Assessing a Speaker for Fast Speech in Unit Selection Speech Synthesis

Donata Moers[1], *Petra Wagner*[2]

[1] Institut für Kommunikationswissenschaften, Abt. Sprache und Kommunikation, Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn, Germany

[2] Fakultät für Linguistik und Literaturwissenschaft, Universität Bielefeld, Bielefeld, Germany

dmo@ifk.uni-bonn.de, petra.wagner@uni-bielefeld.de

Abstract

This paper describes work in progress concerning the adequate modeling of fast speech in unit selection speech synthesis systems, mostly having in mind blind and visually impaired users. Initially, a survey of the main characteristics of fast speech will be given. Subsequently, strategies for fast speech production will be discussed. Certain requirements concerning the ability of a speaker of a fast speech unit selection inventory are drawn. The following section deals with a perception study where a selected speaker's ability to speak fast is investigated. To conclude, a preliminary perceptual analysis of the recordings for the speech synthesis corpus is presented.

Index Terms: speech synthesis, unit selection, fast speech

1. Introduction

Especially the blind and visually impaired prefer a fast speech output when using a speech synthesis system [1, 2, 3]. However, up to now fast speech is inadequately implemented in unit selection synthesis systems. Architectures like formant or diphone synthesis are able to produce synthetic speech at a fast speech rate, but the generated speech does not reflect the characteristics of natural fast speech.

The phonetic characteristics of natural fast speech are found to be very different from those of speech produced at "normal" speech rates. The faster somebody speaks the less intelligible his utterances become. For the most part this is up to the increasing overlap of single segments when speaking rate increases. The articulatory targets important for a clear pronunciation are no longer reached [4, 5]. In vowels, this mainly becomes manifest in shorter duration and a change in characteristic formant frequencies [6, 7]. Consonants are assimilated more often and are even changing their consonantal category. Their intensity decreases as well as their realization becomes incomplete. Sometimes, they are even elided completely [6, 8]. Larger units like syllables or intonation phrases are affected likewise: the duration of syllables is shortened and the total number of stressed syllables decreases [9, 10].The number and strength of phrase boundaries declines [10] and the fundamental frequency contour becomes flatter [11]. Speaking generally, coarticulation, reduction, assimilation and elision are augmenting throughout fast speech.

In order to model fast speech in speech synthesis, there are several options. The first is to accelerate linearly the "normal" speech by means of duration manipulation. The second is to mimic certain prosodic features typical for fast speech such as fewer and shorter pauses, flatter intonation contour and decreased strength and number of prosodic boundaries. Previous studies indicate that these approaches lead to different results in perception experiments. E. g. artificially produced fast words whose temporal pattern was equivalent to natural fast speech were judged to be less intelligible than artificially produced fast words which were simply linearly compressed. The less the stimulus deviated from its canonical form the better the word was understood by listeners [12]. This indicates that a clear pronunciation is still preferred over a synthesis that includes typical phonetic characteristics of natural fast speech such as reductions, elisions and strong coarticulation.

Furthermore, in a comparison of two synthesis architectures where a linear tempo manipulation is easily performed, i. e. formant or diphone synthesis, blind listeners preferred the less natural sounding formant synthesis over diphone synthesis with regards to intelligibility in very fast speech [3]. This indicates that the fast and smooth acoustic transitions in natural speech are also important for the intelligibility of synthetic speech. Such transitions are not treated adequately by traditional diphone concatenation synthesis but can be modeled by a formant synthesis. Since discontinuities pose a problem for concatenative synthesis in general and unit selection synthesis in particular, Breuer [13] suggested to simply treat certain phone sequences which are prone to heavy coarticulation as atomic in the sense that they are regarded as two or more phones, but one indivisible synthesis unit. This approach might lead to a possible solution to model fast synthetic speech both naturally – by using prerecorded concatenation units – and intelligibly – by including typical smooth transitions in heavily coarticulated contexts.

Taking into account the aforementioned preconditions, the main focus of the project is the definition of robust directives which should be obeyed when building a unit selection synthesis for the visually impaired that can produce fast or very fast speech in an acceptable quality regarding both intelligibility and naturalness. Thus, the approach chosen here includes the creation of an independent inventory inherently showing all segmental and suprasegmental characteristics of natural fast speech and at the same time avoiding too heavy reduction and coarticulation for the benefit of intelligibility.

