Is Intelligibility Still the Main Problem? A Review of Perceptual Quality Dimensions of Synthetic Speech

Florian Hinterleitner[1], *Christoph R. 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, cno@tf.uni-kiel.de,
sebastian.moeller@telekom.de

Abstract

In this paper, we present a comparative overview of 9 studies on perceptual quality dimensions of synthetic speech. Different subjective assessment techniques have been used to evaluate the text-to-speech (TTS) stimuli in each of these tests: in a semantic differential, the test participants rate every stimulus on a given set of rating scales, while in a paired comparison test, the subjects rate the similarity of pairs of stimuli. Perceptual quality dimensions can be derived from the results of both test methods, either by performing a factor analysis or via multidimensional scaling. We show that even though the 9 tests differ in terms of used synthesizer types, stimulus duration, language, and quality assessment methods, the resulting perceptual quality dimensions can be linked to 5 universal quality dimensions of synthetic speech: (i) *naturalness of voice*, (ii) *prosodic quality*, (iii) *fluency and intelligibility*, (iv) *disturbances*, and (v) *calmness*.

Index Terms: text-to-speech (TTS), perceptual quality dimensions, evaluation

1. Introduction

Even though the quality of modern TTS systems has reached a level of quality that no longer reminds listeners of robot-like voices but of real human speakers, different degradations still diminish the overall quality impression: most PSOLA-based diphone synthesizers lead to artificial voices due to frequent concatenations of speech units, HMM-synthesizers can generate natural-sounding but also "noisy" speech, and the quality of unit-selection systems mainly depends on the size of the used speech corpus, how well the units fit together and how well this corpus fits to the text that is to be synthesized. These impairments all sound differently, thus they degrade speech along different perceptual dimensions. Hence, the quality of synthetic speech is of multidimensional nature.

Several listening tests have been carried out over the past years in order to reveal the inherent perceptual quality dimensions of synthetic speech. As a result, a variety of different dimensions appear to exist. In one study [1] the dimensions were labeled (i) *prosody* and (ii) *segmental*, the next study [2] yielded the dimensions (i) *naturalness* and (ii) *intelligibility*, and another study [3] resulted in the dimensions (i) *naturalness of voice*, (ii) *temporal distortions*, and (iii) *calmness*. Given the different synthesizers that were used, the variations in stimulus duration, and the diverse assessment methodologies, the ambiguity is not surprising.

In this paper, we present a comparative overview of perceptual quality dimensions which resulted from 9 studies on TTS quality, and we will show that these dimensions can be attributed to a unifying set of dimensions. In Section 2, we introduce the two different approaches to multidimensional analysis for speech signals that were used in the 9 studies. Test details are given in Section 3. Section 4 highlights the similarities and differences between the studies. In Section 5, we compare the quality dimensions of all studies and introduce a set of 5 universal TTS-quality dimensions to which all other dimensions can be linked. Finally, in Section 6 we conclude the results and give a perspective to future work.

2. Multidimensional analysis

The two main approaches to analyzing perceptual quality dimensions with the help of human listeners are discussed in the following.

In a semantic differential (SD), pre-defined attribute scales are used to measure the auditory impression of the listeners. This guarantees a direct relationship between the used attribute scales and the derived quality dimensions. Therefore, the results are usually easy to interpret. On the downside, the ratings of the test participants are always limited to the set of presented scales. If a quality impression can not be expressed by any of the presented scales, this information will be lost. Thus, it is crucial to carefully choose a set of scales for the listening test. To reduce the influence of the test designers to a minimum, a suitable set of scales can be developed through several pretests, i.e., the goal of the first pretest is to collect attributes and corresponding attribute scales which describe the auditory impression of the listeners; in a second pretest, this set of attribute scales can be reduced to a final selection of scales.

In comparison, the multidimensional scaling (MDS) approach with paired comparison (PC) testing is solely based on the perceptual impression of the listener and not on any given rating scales. Participants are instructed to rate the similarity of one feature of pairs of speech signals, e.g., similarity in naturalness. Therefore, every stimulus in a set of n stimuli has to be compared to all remaining $n-1$ stimuli. The outcome is a matrix that represents the similarity between all stimuli [4]. Via an MDS algorithm, the dimensionality of this matrix can then be reduced until the solution is interpretable but still represents the observed stimulus space. However, since a complete comparison of all stimuli leads to $\frac{n(n-1)}{2}$ comparisons and a listening-test duration of several hours per subject, this approach is hardly deployable with larger sets of objects. For these cases, Tsogo [5] proposed a sorting task. Here, subjects are instructed to build groups of stimuli that are similar to each other while being different from the stimuli in other groups.

This yields one incidence matrix per subject from which a similarity matrix can be derived that can be further processed as described above. Even though the MDS approach has the advantage that the participants' ratings are not influenced by given rating scales, its major drawback is the interpretability of the resulting dimensions. MDS dimensions as such give no indication on their interpretation, thus, additional knowledge about the nature of the stimuli has to be obtained, e.g., via expert listening, rating scales or measures derived from the synthesis system.

3. Subjective TTS evaluations

This section gives an overview of the 9 different TTS databases as well as an interpretation of the resulting perceptual quality dimensions.

3.1. Test 1

In 1995, Kraft and Portele [1] evaluated five German-speaking TTS systems in an auditory listening test. The database consisted of stimuli produced by 2 formant synthesizers (male voices) and 3 diphone/demisyllable synthesizers (2 female voices, 1 male voice). The 44 subjects were instructed to rate the stimuli on 8 presented absolute category rating (ACR) scales with 5 to 6 categories. 6 familiar and unfamiliar passages were synthesized with a total duration of about 100 words. A subsequent Principal Component Analysis (PCA) with Promax rotation revealed 2 factors which were connected to (i) *prosodic and long term attributes* and to (ii) *segmental attributes*. Even though, the first dimension was linked to prosody it also comprises attribute scales that are specific to the voice of the systems, such as naturalness and pleasantness.

3.2. Test 2

In [6], a pilot study was conducted in order to unveil the perceptual quality dimensions of the Festival synthesizer [7]. 8 sentences from the TIMIT database [8] were chosen and synthesized with an English-speaking female voice. The stimulus duration varied from 1.9 to 4.1 seconds. 8 native speakers of English which were all experienced with listening to synthetic speech took part in a paired comparison (PC) test. They were instructed to rate whether the two presented stimuli were *similar* or *different* in terms of naturalness. The responses were compiled into a dissimilarity matrix which was then processed via an MDS analysis. The resulting dimensions were interpreted through visual and auditory analysis of the configuration of the stimulus space. The first dimension represents (i) *prosodic cues* which reflect the appropriateness of duration and intonation. The second dimension is linked to (ii) *segmental and unit-level cues*. It describes the appropriateness of units selected for synthesis as well as the number of selected units.

3.3. Test 3

To test the reliability and validity of the test method proposed in the ITU-T Rec. P.85 [9], Viswanathan et al. [2] conducted a series of 5 consecutive listening tests. In the final study, stimuli produced by 5 English-speaking TTS systems were evaluated on 9 5-point ACR scales. Additionally, participants were instructed to also rate the overall quality and the acceptability of the systems. The investigated systems used either phones or sub-phone units for concatenative synthesis. The synthesizers included algorithmic variations for pitch and duration generation. The stimuli were rated by 128 naïve test participants. A factor analysis revealed 2 factors: Dimension 1 is related to the extent to which speech is similar to natural human speech and was thus labeled (i) *naturalness*; Dimension 2 describes how well the content of the signal can be understood, hence it can be assigned to the (ii) *intelligibility* of the signal.

3.4. Test 4

In [10], speech material from 6 German "off-the-shelf" TTS systems was evaluated. The stimuli were created by diphone-based synthesizers using the pitch-synchronous-overlap-add (PSOLA) technique and unit-selection systems. A total of 10 speech samples have been generated per TTS system, half for male speakers and half for female ones. The synthesized speech samples have an average duration of 12s and consist of two utterances separated by a silence interval of approximately 2s. The listening test closely followed the ITU-T Rec. P.85 [9]. Thus, besides the rating of the stimuli on 8 ACR scales, the 17 test participants were also given a parallel task. As suggested in P.85, the listening test also included natural speech reference files. A subsequent Principal Axis Factor (PAF) analysis with Promax rotation revealed 2 dimensions. The first dimension consists of scales concerning the naturalness of the synthesized voice as well as prosodic attributes of the signal. The second dimension comprises scales that cover the fluency and intelligibility of the signal. Thus, dimension 1 was labeled (i) *naturalness and prosody* while dimension 2 was named (ii) *intelligibility*.

