

Assessing the Adequate Treatment of Fast Speech in Unit Selection Speech Synthesis Systems for the Visually Impaired

Donata Moers, Petra Wagner, Stefan Breuer

Institut für Kommunikationswissenschaften, Abteilung Sprachliche Kommunikation
Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn, Germany
{dmo,pwa,sbr}@ifk.uni-bonn.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 phonetic characteristics of fast speech will be given. From this, certain conclusions concerning an adequate modeling of fast speech in unit selection synthesis will be drawn. Subsequently, a questionnaire assessing synthetic speech related preferences of visually impaired users will be presented. The last section deals with future experiments aiming at a definition of criteria for the development of synthesis corpora modeling fast speech within the unit selection paradigm.

1. Introduction

The option of making a synthesizer “talk fast” is elementary for users who are crucially dependent on their synthesis system in many everyday tasks such as browsing the web, reading emails, reading newspapers etc. and who hardly have any alternative to synthetic speech, i.e. visually impaired or blind users. While reading a web page, the – not visually impaired – user will usually concentrate on certain text passages, e.g. headlines and skip everything that appears to him/her as less interesting. This selective attention leads to a fast reading of least important parts or while a decision is being made whether the text passage currently read is interesting at all. The visually impaired user may want to have a similar option – the possibility to “skim through”. An optionally fast, or even very fast synthesis system is therefore often preferred by this user group.

The phonetic characteristics of fast speech are found to be very different from those of speech produced at “normal” speed. In order to model fast speech during synthesis, the engineer has several options. It is possible to either accelerate the “normal” speech linearly with the help of duration manipulation, to mimic certain prosodic features typical for fast speech such as pauses, intonation and strength of prosodic boundaries or to create an independent inventory inherently showing all segmental and suprasegmental characteristics of fast speech. Previous studies indicate that the different approaches lead to different results in perception experiments. E.g. 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 of the word in normal speech the better the word was understood by the listeners [1]. 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 synthesis and diphone synthesis, blind listeners preferred the less natural sounding formant over diphone synthesis with regards to intelligibility in very fast speech [2]. This indicates that the fast and smooth acoustic transitions in natural fast speech are 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 in particular, Breuer [3] 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.

However, a lot of questions concerning the proper treatment of fast speech in unit selection synthesis remain. Taking into account the aforementioned preconditions, the main focus of the – ongoing – project presented here is the definition of robust directives which should be obeyed when building a unit selection synthesis for the visually impaired which can produce fast or very fast speech in an acceptable quality.

2. Phonetic Characteristics of Natural Fast Speech

As stated in the introduction, the characteristics of fast speech differ from those produced at “normal” tempo. Hence, in this section a short overview of the general phonetic characteristics of naturally fast speech is given.

Fast speech differs from “normal” speech both in quality/quantity of vowels and in quality/quantity of consonants. Suprasegmental features like accents, phrase boundaries and the pause durations are also affected by a change in speaking rate. The course of the fundamental frequency is strongly influenced by tempo acceleration. How these differences come about and whether speakers are able to avoid them – because this might be an important option for a synthesizer as well – will shortly be described in the following paragraphs.

2.1. Vowels

Vowels can roughly be described as to consist of three parts: the onset at the beginning of a vowel, which includes the

formant movements (transitions) from the preceding sound, the so called steady state almost covering the greatest part in the middle of the vowel, where the formant frequencies stay stable, and the offset, which includes the transitions to the following sound. These transitions from and to another sound are characteristic for certain combinations of sounds and thus important for their correct identification [4].

When speaking faster, vowels are shortened in duration. This process mostly affects the steady state, which is logical since the transitions are very important for the vowel's perceptual identification and may therefore not be curtailed or even left out.

There is not only pure vowel shortening when speaking faster. Another important effect is vowel reduction. Here, reduction refers to a shift of the formant frequencies towards the neutral vowel in the middle of the vowel space [5]. One can assume that this reduction is the consequence of the limited movement velocity of the articulators and/or increasing coarticulation of segments. It is still a matter of ongoing discussion whether the shift of the formant frequencies is a directed movement towards the neutral vowel or simply a consequence of mutual influence between neighboring segments. However, it is questionable whether these phenomena can or should be regarded separately at all. Nevertheless, both of them affect the produced vowel quality and consequently have an impact on the listeners' perception.

