Slowing down synthetic speech by prosodic phrasing

Jürgen Trouvain

Institute of Phonetics, University of the Saarland, Saarbrücken,
Germany and Phonetics Consultancy Trouvain, Saarbrücken, Germany

Corresponding addresses:
University of the Saarland
FR. 4.7: Phonetics
66041 Saarbrücken
Germany

e-mail: trouvain@coli.uni-sb.de
tel.: +49 - (0)681 – 302 – 46 94
fax: +49 - (0)681 – 302 – 46 84

and

Phonetik-Büro Trouvain
Gustav-Bruch-Str. 46
66123 Saarbrücken

e-mail: trouvain@phonetik-buero.de
tel.: +49 - (0)681 – 935 75 78
fax: +49 - (0)681 – 935 75 79

Abstract

In most speech synthesizers, tempo control is performed by linear time scaling although tempo change in human speech is non-linear in nature. In perception experiments with a German speech synthesizer it was found that the versions with adjusted prosodic breaks and pauses are preferred over the linear versions for a "very slow" rate but not for "rather slow". The results are discussed with regard to benefits for various user groups one the one hand, and a possible general preference for slower tempo as a default setting on the other hand.

Keywords: speech synthesis, prosody, speech rate, evaluation

1. Introduction

In synthetic speech listeners may have different preferences with respect to the speech tempo. Various criteria can play a role such as experience with synthetic speech, familiarity with the voice, age of the listener, language proficiency of the listener, degree of possible hearing deficiencies, density of information, type of spoken text, duration of exposure to synthetic speech, and individual tempo preference.

For reasons of adaptation it can be assumed that persons who are confronted with synthetic speech for the first time would prefer slower synthetic speech than the default tempo. In contrast, people working with a speech synthesizer every day would advocate faster speech rates.

At present, if tempo in speech synthesizers is made adjustable it is usually performed linearly: the segmental and prosodic structures are kept constant, while only the segment durations are proportionally changed to the desired zooming factor. The result is similar to (but not the same as) a speech file played back with a lower sampling rate while retaining pitch characteristics. In contrast to such a *linear*, or uniform, manipulation of the temporal structure, the changes observable in humans' tempo-changed speech can be characterised as *non-linear*, or non-uniform. The number of pitch accents and prosodic breaks can be altered as well as the number and the mean duration of pauses. The duration of a segment is changed according to its degree of elasticity, i.e. very elastic segments such as long vowels are stretched and shrinked to a higher degree than less elastic segments such as short vowels. Sub-segmental timing can be affected in terms of the duration of steady states, a target undershoot, the degree of coarticulatory overlap, and the degree of articulatory velocity.

In the experiments summarized here (for details see Trouvain, 2002a,b), the hypothesis that slowed down synthetic speech with non-linear changes found in human speech would be preferred by listeners over linear methods is tested. As a first step, the speech tempo model applied here is restricted to prosodic phrase breaks with implications for pausing and, to a lesser extent, for phrase-final lengthening. In this way the number, the locations and the durations of pauses are controlled. Listening tests with stimuli generated by a German speech synthesizer are described and the results interpreted.

2. Tempo Control in Speech Synthesizers

Apart from linear time-scale modifications there have been several attempts to slow down the tempo of synthetic speech in a non-linear way (see Table 1). It is remarkable that only one model was actually tested with listeners. The others are either grounded on formal assumptions based on observations of natural speech, or they depend on speech production data with an evaluation of the model against these production data. Furthermore, none of the above-mentioned models considered *all* levels of non-linear changes mentioned in the previous section. As a consequence it would seem obvious a) to consider *all* levels in the model, and b) to perform perception tests.

Table 1. Approaches of non-linear control of slow tempo in speech synthesis. Language, evaluation method (production data or perception test), and considered levels of observed phenomena are indicated.

Study	Klatt (1979)	Kohler (1990)	Bartkova (1991)	Monaghan (1991)	Higginbotham et al. (1994)	Zellner-Keller (in press)
Language	American English	German	French	British English	American English	French
Evaluation	-	-	Data	-	Perception	Data
Prosodic breaks	x		x	x	(x)	x
Pitch accents				x		
Segments & syllables						x
Pause durations	x		x		x	x
Segment durations		x	x			x
Sub-segmental timing						

However, arguments against such an all-or-none model test are that a) the test result cannot explain which aspect of modeling accounts for the hypothesized better performance, b) it cannot be assured that all

presented aspects can be appropriately modeled, and c) a simple copy of natural speech phenomena to synthetic speech does not guarantee the listeners' acceptance.