2. H&H theory

As coarticulation as well as reduction affect the intelligibility of natural speech adversely one has to ask if there basically exists a possibility to avoid these phenomena when speaking in a normal or even in a fast rate. An answer to this question is given by the Hypo- and Hyperspeech theory (H&H theory) by Lindblom [14]. It claims that despite the continuous course of speech and the coarticulation and reduction effects involved a speaker should be able to realize a sufficient contrast while speaking. This is necessary to be understood by a listener and thus to communicate successfully.

While talking, the speaker has to choose between articulatory effort on the one hand and reaching the communicative goal on the other hand. So the speech output is influenced by economic as well as by communicative factors. The economic reasons get manifested in less carefully articulated speech (hypospeech), the communicative goal in very clearly articulated (hyper-) speech. Lindblom himself describes the situation as follows: „Hence speakers are expected to vary their output along a continuum of hyper- and hypospeech" [ibid.: 403]. Consequently, speakers should be able to speak both fast and clear if they increase articulatory effort. In order to build a useful synthesis inventory to model fast speech, a speaker needed to be found who was able to realize this speaking style best.

3. Speaker requirements

Research in unit selection speech synthesis has shown that the quality of the synthetic speech for the most part is determined by the inventory speaker. Skilled speakers who learned to speak with consistent voice quality and high articulatory precision over a long period will generally produce an inventory at higher quality than untrained speakers [15].

If the inventory is based on fast speech the emerging problems of articulatory precision and consistent voice quality would presumably increase. Assuming that untrained speakers will reduce the articulatory precision for the benefit

of economic reasons to a greater extent than skilled speakers the inventory speaker should fit the following criteria:

- He/she should be a skilled speaker who is able to speak both very fast and very clearly. Previous studies for German [16] and Dutch [17] showed a maximum speaking rate at approximately 8 syllables/second when the speech was still highly intelligible. This rate is the set target for the fast speech rate inventory which is to be developed here.
- The speaking experiences of the speaker should not emanate from one specific domain. He/she should not use a specific speaking style like news anchor or auction house style because these speaking styles are nontransferable to other domains.

Based on these requirements the search for a suitable speaker began with a group of 9 voluntary people who had done corpus recordings for speech synthesis before or had other speaking related experiences as a radio presenter or similar. Prerecordings were carried out, based on different tasks. The tasks included the reading of 5 German sentences in normal and fast speech rate as well as the realization of 2 additional sentences containing some English phrases, also in two rate conditions. Altogether, there were 6 female and 3 male candidates whose speech was judged by 12 phonetically trained listeners. The individual speaker's fastest possible articulation rate, the perceptual clarity concerning fast speech and their individual voice characteristics were assessed. The sustainment of voice quality and voice intensity as well as accuracy of articulation and naturalness of intonation and pronunciation – the latter foremost in the fast speech version – were the most important judgment criteria. They are known to be the best guarantee for a high degree of naturalness in unit selection speech synthesis [15].

This way, the presumably most suitable speakers (2 female and 1 male speaker) for a fast speech inventory were determined. After a second run of assessment, one of the female speakers turned out to be the most able to speak very clearly at maximum speaking rate and subsequently was chosen for corpus recordings.

4. Speaker evaluation

As during the prerecordings no special attention was paid to the precision of articulation, again recordings at both normal and fast speech rate were carried out. These recordings were based on a text which already was used in the BonnTempo-Corpus [16]. The text derived from the narrative *Selbs Betrug* by B. Schlink (1994) and included 4 main and 3 subclauses containing 76 syllables. At the beginning, the speaker was told to read the text three times in a normal speech rate. Afterwards, she had to read the text again three times as fast as possible. To prove that she was indeed able to speak both fast and clear three further recordings were carried out. At this, the speaker was asked to intentionally enhance the articulatory effort and to speak particularly clear. The speech rate was intended to increase for each of the three fast versions in both the fast and the fast and clear condition. Thus, there were six samples of fast speech to analyze, three consisting of simply fast speech and three consisting of both fast and very clear speech, respectively.