3.5. Test 5

In [11], Mayo et al. pursued the investigations described in Section 3.2. 24 sentences from the TIMIT corpus were selected and synthesized with an English-speaking female voice by the diphone-based Festival speech-synthesis system. The average duration of the stimuli was 2.7s. 30 participants took part in a PC test, where they were instructed to rate the similarity of a pair of stimuli in terms of naturalness. Two types of acoustic analysis were carried out: the automatic analysis consisted of measures that were computed by Festival during the synthesis process (e.g., target and join costs) and measures that were derived from those features (e.g., total cost, target costs of different types of diphones); the manual analysis included comparisons with natural speech files (e.g., number of transcription/pronunciation errors per synthetic utterance).
A subsequent MDS analysis yielded 3 dimensions. Through visual, auditory and cluster analysis these dimensions could be linked to (i) *overall join quality/quantity*, (ii) *join distribution and detectability*, and (iii) *unit appropriateness and prosody*. In our view the first two dimensions are thus connected to segmental attributes that concern the fluency and the intelligibility of the speech signal, while the third dimension represents global characteristics that describe the prosodic quality of the signal.

3.6. Test 6

Test 6 was part of an extensive study [12] in which the inherent quality dimensions of state-of-the-art TTS systems were investigated. 16 German-speaking synthesizers (formant synthesizers, PSOLA-based diphone synthesizers, unit-selection systems, and HMM-synthesizers) were used to generate 2 samples for each of the 30 different configurations[1] of synthesizers. The average duration was 10s. All stimuli were rated by 30 participants on 16 continuous scales (CS) that were developed dur-

[1] A configuration denotes a specific combination of one voice and one synthesis system

Table 1: Comparison of the main characteristics of the different test setups for Test 1-9.

	TEST 1	TEST 2	TEST 3	TEST 4	TEST 5	TEST 6	TEST 7	TEST 8	TEST 9
Year	1995	2005	2005	2007	2011	2011	2011	2012	2012
Language	German	English	English	German	English	German	German	German	English
Synthesizer type:									
Formant	✓					✓		✓	
Concatenative	✓		✓	✓		✓		✓	✓
Unit-selection		✓		✓	✓	✓	✓	✓	✓
HMM						✓		✓	✓
Number of systems	5	1	5	6	1	16	2	20	10
Number of configurations	5	1	5	12	1	30	5	57	10
Stimuli per configuration	6	8	9	5	24	2	8	1	22
Quality assessment via	ACR	PC	ACR	ACR	PC	CS	CS	CS	CS
Number of scales	8	-	9	7	-	16	9	16	7
Length of stimuli	100 words	1.9-4.1s	20-25 words	12s	2.7s	10s	55s	5s	45s

ing two extensive pretests. A subsequent PAF with Promax rotation revealed 3 perceptual dimensions. The first and most broad dimension comprises scales like accentuation, naturalness, rhythm, and pleasantness. Thus, it was labeled (i) *naturalness*. The second dimension consists of scales that specify (ii) *disturbances* in the signal, e.g., hissing and noise. The last dimension is related to (iii) *temporal distortions*, e.g., concatenation artifacts which occur in unit-selection synthesis. Additionally, the scale (iv) *speed* appeared to be a supplementary dimension.

3.7. Test 7

In [13], a pilot study was conducted to find a suitable set of attribute scales for the quality assessment of TTS in audiobook reading tasks. 2 German-speaking unit-selection synthesizers with female and male voices were used to synthesize book passages from 8 different books. The passages had an average duration of 55s and were chosen with the intention to cover a broad variety of different writing styles. Attribute scales from the P.85 questionnaire as well as scales that were developed especially for the evaluation of TTS-read audiobooks were used in this test. A PAF analysis with Promax rotation yielded 2 dimensions: the first dimension includes scales like voice pleasantness and listening effort and is thus related to the (i) *listening pleasure*; the second dimension comprises scales like intonation and speech pauses, hence it reflects the (ii) *prosody & rhythm* of the speech signal.

3.8. Test 8

This database was gathered during a study [3] which aimed to complement and to expand the results from Section 3.6. Therefore, 30 female and 27 male stimuli with an average duration of 5s were generated from the same utterance by different configurations of German-speaking TTS systems (formant synthesizers, PSOLA-based diphone synthesizers, unit-selection systems, and HMM-synthesizers). The stimuli were evaluated by 40 naïve test participants in a sorting task. The resulting dissimilarity matrix was processed via an MDS analysis and yielded 3 perceptual quality dimensions.
In a post-test, all stimuli were rated on the same 16 CS as described in Test 6. 12 test participants (5 expert listeners from the Quality and Usability Lab of the TU Berlin and 7 naïve subjects) took part in the test. Subsequently, the 3 quality dimensions were interpreted by means of expert listening and the ratings on the 16 attribute scales: dimension 1 describes voices with personality and charisma and was thus labeled (i) *naturalness of voice*; the second dimension is related to concatenation artifacts as well as the prosody of the signal, hence it is describes (ii) *temporal distortions*; the third dimension distinguishes between relaxed and slow speaking TTS systems and synthesizers which generate stressed and restless sounding voices, therefore it was labeled (iii) *calmness*.

3.9. Test 9

This database [14] was gathered within the scope of the TTS-audiobook-reading task of the Blizzard Challenge 2012 [15]. The results from the pilot study in Section 3.7 were the basis of the experimental setup. 10 male English-speaking synthesizers were used to synthesize book passages from 13 different books. The passages had an average duration of 45s. As in Test 7, the passages were selected with the aim to cover different writing styles. The recommendations from [13] lead to several changes in the selection and the labelling of the attribute scales. A PAF analysis with Promax rotation yielded 2 dimensions which mainly confirmed the dimensions (i) *listening pleasure* and (ii) *prosody and rhythm* from Test 7.

4. Similarities and differences

This section outlines the similarities and differences of the studies presented in the previous section and their impact on the resulting quality dimensions. An overview of all relevant characteristics of the experimental setup is discussed in the following and can be seen in Table 1.
As mentioned in Section 2, the quality-assessment method has a major influence on the resulting quality dimensions. The ratings in listening tests that use attribute scales to assess quality are always limited on the presented scales. Thus, characteristics that cannot be expressed by any of the scales are not captured. However, the experimental setups in Test 6 and 8, where 16 scales

Table 2: Perceptual quality dimensions of synthetic speech.

Dimensions	Relevant Scales	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6	Test 7	Test 8	Test 9
Naturalness of Voice	*naturalness voice pleasantness*	Prosody		Naturalness	Naturalness and Prosody		Naturalness	Listening Pleasure	Naturalness of Voice	Listening Pleasure
Prosodic Quality	*stress rhythm prosody intonation*		Prosodic Cues	Intelligibility		Unit Appropriatness and Prosody		Prosody & Rhythm	Temporal Distortions	Prosody & Rhythm
Fluency and Intelligibility	*fluency intelligibility bumpiness polyphony*	Segmental	Segmental or Unit Level Cues		Intelligibility	Overall Join Quality/Quantity Join Distribution and Detectability	Temporal Distortions			
Absence of Disturbances	*hissing noise rasping disturbances*						Disturbances			
Calmness	*speed tension*						Speed		Calmness	

were presented, are more likely to give a deeper insight into the perceived quality. Nevertheless, one cannot be certain that naïve listeners which do not have detailed knowledge about the quality characteristics of speech, all understand the wording of the scales in the same way. In contrast, the PC test and the sorting task with subsequent MDS analysis bypass this constraint, but there is no information on the interpretation of the resulting stimulus space.
Moreover, the resulting quality dimensions also depend on the different types of synthesizers that were part of the test database. Thus, synthesizer-specific characteristics, e.g., the noise of HMM-synthesizers or the sonic glitches of concatenative systems, can naturally only be assessed if these types of systems are part of the study. Accordingly, studies that only feature formant synthesizers and diphone based concatenative systems, e.g., as in Test 1, are most likely to lead to different dimensions than studies that only assess unit-selection synthesizers, e.g., as in Test 2 and 5.
Furthermore, the duration of the generated stimuli also affects the perceived quality. The stimuli from the audiobook-reading tasks in Test 7 and 9, with durations of 55s and 45s, respectively, could bring other quality aspects into focus than stimuli from a different use case. In addition, very short stimuli, as in the Tests 2 and 5, can be difficult to judge in terms of voice or prosodic quality.

5. Results

The differences in the quality assessment methods, the synthesizer types used in the tests, and the different stimulus durations in most of the studies indicate ambiguous results. In the following, we present a comparative overview of the perceptual quality dimensions resulted from the studies in Section 3 and show that these dimensions can be linked to 5 universal perceptual quality dimensions of synthetic speech which are:

- **naturalness of voice**
- **prosodic quality**
- **fluency and intelligibility**
- **absence of disturbances**
- **calmness**

5.1. Naturalness of voice

As can be seen in Table 2, the dimension *naturalness of voice* is part of the outcome of most studies, with the exception of the MDS experiments (Test 2 and 5). However, this can be explained considering the stimuli from those tests: they were all generated from the same voice by the Festival synthesizer. Thus, none of them differed in voice characteristics. Even though the first dimension in the two audiobook tests was labeled *listening pleasure*, which seemed more suitable for this use case, it actually represents the character of the voice.

5.2. Prosodic quality

Due to the overlap of the first two dimensions in some studies (Test 1, 4, and 6) the second dimension seems to be more vague. Test 7 and 9, on the other hand, show that these dimensions can indeed be regarded as independent dimensions, even though they are somewhat correlated [13] [14]. The prosodic dimension can also be retrieved in Test 2 and 5, where the test participants did not perceive a dimension concerning the voice of the signal.

5.3. Fluency and intelligibility

The third prominent dimension covers *fluency and intelligibility* and it can be found in all studies except the audiobook experiments. This dimension captures segmental artifacts that are characteristic for synthesizers that concatenate smaller units like diphones. Considering the requirement for high-quality voices in audiobook reading tasks with very few glitches in concatenation, this dimension is indeed not prominent. The MDS study by Mayo from 2011 (Test 5) shows that this dimension can be further split up, at least for unit-selection synthesizers. On the contrary, the overlap of the prosody and the fluency/intelligibility dimensions in the MDS experiment (Test 8) shows that these two dimensions are hard to distinguish for naïve listeners.

5.4. Absence of disturbances

The dimension *absence of disturbances* could only be retrieved from the extensive experiments in Test 6. This is most likely due to the fact that the presented scales were developed with the help of speech and audio experts which might focus on various types of degradations. Even though the test participants could clearly distinguish, e.g., the grade of noise and hiss in the signal, these degradations were obviously less important to them than issues concerning the voice or the prosody of the signal. Nonetheless, this dimension can be useful to assess the quality of HMM synthesizers or systems that concatenate coded speech units which can produce noisy speech signals.

5.5. Calmness

Finally, the dimension *calmness* was found in Test 6 and 8. This dimension however appears to be less important since most of the speech synthesizers run at a similar speech rate. Nonetheless, when assessing the quality of fast synthesizers, like they are deployed in reading devices for the blind, this quality aspect can play a crucial role.

6. Conclusions and future work

Even though this study combined the results of 9 different experiments, further research will be needed to confirm the 5 resulting dimensions. Especially the relevance of the dimensions 4 and 5 should be further investigated. Nonetheless, including scales marked as relevant in Table 2 in future listening tests is expected to provide a more complete view on the perceptual aspects of the systems.
Furthermore, this study can be a basis for changes in the often criticized evaluation protocol P.85. Being developed in 1995, way before the era of high-quality synthesizers of today, this recommendation could be revised considering the results from this study.
As a concluding remark, we can say that the three major quality dimensions of synthetic speech are (i) *naturalness of voice*, (ii) *prosodic quality*, and (iii) *fluency and intelligibility*. Thus, even though the intelligibility of TTS systems substantially increased over the past decade, this dimension is still important for the perceptual quality.

7. 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.

8. References

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

[2] M. Viswanathan and M. Viswanathan, "Measuring Speech Quality for Text-to-Speech Systems: Development and Assessment of a Modified Mean Opinion Score (MOS) Scale," *Computer Speech and Language*, vol. 19, pp. 55–83, 2005.

[3] 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)*, pp. 240–245, 2012.

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

[5] L. Tsogo, M. Masson, and A. Bardot, "Multidimensional Scaling Methods for Many-Objects Sets: A Review," in *Multivariate Behavioral Research*, vol. 35, no. 3, 2000, pp. 307–319.

[6] 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)*, pp. 1725–1728, 2005.

[7] "The Festival Speech Synthesis System." [Online]. Available: http://www.cstr.ed.ac.uk/projects/festival/

[8] J. S. Garofolo, *Getting Started with the DARPA TIMIT CD-ROM: An Acoustic Phonteic Continuous Speech Database*, National Institute of Standards and Technology (NIST), Gaithersburgh, MD, 1988.

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

[10] K. Seget, *Untersuchungen zur auditiven Qualität von Sprachsyntheseverfahren (Study of Perceptual Quality of Text-to-Speech Systems)*. Diplomarbeit, Lehrstuhl für Netzwerk- und Systemtheorie, Christian-Albrechts-Universität Kiel, 2007.

[11] C. Mayo, R. A. J. Clark, and S. King, "Listeners' Weighting of Acoustic Cues to Synthetic Speech Naturalness: A Multidimensional Scaling Analysis," *Speech Communication*, vol. 53, pp. 311–326, 2011.

[12] 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.

[13] 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.

[14] F. Hinterleitner, C. Norrenbrock, and S. Möller, "Perceptual Quality Dimensions of Text-To-Speech Systems in Audiobook Reading Tasks," *Proc. of the 24th Konferenz Elektronische Sprachsignalverarbeitung (ESSV), Bielefeld, (Germany)*, 2013.

[15] "The Blizzard Challenge 2012." [Online]. Available: http://www.synsig.org/index.php/Blizzard_Challenge_2012