For these reasons it was decided to start with a rather simple non-linear tempo model. It is commonly assumed that changes in speech rate are predominantly changes in pausing with a more or less constant articulation rate (Goldman Eisler, 1968; Künzel, 1997). Based on this assumption, the model aims at changing the *duration* and *number* of pauses. This in turn, requires the prediction of the *location* of pauses to be added or to be skipped. Pauses in read speech are usually linked with prosodic phrase breaks. The prediction of prosodic phrase structure in TTS systems is primarily based on punctuation and/or syntactic analysis. Thus, a prediction of inserted or skipped breaks/pauses must be handled at the relevant stage of linguistic analysis. There are different views on the diversity of prosodic break strength. The strength of the prosodic breaks influences their realizations. A higher-level break may be marked by a longer pause, increased phrase-final lengthening, and a more distinct F0 movement.

For modeling slowed down speech it is proposed to insert minor prosodic breaks *in addition* to the default breaks. Additional breaks will result in more pauses and increased final lengthening. For reasons of simplicity, a new break shall occur after each syntactic noun phrase and after each syntactic adjective phrase. This procedure is similar to the ones used in Klatt (1979) and Bartkova (1991), but different to Higginbotham, Drazek, Kowarsky, Scally, & Segal (1994) where a pause is inserted after each word. The duration of pauses should be considerably changed according to the desired tempo.

3. Listening Tests

To compare different tempo adaptation methods it was decided to compare versions with the same text and the same total duration for a given tempo. All versions to be compared shall show the same total duration. For two listening tests, a news paragraph (42 words and 36 words, respectively) has been synthesized with the German TTS synthesis system "Mary" (Schröder & Trouvain, 2003). Versions with four different speeds were generated with reference to the default output (including minor incorrect forms): "very slow" (140% of the default duration; 3.66 syllables/sec), "rather slow" (120%; 4.27

syllables/sec), and two fast versions not reported here. For each tempo, versions according to the two methods were generated: 1) a purely linear time-scaled version with preserved F0 characteristics, 2) a hybrid version generated in three steps (step 1: adjusting prosodic breaks; step 2: adjusting pause duration according to break level; step 3: linear time-scaling of the remaining signal).

For test 1, the model predicted changes in pause durations for each level of prosodic break. Furthermore, new minor breaks were inserted at syntactically appropriate locations leading to more pauses and more phrase-final lengthening. For test 2, the model has been refined ("adjusted model 2"): pauses, which were judged to be too long, have been shortened, the number of pauses have been reduced (but there were still more than in the default case), and the factor for phrase-final lengthened accented syllable rhymes has been increased.

Each of the eight stimuli (4 tempo; 2 orders) consisted of a linear–adjusted pair and were presented in a forced choice preference via loudspeakers in a quiet office to German native-speaking subjects (test 1 n = 15, test 2 n = 10). For a more detailed description of the audio material see Trouvain (2002c).

The hypothesis was that the "adjusted" versions are always preferred over the "linear" versions. The results presented in Table 2 confirm this hypothesis in the first test partially for "very slow", but not for "rather slow". Although there is an improvement from 17% in test 1 to 40% preferences in test 2, the adjusted model is not superior over the linear one in "rather slow".

Table 2. Results for the comparison of methods for two slow rates as percentage of preferences.

Tempo	Test	Linear model	Adjusted model 1	Adjusted model 2
Very slow	Test 1	17%	83%	
	Test 2a	70%	30%	
	Test 2b	20%		80%
	Test 2c		10%	90%
Rather slow	Test 1	83%	17%	
	Test 2	60%		40%

Although for "very slow" the result has been replicated with the second adjusted model (test 2b), the reverse picture surprisingly appeared for the first adjusted model (test 2a), which is in clear contrast to the result in test 1. The weak performance of model 1 is also reflected in the direct comparison of both adjusted models in test 2c.

4. Summary and Discussion

The perception tests described here asked for the preference between a linearly and a non-linearly tempo-modified version using a German diphone synthesizer. It has been shown that there is a benefit from an approach considering the structure and the realization of prosodic phrases. Slowing down seems to be sufficiently modeled by pauses as markers of phrase breaks featuring a) longer *relative* pause duration, b) an increased number of pauses, c) appropriate pause locations, with a *moderately* slower articulation rate as a consequence. This method could be promising to improve synthesizers for various applications listed at the beginning of the article.

