Integrating a Fast Speech Corpus in Unit Selection Speech Synthesis: Experiments on Perception, Segmentation, and Duration Prediction

Donata Moers[1,2], *Petra Wagner*[2], *Bernd Möbius*[1,3], *Filip Müllers*[1], *Igor Jauk*[1]

[1] Institut für Kommunikationswissenschaften, Abt. Sprache und Kommunikation, Rheinische Friedrich-Wilhelms-Universität Bonn, Germany
[2] Fakultät für Linguistik und Literaturwissenschaft, Universität Bielefeld, Germany
[3] Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Germany
dmo@ifk.uni-bonn.de, petra.wagner@uni-bielefeld.de, moebius@ifk.uni-bonn.de, fmu@ifk.uni-bonn.de, ija@ifk.uni-bonn.de

Abstract

This paper examines viable paths for integrating a fast speech corpus into a unit selection synthesis system. After selecting a suitable speaker, two inventories were recorded: one at normal and one at fast speech rate articulated as accurately as possible. A perceptual evaluation showed that for ultra fast speech rate, stimuli generated from fast utterances were judged to be as intelligible as stimuli generated from normal rate utterances; moreover, they were clearly preferred with respect to naturalness. Based on the results of an automatic phone segmentation, which produced only marginal differences in label timing accuracy, CART based duration prediction models for both corpora were built. Prediction accuracy was very similar. We conclude that automatic phone segmentation and CART based duration prediction are applicable to both normal and fast rate recordings.

Index Terms: fast speech, unit selection, duration prediction

1. Introduction

Many people with severe visual disabilities use speech synthesis as part of their everyday life. These listeners often prefer fast speech output [1, 2]. But also expressive speech synthesis and spoken language interfaces may require an integration of fast speech into synthesis systems. Architectures like formant or diphone synthesis are able to produce synthetic speech at (ultra) fast speech rates, but the generated speech does not sound very natural. Unit selection synthesis systems are capable of delivering more natural output, but fast speech has not been adequately implemented into such systems to date. In order to investigate the modeling of fast speech in unit selection synthesis, we first decided to create two independent but, in terms of content, identical unit selection inventories: one at normal and one at fast speech rate.

The phonetic characteristics of natural fast speech differ from those of speech produced at normal speech rates. The faster somebody speaks the less intelligible his/her utterances become. This is due to the increasing overlap of articulatory gestures when speaking rate increases; the articulatory targets important for a clear pronunciation are no longer reached [3]. In vowels, this manifests itself in a shorter duration and a change in characteristic formant frequencies [4]. Consonants are assimilated more often, their intensity decreases and their realizations become incomplete. Some are changing their consonantal category or are even elided completely [4, 5]. Larger prosodic units like syllables or intonation phrases are affected as well: Syllable durations are shortened and the total number of stressed syllables decreases, the number and strength of phrase boundaries declines and the fundamental frequency contour is flattened [6, 7].

Strong coarticulation, reduction and other deviations from the clear canonical form affect the intelligibility of natural speech adversely [8, 9]. Hence, these phenomena are undesirable in speech synthesis and need to be avoided during corpus recordings, if possible. As a consequence of the Hypo- and Hyperspeech theory (H&H theory) [10], speakers can be expected to speak both fast and clear if they increase their articulatory effort in line with speech tempo. For the work reported here, a speaker was selected who was able to produce this speaking style (both fast and clear) in an optimal way. Thus, undesirable phenomena like reduction and strong coarticulation were avoided as much as possible during corpus recordings [11].

Inventory preparation is one of the most time consuming steps during the development of new corpora for unit selection synthesis. In order to build a useful and manageable inventory for fast speech, we first investigated whether fast speech utterances articulated as accurately as possible have a perceptual disadvantage compared to normal speech rate utterances by speeding up both versions linearly to ultra fast speech rates. Moreover, we evaluated whether automatic phone segmentation is applicable not only to normal but also to fast speech rate recordings. As it is still unclear, whether it is reasonable to also build a separate duration prediction model on the basis of fast speech rate recordings, we constructed CART based segmental duration prediction models for both normal and fast speech rate utterances separately.

