TOWARDS A BETTER UNDERSTANDING OF TTS SYNTHESIS: SUBJECTIVE QUALITY AND ITS INSTRUMENTAL ASSESSMENT

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

[1] *Digital Signal Processing and System Theory, Christian-Albrechts-University of Kiel, Germany*

[2] *Quality and Usability Lab, Deutsche Telekom Laboratories, TU Berlin, Germany*

cno@tf.uni-kiel.de, florian.hinterleitner@telekom.de

Abstract: The purpose of this contribution is to give new insights into instrumental quality estimation of text-to-speech (TTS) signals. Two main aspects are in our focus: (1) What makes up the subjective quality of TTS signals from the naïve listener perspective ? (2) How can the subjective quality be measured instrumentally? Regarding the first question typical impairments in TTS signals are identified based on a newly assembled German auditory test database comprising 14/15 state-of-the-art TTS systems for female/male voices. The results of a full semantic differential are described with emphasis on the potential to describe the quality space by means of a small number of quality dimensions. The second question addresses the development of a suitable feature set for signal-based estimation of subjective quality. We take up the idea of auditory-inspired modulation features which have been shown to represent most of the articulatory information conveyed in speech signals. The potential for robust instrumental quality diagnosis is discussed.

1 Introduction

The perspective of an engineer who is about to construct a useful system is most likely divergent to the perspective of the costumer who might like to buy this system. This discrepancy applies especially to the area of speech-processing research where the subjective quality of an altered speech signal ought to be assessed by means of a proper listening test or via a suitable instrumental quality estimator. Clearly, both approaches are linked in that the latter should best reproduce any listening test result, given the same conditions. For TTS signals such a link has yet to emerge. Since the quality of TTS systems is about to reach a level suitable for mass-market applications (e.g. email and SMS readers), we wish to bring light into the costumer perspective in the sense of perceptual quality in Section 2. Furthermore, implications for instrumental quality measurement are discussed in Section 3. The paper closes with a conclusion and an outlook in Section 4.

2 Perceptual quality

A person listening to a speech signal is easily able to decide right away whether there is a human being speaking or a computer which synthesizes an orthographic text. It may be assumed that this involves some rating about the naturalness of the signal which is unconsciously produced

in the brain. This rating effectively defines the perceptual quality. There are various techniques to investigate the mechanisms behind this process. In any case, some listening-test procedure is necessary where one would not just ask for the overall quality of a signal, rather for ratings of certain aspects which are assumed to be relevant for the quality impression. We expect that such ratings are not as vulnerable to subjective biases than overall quality ratings. Circumventing the problem of choosing some set of attributes from which the test designer (possibly the TTS engineer!) thinks that they capture the quality space of the average costumer, we tried to identify all relevant attributes in a separate listening test carried out prior to a multidimensional analysis. This is the starting point for the semantic differential (SD) which has been carried out in [1]. A short summary follows in the next subsections, accompanied by some signal examples.

2.1 TTS database

10 German sentences were prepared such that none of them contained words which are off the scope of German dictionaries (e.g. proper names). The spoken length was about 10 s to avoid listener fatigue during the listening tests. 14/15 different state-of-the-art TTS systems were used for synthesis with female/male speakers, for some of them with up to 6 different voices. Overall 350/280 different test stimuli were produced. Furthermore, the sentences were read by 4/4 amateur and 4/4 professional speakers and recorded to provide a natural reference. All speech files were downsampled to 16 kHz and level normalized to -26 dBov.

2.2 Semantic Differential

Based on the TTS database described in the previous subsection, 3 separate listening tests were conducted in order to yield a thorough description of the quality space: In a first pretest, 2179 attributes were collected from 12 expert listeners (4 female, 8 male) which were then manually condensed to 28 attribute scales (antonym pairs) by choosing the most frequently named and dropping obviously redundant synonyms. In the second pretest, 9 expert listeners (3 female, 6 male) and 13 naïve listeners (8 female, 5 male) were instructed to rate the test signals using only those attribute scales derived from pretest 1 which they found relevant for expressing their quality impression. Subsequently, scales that correlated highly with others or were used rather rarely were excluded from further analysis. With the aid of a Principal Component Analysis (PCA) 16 attribute scales were identified which were used for the final main test. Here, 30 naïve listeners rated 30 stimuli per gender, with each synthesizer configuration represented by 2 stimuli. In Table 1 the 16 attribute scales used in the main test are given.

