthetic speech in everyday applications like email readers and telephone information systems, there are still numerous impairments that corrupt different quality aspects of the generated voice: Due to the frequent concatenations of speech units, most diphone synthesizers sound artificial, HMM speech synthesis can lead to natural-sounding but very noisy speech, and the quality of unit-selection synthesizers not only depends on the degree to which the speech units fit together but also on how well the speech units fit to the text that should be synthesized. These impairments sound differently, i.e., they degrade the perceived quality along different perceptual dimensions. Thus, the quality of synthetic speech is of multidimensional nature and has to be described by different perceptual dimensions.

Several listening tests have been carried out to analyze the underlying perceptual quality dimensions of synthetic speech:

Kraft et al. [1] examined five speech-synthesis systems. A factor analysis revealed one factor that covers prosodic and long term attributes and one factor representing segmental attributes. Since this study is from 1995, state-of-the-art synthesis methods like unit-selection synthesis and HMM-synthesis were not investigated.

In 2005, Mayo et al. [2] evaluated TTS stimuli generated by eight different versions of the Festival [3] speech synthesizer. A multidimensional scaling revealed three dimensions which represent the appropriateness of prosody as well as the appropriateness and the number of selected units. However, since only the performance of one type of unit-selection system was analyzed, those results could not be generalized.

In [4] the authors analyzed synthetic speech files generated by 14/15 German TTS systems with female/male voices. In a first pretest, a broad basis of attributes that describe auditory features of synthetic speech was collected. These attributes were condensed into 28 scales that were used to evaluate the same stimuli in a second pretest. The 16 most important scales were used in the main test in which 60 stimuli produced by 14 female and 15 male TTS systems (2 stimuli per system) were evaluated. A subsequent factor analysis yielded three perceptual quality dimensions that were labeled: naturalness, disturbances, and temporal distortions.

In our view, the latter study clearly led to a better understanding of perceptual quality of synthetic speech. The results are relevant for the development and optimization of modern TTS systems because the authors analyzed TTS stimuli generated by all popular speech synthesis methods (diphone concatenation, unit-selection synthesis, HMM-synthesis) on scales that most likely cover all common TTS artifacts. Moreover, the study yielded three quality dimensions that were easily interpretable and generally intelligible. Nevertheless, the method used in [4] has some drawbacks:

1. During the two pre-tests the attributes were mainly solicited from audio and speech experts as test participants. This fact should have guaranteed that all relevant artifacts are captured and reproduced in the derived rating scales. How-

ever, the disadvantage of this approach is that experts also perceive degradations that are not as relevant for the quality impression of normal listeners. Though naïve listeners might be able to discern stimuli with respect to specific degradations (i.e. on corresponding attribute scales), this score does not necessarily affect their quality impression. Hence, through the process of a factor analysis this could have lead to dimensions that have no significant influence on the overall quality.
2. Even though the authors' intention was to keep the developed scales as simple as possible, one cannot be certain that the naïve listeners of the main test understood all scales in the same way.
3. Furthermore, methods that assess quality on global scales always limit the test participants' rating to the presented scales.
Therefore, it is crucial to investigate if those three quality dimensions could also be derived through a listening test that is not based on given rating scales, but rather on the unrestricted perceptual quality impression of the listeners itself.
To overcome these drawbacks, in this paper we present further research on the above mentioned dimensions. A multidimensional scaling was carried out to review the results of [4] and to gain deeper insight into the dimension *naturalness* which seemed to be too broad to us in order to be useful for system optimization.
Section 2 provides an overview of the test database. In Section 3 the main principle of multidimensional scaling and the sorting task that was chosen for the listening tests are explained. The results of the experiments are analyzed in Section 4. Finally, Section 5 discusses the outcome and gives a perspective to future work.

2. TEST DATABASE

To guarantee a fair comparison of the test stimuli, one German sentence with a synthesized length of approximately 5s was chosen for the listening test. 20 different TTS systems were selected (some of them with different voices) and used to synthesize this utterance. All in all, we used 16 female systems in 30 different configurations (*configuration* denotes a specific combination of TTS system and voice) and 19 male systems in 27 different configurations. The TTS systems that were used to generate the stimuli are the following ones (the number of female/male voices for each synthesizer are marked in brackets):
Acapela Infovox3 (2/1), AT&T Natural Voices (1/1), atip Proser (2/1), BOSS (2/0), Cepstral Voices (1/1), Cereproc CereVoice (1/1), DRESS (4/4), ESpeak (0/1), Fonix Speech FonixTalk (2/2), IVONA (1/1), Loquendo TTS (1/1), MARY bits (2/2), MARY hmm-bits (2/2), MARY MBROLA (2/3), Meridian Orpheus (0/1), NextUp Talker (1/1), NextUp TextAloud (0/1), Nuance RealSpeak (4/1), SVOX (2/1), SyRUB (0/1).
Thus the database contains stimuli generated by diphone synthesizers (DI), unit-selection synthesizers (US), and one HMM-synthesizer (HMM). For some of the commercial systems we have no information on their synthesizer type (n/a). All speech files were downsampled to 16 kHz and level normalized to -26 dBov using the speech-level meter [5].

3. SORTING TASK AND MULTIDIMENSIONAL SCALING

This section describes the main principle of multidimensional scaling (MDS) as well as a method suitable for large object sets. To simplify the interpretation of the dimensions extracted by the MDS algorithm, a post-test was carried out in which all stimuli were analyzed on 16 attribute scales [4].

3.1. Multidimensional scaling

The main idea of multidimensional scaling is to identify orthogonal perceptual dimensions without prior knowledge about the nature of the stimuli, by asking test participants to scale the dissimilarities between pairs of stimuli. The dissimilarities between stimuli can then be transformed into a stimulus space in which the between-point distances correspond to the dissimilarities between stimuli.
Dissimilarities are usually derived in listening tests in which each stimulus in a set of n stimuli is compared with all other $n-1$ stimuli. Subjects rate the similarity of two stimuli on a scale with the end points "very similar" and "not similar at all". The outcome is a matrix that represents the similarity between all stimuli [6].
Via an MDS algorithm, the dimensionality of this matrix can be reduced until the solution is interpretable but still represents the observed stimulus distances. As a badness-of-fit measure for the MDS representation, Kruskal [7] introduced the Stress function (low Stress values indicate a better fit).
The downside of paired comparison tests is that a complete comparison of all stimuli leads to $\frac{n(n-1)}{2}$ comparisons. With large sets of objects the amount of comparisons reaches a level that is not suitable for the assessment in listening tests. With the database described in Section 2 this would yield 435/351 comparisons for female/male stimuli and a test duration per subject of over two hours. Therefore, we introduce a method in the following paragraph to derive dissimilarity matrices without a full paired-comparison test.

3.2. Sorting task

Tsogo [8] proposed to use a sorting task when dealing with large object sets. Subjects are instructed to build groups of stimuli that are similar to each other while being different from stimuli in other groups. This results in one $n \times n$ incidence matrix per subject containing zeros and ones representing unsimilar and similar objects. Adding the matrices of all subjects together yields one similarity matrix from which

Table 1: Pearson correlation between rotated factor scores and attribute scale ratings.

	DIMENSION 1		DIMENSION 2		DIMENSION 3	
SCALES	FEMALE	MALE	FEMALE	MALE	FEMALE	MALE
accentuation	.84	.74	.76	.83	-.53	
bumpiness	-.61		-.81	-.79		
clink	-.67	-.60	-.56		.62	
distortions	-.72	-.77	-.77	-.70		
disturbances	-.63	-.69	-.77	-.69		
fluency	.59	.53	**.88**	**.78**	-.53	
hiss	-.54					
intelligibility	**.83**	**.87**	.73	.64		
naturalness	.85	.85	.77	.88		
noise						
pleasantness	**.88**	**.87**	.78	.78		
polyphony	-.64	-.63	-.77	-.58	.54	
rasping sound	-.56	-.63				
rhythm	.80	.76	**.86**	**.84**	-.55	
speed					**.54**	**.65**
tension	-.71	-.51	-.59	-.59	.68	.54
overall impression	.90	.89	.76	.76		

Note: for better readability correlations with $|R| < .50$ are suppressed; the correlations for the selected scales are in bold

one can easily derive a dissimilarity matrix as an input for the MDS algorithm.

To avoid that test participants sort stimuli with respect to gender, the test consisted of 2 sessions, one with female and one with male stimuli. Stimuli had to be sorted in up to eight groups with a minimum of two stimuli per group. 40 naïve subjects aged between 19 and 37 (20 female, 20 male, mean age: 25.6 years) took part. All of them were native German speakers, none had any known hearing disabilities. All subjects were paid for their participation. The stimuli were presented via headphones (AKG K601) and a high-quality sound device (RME Hammerfall DSP Multiface II) in a soundproof booth.

3.3. Post-test

MDS dimensions as such give no indication on their interpretation unless the stimuli are analyzed along the identified dimensions via expert listening or an additional auditory test. Thus, an interpretation is often a vague and moreover a highly subjective task. Therefore, we opted for a post-test in which all stimuli were rated on the 16 attribute scales from [4].

The stimuli were presented in two groups (one with female, one with male stimuli) in randomized order. 5 expert listeners from Telekom Innovation Laboratories and 7 naïve subjects aged between 23 and 31 (5 female, 7 male, mean age: 27 years) took part in the post-test. All of them were native German speakers without any known hearing disabilities. The stimuli were presented via headphones (Sennheiser HD 485) and a high-quality sound device (Roland Edirol UA-25) in a quiet listening environment.

4. PERCEPTUAL DIMENSIONS

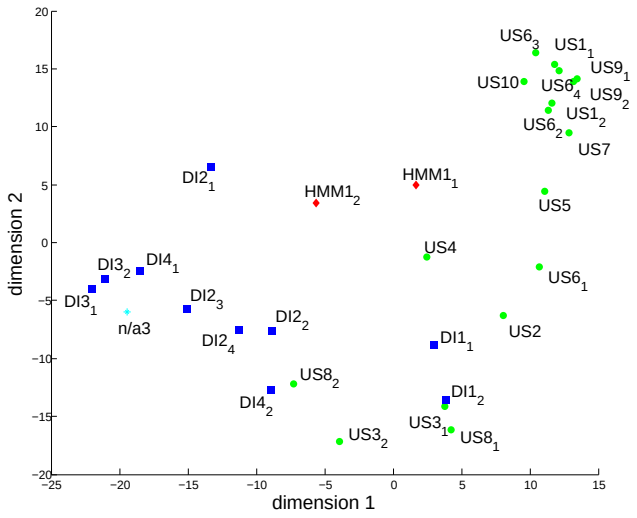
4.1. Data analysis

Via a non-metric MDS [9] three dimensions were extracted for both female and male stimuli. The statistical fit parameter Stress1 for the female/male solution reached values of 0.07/0.06 and was thus far below the Stress1 values for random data as reported in [10]. To maximize the variance in each dimension the stimulus space was Varimax-rotated.

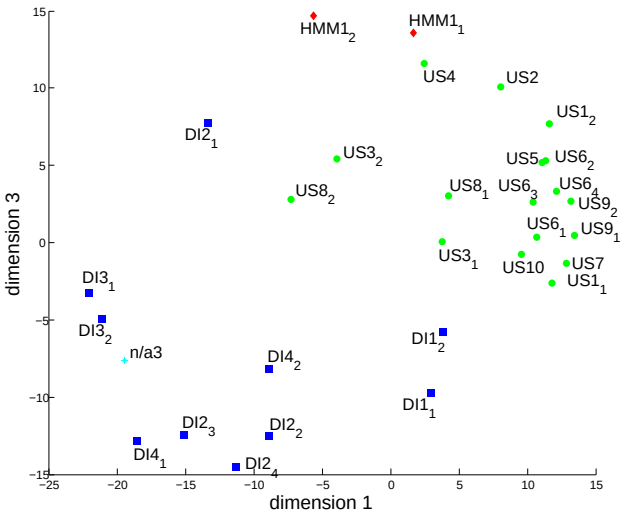
In order to ensure a meaningful interpretation, the Pearson correlation coefficient R between the factor scores and the median values of the 16 attribute scales from the post-test was computed. Among others, dimension 1 correlated highly with the scales *pleasantness* ($|R| \geq .80$) and *intelligibility* ($|R| \geq .80$), dimension 2 correlated highly with *rhythm* ($|R| \geq .80$) and *fluency* ($|R| \geq .70$), and dimension 3 reached the highest correlation with *speed* ($|R| \geq .50$). Dimensions 1 and 2 both also correlated highly with the scale naturalness ($|R| \geq .80$). But since this scale does not help to discern between both dimensions it was not taken into account for the following optimization.

In the next step the point configurations for both female and male stimuli were further rotated in a way which maximized the correlation between each dimension and the scale(s) mentioned in the previous paragraph. The correlations between the optimized rotated dimensions and the attribute scales can be seen in Table 1.

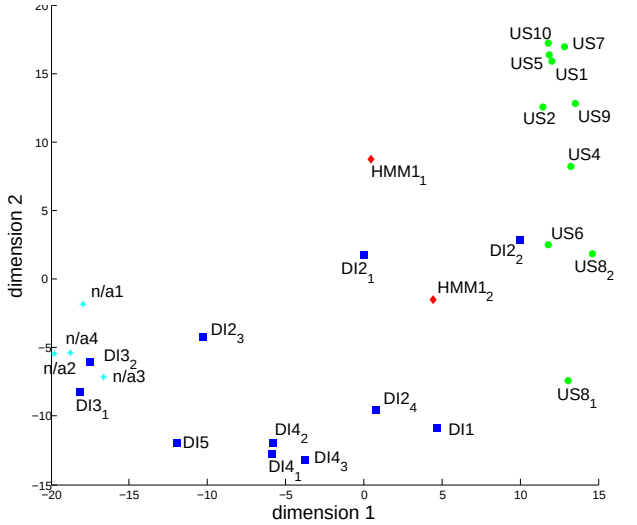
As can be seen, dimension 1 not only correlates highly with *pleasantness* and *intelligibility* but also with the scales *naturalness* and *accentuation* for female and male data. The



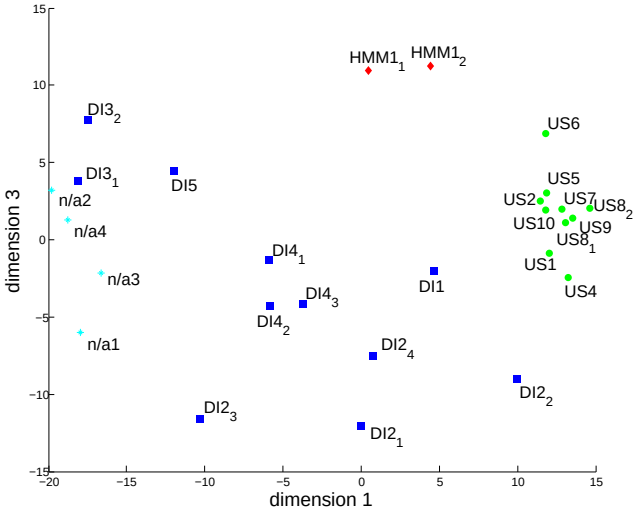
(a) Dimension 1 and 2 for female systems.



(b) Dimension 1 and 3 for female systems



(c) Dimension 1 and 2 for male systems



(d) Dimension 1 and 3 for male systems

Fig. 1: Factor values for female and male data.

correlation with *overall impression* is the highest for this dimension. Dimension 2 achieves high correlations with *rhythm* and *fluency* but also with *bumpiness* and *naturalness*. The correlations between dimension 3 and the attribute scales are much lower than for the other two dimensions. However, it is the only dimension that correlates with the scale *speed* with $|R| \geq .50$. Moreover, it correlates with the scales *tension* and the female data achieves a correlation of $R = .62$ with *clink*.

4.2. Interpretation

The 2D-mapping of the stimuli in the perceptual space of dimensions 1 and 2 and dimensions 1 and 3 can be seen in Figure 1. The two upper subfigures (Figures 1a and 1b) display the perceptual space for the female stimuli while the lower subfigures (Figures 1c and 1d) show the results for the male stimuli. Subscripted indices indicate different voices; US synthesizers are marked with green dots, DI synthesizers with blue squares, HMM synthesizers with red diamonds, and synthesizers of an unknown type are marked with a cyan asterisk. For the interpretation of the extracted dimensions all stimuli were sorted according to their value in each dimension. The auditory impression of the sorted stimuli along with the correlations from Table 1 served as indications for the interpretation of the dimensions. The voice of high-ranked stimuli in dimension 1 sounded very human-like even if the speech was somehow distorted. These stimuli can be described as voices with personality and charisma, thus this dimension was labeled **naturalness of voice**. The auditory impression for stimuli sorted along dimension 2 confirms the high positive correlations with *rhythm* and *fluency* and the high negative correlation with *bumpiness*. Stimuli with low values in this dimension lacked natural sounding prosody. Therefore this dimension can be associated with **temporal distortions** (low values indicate severe temporal distortions). Stimuli with high values in dimension 3 were slowly speaking and relaxed while voices with low values sounded more stressed and restless. This impression is confirmed by the correlations from Table 1. Thus dimension 3 describes the **calmness** of the voice.