4.1. Acoustic evaluation

As a first step, an analysis of the acoustic characteristics of the different fast rate versions was performed. The question was if by means of these characteristics it would become apparent that the speaker was indeed able to avoid undesirable effects like coarticulation and reduction in fast speech. In detail, the following phenomena were analyzed:

- Shortening and reduction of vowels
- Schwa elision
- Syllabification of consonants
- Assimilation of consonants
- Incomplete closure and/or incomplete plosive bursts
- Changes in Voice Onset Time
- Reduced intensity of fricatives
- Reduced number of stressed syllables
- Reduced number and duration of pauses
- Reduced number and intensity of phrase boundaries
- Flattened fundamental frequency contour

As this was an extensive analysis a detailed description is not included here. Nevertheless, the results of the acoustic evaluation in general showed that – in line with the H&H theory – all of the phenomena specified above occurred more rarely in the fast and clear speech utterances than in the simply fast utterances, which were realized without any exceptional articulatory effort.

4.2. Perceptual evaluation

Subsequently, a perceptual evaluation of the different fast versions was conducted. Therefore, excerpts (cf. Figure 1a, 1b) of the different recordings were selected which were expected to show both distinct coarticulation and reduction effects. A perception experiment was created consisting of nine subtest. Each of the subtests contained the same excerpts of the different fast as well as both fast and clear versions. The excerpts were compared pairwise and judged by phonetically skilled (n = 10) as well as by phonetically untrained (n = 13) listeners. It was anticipated that the explicitly clearly articulated utterances would be judged as being more comprehensible than the fast but not clearly spoken versions.

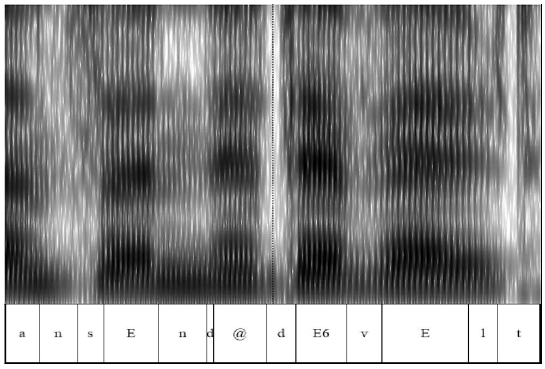

Figure 1a: *Spectrogram showing the excerpt "ans Ende der Welt" (to the end of the world) of a clear fast speech version.*

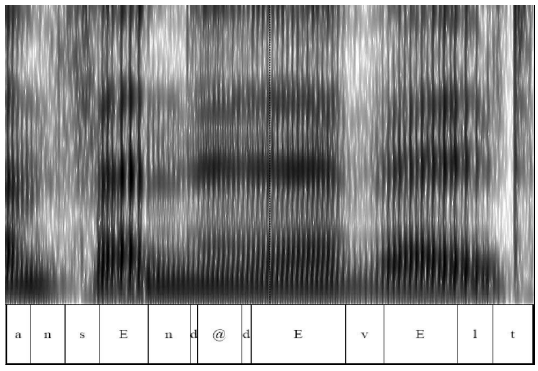

Figure 1b: *Spectrogram showing the excerpt "ans Ende der Welt" (to the end of the world) of an unclear fast speech version.*

The excerpts were chosen in the way that the content was still intelligible. However, to avoid problems in comprehension the text of each excerpt was displayed at the beginning of each subtest. Furthermore, the subjects had the possibility to repeat the stimuli up to three times. Altogether, they were presented 135 stimuli, each of them consisting of a pair of the same excerpt deriving from different versions. The subjects were instructed to choose from each pair the realization which

was pronounced more clearly or rather the one they understood better. The experiment was conducted in a quiet environment, stimuli were presented via earphones.

Table 1: *Speech rate (syllables per second), mean value of similarly fast versions, gained points, scaled number of points.*

Version	Speech rate	Mean value	Gained points	Scaled points
clear03	7,25	7,30	670	674,76
unclear01	7,35	7,30	568	563,96
clear01	7,53	7,69	701	716,21
unclear03	7,85	7,69	342	334,89
clear02	8,26	8,32	576	580,16
unclear02	8,38	8,32	247	245,24

Because of the varying intended speech rates, first of all the speech rate was defined precisely in syllables per second for each version (cf. Table 1). After that, for each pair of similarly fast versions the arithmetic mean value was calculated. To measure the difference in intelligibility between the different versions, each excerpt which was judged as articulated more clearly or rather more comprehensible received one point. In order to account for the varying tempo the total number of gained points was divided by the exact speech rate and then multiplied by the mean value of the corresponding pair. In doing so, a normalized value was obtained which gave an account of the "better to understand and/or more clearly" judgments relative to the speech rate.