2.3 Quality dimensions

The most interesting task is now to evaluate the remaining redundancy of the attributes from Table 1 and hence to reduce the quality space further. In [1] a principal factor analysis with 3 factors was carried out. We tested several methods for dimensionality reduction. For comparison purposes, we give the rotated factor matrix in Table 2 which resulted from a PCA with subsequent Varimax rotation. Only the first 4 components are considered for analysis accounting for 64.73 % of the data variance. The blanks in Table 2 indicate loadings below 0.4 and were dropped for better readability. Significant loadings above 0.6 are in bold. From the matrix, attributes can be grouped into 4 dimensions which are indicated by separate table blocks: (1) The first dimension is denoted as **naturalness**. Most attributes in this category reflect suprasegmental quality issues, where a proper prosodic structure (e.g. RHYT) is crucial for high naturalness. However, attributes BUMP and DSTO suggest that segmental quality, i.e. articulation, might

ACRONYM	ENGLISH	GERMAN
ACT	unnatural accentuation vs. natural accentuation	unnatürliche Betonung vs. natürliche Betonung
NAT	artificial vs. natural	künstlich vs. natürlich
RHYT	unnatural rhythm vs. natural rhythm	unnatürlicher Rhythmus vs. natürlicher Rhythmus
PLT	unpleasant vs. pleasant	unangenehm vs. angenehm
TENS	tense vs. calm	angespannt vs. ruhig
BUMP	bumpy vs. not bumpy	holprig vs. nicht holprig
DSTO	distorted vs. undisorted	verzerrt vs. nicht verzerrt
HISS	hissing vs. not hissing	zischend vs. nicht zischend
NOIS	noisy vs. not noisy	verrauscht vs. rauschfrei
RASP	raspy vs. not raspy	kratzig vs. nicht kratzig
DSTU	undisturbed vs. disturbed	gestört vs. ungestört
CLIN	clinking vs. not clinking	klirrend vs. nicht klirrend
POLY	several voices vs. one voice	mehrstimmig vs. einstimmig
COMP	unintelligible vs. intelligible	unverständlich vs. verständlich
FLUE	interrupted vs. continous	unterbrochen vs. kontinuierlich
SPE	fast vs. slow	schnell vs. langsam

Table 1 - Attribute scales used in the main test. Since the test language was German we also give the corresponding translations.

also play a major role here. (2) The second dimension is denoted as **disturbances** and mainly represents the quality of the *sound*. (3) The third dimension **temporal distortions** mainly captures the attribute POLY, linked with the impression of several voices speaking at the same time (see Section 2.4). (4) The fourth dimension is almost exclusively dependent on the perceived **speed** (SPE). Strictly speaking, this is not a quality dimension, yet it is considered a crucial attribute of synthetic speech and is thus kept for further analysis.

	FACTOR LOADINGS			
	1	2	3	4
ACT	**0.86**			
NAT	**0.85**			
RHYT	**0.84**			
PLT	**0.77**			
BUMP	**0.68**			
TENS	0.60			0.46
DSTO	0.56	0.41		
HISS		**0.77**		
NOIS		**0.74**		
RASP		**0.68**		
DSTU	0.42	0.58	0.36	
CLIN		0.47		0.40
POLY			**0.87**	
COMP	0.48		**0.61**	
FLUE	0.51		0.56	
SPE				**0.88**

Table 2 - PCA factor matrix from the main test data of the SD after Varimax rotation. The matrix contains loadings of the first 4 dimensions where blanks indicate loadings below 0.4.

2.4 Signal examples

To highlight the practical meaning of the results from the SD, we give some signal examples[3] which reflect the quality space in some core aspects. Samples are cut from the sentences used in the SD and natural samples are provided for comparison. For a given scale, a mean opinion score (MOS) of a signal is considered as significantly poor if it is below a threshold. This threshold is evaluated as the per-gender mean of all available MOS values of that scale minus the standard deviation.

