Tempo Control in Speech Synthesis by Prosodic Phrasing

Jürgen TROUVAIN
Institute of Phonetics
University of the Saarland
Saarbrücken, Germany
trouvain@coli.uni-sb.de

Abstract

Tempo control in most speech synthesisers is performed by linear time-scaling although tempo change in human speech shows a non-linear nature. In a perception experiment with a German speech synthesiser it was found that the versions with adjusted prosodic breaks and pauses are preferred over the linear versions for two fast rates and particularly for "very slow". However, the model for "rather slow" needs a refined syntax-prosody mapping.

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, hearing deficiencies, density of information, type of spoken text, duration of synthetic speech or simply the individual tempo preference in general. 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 synthesiser every day would advocate faster speech rates.

At present, if tempo in speech synthesisers is made adjustable then it is usually performed linearly: the segmental and prosodic structures are kept constant, just 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 or a higher 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. It includes assimilations, reductions, deletions of segments and syllables as well as the shift of syllable boundaries. 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 along its degree of elasticity. 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 experiment described here, the assumption is tested that synthetic speech with slow or fast tempo oriented to non-linear changes of human speech would be preferred by listeners over linear methods. 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 synthesiser are described and the results interpreted.

2 Tempo Control in Speech Synthesisers

Apart from linear time-scale modifications there have been several attempts to scale the tempo of synthetic speech in some non-linear way (see table 1). It is remarkable that only two models were actually tested with listeners. The others are either grounded on formal assumptions based on obervations 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. 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 tempo control in speech synthesis (except * for recorded, non-synthesised speech). Language (Am E = American English; Br E = British English; French, German), tempo (sl = slower; fa = faster), evaluation method (*prod*uction data or *perc*eption test), and considered levels of observed phenomena.

study	Klatt (1979)	Kohler (1990)	Monaghan (1991)	Bartkova (1991)	Higginbotham et al. (1994)	Covell et al. (1998)*	Zellner-Keller (in press)
language	Am E	German	Br E	French	Am E	Am E	French
tempo	sl/fa	sl/fa	sl/fa	sl/fa	sl	fa	sl/fa
evaluation	-	-	-	prod	perc	perc	prod
prosodic breaks	x		x	x	x		x
pitch accents			x				
segments & syllables							x
pause durations	x			x	x	x	x
segment durations		x		x		x	x
sub-segmental timing						x	

However, arguments against such an all-or-none model test are that a) the test results cannot explain which aspect of the model accounts for the hypothesised better performance, b) it cannot be assured that all presented aspects can be appropriately modelled, 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). Based on this assumption the model aims at the change of *duration* and the *number* of pauses. This again 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 this stage of linguistic analysis. There are different views on the diversity of prosodic break strength. The strength of the prosodic break may affect the realisations of the breaks. A stronger break may be marked by a longer pause, more phrase-final lengthening and a more distinct F0 movement.

Our simple model proposes the following: for slowing down minor prosodic breaks are inserted in addition to the default breaks. Additional breaks will result in more pauses and more 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 one used in Klatt (1979) (and repeated in Allen et al., 1987) and Bartkova (1991), but different to Higginbotham et al. (1994), where a pause is inserted after each word. The duration of pauses should be considerably changed according to the desired tempo. For speeding up predicted breaks shall be skipped with the result of fewer pauses and less phrase-final lengthening.

3 Listening Tests

Regarding 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. A news paragraph (3 sentences; 42 words) has been synthesised with the German TTS synthesis system "Mary" (Schröder & Trouvain, 2001). In "Mary", phrase-final lengthening is treated within the duration prediction model which is based on the additive-multiplicative rules of Klatt (1979). The intrinsic duration of syllable nuclei (usually vowels) is multiplied with factor 1.4 in minor phrase position and with factor 0.6 in non major phrase position; the intrinsic duration of coda consonants is multiplied by factor 1.1 in minor phrase final position.

Versions with 4 speeds were generated with reference to the default output (including minor incorrect forms): "very slow" (140% of the default duration), "rather slow" (120%), "rather fast" (80%), "very fast" (60%).

Pause durations were assigned according to the break strength and the envisaged tempo as can be seen in table 2. In "Mary", 7 levels of prosodic breaks are differentiated which correspond to those in Price et al. (1991) and the ToBI conventions. The break levels are labelled from "0" to "6" (with levels "1" and "5" being neglected). Level "2" is only considered for the slow versions of this study.

Table 2. Pause durations in msec according to prosodic break strength and envisaged tempo.

	very fast	fast	default	slow	very slow
break	60%	80%	100%	120%	140%
[2]	-	-	-	120	200
[3]	20	80	120	200	410
[4]	50	100	200	410	700
[6]	100	200	410	700	1000

Table 3. The second sentence extracted from the two *very slow* versions (A = linear; B = adjusted). For each stretch of text (upmost line) and prosodic breaks (upper line for A & B) duration of pause and articulation phases in msec are given (bottom lines). In cases where a break [2] is indicated for the *adjusted* version there is no break "-" in the *linear* version.

		Die Partei		teilte in Düsseldorf		und Berlin mit,		die Liste		sei am 10. April		eingetroffen.	
A	[6]		[2]		-		[4]		-		-		[6]
	653	742	249	1401	0	1103	312	595	0	1573	0	1012	634
B	[6]		[3]		[2]		[4]		[2]		[2]		[6]
	1090	541	494	1193	237	754	792	431	221	1200	210	737	1090

For each tempo, versions according to two methods were generated: purely *linear* time-scaled versions with preserved pitch characteristics, and hybrid versions with *adjusted* break prediction. Step 1 consisted of a prosodic re-phrasing for the *adjusted* versions, as explained in the previous section; in step 2, pause durations according to table 2 were assigned to the adjusted versions; finally, in step 3, a linear time-scaling took place for all versions so that for each tempo category the versions of both methods showed exactly the same duration. This resulted in 8 stimuli containing *linear–adjusted* pairs[i]. Table 3 shows an example sentence for the very slow versions. There, the effects of the inserted and promoted breaks for the durations of the pause and articulatory phases is visible.

In a forced choice preference test 8 stimuli (4 tempos; 2 orders) were randomly presented via loudspeakers in a quiet office to 15 German native-speaking subjects.

The first hypothesis is that the *adjusted* versions are always preferred over the *linear* versions. It is additionally expected that the break/pause effect is more distinct at slower rates since a slower reading style is usually characterised by more pauses. The results presented in table 4 confirm both hypotheses for three speech rates with the exception for *rather slow* (120%).

There are at least three assumptions to explain this outlier. In both slow versions, the number of pauses was more than doubled. It might be for the *adjusted* 120% version that the "interruption" of normally speeded speech by many pauses left a "choppy" impression and was for this reason not useful for a reasonable information chunking. Obviously what seems good for *very* slow rates must not be good for *rather* slow rates. A more moderate increase of the number of pauses seems appropriate. Some subjects reported that some pause locations were felt as disturbant. 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.

Table 4. Number of subjects (n=15) with consistent preferences for the linear method, the version with the adjusted breaks and those subjects showing inconsistencies.

		linear	adjusted	inconsist.
very slow	140%	0	10	5
rather slow	120%	5	0	10
rather fast	80%	0	8	7
very fast	60%	2	5	8

Despite the good performance of the simple break/pause model in this test, non-linear speech tempo adjusting for faster rates clearly needs further modifications. In a next step, de-accenting could be applied with the effect of less accentual lengthening. De-accenting could also

counteract the impression of over-accenting whereas phonemic reductions as well as spectral reductions could oppose the often felt segmental hyper-articulation. Further benefits can be expected from modelling the segment durations considering the different degrees of sound segment elasticity.

In contrast to speeding up, slowing down seems to be sufficiently modelled by a longer *relative* pause duration (reflected in pause-to-articulation ratio) at more pause locations with consequently a moderately slower articulation rate. Too slow articulation can enhance the sometimes reported effect of boredom (cf. the articulation rates of both *very slow* versions: 3.86 syll/sec for *linear* vs. 4.98 syll/sec for *adjusted*; speaking rate including pauses 3.48 syll/sec for both versions). Although the described mechanism was shown to work for *very slow*, the *rather slow* tempo apparently needs a refined break/pause prediction model. A prerequisite for such a refinement is to find out more about the factors determining the relation between word boundaries and the strength of prosodic breaks.

References

Allen, J., Hunnicutt, S., & Klatt, D. (1987) *From Text to Speech: The MITalk System.* Cambridge University Press, Cambridge.

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

Covell, M., Withgott, M., Slaney, M. (1998) MACH1: Nonuniform time-scale modification of speech. Proc. ICASSP Seattle.

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 Frontiers of Speech Communication Research, Lindblom, B. & Öhmann, S., eds, Academic Press, London New York San Francisco, pp. 287-299.

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

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.

Price, P.J, Ostendorf, S., Shattuck-Hufnagel, S. & Fong, C. (1991) *The use of prosody in syntactic disambiguation*. