Synthesizing Fast Speech by Implementing Multi-Phone Units in Unit Selection Speech Synthesis

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

[1] Division of Language and Speech Communication, University of Bonn, Germany
[2] Fakultät für Linguistik und Literaturwissenschaft, University of Bielefeld, Germany
[3] IMS, University of Stuttgart, Germany
{dmo,ija,bmo}@ifk.uni-bonn.de, petra.wagner@uni-bielefeld.de

Abstract

This paper presents a new approach to synthesizing fast speech in unit selection synthesis. After recording two inventories - one at normal and one at fast speech rate articulated as accurately as possible - speech was synthesized from both corpora independently. Since fast speech differs from normal rate speech in terms of acoustic characteristics, the concept of multi-phone (phoxsy) units [1] was implemented and used to synthesize speech at both speaking rates again. A perceptual evaluation showed that phoxsy units enhanced the intelligibility especially for fast synthetic speech significantly.
Index Terms: fast speech, unit selection, phoxsy units

1. Introduction

Using speech synthesis as part of their daily life, many people with severe visual disabilities often prefer fast speech output [2, 3]. Architectures like formant or diphone synthesis are able to produce synthetic speech at 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 model fast speech in speech synthesis, there are several options. The first is to accelerate the normal rate speech by means of duration manipulation. The generated speech often shows artifacts known to appear when using such algorithms [4] and does not sound natural. The second option is to mimic certain prosodic features typical for fast speech such as fewer and shorter pauses or decreased strength and number of prosodic boundaries. Previous studies indicate that this approach leads to a decreasing intelligibility of fast speech. A clear pronunciation was preferred over a synthesis that showed typical phonetic characteristics of natural fast speech [5]. Therefore, our approach includes the creation of an independent unit selection inventory for fast speech inherently showing segmental and suprasegmental characteristics of natural fast speech to enhance naturalness. At the same time, too heavy reduction and coarticulation typical for natural fast speech have to be avoided for the benefit of intelligibility.

The fast and smooth acoustic transitions occurring in natural speech are important for the intelligibility of synthetic speech [5]. Such transitions are not treated adequately by traditional diphone concatenation synthesis but can be modeled by formant synthesis. Corresponding to this, blind listeners prefer the less natural sounding formant synthesis over diphone synthesis with regards to intelligibility in very fast speech [2]. Since the acoustic transitions of subsequent segments play a vital role in the intelligibility of speech, the discontinuities added to the speech chain during concatenation must be minimized. As a consequence, Breuer and Abresch [1] suggested to treat 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 is taken up in the present study. It might lead to a possible solution for modeling fast synthetic speech both naturally – by using prerecorded concatenation units – and intelligibly – by including typical smooth transitions in heavily coarticulated contexts in order to achieve synthetic speech that is both maximally natural and maximally fast.

2. Phoxsy units

In the field of unit selection synthesis, it is well known that linguistically motivated units like phones do not provide optimal properties for concatenation. The main disadvantage of this type of units is the disregard of acoustic and auditive continuity. Phoxsy units (*pho*ne extensions for *sy*nthesis) are defined to systematically avoid concatenation points in the signal at positions where they are highly undesirable [1, 6]. Basically, they are sequences of phones prone to heavy coarticulation with fluent transitions and phonetically non-existing boundaries. Table 1 lists possible phone combinations defined as phoxsy units by [1, 6]. The "IPA" column shows the unit definitions transcribed in the International Phonetic Alphabet. The "BOSS-SAMPA" [6] column shows the way how the units have been processed before phoxsy definition, whereas the "phoxsy" column shows the new unit definitions in BOSS-SAMPA notation, which is a modified X-SAMPA notation.

IPA	BOSS-SAMPA	phoxsy
ʔ + vowel	? + vowel	? + vowel
h/ɦ + vowel	single phones	h + vowel
j+ vowel	single phones	j + vowel
ʊ/v + vowel	single phones	v + vowel
ʁ/ʁ/ɾ/r + vowel	single phones	r + vowel
l + vowel	single phones	l + vowel
ən/n	@n	@n
əm/m	single phones	@m
əl/l	single phones	@l
j/ʊ/v/ʁ/ʁ/ɾ/r/l + ən	single phones	j/v/r/l + @n
j/ʊ/v/ʁ/ʁ/ɾ/r/l + əm	single phones	j/v/r/l + @m
j/ʊ/v/ʁ/ʁ/ɾ/r/l + əl	single phones	j/v/r/l + @l
ts	ts	ts
pf	pf	pf

Table 1: *Unit definitions in IPA, in BOSS-SAMPA and as phoxsy units.*

Breuer and Abresch [1] have shown that the usage of phoxsy units in unit selection speech synthesis improves the quality of the synthesized speech.

2.1. Implementation

Taking the findings of [1] into account, phoxsy units were implemented as an independent multi-phone unit level in the Bonn Open Speech Synthesis System (BOSS) [7] in order to provide a robust and accessible usage. As a side effect, considerable runtime improvements compared to the use of lower-level units like phones were expected.

The modular architecture of BOSS [7] allowed an unproblematic integration of the new multi-phone level into the existing system. The BOSS tool blf2xml, which extracts information from the BOSS Label Format (BLF) files [8] and creates an XML database, has been extended in order to recognize phoxsy units using the BOSS_FSA class (a finite state automaton). The tool also inserts the units into a XML database. Other BOSS tools have also been adapted to calculate additional information like context classes, phrasing information, and MFCCs for phoxsy units and to add them to the XML database. The tool blfxml2db inserts the new multi-phone level into the MySQL database, calculating the unit index. The unit index is a unique number which identifies every unit in the corpus. Mapping tables provide links between the units of two adjacent levels. These levels are arranged hierarchically from words over syllables to phones and halfphones. The phoxsy multi-phone level is implemented as an intermediate level between syllables and phones. In order to maintain the hierarchy of the unit levels, this implicates that a complete coverage of the corpus by phoxsy units is necessary. The syllable map has been adapted in order to provide links between syllables and phoxsy units instead of syllables and phones. A phoxsy unit map has been created in order to provide links between phoxsy units and phones. A new preselection file for multi-phone unit preselection has been created. The BOSS_Unitselection class has been adapted and a new level PHOXSY has been added to the BOSS_Node class. BOSS_Transcription has also been adapted to identify and insert phoxsy units into the internal system communication structure. It uses the same mechanism as the blf2xml tool.

3. Corpus Recordings

The phonetic characteristics of natural fast speech differ from those of speech produced at normal speech rates. Due to the increasing overlap of articulatory gestures when speaking rate increases, the utterances of a speaker become less intelligible. The articulatory targets important for a clear pronunciation are no longer reached [9]. Strong coarticulation, reduction and other deviations from the clear canonical form affect the intelligibility of natural speech adversely [5, 10, 11]. Hence, these phenomena are undesirable in speech synthesis and need to be avoided during corpus recordings.

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 generally produce an inventory at higher quality than untrained speakers [12]. If the inventory is based on fast speech the emerging problems of articulatory precision and consistent voice quality will presumably increase. Assuming that untrained speakers will reduce the articulatory precision for the benefit of economic reasons to a greater extent than skilled speakers, a skilled speaker who was able to produce the required speaking style - both fast and clear - in an optimal way [13] was selected for inventory recordings.

To investigate the modeling of fast speech in unit selection synthesis, two independent but, in terms of linguistic content, identical unit selection inventories were created: one at normal and one at fast speech rate. Text materials consisted of 400 sentences which were selected randomly from the BITS Corpus [14] for German. Phonological balance was not taken into account. The 400 sentences were recorded in the following 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. The version articulated both most clearly and fast was selected by the phonetically trained persons and included in the fast speech corpus. Recordings at normal speech rate took approximately 10 hours, recordings at fast speech rate took nearly twice as long. This way, two unit selection corpora were created: one at normal speech rate and one at fast speech rate articulated as accurately as possible.

4. Synthesizing Normal Rate Speech

After corpus preparation and implementation [15], 15 sentences each containing at least three phoxsy units were synthesized on the basis of the normal speech rate corpus by using different strategies:

- using only phones for synthesis
- using only phoxsy units for synthesis
- using all unit levels excluding phoxsy units for synthesis
- using all unit levels including phoxsy units for synthesis

As a pairwise comparison between all four versions of a sentence would have exceeded a reasonable amount of judgments, it was decided to split the test sentences into two subtests, one compairing stimuli generated from a single unit level (phones only versus phoxsy units only) and another subtest comparing stimuli generated from all unit levels excluding and including phoxsy units, respectively.

The first experiment was a pairwise comparison between stimuli synthesized by using only phones and stimuli synthesized by using only phoxsy units. 23 subjects took part in the experiment. All of them were naive listeners and not experienced in using speech synthesis. Subjects were asked to indicate which version of the sentence was more intelligible and which version sounded more natural. Each of the 15 sentences was presented twice to the listeners to assess the reliability of judgments. Figure 1 shows the “more intelligible” and “more natural” judgments for stimuli consisting of phones only (dark grey) and for stimuli consisting of phoxsy units only (light grey). Results showed a significant difference (χ^2, $p < 0.05$) for intelligibility judgments with phoxsy units being rated as more intelligible. For naturalness judgments, no significant difference between the two versions was found.

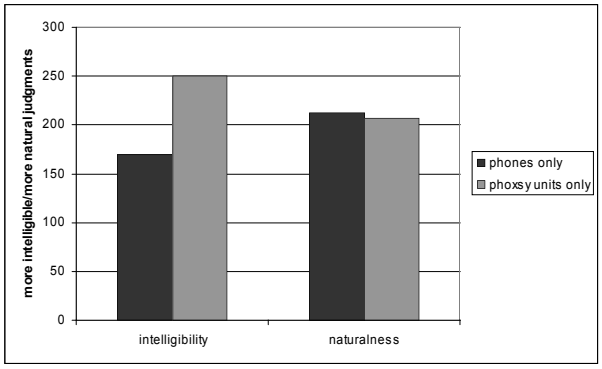

Figure 1: *Total number of preferred versions in terms of intelligibility and naturalness judgments for normal rate stimuli consisting of phones (dark grey) or phoxsy units (light grey) only.*

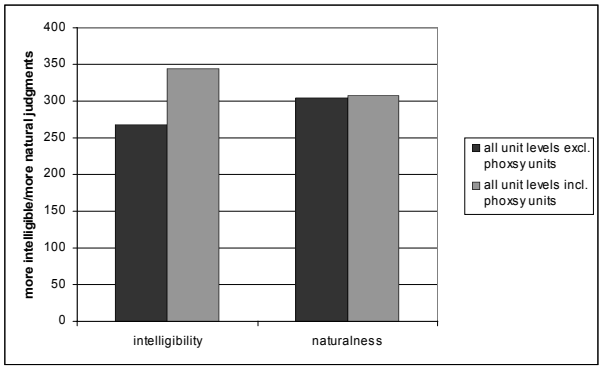

Figure 2: *Total number of preferred versions in terms of intelligibility and naturalness judgments for normal rate stimuli consisting of units from all levels beside phoxsy units (dark grey) or including phoxsy units (light grey).*

The second perceptual evaluation was a pairwise comparison between stimuli synthesized from all unit levels excluding phoxsy units and stimuli synthesized from all levels including phoxsy units. 14 subjects took part in the experiment. All of them were naive listeners and not experienced in using speech synthesis. Subjects were asked to indicate which version of the sentence was more intelligible and which version sounded more natural. Each of the 15 sentences was presented twice to the listeners. Figure 2 shows the "more intelligible" and "more natural" judgments for stimuli consisting of all unit levels excluding phoxsy units (dark grey) and for stimuli consisting of all levels including phoxsy units (light grey). Again, results showed a significant difference (χ^2, $p < 0.005$) for intelligibility judgments. For naturalness judgments, no significant difference between the two versions was observed.

The evaluation of the synthesis of normal rate speech showed a significant advantage in intelligibility for stimuli generated from phoxsy units only and stimuli generated from all unit levels including phoxsy units compared to the conditions where phoxsy units were left out for synthesis. Thus, the results presented by [1] were confirmed.

5. Synthesizing Fast Rate Speech

Strong coarticulation and reduction are not totally avoidable during the production of fast speech, even if it is produced with high precision and enhanced articulatory effort. Since phoxsy units are defined as sequences of phones prone to heavy coarticulation, their use may have a considerable impact on the intelligibility and naturalness of synthesized fast speech. On the one hand, a possible effect of using phoxsy units may be a degrading intelligibility of speech synthesized from natural fast speech including natural coarticulation and reduction phenomena. Alternatively, multi-phone units may enhance the intelligibility and/or naturalness of fast speech synthesized from a fast speech inventory because they provide more contextual information than single phones and therefore cover for coarticulation and/or reduction phenomena.

Again, 15 sentences containing at least three phoxsy units were synthesized by applying different strategies:

- using only phones for synthesis
- using only phoxsy units for synthesis
- using all unit levels excluding phoxsy units for synthesis
- using all levels including phoxsy units for synthesis

For fast speech, the first experiment was a pairwise comparison between stimuli synthesized from the fast speech inventory by using only phones for synthesis and stimuli synthesized by using only phoxsy units. 22 subjects took part in the experiment. As before, all of them were naive listeners and not experienced in using speech synthesis. Subjects were asked to indicate which version was more intelligible and which sounded more natural. Each of the 15 sentences was presented twice to the listeners. Figure 3 shows the "more intelligible" and "more natural" judgments for stimuli consisting

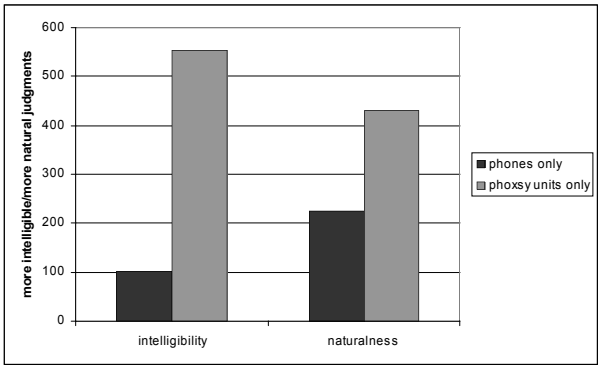

Figure 3: *Total number of preferred versions in terms of intelligibility and naturalness judgments for fast rate stimuli consisting of phones (dark grey) or phoxsy units (light grey) only.*

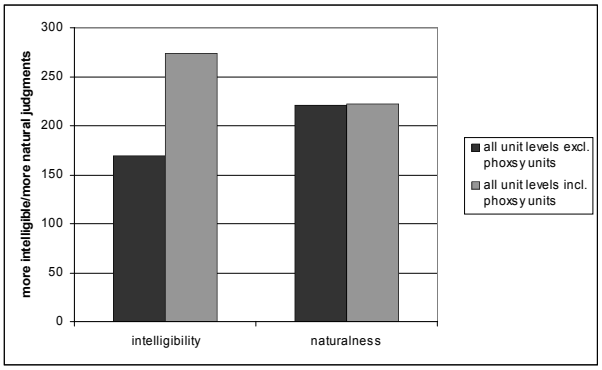

Figure 4: *Total number of preferred versions in terms of intelligibility and naturalness judgments for fast rate stimuli consisting of units from all levels beside phoxsy units (dark grey) or including phoxsy units (light grey).*

of phones only (dark grey) and for stimuli consisting of phoxsy units only (light grey). Results showed a significant difference (χ^2, $p < 0.001$) for intelligibility judgments as well as for naturalness judgments (χ^2, $p < 0.001$).

The second experiment was a pairwise comparison between stimuli synthesized by using all unit levels excluding phoxsy units and stimuli synthesized by using all unit levels including phoxsy units. 15 subjects took part in the experiment. All of them were naive listeners and not experienced in using speech synthesis. Subjects were asked to indicate which version of the sentence was more intelligible and which version sounded more natural. Figure 4 shows the "more intelligible" and "more natural" judgments for stimuli consisting of all unit levels excluding phoxsy units (dark grey) and stimuli synthesized by using all unit levels including phoxsy units (light grey). Results showed a significant difference (χ^2, $p < 0.001$) for intelligibility judgments. For naturalness judgments, no significant difference between the two versions was found.