Example 1 (Fig. 1) gives 2 versions of the phrase "Kannst du mir sagen". Figure 1(a) shows the synthesized version and Figure 1(b) the natural version, both for male voices. Above the waveforms, we give the corresponding (interpolated) curves of the fundamental frequency F_0, estimated using the standard configuration of the phonetics software PRAAT [2]. The synthesized version provokes the typical auditory impression of a robotic voice: Unnatural rhythm (i.e strongly accentuated transition at 0.3 s, between /kanns/ and /t/) and a rather flat speech melody (ΔF_0 = 30 Hz). Furthermore, a common TTS artefact appears at instant 0.95 s at the boundary between the 2 syllables /sa/ /gen/. The pronounced energy gap gives rise to the impression of the syllables spoken separately and overly accentuated due to missing coarticulation. This is also documented by an invalid pitch estimation ($F_0 = 0$ Hz). Furthermore, an artificial "acceleration" within the word /sagen/ can be noted. The signal (complete sentence) was rated as significantly poor on scales BUMP and FLUE. In comparison, we note that the natural version in Figure 1(b) exhibits a faster tempo and a smoother pitch curve with increased range ($\Delta F_0 = 50$ Hz).

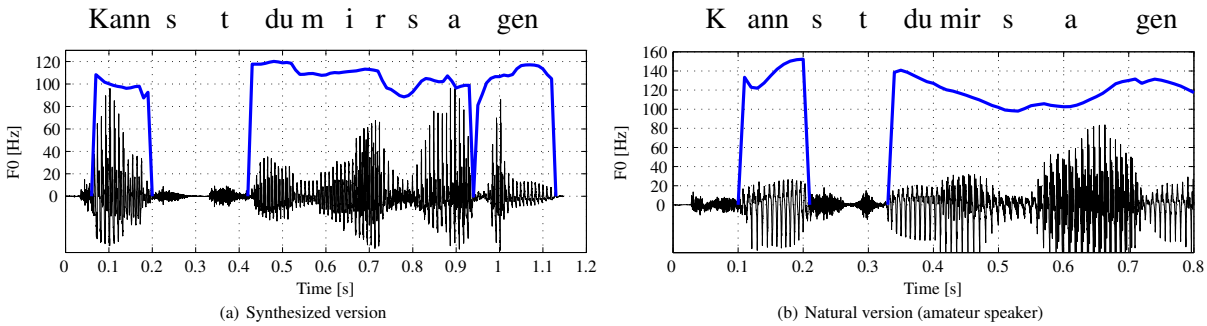
Example 2 (Fig. 2) gives 2 versions of the German word "Möglichkeit", both for male voices. Again, the synthesized version in Figure 2(a) suffers from inappropriate syllable prominence, yet in a different way. The vowels /ö/ (0.1 s) and /i/ (0.32 s) are hold overly long and give rise to the impression of a speech disorder, again accompanied by a flat pitch curve ($\Delta F_0 = 30$ Hz). The voice quality is also low, however this is not visible in the plot. The whole signal was rated as significantly poor in the dimension *naturalness* (e.g. RHYT, BUMP, DSTO). Interestingly, the impression of overall speed (SPE) is not significantly slower; it appears that, apart from the overall phone rate, the duration of consonants and pauses mainly defines the perceived speed. Figure 2(b) gives the version of a natural speaker for comparison ($\Delta F_0 = 50$ Hz).

Example 3 (Fig. 3) gives 2 versions of the utterance "Ich hasse es wenn es", using female voices. The auditory impression of the synthesized version in Figure 3(a) is dominated by a number of jumps, most pronounced at time 0.55 s (schwa-vowel transition from /hasse/ to /es/) and at time 1.05 s (nasal-vowel transition from /wenn/ to /es/). Such jumps are often due to missing pauses (at 0.55 s) and slight pitch jumps (at 1.05 s). The higher energy of the unvoiced sections (i.e. the fricatives /s/) relative to the voiced sections does not play a quality-defining role here. As expected, dimensions *naturalness* (e.g. RHYT, BUMP, and FLUE) and *temporal distortions* are rated as significantly poor. The POLY scale received the second worst rating from all female stimuli. This scale is mainly influenced by unnatural jumps which are not realizable by human speakers due to the bounded speed of the vocal-tract movement. This produces the impression of another person speaking at the same time. Figure 3(b) shows the natural version. The pitch range is again much higher ($\Delta F_{0,nat} = 120$ Hz vs. $\Delta F_{0,syn} = 65$ Hz).

3 Instrumental quality estimation

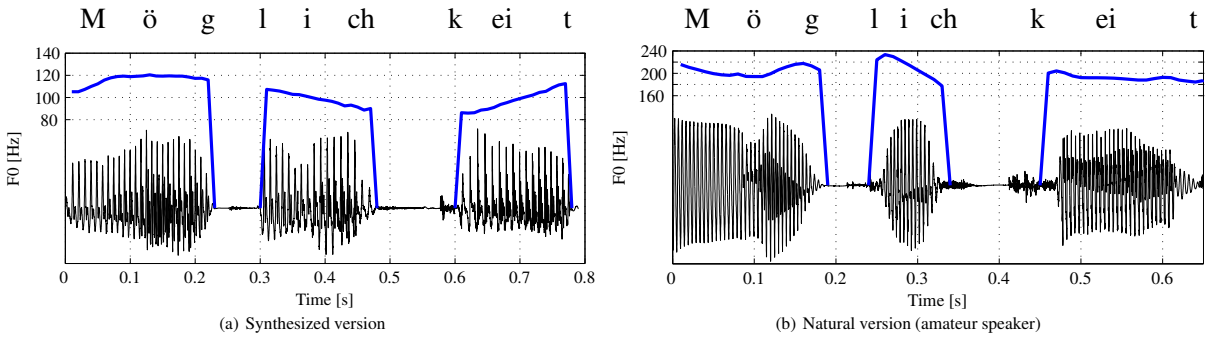
Recalling the notes on perceptual quality from the previous section, the instrumental quality estimation of a TTS signal is related to its similarity to a natural version, spoken by any hu-

[3] All examples are available for listening on *http://www.dss.tf.uni-kiel.de/publications.*



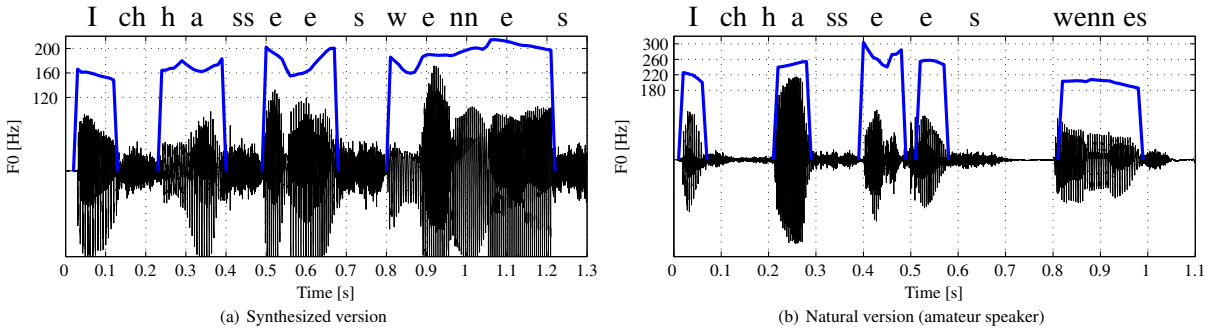
(a) Synthesized version (b) Natural version (amateur speaker)

Figure 1 - Two versions of the German phrase "Kannst du mir sagen", both for male voices. Above the waveforms the corresponding pitch curve (F_0) is given for which the y-scale applies.



(a) Synthesized version (b) Natural version (amateur speaker)

Figure 2 - Two versions of the German word "Möglichkeit", both for male voices. Above the waveforms the corresponding pitch curve (F_0) is given for which the y-scale applies.



(a) Synthesized version (b) Natural version (amateur speaker)

Figure 3 - Two versions of the German phrase "Ich hasse es wenn es", both for female voices. Above the waveforms the corresponding pitch curve (F_0) is given for which the y-scale applies.

man speaker. Two main questions follow from this: (1) What signal-based features capture the relevant quality aspects? (2) How can one provide a reference for naturally spoken speech? The first question imposes presumably the biggest challenge. This is because most subjective quality aspects of TTS (see Section 2.4) join the lack of reliable and analytical models for a purely signal-based description. Even apparently simple tasks are charged with full psychoacoustic complexity. For example, in [3] several spectral distance measures were tested in order to detect audible concatenation errors within single words. However, an instrumental classifier achieved only a correct prediction rate of 37%. It can be inferred that only a sufficient number of different features, which all suffer some unreliability, allows for an instrumental quality estimator of reasonable reliability. In the authors' view, this approach is the basis of the current

non-intrusive algorithms recommended in ITU-T Rec. P563 [4] and ANIQUE+ [5], specified to estimate the quality of telephone-bandwidth transmission-degraded speech signals. The method involves the evaluation of several quality-sensitive parameters which are weighted with regressive coefficients to yield the estimated MOS. The mortgage of this concept is often missing information about the (reference) behaviour of individual parameters. Another way to overcome this gap is by means of statistical feature modeling (Gaussian mixtures, hidden Markov models) [6], where the application of an MFCC-based feature set yielded promising results in case of TTS signals. Yet, the instrumental estimation of restricted quality aspects as in Section 2, not just overall quality, first requires the study of more specific feature sets. Thus, we summarize the quality aspects to be measured from an engineering viewpoint in the following section. Afterwards, we address a main approach used in ANIQUE+ [5], namely the detection of unnatural modulation components in speech for non-intrusive quality diagnosis.

3.1 Physical features of TTS quality

Recalling the results of the SD from Section 2 we identify 3 quality aspects in TTS which need to be measured:

- *Prosody quality* with emphasis on intonation, rhythm, accentuation and speed imposes the biggest quality issue in TTS synthesis.
- *Voice and sound quality*. This refers to any signal processing which deteriorates the naturalness of the vocal source. Sound quality is associated with "add-on" processing artefacts not perceived as being part of the speech production.
- *Unnatural jumps and missing pauses*. According to the factor analysis this type of concatenation error is perceived separately from the remaining aspects of prosody.

3.2 Modulation spectrum

It is well known that the physiological perception of speech signals can be analyzed within the scope of modulation theory. In short, the vocal source performs prevalent frequency modulation (pitch) and the vocal tract movement performs prevalent amplitude modulation at a reasonably slow rate (0.25-20 Hz) [8, 9]. Speech intelligibility arises mainly from the latter [7] and will be analyzed by means of the speech-envelope spectrum (SES) which is essentially adopted from [8].

3.2.1 Method

The signals are octave-band filtered with center frequencies between 0.125 and 4 kHz. This range is assumed to be suffcient for modulation analysis. Filter outputs are squared, low-pass filtered, and downsampled to 200 Hz yielding temporal intensity envelopes of the signal. Each envelope is then subjected to a $\frac{1}{3}$-octave-band filter bank with center frequencies between 0.25 and 48 Hz. We extended the frequency range against [8] since TTS signals might well contain modulation components at frequencies beyond those present in natural speech. Modulation amplitudes are given by the output of these filters normalized by the mean of the resampled, squared waveform of the corresponding octave-filter output. Averaging across the $\frac{1}{3}$-octave bands yields the final SES. Note that this method does not employ the use of the Fourier transform as in [9].

3.2.2 Results

We calculate natural reference SES for both genders separately by averaging the per-signal SES over the natural recordings (40 stimuli for each gender), see Figure 4. Both SES are peaked at 4 Hz modulation frequency (f_{mod}) which matches to the average syllable rate of natural speech [9]. The SES of the 3 examples from Section 2.4 all have similar shapes, however with some interesting deviations. The SES of Example 1 (increased bumpiness) exhibits reduced low frequency ($f_{mod} < 6$ Hz) content. For the higher frequencies the SES matches the reference quite well apart from an atypical rise for $f_{mod} > 20$ Hz. The SES of Example 2 (overlengthy vowels) is much flatter than the reference, with a reduced ratio between high frequency ($f_{mod} > 4$ Hz) and low frequency ($f_{mod} \leq 4$ Hz) modulation indices. Example 3 stands out by chronic signal jumps (concatenation errors). The corresponding SES is also flatter, but with reduced low-frequency content ($f_{mod} < 10$ Hz).

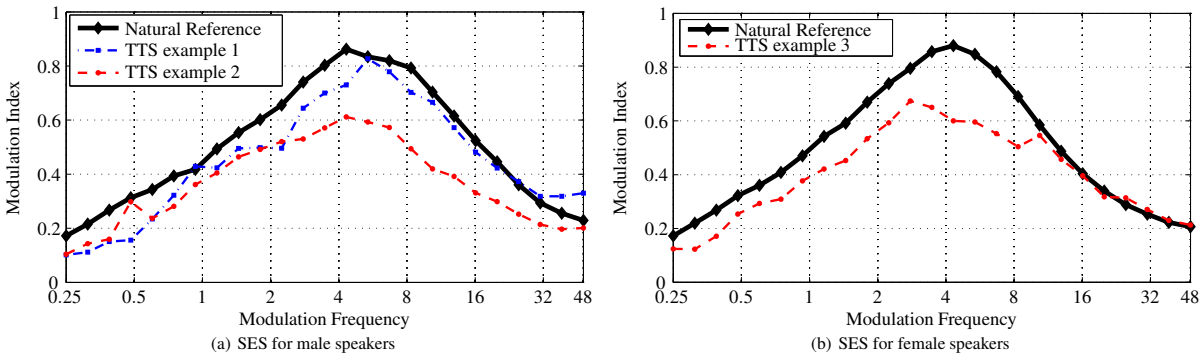


Figure 4 - Speech-envelope spectra (SES) for the signal examples from Section 2.4. Complete sentences were used for analysis. The natural reference represents the average SES for (a) male and (b) female speakers.