A closer look at Figures 1a and 1c reveals a clustering effect: most US systems build a cluster in the upper right corner while most DI synthesizers are in the lower left corner. Therefore, the only systems that achieve high values in dimension 2 are US synthesizers. However, not all US stimuli that sound very human-like also reach high values in dimension 2 (US2, $US6_1$ in Figure 1a and $US8_1$ in Figure 1c).

We expect dimension 2 to be mainly linked to concatenation artifacts. Thus, we assumed that especially DI synthesizers that connect lots of small speech units would score low values in this dimension. In this context it is interesting to see that the stimuli in Figure 1a with the lowest values in dimension 2 are also US synthesizers ($US3_1$, $US3_2$, $US8_1$). Nonetheless, these are two non commercial, scientific systems that apparently have major issues with the pitch contour at the junctions between units. Strikingly, the stimulus $US8_1$ in Figure 1c achieves one of the highest ratings of male stimuli in dimension 1, still its overall impression rating was only 3.25. The main reason for that is the low score in the dimension *temporal distortions*.

In Figures 1b and 1d the clustering effect for the male stimuli can be seen even more clearly. The US systems sound very human-like and show similar speech rates. The DI systems span across the whole co-domains of both dimensions. However, DI stimuli with natural voices sound stressed ($DI2_2$ in Figure 1d) while DI stimuli with low values in dimension 1 sound calmer ($DI3_2$ in Figure 1d). The highest values in dimension 3 and therefore the system with the lowest perceived speech rate is the HMM synthesizer for female as well as for male voices. Considering the clusters of TTS systems the figures also show that the systems of unknown type are most likely all DI synthesizers .

Furthermore, it can be stated that most of the stimuli that were produced by the same TTS system build clusters regardless of the voices gender, e.g DI3 and HMM1. Nonetheless, some of the stimuli of one system score very differently in one dimension (US6 in Figure 1a, DI2 in Figure 1c). Moreover, while high values in the dimensions 1 and 2 indicate a better overall impression, Figure 1b and 1d denote that the highest-ranked stimuli on the scale overall impression ($US6_2$, $US6_3$, $US1_1$ in Figure 1b and US7, US10, US5 in Figure 1d) show medium values in dimension 3. Therefore, a medium speech rate seems to achieve the best listening impression.

5. DISCUSSION AND FUTURE WORK

An auditory experiment with 20 different TTS systems has been carried out. Participants grouped stimuli according to their similarity. A subsequent MDS yielded three dimensions that capture *naturalness of voice*, *temporal distortions*, and *calmness* of the signals.

On first sight these results differ slightly from the previous multidimensional analyses presented in [4].

A closer look reveals that the dimension *naturalness of voice* specifies the broader dimension *naturalness*. The results show that this quality dimension apparently covers how human-like the synthetic voice sounds. This dimension gives an answer to the question: does the listener have the impression that this TTS signal has been produced by a human being or does it sound like it was produced by a computer?

Dimension 2 clearly corresponds to the same quality impression as the dimension *temporal distortions* from the previous study [4]. It is evidently linked to prosodic characteristics of speech signals [11]. Synthesizers with low scores in this dimension often lack natural-sounding prosody. This is often a result of failures during the concatenation of speech units in diphone and unit-selection synthesis. For the female stimuli this effect cannot only be witnessed for DI systems but also for the two synthesizers US3 and US8.