2.2. Consonants

Like vowels, consonants are influenced by the acceleration of speaking rate. Like vowels, they are shortened in duration. However, due to the fact that most consonants do not possess a steady state which can be compressed without losing the segment's main characteristics, consonantal shortening is much less pronounced compared to vowels.

Hence, different types of consonants are affected differently by speech rate acceleration. E.g., plosives *become weaker*, which means, that the closures are not complete resulting in a lack of pressure. This leads to plosive bursts performed with less intensity. In consequence, the acoustic characteristics of plosives are more similar to approximants in fast speech [5]. A similar kind of weakening happens to fricatives too: The centers of gravity in their noise spectra show less intensity. Being a combination of plosive and fricative, affricates turn into pure (reduced) fricatives when speaking faster [6].

Another phenomenon occurring in fast speech is the syllabification of consonants. Due to reduction and finally elision of vocalic segments, consonants may become the syllable nucleus. This is accompanied by a duration prolongation of the respective syllabified consonant [7].

Furthermore, the phonetic distinction between voiced and unvoiced consonants is influenced by an increase in speech rate. Since voice onset time (VOT) is decreased, its function as a perceptual cue to distinguish between voiced and unvoiced plosives is neutralized [8].

The effects accumulated above are partly a result of an increasing gestural overlap between subsequent segments in fast speech. The segments have to be articulated in a smaller temporal frame and are therefore produced with more interference, often referred to as coarticulation. Another factor is – similar to vowels – reduction. Due to the fact that the articulators are limited in their movement velocity, the

articulators do not reach the optimal target position for each segment. Therefore, the segments as well as the transitions from one to another are not produced as clearly as in speech uttered at normal speed. On the segmental level, these phenomena lead to elision, reduction and assimilation processes, but it is highly context dependent whether or not the phenomena do occur or not.

2.3. Suprasegmental Duration

Apart from phone-specific effects, it has been shown that larger entities, such as the syllable, also behave differently under variations of speaking rate. E.g., unstressed syllables show a stronger shortening in fast speech than stressed ones [9], [10] which actually increases the difference in duration between stressed and unstressed syllables [11], [12]. An investigation in American English indicated that the proportion of stressed syllables decreased from nearly 75 % in normal speech tempo to less than 50 % in fast speech [13]. Anyway, the duration of stressed syllables or even stressed vowels in a stress group stayed stable, despite the increasing number of unstressed syllables.

Nooteboom [14] stated that the vocalic part of a syllable is more variable in fast speech than the consonantal part. But it was also shown, that the syllable internal proportion into 1/3 consonantal and 2/3 vocalic part stays almost stable across different speech tempos [15]. The average number of phones per syllable decreases as speaking rate increases. In addition, the elasticity hypothesis of Campbell and Isard [16] states that the relative duration of the syllable constituents is adjusted to the temporal frame of the syllable by scaling the intrinsic duration according to the temporal demands. Different factors have an influence on this scaling, among them the number of phones in the syllable, the position of the syllable in the phrase, the stress assigned to the syllable and the content of its parent word [ibid.].

2.4. Prosodic Organization

2.4.1. Pauses and Phrase Boundaries

When speaking faster, one of the first and easiest things to do in order to minimize the time for speech production is to decrease or even delete the pauses between utterances or phrases. Thus, there are fewer and shorter pauses in fast speech. The number of phrases decreases as well as prosodic boundaries are omitted or at least reduced [17], [18]. Monaghan [18] also showed also that in fast speech accents are left out and only the most important information remains accented.

2.4.2. Fundamental frequency

In fast speech, fundamental frequency excursions are less pronounced, the intonation contour becomes flatter and the pitch range is reduced. Due to its monotony, this speaking style can give the listener the impression of tediousness [17].