3.2.3 Discussion

As a matter of fact, these examples demonstrate that SES analysis provides information about the temporal structure of a speech signal. A review of all SES from the database revealed that all signals with severe chronic temporal distortions showed large deviations from the reference SES. Nevertheless, some outliers suggest some nonlinear relation with perceived quality. Hence, further study of the SES is necessary. Interestingly, we found the sum of modulation indices in the range below 3 Hz modulation frequency to correlate with perceived speed (SPE). Per-stimulus Pearson correlation coefficients attained R = 0.53 and R = 0.70 for male and female stimuli, respectively. The evaluation of some modulation energy ratio as used in [5, 9, 6] did not turn out to be useful. This is plausible concerning the relation to speed, since higher modulation frequencies or their ratio to the lower frequencies might rather indicate aspects of accentuation and intelligibility [9, 8], not speed. This result is in line with [7], where it was shown that decreasing speech tempo causes only a shift of the SES towards lower frequencies. In summary, the SES does not show sufficient accuracy for robust quality indication of TTS signals, at least in the case of the diverse TTS database used in the experiment. But the SES is potentially useful for analysis of less variate databases. For example, when modifications of a given TTS system shall be evaluated using a large data set, the SES can help to identify distorted signals.

4 Conclusion and Outlook

Based on a newly assembled TTS database, steps towards an established link between subjective and instrumental quality have been presented. A full semantic differential revealed 4 main quality dimensions: *naturalness, disturbances, temporal distortions* and *speed*. The modulation spectrum of speech (SES) allows for analysis of the temporal structure of speech, where heavily distorted signals show large deviations from the natural mean SES. However, some outliers suggest a closer study of the different calculation methods of the modulation spectrum, proposed in the literature [5, 9, 8, 6]. Furthermore, the joint evaluation of temporal structure and fundamental frequency opens a reasonable perspective towards TTS-specific measurement of prosodic quality.

5 Acknowledgements

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) under grants HE 4465/4-1 and MO 1038/11-1.

References

[1] HINTERLEITNER, F., MÖLLER, S., NORRENBROCK, C., HEUTE, U.: Perceptual Quality Dimensions of Text-to-Speech Systems. In: *Proc. 12th Int. Conf. on Spoken Language Process. (Interspeech 2011)*, Florence, Italy, 2011

[2] BOERSMA, P. AND WEENIK, D.: PRAAT, software for speech analysis and synthesis. *http://www.fon.hum.uva.nl/praat*, 2005.

[3] STYLIANOU, Y., SYRDAL, A.K.: Perceptual and Objective Detection of Discontinuities in Concatenative Speech Synthesis. In: *Proc. Int. Conf. Acoust., Speech, Signal Proc.*, vol. 2, pp.837-840, 2001.

[4] ITU-T REC. P.563: Single-ended Method for Objective Speech Quality Assessment in Narrow-band Telephony Applications. *Int. Telecomm. Union*, Geneva, 2004.

[5] ANSI ATIS 0100005-2006: Auditory non-intrusive quality estimation plus (ANIQUE+): Perceptual model for non-intrusive estimation of narrowband speech quality. *American National Standards Institute* , Washington DC, 2006.

[6] FALK, T.H.: Blind Estimation of Perceptual Quality for Modern Speech Communications. *Ph.D. thesis, Queen's University, Kingston, Ontario, Canada*, 2008.

[7] STILP, C.E., KIEFTE, M., ALEXANDER, J.M., KLUENDER, K.R.: Cochlea-scaled spectral entropy predicts rate-invariant intelligibility of temporally distorted sentences. In: *Journal of the Acoustical Society of America*, 128(4), pp. 2112-2126 , 2010.

[8] HOUTGAST, T., STEENEKEN, H.J.M.: A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. In: *Journal of the Acoustical Society of America*, 77(3), pp. 1069-1077 , 1985.

[9] FALK, T.H., CHAN, W.-Y. AND SHEIN, F.: Characterization of atypical vocal source excitation, temporal dynamics and prosody for objective measurement of dysarthric word intelligibility. In: *Speech Communication* [In Press, Corrected Proof], 2011.