The evaluation of fast speech synthesized from the fast speech inventory showed a significant advantage in both intelligibility and naturalness for stimuli generated from phoxsy units only. For stimuli generated from all unit levels excluding or including phoxsy units respectively, a significant difference was found for intelligibility judgments. Hence, multi-phone units are not only applicable to enhance the intelligibility of speech synthesized from a normal rate inventory, but also improve the intelligibility and to some extent the naturalness of fast speech synthesized from an independent fast speech inventory.

6. Discussion

Phoxsy units [1] were implemented as an independent multi-phone unit level in the Bonn Open Speech Synthesis System (BOSS) [8] in order to provide a robust and accessible usage. An evaluation of the synthesis of normal rate speech showed a significant advantage in intelligibility for stimuli generated by using only phoxsy units compared to stimuli synthesized by using only phones. For stimuli generated from all unit levels including phoxsy units, the intelligibility judgments also showed a significant advantage for this stimuli compared to stimuli generated by leaving out phoxsy units for synthesis. However, this significance was not as high as for the single unit condition. For naturalness judgments, no significant difference between the two versions in both single unit and all unit levels condition was found. The results presented by [1] were confirmed for normal rate speech.

For the synthesis of fast rate speech, an independent fast speech inventory was recorded where the fast speech was articulated as accurately as possible. Here, the use of phoxsy units has a considerable impact on both the intelligibility and naturalness of the synthesized speech since this multi-phone units provide more contextual information than single phones and therefore cover for coarticulation and reduction phenomena which may cause a degrading intelligibility. As expected, a perceptual evaluation showed a significant advantage in intelligibility and naturalness for stimuli generated by using only phoxsy units compared to stimuli synthesized by using only phones. For stimuli generated from all unit levels excluding or including phoxsy units respectively, a significant difference was only found for intelligibility judgments. Therefore, phoxsy units are not only applicable to enhance the intelligibility of synthesized speech at normal speaking rate, but also to improve the intelligibility and to some extent the naturalness of fast speech synthesized from an independent fast speech inventory.

7. Conclusions

The aim of the investigations presented here was the evaluation of a new approach to the synthesis of fast speech in unit selection speech synthesis. In line with results reported in the literature for normal rate speech [1], we showed that the implementation of multi-phone (phoxsy) units enhanced the intelligibility of synthesized speech significantly, especially when used for the synthesis of fast speech from an independent fast speech inventory. Further investigations will include the evaluation of the synthetic speech generated for the present study by visually impaired listeners as well as the evaluation of utterances generated from the different speech rate corpora and accelerated to varying (fast) speech rates by both listener groups.

8. References

[1] Breuer, S., Abresch, J., "Phoxsy: Multi-phone Segments for Unit Selection Speech Synthesis". In Proc. of Interspeech 2004 – ICSLP, Jeju Island, Korea, 2004.

[2] 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.

[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] 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.

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

[6] Breuer, S., "Multifunktionale und multilinguale Unit-Selection-Sprachsynthese". University of Bonn. Bonn, 2009. http://hss.ulb.uni-bonn.de/2009/1650/1650.htm

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

[8] Breuer, S., Abresch, J., Wagner, P. and Stöber, K., "BLF - ein Labelformat für die maschinelle Sprachsynthese mit BOSS II". In Hess, W., Stöber, K. (ed.), Tagungsband Elektronische Sprachsignalverarbeitung ESSV; Studientexte zur Sprachkommunikation. Bonn, 2001.

[9] Goldman-Eisler, F. (1961): The significance of changes in the rate of articulation. Language and Speech. Vol. 4, S. 171 – 174.

[10] 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.

[11] 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.

[12] Maus, V., "Zur Frage der Eignung von Sprechern als künstliche 'Stimme' in der konkatenativen Sprachsynthese." Master's Thesis. University of Bonn, 2004.

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

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

[15] Moers, D., Wagner, P., Möbius, B., Müllers, F. and Jauk, I., "Integrating a Fast Speech Corpus in Unit Selection Speech Synthesis: Experiments on perception, segmentation and duration prediction". In Proceedings of Speech Prosody 2010 Chicago, IL, 2010.