2.5. Semantic and Pragmatic Influences on Rate

As already mentioned, stressed syllables are shortened less than unstressed syllables in fast speech. They remain nearly stable concerning their degree of accentuation if the information they carry is important for comprehension.

Therefore accentuated syllables in content words, that tend to have a higher information content compared to function words, remain stable with an accelerating speech rate [20]. Consequently, content words are less reduced than function words as well.

Similar to tempo changes in a musical piece, speakers vary their tempo within an utterance relative to the linguistic context [21]. Quené [22] found that the Just Noticeable Difference (JND)¹ for human speech adds up to 2.5 % to 5 % difference in speech rate relative to the fundamental rate. Professional speakers produced a variation up to 4 % depending on the degree of novelty of the information in the relevant utterance. Tempo changes which are above the JND threshold are obviously relevant for communication. A speaker may express the relevance of an utterance in a greater context simply by changing the tempo and listeners can interpret a change of speaking tempo as a sign for the importance of what is said.

2.6. Speaking Strategies

Despite the continuous speech flow accompanied by coarticulation, a sufficient contrast between neighboring segments is both necessary and achievable in successful human communication. According to Lindblom's theory of hyper- and hypoarticulation (H&H theory) [23] a contrast is sufficient if it allows the listener to discriminate the signal to the extent necessary to identify the intended item in his mental lexicon. In contrast, the speaker produces speech earmarked and future-oriented. This causes a dilemma because on the one hand the speaker tries to communicate with as little effort as possible. *Hypospeech*, a somewhat more slurry pronunciation style, is the result of this economic constraint. On the other hand the speaker wants to reach a communicative goal, he therefore needs to maintain the phonetic contrast necessary for comprehension. Thus, in situations where comprehension might be more difficult (e.g. in a loud environment) or absolutely essential (e.g. when giving driving instructions) speakers tend to use *hyperspeech*, a very exact pronunciation style. Lindblom describes this phenomenon as follows: „speakers are expected to vary their output along a continuum of hyper- and hypospeech“. To be understood by a listener the speaker's (speech)-signals need to feature a sufficient contrast for the listeners' lexical access. For fast speech, we would normally expect speakers to use hypospeech while speaking fast – due to economy. However, speakers may be well able to speak both fast and clear (hyperspeech) if the situation requires this – within certain articulatory constraints.

2.7. Perception

As explained above, the main problem during the perception of natural fast speech is the omission of several acoustic characteristics which are necessary for the correct identification of what has been said. In contrast, it has been shown that if natural speech was compressed up to 65 % of its original duration it was still “perfectly intelligible” [1]. Obviously, the natural acoustic transitions keep the speech intelligible even at fast tempo but the content needs to be semantically or pragmatically predictable to be understood. Even if the temporal compression is further intensified and the

¹ Just noticeable difference is the smallest difference in a specified modality of sensory input that is detectable by a human being. [27]

compressed utterances have only 35 % of their original duration, they remain comprehensible in the majority of cases (53 %) [24].

2.8. Conclusions and Implications for Fast Synthetic Speech

Speakers follow certain strategies when speaking fast, they reduce vowels and consonants, flatten the fundamental frequency contour and try to minimize duration of pauses and of segments that can be contracted best, i.e. vowels. This process may lead to a loss of distinctiveness and consequently comprehension. However, speakers obey certain rules in order to keep the communication chain working: Semantically important elements of speech are compressed/reduced less than unimportant ones. Nevertheless, with a lot of effort, speakers are well able to speak both clear and fast.

It is possible that a modeling of these speaker strategies may increase naturalness of synthetic speech. Furthermore, it is possible that a stronger contrast between clearly spoken, semantically important and slurrily spoken, less important elements may even increase comprehension of fast synthetic speech, since it draws the attention to the main content of an utterance.

Furthermore, we know that the acoustic transitions of subsequent segments play a vital role in the intelligibility of (fast) speech. The discontinuities added to the speech chain during concatenation must therefore be minimized. This can be achieved straightforwardly by combining phones which are prone to heavy coarticulation into indivisible synthesis units.

We therefore aim to integrate the insights of H&H theory and flexible approaches to inventory creation for unit selection synthesis in order to achieve synthetic speech that is both maximally natural and maximally fast.

3. Preliminary Evaluation

The goal of our present study is to determine an optimal strategy for modeling fast synthetic speech for the visually impaired user. A fundamental problem is the circumstance that preferences – especially of the blind or otherwise visually impaired people – are not investigated as much as it would be necessary for designing an optimal inventory for a fast unit selection speech synthesis.

When starting work for the project some questions came up: What do the blind or visually impaired people aim for concerning speech synthesis? Do they really prefer a monotonous fast synthesis being prosodically relatively close to natural fast speech as suggested in [19]? Or do they not mind a lack in naturalness as long as acoustic transitions important for segment identification are adequately modeled as in formant synthesis [2]? Is it important that the information bearing units are less compressed/reduced than the words carrying less semantic load? What kind of speech quality do they prefer?

The literature concerning these problems proved to be very poor and so it was decided to start a survey among the prospective users. A questionnaire was designed which includes questions about the users'

- fields of synthesis applications
- used or preferred speech synthesis devices
- global preferences concerning speech tempo
- preferred speech rate when listening to synthetic speech

A second part of the questionnaire deals with several detailed questions related to

- the preferred or desired intelligibility
- the preferred intonation and prosody of fast speech
- the users' desire for an even faster output than what is currently possible
- preferences concerning the tradeoff between naturalness, liveliness and the possibility to have a synthesizer talk very fast.

3.1. (Expected) Results

Due to the fact that at the time of writing this paper the questionnaire has just been released to the public, there are no results available. Nevertheless the following section contains some information concerning the expected outcome. During the workshop, detailed results of the survey will be presented.

4. Further experiments

Based on our previous investigations (cf. 2.) and the outcome of the questionnaire (cf. 3.), we are currently setting up a series of perception experiments aimed to determine an optimal strategy for building a unit inventory that enables us to model fast synthetic speech. The synthetic quality should be especially suited for applications used by the visually impaired. Below we describe the different steps currently undertaken to gather stimuli containing the different articulatory and acoustic features under examination. Then, the anticipated experimental setup is explained. Of course, these are still subject to amendments based on the prospective survey's results.

4.1. Recordings of Synthesis Units

According to the H&H theory, speakers are able to speak both fast and clear if they increase effort. In order to build a useful synthesis inventory that models fast speech, a speaker needed to be found who was able to realize this speaking style best. To determine a competent inventory speaker, preliminary recordings of 9 volunteers were carried out. These recordings were rated by 12 phonetically trained people. They assessed the individual speakers fastest possible articulation rate, their perceptual clarity during fast speech and their individual voice characteristics. Based on these parameters, the presumably most suitable speaker for a fast inventory of a unit selection speech synthesis system was determined.

During inventory creating, the selected speaker read a subset (400 sentences) of the language material contained in the BITS-Corpus [25]. The BITS-Corpus was simply chosen due to its availability and its phonologically balanced design fulfilling the general criteria of unit selection speech synthesis systems.

The sentences are recorded in 2 conditions:

- "normal" speech rate (4 to 5 syllables per second)
- maximum "clear" speech rate (6 to 8 syllables per second)

All recordings were conducted in a sound treated recording studio of our institute. Due to the fact that not all recordings can be done in only one session a strict monitoring of speaking rate, phrasing and intensity is necessary. Prior to each session and within each session, several reference

sentences are presented to the speaker in order to (re)adjust her performance and speaking style. The reference sentences are recordings of the first recording session. Special attention is paid to an adjustment of speaking rate, phrasing and 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 [26].

All recordings are labeled automatically and corrected manually. Thus, we create two unit selection inventories: one in normal speech rate and one in fast speech rate. In order to assess the general quality of the normal rate inventory and make sure it fulfills the baseline criteria of an acceptable unit selection corpus, the normal rate inventory will be compared with the performance of the existing BITS-inventory. This assessment will be performed by generating and comparing identical sentences from the two different inventories.

4.2. Stimuli and Experimental Setup

As stimuli, different sentences will be generated from the two inventories recorded previously. The stimulus sentences have also been recorded but have not been included in the inventory. Thus, we have templates for further manipulations and comparisons. The first sentence will be generated from normal rate units, the second from fast rate units. A third sentence will be mixed: content words generated from the normal rate units and function words generated from the fast rate units. The motivation for these three groups is that it is still unclear whether listeners prefer fast synthetic speech generated from fast units (most natural?), compressed normal units (most intelligible?) or a mixture of both, trying to mimic the speaking strategies explained by the H&H-theory.

The sentences which are partly or completely generated from the normal rate units presumably will have to be largely manipulated concerning their duration and f_0 based on the prerecorded template. It is expected that the sentences which have been generated from the fast rate units will require a comparatively marginal manipulation. This manipulation may create another variable influencing the results of the perception experiments.

There are three groups of stimulus sentences which will be evaluated pairwise in preference tests:

Stimulus Group 1:

- Generated from normal rate units
- Presumably little coarticulation
- Presumably massive prosodic manipulation

Stimulus Group 2:

- Generated from fast rate units
- Presumably massive, but typical coarticulation
- Presumably little prosody manipulation

Stimulus Group 3:

- Generated from normal and fast rate units
- Presumably little coarticulation in content words and massive coarticulation in function words
- Presumably some prosody manipulation

Additionally, stimuli representing a normal speech rate will be generated from the two inventories. These sentences represent a crosscheck. Here, we expect that the sentences generated from the normal rate units are judged much better

than that generated from the fast rate units. On the one hand, the fast rate units will have to be massively manipulated, on the other hand they will cause intelligibility problems for the listeners due to their strong pertinent coarticulation and reduction.

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. their 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 consists 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 fast speech rate, maybe even not intelligible for the visually unimpaired. Furthermore, it is assumed that the fast versions of the sentences where the content words are synthesized from the normal rate units are preferred because the important information is more intelligible and easy to understand.

5. Conclusions

Our paper comprises phonetic knowledge concerning fast speech, discusses implications for its most adequate modeling in concatenation based synthesis applications aimed at visually impaired users and presents a research strategy to investigate this problem further. 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.

Since our paper described work in progress, only very preliminary results are presented, but first results with regards to the – formerly poorly investigated – tempo related synthesis preferences of visually impaired users will be reported during the workshop.

6. Acknowledgements

The authors would like to thank our speakers, the IfK staff for technical help and David Frauenkron for support during the recordings. Special thanks go to the people who encouraged the work which was presented here.