Finally, dimension 3 (*calmness*) corresponds to the dimension *speed* that was also detected previously but seemed to be of minor importance. Apparently, the impression of fast speech also invokes a feeling of stress and restlessness while a slow speaker sounds relaxed and in some cases even a bit drowsy.
A wide difference can be noticed concerning the dimension *disturbances*. While this dimension was highly relevant in the previous study [4], the MDS experiment did not capture a perceptual importance for the participants of this experiment. Though listeners could clearly distinguish, e.g., the grade of noise and hiss in the signal through the attribute scales presented in the post-test, this cannot be stated for the sorting task. We assume that this effect is due to the nature of most TTS signals: even though the quality improved dramatically over the years there are still major issues that catch the attention of listeners. Those issues mainly concern naturalness and prosody (dimension 1 and 2) and become so important that minor problems like disturbances are masked by the degradations in the first two dimensions. Moreover, listeners are used to impairments like noise and hiss through, e.g., coding and transmission artifacts in cell phone or IP-based communication and might thus be more tolerant with respect to disturbances in the signal. The semantic differential from the previous study clearly yielded very analytic information while the MDS resulted in dimensions that are completely unaffected (e.g., by attribute scales) and thus represent the relevant dimensions that were perceived by the listeners.
Since dimension 3 is less prominent than the other two, a constant speech rate between stimuli can be a precondition for future listening tests. Moreover, a semantic differential developed specially for dimensions 1 and 2 would certainly lead to a better understanding of those dimensions and might even result in new subdimensions.
The analysis of the results with respect to the types of TTS systems (diphone, unit-selection, HMM-synthesis) brought interesting insights concerning, e.g., concatenation techniques and speech rate. From a system-developer's point of view it would have been also interesting to analyze the results with respect to the inventory size or the amount of training data. But, unfortunately, this information is unknown to us especially for all of the commercial systems. Moreover, we plan to develop attribute scales that specifically capture the quality aspects of dimensions 1 and 2. Furthermore, we are currently using these results to develop a dimension-based quality predictor.

6. ACKNOWLEDGEMENTS

The present study was carried out at Telekom Innovation Laboratories, 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 Steffen Werner from Daimler AG, Jan de Moortel from Nuance, Donata Moers from University of Bonn, and Guntram Strecha from University of Dresden for their support.

7. REFERENCES

[1] V. Kraft and T. Portele, "Quality Evaluation of Five German Speech Synthesis Systems," *Acta Acustica 3*, pp. 351–365, 1995.

[2] C. Mayo, R. A. J. Clark, and S. King, "Multidimensional Scaling of Listener Responses to Synthetic Speech," *Proc. of the 6th Annual Conference of the International Speech Communication Association (Interspeech 2005)*, pp. 1725–1728, 2005.

[3] Alan W. Black and Paul A. Taylor, "The Festival Speech Synthesis System: System Documentation," Tech. Rep. HCRC/TR-83, Human Communication Research Centre, University of Edinburgh, Scotland, UK, 1997, Avaliable at http://www.cstr.ed.ac.uk/projects/festival/manual/.

[4] 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)*, pp. 2177–2180, 2011.

[5] ITU-T Rec. P.56, *Objective Measurement of Active Speech Level*, International Telecommunication Union, Geneva, 1993.

[6] I. Borg and Groenen P., *Modern Multidimensional Scaling - Theory and Applications, 2nd edition*, Springer Series in Statistics, New York, 2005.

[7] J. Kruskal and M. Wish, "Multidimensional Scaling," in *Quantitative Applications in the Social Sciences*. 1978, vol. 07-11, Sage.

[8] L. Tsogo, M.H. Masson, and A. Bardot, "Multidimensional Scaling Methods for Many-Objects Sets: A Review," in *Multivariate Behavioral Research*, 2000, vol. 35, pp. 307–319.

[9] G. A. F. Seber, *Multivariate Observations*, J. Wiley & Sons, New York, 1984.

[10] K. Sturrock and J. Rocha, "A Multidimensional Scaling Stress Evaluation Table," in *Field Methods*, 2000, vol. 12(1), pp. 49–60.

[11] C.R. Norrenbrock, F. Hinterleitner, U. Heute, and S. Möller, "Quality Analysis of Macroprosodic F0 Dynamics in Text-To-Speech Signals," *Proc. of the 13th Annual Conference of the International Speech Communication Association (Interspeech 2012)*, 2012 (accepted).