Taking these considerations into account, the main goal of our research is the integration of a fast speech corpus into the unit selection synthesis system BOSS [12] to produce fast speech in an acceptable quality based on a special fast speech inventory. In the future, robust guidelines for integrating fast speech corpora into unit selection synthesis systems are expected to result from the approach presented here.

2. Corpus development and evaluation

2.1. Corpus recordings

Two speech corpora were recorded. Text materials consisted of 400 sentences which were selected randomly from the BITS Corpus [13] for German. This corpus was chosen because its phonologically balanced design meets the general criteria of unit selection speech synthesis systems. For the randomly chosen subcorpus, phonological balance was not taken into account. The selected 400 sentences were recorded in two conditions:

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

All recordings were made in a sound treated studio. Because the recordings could not be performed in a single session, a strict monitoring of speaking rate and speaking style including accentuation, phrasing and intensity was required. As a consequence, several reference sentences were presented to the speaker repeatedly in order to (re)adjust her performance, before each session as well as within the sessions. The reference sentences were recordings from the first session. The speaker generally followed the strategy of approaching the fastest speaking rate by repeated, accelerated renditions of a sentence. Thus, fast versions of one sentence were recorded repeatedly in succession, accelerating tempo and enhancing articulatory effort each time, until the optimal combination of tempo and precision was reached. Two phonetically trained persons supervised the recordings.

Recordings at normal speech rate took approximately 10 hours, recordings at fast speech rate took nearly twice as long. Subsequently, the version articulated both most clearly and fast was perceptually selected by a phonetically trained person and included in the fast speech corpus. This way, two unit selection corpora were created: one at normal speech rate and one at fast speech rate articulated as accurately as possible.

2.2. Perceptual evaluation of corpus recordings

Janse [8] reported that 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 linearly compressed. The less the stimulus deviated from the canonical form the better the word was understood by listeners. Taking these findings into account we evaluated whether the normal rate sentences were judged to be more intelligible than the fast rate sentences when produced at the same speech rate. Therefore, the normal rate sentences were sped up linearly by means of TD-PSOLA until they met the higher speech rate of the corresponding natural fast rate 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 second step was the acceleration of both normal and fast rate utterances to an even faster and therefore unnatural speech tempo. Thus, sentences generated from the normal rate recordings had to be modified more strongly with respect to their duration, whereas sentences generated from the fast rate speech required a comparatively smaller duration manipulation. In this ultra fast condition, stimuli generated from fast speech utterances were expected to be judged at least as intelligible as the stimuli generated from normal rate utterances and at the same time sounding more natural than these.

The experiment included 20 sentences randomly chosen from the corpus. Altogether, 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 experiment was conducted in a quiet environment and stimuli were presented via earphones. 11 subjects participated in the experiment.

First, subjects were instructed to choose from each pair the realization which was pronounced more clearly. Afterwards, they were asked to indicate the more natural sounding utterance of each pair. As expected, in the first part of the experiment the stimuli generated from normal speech rate utterances were judged to be more intelligible than the natural fast ones (χ^2, $p < 0.05$). This advantage of the normal rate utterances disappears in the ultra fast condition (Figure 1). There is even a slight but not significant tendency to prefer the stimuli generated from natural fast speech. However, the natural fast stimuli are clearly preferred with respect to naturalness (χ^2, $p < 0.0001$). These results confirm our initial hypotheses.

2.3. Automatic phone segmentation

The quality of synthesized speech largely depends on labeling accuracy [14]. If recordings are based on fast speech, using the same segmentation algorithm for both normal and fast speech rate recordings might result in a considerably increasing amount of incorrect labels for the fast speech utterances. If so, automatic phone segmentation would not be applicable to our fast speech rate recordings although they were articulated as accurately as possible.

To evaluate the performance of an HTK based aligner [15] used for this task, 49 sentences of each subcorpus were manually corrected on the basis of automatically generated labels. Manual labeling was done by only one person to maximize consistency. Each phone was listened to several times to minimize effects from neighbouring phones. Label timing errors were then calculated for all phones in normal and fast speech rate by subtracting the manual label time from the automatic label time. When the result was positive the automatically generated label was set too late with respect to the manual label; when the result was negative the automatic label was set too early. Frequency distributions of label timing errors for both the normal and fast rate utterances are shown in Figure 2.

We found a 90% accuracy for the normal speech rate utterances and a 91% accuracy for the fast speech rate utterances

Speech rate	< 5	< 10	< 15	< 20
Normal	42.75%	68.82%	83.02%	90.44%
Fast	42.62%	65.67%	81.97%	90.79%

Table 1: *Percentage of boundaries within different tolerance intervals for normal and fast speech rate. Tolerance intervals are in ms.*

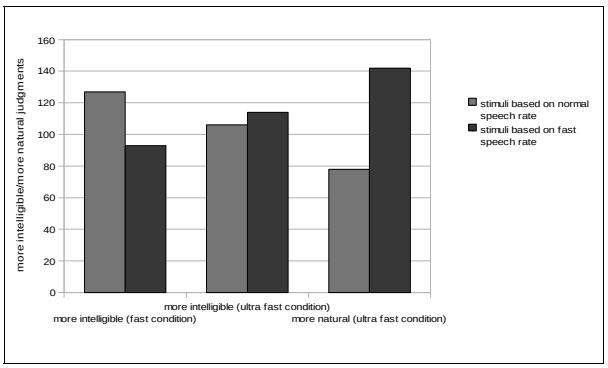

Figure 1: *Intelligibility judgments for fast and ultra fast stimuli, naturalness judgments for ultra fast stimuli.*

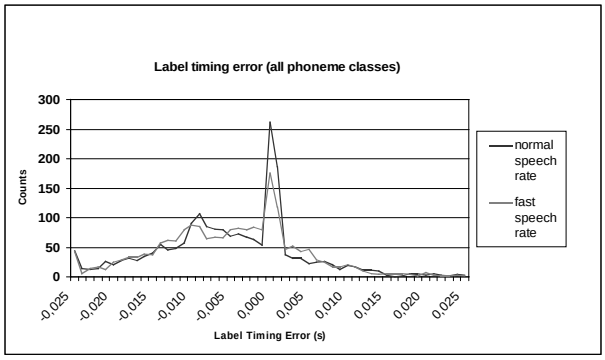

Figure 2: *Frequency distribution (counts) of label timing errors for normal and fast speech rate (s).*

within a 20 ms interval, which is in line with an average agreement of 94% between human labelers reported in the literature [16]. Decreasing the interval to 15 ms, an 83% accuracy for the normal and an 82% accuracy for the fast speech rate utterances are reached. Within a 10 ms interval, the accuracy was 69% for normal and 67% for fast speech rate. A 43% accuracy was found within a 5 ms tolerance interval for both normal and fast rate utterances (Table 1). Thus, differences in label timing accuracy between the normal and the fast rate sentences were marginal across all tolerance intervals.

The results indicate that automatic phone segmentation is a technique not only applicable to corpus recordings at normal speech rate but also to recordings at fast speech rate, at least if they are conducted with high precision and enhanced articulatory effort.

2.4. Duration prediction

The duration of phonetic segments is an important prosodic factor in the production of natural sounding synthetic speech [17]. Considering the results of Janse [8] as well as our results (sec 2.3), we decided to create segmental duration prediction models by building CART based regression trees [18] for the normal and the fast speech corpus separately, taking into account important phonetic and prosodic features influencing segmental duration.

According to Janse, one might expect that the duration prediction models show significantly higher correlations between observed and predicted durations with normal rate test data compared to fast rate test data. Fast speech may show more variability as it is presumably produced by using different and/or changing articulation strategies even when it is articulated with enhanced articulatory effort. Furthermore, speakers probably have more practice at their normal speech rate production which according to Pierrehumbert [19] may lead to entrenchment effects resulting in decreased variability. On the other hand, having more practice at normal speech rate productions may also lead to an increase of variability for this speaking style. Due to more exemplars available for normal speech rate productions, one might expect greater articulatory variation at normal speech rate. Hence, the correlation between observed and predicted duration for fast rate speech would be at least as high as the correlation for normal rate speech despite all adverse influences like coarticulation and reduction, which are not completely avoidable even in clear and fast speech.

The tool used for the construction of the CART based duration prediction models was *wagon* from the Edinburgh Speech Tools [20]. The feature set applied was adapted to the requirements of our unit selection synthesis system BOSS [21]:

- phoneme identity
- phoneme duration
- preceding phoneme
- following phoneme
- second following phoneme
- phrase position
- syllabic stress

The phoneme itself is the feature whose duration is to be predicted. The phoneme durations extracted from the particular corpus were the training data. The position in the phrase can be *initial*, *medial* or *final*. Syllabic stress can have the values *primary*, *secondary* or *none*. The phoneme itself, its duration, the preceding and following phoneme as well as the syllabic stress can be extracted directly from the corpus (assuming that the corpus is already segmented and preprocessed by BOSS-Tools [21]); the second following phoneme and the phrase position have to be calculated during processing.

Corpus speech rate	RMSE	Correlation	Mean (abs) Error
Normal	39.66	0.8018	20.16 (34.16)
Fast	23.97	0.7749	12.37 (20.53)

Table 2: *CART duration prediction results for two corpora (RMSE and Mean (abs) error are in ms).*

Normal speech rate corpus (Dataset of 18487 vectors of 7 parameters)	Fast speech rate corpus (Dataset of 18240 vectors of 7 parameters)
1. phone identity: 0.4734	1. phone identity: 0.4736
2. pos. in phrase: 0.6750	2. pos. in phrase: 0.6649
3. follow. phoneme: 0.7862	3. follow. phoneme: 0.7559
4. prec. phoneme: 0.8000	4. prec. phoneme: 0.7681
5. syllabic stress: 0.8009	5. syllabic stress: 0.7738
6. 2nd follow. phon.: 0.8018	6. 2nd follow. phon.: 0.7749

Table 3: *Feature ranking comparison (stepwise) of CART results for the two corpora.*

Results show that the correlation between observed and predicted durations for the fast rate utterances is 0.78 whereas the correlation for the normal rate utterances is 0.80 (Table 2). This is only a slight difference; both correlations are similar to results reported for the prediction of segmental duration in normal speech rate in other languages [22]. RMSE and Mean (abs) Error are smaller for the fast speech corpus. We attribute this to the fact that overall segment durations are smaller for fast speech.

Looking at the feature ranking generated by the *stepwise* option of *wagon* [20] (Table 3), major differences between the two duration prediction models become apparent. The most important feature for the normal rate speech is the phoneme itself; for the fast speech, it is the following phoneme. This might be due to stronger coarticulatory effects attributed to increased articulatory overlap in fast speech. Syllabic stress surprisingly shows only marginal differences for normal vs. fast speech. Since the total number of stressed syllables and their durations generally decrease in fast speech, we expected that syllabic stress would show a higher impact on the correlation between observed and predicted durations for normal rate speech than for fast rate speech. This is not the case, which might be attributed to the enhanced articulatory effort during fast speech production.

In general, the duration prediction models show similar correlations between observed and predicted durations for both normal and fast rate speech. Thus, building a CART based segmental duration prediction model is applicable to normal as well as fast speech corpus recordings.

3. Discussion

The perceptual evaluation of corpus recordings confirmed the results reported by Janse [8] for fast speech tempo (8 syllables per second). In the ultra fast condition (16 syllables per second) there was a slight tendency for listeners to prefer stimuli generated from fast sentences with respect to intelligibility and a significant preference with respect to naturalness. However, the normal rate versions might have suffered more strongly from the modification of the speech signal imposed by TD-PSOLA [23], which in turn might influence the naturalness adversely. Nevertheless, we decided to use this algorithm here because it is still generally applied in speech syn-

thesis systems. An alternative approach would be the application of other acceleration algorithms, e.g. non-linear time scaling.

The automatic phone segmentation showed only marginal differences in label timing accuracy for the two corpora. We concluded that automatic phone segmentation is a technique applicable to recordings at both normal and fast speech rate, at least if the latter is performed with high precision and enhanced articulatory effort. However, the absolute count of label timing errors around 0 ms is considerably higher for normal than for fast speech rate which might be attributed to increased coarticulatory effects and reduction at fast speech rate. In addition, since segment durations at fast speech rate are shorter overall, one might ask if the tolerance intervals chosen here are appropriate to judge the accuracy of fast speech label timing.

CART based duration prediction models considering important phonetic and prosodic features influencing segmental duration showed that the correlation between observed and predicted duration was comparable for recordings at both speech rates. Thus, this technique is also applicable for normal as well as fast and clear speech rate corpus recordings. Nevertheless, there are slight differences in results between normal vs. fast speech rate. The model may require a refinement of the features to enhance the correlation between segmental duration and features predicting duration patterns of fast speech.

4. Conclusions

The aim of the investigations presented here was the integration of a fast speech corpus into the unit selection speech synthesis system BOSS. We have shown that fast speech articulated as accurately as possible has no perceptual disadvantage compared to normal rate speech when accelerated to ultra fast speech rates. Furthermore, we showed that automatic phone segmentation and CART based duration prediction are applicable to both normal and fast speech rate recordings. Further investigations will include the perceptual evaluation of synthesized utterances accelerated to varying speech rates based on the different speech rate corpora.

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

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

[4] 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.

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

[6] 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.

[7] 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.

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

[9] Krause, J.C. and Braida, L.D., "Investigating Alternative Forms of clear speech: the effects of speaking rate and speaking mode on intelligibility", Journal of the Acoustical Society of America 112: 2165–2172, 2002.

[10] 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.

[11] Moers, D. and Wagner, P., "Assessing a Speaker for Fast Speech in Unit Selection Speech Synthesis", Proc. Interspeech 2009, Brighton, 2009.

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

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

[14] Kominek, J., Bennett, C., and Black, A.W., "Evaluating and Correcting Phoneme Segmentation for Unit Selection Synthesis", Proc. Eurospeech 2003, Geneva, 2003.

[15] Jauk, I., "Anleitung zur Einbindung eines Korpus in BOSS 12.11.2009", http://www.sk.uni-bonn.de/forschung/phonetik/sprachsynthese/boss/BOSS_Korpuserstellung.pdf

[16] Adell, J. and Bonafonte, A, "Towards Phone Segmentation for Concatenative Speech Synthesis", Proc. 5th ISCA Workshop on Speech Synthesis (SSW5), Pittsburgh, 2004.

[17] Carlson, R., Granström, B. and Klatt, D., "Some Notes on the Perception of Temporal Patterns in Speech", in Lindblom, B. and Öhman, S. [Ed], Frontiers of Speech Communication Research, London: Academic Press, 1979.

[18] Breiman, L. et al., "Classification and Regression Trees", Belmont: Wadsworth, 1984.

[19] Pierrehumbert, J.B.; "Exemplat dynamics: Word frequency, lenition and contrast", in J. Bybee & P. Hopper (eds.), Frequency effects and emergent grammar, Amsterdam: John Benjamins, 2000.

[20] King, S., Black, A.W., Taylor, P., Caley, R. and Clark, R., "Edinburgh Speech Tools Library. System Documentation Edition 1.2, for 1.2.3 24th Jan 2003", http://www.cstr.ed.ac.uk/projects/speech_tools/manual-1.2.0/

[21] Breuer, S. et al., "Bonn Open Synthesis System (BOSS) 3 Documentation and User Manual December 20, 2005", http://www.sk.uni-bonn.de/forschung/phonetik/sprachsynthese/boss/BOSS_Documentation.pdf

[22] Klessa, K., Szyma´nski, M., Breuer, S. and Demenko, G., "Optimization of Polish Segmental Duration Prediction with CART", Proc. 6th ISCA Workshop on Speech Synthesis (SSW-6), Bonn, 2007.

[23] 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.