7. References

- [1] Janse, E. (2003): *Word perception in natural-fast and artificially time-compressed speech*. Proceedings 15th ICPhS. Barcelona. pp. 3001 - 3004.
- [2] Trouvain, J. (2006): *Subjektive Verständlichkeit von Computerstimmen bei verschiedenen Geschwindigkeiten. Eine Pilotstudie mit zwei Benutzergruppen*. Saarbrücken 2006.
- [3] Breuer, S.; Abresch, J. (2004): *Phoxsy: Multi-phone Segments for Unit Selection Speech Synthesis*. In: Proceedings ICSLP. Jeju.
- [4] Martínez, F.; Tapias, D.; Alvarez, J.; León, P. (1997): *Characteristics of slow, average and fast speech and their effects in large vocabulary continuous speech recognition*. Proceedings Eurospeech. Rhodes, Greece.
- [5] Kohler, K.J. (1990): *Segmental reduction in connected speech in German: Phonological facts and phonetic explanations*. In: Hardcastle, W.J.; Marchal, A. (eds.): *Speech Production and Speech Modelling*. Dordrecht : Kluwer. pp. 69 - 92.
- [6] van Son, R. J. J. H.; Pols, L. C. W. (1996): *An acoustic profile of consonant reduction*. Proceedings ICSLP. Philadelphia. pp. 1529 – 1532.
- [7] Roach, P.; Sergeant, P.; Miller, D. (1992): *Syllabic consonants at different speaking rates: A problem for automatic speech recognition*. *Speech Communication*. Vol. 11, pp. 475 - 479.
- [8] Kessinger, R.H.; Blumstein, S.E. (1998): *Effects of speaking rate on voice-onset time and vowel production: Some implications for perception studies*. *Journal of Phonetics*. Vol. 26, pp. 117-128.
- [9] Peterson, G.E.; Lehiste, I. (1960): *Duration of syllable nuclei in English*. *JASA*. Vol. 32, S. 693 - 703.
- [10] Gopal, H.S. (1990): *Effects of speaking rate on the behaviour of tense and lax vowel durations*. *Journal of Phonetics*. Vol. 18, pp. 497 - 518.
- [11] Delattre, P.C. (1966): *A comparison of syllable length conditioning among languages*. *Int. Review of Applied Linguistics*. Vol. 4, pp. 183 - 198.
- [12] Hoquist, C.E. (1983): *Syllable duration in stress-, syllable- and mora-timed languages*. *Phonetica*. Vol. 40, S. 203 - 237.
- [13] Crystal, T.H.; House, A.S. (1990): *Articulation rate and the duration of syllables and stress groups in connected speech*. *JASA*. Vol. 88, pp. 101 - 112.
- [14] Nootboom, S. (1972): *Production and perception of vowel duration: A study of durational properties of vowels in Dutch*. PhD thesis. Rijksuniversiteit Utrecht.
- [15] Kuwabara, H. (1997): *Acoustic and Perceptual Properties of Phonemes in Continuous Speech as a Function of Speaking Rate*. In: Proceedings Eurospeech. Rhodes, Greece.
- [16] Campbell, W.N.; Isard, S.D. (1991): *Segment durations in a syllable frame*. *Journal of Phonetics*. Vol. 19, pp. 37 - 47.
- [17] Fougeron, C.; Jun, S. (1998): *Rate effects on French intonation: prosodic organization and phonetic realization*. *Journal of Phonetics*. Vol. 26, pp. 45 - 69.
- [18] Monaghan, A. (2001): *An Auditory Analysis of the Prosody of Fast and Slow Speech Styles in English, Dutch and German*. In: Keller, E.; Bailly, G.; Monaghan, A. et al. (eds.): *Improvements in Speech Synthesis*. Chichester. pp. 204 - 217.
- [19] Fellbaum, K. (1996): *Einsatz der Sprachsynthese im Behindertenbereich*. In: Fortschritte der Akustik. DAGA'96, Oldenburg : DEGA. pp. 78-81.
- [20] Schindler, F. (1975): *Faktoren phonetischer Performanz. Instrumentalphonetische Versuche zur akustischen Bestimmung des Ausprägungsgrades von Eigenschaften des lautsprachlichen Signals*. *Zeitschrift für Dialektologie und Linguistik*. Beihefte. Neue Folge Nr. 14 der Zeitschrift für Mundartforschung. Franz Steiner, Wiesbaden.
- [21] Nootboom, S.; Eefting, W. (1994): *Evidence for the adaptive nature of speech on the phrase level and below*. *Phonetica*. Vol. 51, pp. 92 – 98.
- [22] Quené, H. (2006): *On the just noticeable difference for tempo in speech*. Utrecht 2006.
- [23] Lindblom, B. (1990): *Explaining phonetic variation: A sketch of the H&H-Theory*. In: Hardcastle, W.J.; Marchal, A. (eds.):

A.: Speech Production and Speech Modelling. Dordrecht: Kluwer. pp. 403 - 439.

- [24] Janse, E.; Nootboom, S.; Quené, H. (2003): *Word-level intelligibility of time-compressed speech: prosodic and segmental factors*. Speech Communication, Vol. 41. pp. 287–301.
- [25] Schiel, F.; Draxler, C.; Ellbogen, T.; Jänsch, K.; Schmidt, S. (2006): Die BITS Sprachsynthesekorpora - Diphon- und Unit Selection-Synthesekorpora für das Deutsche.
- [26] Greisbach, R. (1992): Reading aloud at maximal speed. Speech Communication. Vol. 11, pp. 469 - 473.
- [27] Eefting, W.; Rietveld, A. (1989): Just noticeable differences of articulation rate at sentence level. Speech Communication, Vol. 8. pp. 355–361.