However, with the focus on the outcome of the experiment 2a it must be admitted that more testing is needed to check various syntacto-prosodic variations in different text types. Some subjects reported that some pause locations were felt as disturbing. This implies that for slower speech rates not each syntactic break can be treated in the same way for predicting prosodic breaks. Here, a refined syntax-prosody mapping as well as the consideration of rhythmical balances across prosodic phrases is needed (cf. Atterer, 2002). This kind of research is also of great importance for timing in synthetic speech in general, not just for tempo scaling. The greater impact of the prediction of an acceptable prosodic structure compared to the prediction of segment duration has been shown in Brinckmann and Trouvain (2003).

Obviously, the benefit for *very* slow rates must not be beneficial for *rather* slow rates. This finding seems to be counter-intuitive at first glance. Even if some improvements could be reached with a refined model, the phrasing performance of an assumed "rather slow" tempo could not win over the linear adaptation with a default phrasing structure. An explanation of this result can be that listeners of synthetic speech prefer a slower rendition of speech but with a default phrasing, just as listeners of (other) forms of minor-quality speech prefer a slower

tempo (cf. Uchanski, Choi, Braida, Reed, & Durlach, 1996). If this were true, the consequences for speech synthesis developers would be to deliver synthesis systems with a rather slow tempo in the default setting. But there is of course the risk that a slow tempo may trigger the impression of boredom. This requires a more appropriate model of intonation to counteract the often-reported deficiencies of synthetic speech in general.

References

Atterer, M. (2002): Assigning prosodic structure for speech synthesis: a rule-based approach. *Proc. Prosody 2002*, Aix-en-Provence, pp. 147-150.

Bartkova, K. (1991): Speaking rate modelization in French application to speech synthesis. *Proc. ICPhS*, Aix-en-Provence (3), pp. 482-485.

Brinckmann, C. & Trouvain, J. (2003): The role of duration models and symbolic representation for timing in synthetic speech. *International Journal of Speech Technology Research* 6, pp. 21-31.

Goldman Eisler, F. (1968): *Psycholinguistics. Experiments in Spontaneous Speech*. Academic Press, London, New York.

Higginbotham, D. J., Drazek, A. L., Kowarsky, K., Scally, C. & Segal, E. (1994): Discourse comprehension of synthetic speech delivered at normal and slow presentation rates. *Augmentative and Alternative Communication* 10, pp. 191-202.

Klatt, D. H. (1979): Synthesis by rule of segmental durations in English sentences. In: Lindblom, B. & Öhmann, S. (eds.): *Frontiers of Speech Communication Research*, London etc: Academic Press, pp. 287-299.

Kohler, K. (1990): Zeitstrukturierung in der Sprachsynthese. *ITG-Fachberichte* 105, pp. 165-170.

Künzel, H.J. (1997). Some general phonetic and forensic aspects of speaking tempo. *Forensic Linguistics* 4 (1), pp. 48-83.

Monaghan, A.I.C. (1991): Accentuation and speech rate in the CSTR TTS system. *Proc. ISCA Workshop on Phonetics and Phonology of Speaking Styles*, Barcelona, pp. 41/1–5.

Schröder, M. & Trouvain, J. (to appear 2003): The German text-to-speech synthesis system MARY: a tool for research, development and teaching. *International Journal of Speech Technology Research* 6.

Trouvain, J. (2002a): Tempo control in speech synthesis by prosodic phrasing. *Proc. Konvens 2002*, Saarbrücken, DFKI Document D-02-01, pp. 227-230.

Trouvain, J. (2002b): Temposteuerung in der Sprachsynthese durch prosodische Phrasierung. *Proc. Konferenz Elektronische Sprachsignalverarbeitung (ESSV) 2002*, Dresden, pp. 294-301.

Trouvain (2002c): http://www.coli.uni-sb.de/~trouvain/tempo.html

Uchanski, R. M., Choi, S. S., Braida, L. D., Reed, C. M. & Durlach, N. I. (1996): Speaking clearly for the hard of hearing IV: further studies of the role of speaking rate. *Journal of Speech and Hearing Research* 39, pp. 494-509.

Zellner-Keller, B. (in press): Prediction of temporal structures for various speech rates. In Campbell, N. et al. (eds.): *Progress in Speech Synthesis II*. Springer-Verlag, Berlin, Heidelberg.