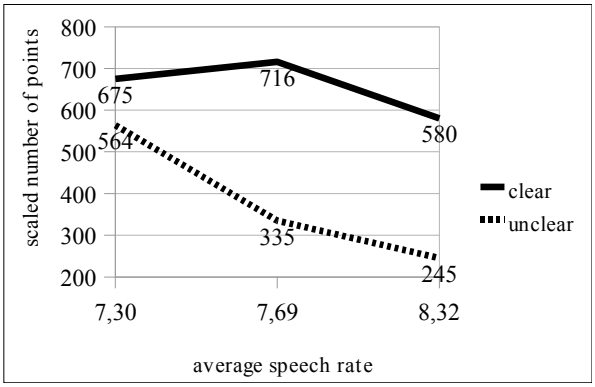

Figure 2: *Judgment in scaled number of points, mapped on speech rate.*

Looking at Figure 2 it already becomes obvious that the fast and intentionally clear articulated versions perform significantly better than the inarticulated versions. A chi-square test confirms these findings ($p < 0.001$). Between the phonetically skilled and the phonetically untrained listeners there was no significant difference in judgment to observe.

4.3. Corpus recordings

After validating that the chosen speaker was indeed able to speak both very fast and very clearly, corpus recordings started. The base of these recordings were 400 sentences which were selected randomly from the BITS Corpus [19]. The BITS-Corpus itself was chosen due to its availability and its phonologically balanced design fulfilling the general criteria of unit selection speech synthesis systems. It was developed especially for diphone and unit selection speech synthesis and comprises more than 1600 sentences in total. The selected 400 sentences were recorded in 2 conditions:

- normal speech rate (4 syllables per second)
- maximum clear speech rate (8 syllables per second)

All recordings were conducted in a sound treated recording studio. Due to the fact that not all recordings could be done in one session a strict monitoring of speaking rate, phrasing and intensity was necessary. Consequently, prior to each session as well as within the sessions, several reference sentences were presented to the speaker in order to (re)adjust her performance. The reference sentences were recordings of the first session. Special attention was paid to the adjustment of speaking rate, phrasing, accentuation style and intensity. To reach the fastest rate of speech possible it has proven useful to guide the speaker to the designated tempo gradually [20]. So, fast versions of one sentence were recorded repeatedly in succession, accelerating the speaking tempo and enhancing the articulatory effort each time, as often as possible.

Thus, two unit selection inventories were created: one in normal speech rate and one in fast speech rate articulated as accurate as possible.

4.3.1. Evaluation of linear compression

As Janse [12] found out, artificially produced fast spoken words whose temporal pattern was equivalent to natural fast speech were judged to be less intelligible than artificially produced fast spoken words which were simply linearly compressed. The less the stimulus deviated from the canonical form the better the word was understood by listeners.

Taking these findings into account the first part of the evaluation was to determine whether the normal speech rate sentences were judged as more intelligible than the fast speech rate sentences when having the same speech rate. Therefore, the normal rate sentences were sped up linearly by means of the TD-PSOLA algorithm until they met the higher speech rate of the corresponding fast sentences. It was expected that in this condition the stimuli based on the normal rate versions were judged to be more intelligible, but maybe not as natural as the unmodified fast versions.

The next step was the acceleration of both versions to an even faster speech tempo, namely twice the tempo of the underlying fast speech rate versions. Thus, the sentences which were generated from the normal rate sentences had to be manipulated more strongly with respect to their duration, whereas the sentences generated from the fast rate speech required a comparatively small durational manipulation. The extensive manipulation of the normal rate versions may create another variable influencing the results of the perception experiments: artifacts which are known to appear when using the TD-PSOLA algorithm [21]. Nevertheless, it was decided to use this algorithm here because it is still generally applied in speech synthesis systems. Thus, the stimuli generated from fast speech sentences were expected to be judged as more intelligible and more natural than the stimuli generated from normal rate sentences.

The experiment included 20 sentences which were randomly chosen from the 400 recorded corpus sentences. In order to create the stimuli for the first part of the experiment the total duration of the normal and corresponding fast rate utterance was measured and the ratio of their duration was calculated. With the aid of this durational factor, the normal rate version was accelerated linearly to the tempo of the fast version. For the second part of the experiment the sentences were accelerated to twice the tempo of the natural fast speech rate. Altogether, the subjects were presented with 40 stimuli, each of them consisting of a pair of the same sentence generated from the two different underlying versions by linear acceleration. The subjects were instructed to choose from each pair the realization which they understood better or which was pronounced more clearly. They were also asked to judge the naturalness of the more intelligible sample. The experiment was conducted in a quiet environment and stimuli were presented via earphones. 11 subjects took part in the experiment.

The approach to the analysis of the results was similar to the one in the perceptual study presented before: the version of the sentence which was judged to be more intelligible received one point. As was expected, in the first part of the experiment the stimuli generated from normal speech rate sentences were judged to be more intelligible than the natural fast spoken ones (χ^2, $p < 0.05$). This advantage of the normal speech rate stimuli disappears in the very fast condition (cf. Figure 3). There even is a slight tendency to prefer the stimuli generated from natural fast speech, albeit not a significant one. However, the natural fast stimuli are clearly preferred with respect to naturalness (χ^2, $p < 0.0001$). These results confirm our initial hypotheses.

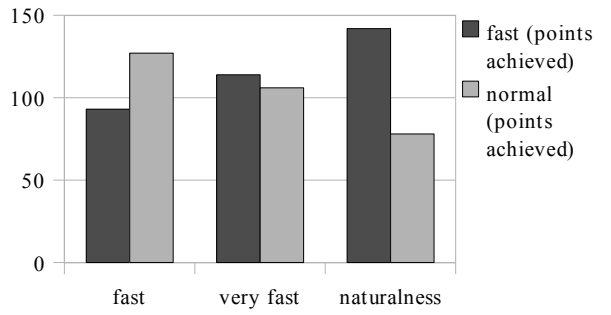

Figure 3: *Intelligibility judgments for fast and very fast stimuli. Naturalness judgments for very fast speech.*

5. Discussion

Due to the results of the acoustic and perceptive evaluation of the preliminary recordings carried out it could be shown that the selected speaker is indeed able to speak very clearly even at fast speech rates. Thereby, she obviously is able to avoid undesirable phenomena like reduction and coarticulation as much as possible. Hence, it became apparent that she is a suitable speaker for a fast spoken unit selection synthesis inventory.

The following evaluation of corpus recordings confirmed the results of Janse [12] for fast speech tempo (8 syllables per second). Stimuli generated linearly from normal rate sentences were judged to be more intelligible than the natural fast spoken ones. In the very fast condition (16 syllables per second), there was no disadvantage of the fast speech rate stimuli; in contrast, there even was a slight tendency to prefer the fast stimuli. However, the very fast stimuli generated from natural fast speech are clearly preferred with respect to naturalness.

The next step of evaluation includes the generation of different utterances by using the two inventories recorded previously as inventory for the unit selection speech synthesis system BOSS [18]. At this, the first sample will be generated from normal rate units and the second sample from fast rate units, both having the same content. The motivation for this is that it is still unclear whether listeners prefer fast synthetic speech generated from fast units (most natural?), compressed normal rate units (most intelligible?) or maybe even a mixture of both, trying to mimic the speaking strategies explained by the H&H theory. Another approach to the investigation of fast speech unit selection synthesis is the application of a different acceleration algorithm, e.g. the non-linear time-scaling algorithm described in [22].

The tests shall be conducted with different listener groups. The first group shall consist of people who are not or only slightly visually impaired (e.g. any impairment can be corrected by wearing glasses or contact lenses). In this group, we expect that the preferred sentences will be the ones generated from the normal rate inventory and that the overall preferred tempo of speech is moderate. A second listener group shall consist of blind or heavily visually impaired people who are reliant on using a speech synthesis system in daily life. Here we expect that these people prefer a very fast speech rate, maybe even not sounding natural anymore and unintelligible for the visually unimpaired.

6. Conclusions

Our paper comprises phonetic knowledge concerning fast speech, discusses speaking strategies and requirements for an inventory speaker deduced thereof. Further, his/her evaluation as well as a first investigation of corpus recordings and acceleration methods are presented. A research strategy to investigate this problem further is considered. If the approach chosen in this investigation proves not to be appropriate to synthesize fast speech in an adequate and acceptable quality other ways of producing fast speech in concatenation based synthesis systems have to be considered.

7. Acknowledgements

The authors would like to thank the speaker, the IfK staff for technical help and all student assistants for support during preparation and accomplishment of the corpus recordings. Special thanks go to the people who particularly encouraged the work which is presented here.

8. References

[1] Moers, D., Wagner, P. and Breuer, S., "Assessing the Adequate Treatment of Fast Speech in Unit Selection Speech Synthesis Systems for the Visually Impaired", Proc. 6th ISCA Workshop on Speech Synthesis (SSW-6), Bonn, 2007.

[2] Fellbaum, K., "Einsatz der Sprachsynthese im Behindertenbereich", in Fortschritte der Akustik. DAGA'96: 78-81, Oldenburg, 1996.

[3] Moos, A. and Trouvain, J., "Comprehension of Ultra-Fast Speech – Blind vs. „Normally Hearing" Persons", in Proc. ICPhS XVI: 677-684, Saarbrücken, 2007.

[4] Goldman-Eisler, F., "The significance of changes in the rate of articulation", Language and Speech 4: 171-174., 1961.

[5] Daniloff, R.G. and Hammarberg, R.E., "On defining coarticulation", Journal of Phonetics 1: 239-248, 1973.

[6] Kohler, K.J., "Segmental reduction in connected speech in German: Phonological facts and phonetic explanations", in Hardcastle, W.J. and Marchal, A. [Ed], Speech Production and Speech Modelling. 69-92, Dordrecht, 1990.

[7] Peterson, G.E. and Lehiste, I., "Duration of syllable nuclei in English", Journal of the Acoustical Society of America 32: 693-703, 1960.

[8] van Son, R. J. J. H. and Pols, L. C. W., "An acoustic profile of consonant reduction", in Proc. ICSLP: 1529-1532, Philadelphia, 1996.

[9] Gopal, H.S., "Effects of speaking rate on the behaviour of tense and lax vowel durations", Journal of Phonetics 18: 497-518, 1990.

[10] Crystal, T.H. and House, A.S., "Articulation rate and the duration of syllables and stress groups in connected speech", Journal of the Acoustical Society of America 88: 101-112, 1990.

[11] Monaghan, A., "An Auditory Analysis of the Prosody of Fast and Slow Speech Styles in English, Dutch and German", in Keller, E. et al. [Ed], Improvements in Speech Synthesis, 204-217, Chichester, 2001.

[12] Janse, E., "Word perception in natural-fast and artificially time-compressed speech", Proc. 15th ICPhS: 3001-3004, Barcelona, 2003.

[13] Breuer, S. and Abresch, J., "Phoxsy: Multi-phone Segments for Unit Selection Speech Synthesis", Proc. ICSLP, Jeju, 2004.

[14] Lindblom, B., "Explaining phonetic variation: A sketch of the H&H-Theory", in Hardcastle, W.J. and Marchal, A. [Ed], Speech Production and Speech Modelling, 403-439, Dordrecht, 1990.

[15] Maus, V., " Zur Frage der Eignung von Sprechern als künstliche 'Stimme' in der konkatenativen Sprachsynthese", unpublished Master's Thesis, Rheinische Friedrich-Wilhelms-Universität Bonn, 2004.

[16] Dellwo, V. and Wagner, P., "Relations between language rhythm and speech rate", Proc. 15th ICPhS: 471-474, Barcelona, 2003.

[17] Janse, E., "Production and Perception of Fast Speech", Dissertation, Universiteit Utrecht, 2003.

[18] Klabbers, E. et al., "Speech synthesis development made easy: The Bonn Open Synthesis System", Proc. of Eurospeech, Aalborg, 2001.

[19] Schiel, F. et al., "Die BITS Sprachsynthesekorpora – Diphon- und Unit Selection-Synthesekorpora für das Deutsche", Proc. Konvens 2006: 121-124, Konstanz, 2006.

[20] Greisbach, R., "Reading aloud at maximal speed", Speech Communication 11: 469-473, 1992.

[21] S.-H. Chen, S.-J. Chen and C.-C. Kuo, "Perceptual Distortion Analysis and Quality Estimation of Prosody-Modified Speech for TD-PSOLA", in Proc. of the ICASSP'06, Toulouse, 2006.

[22] Höpfner, D., "Nichtlinearer Zeitskalierungsalgorithmus für gespeicherte natürliche Sprache", Sprachkommunikation 2008, ITG-FB 211, Aachen, 2008.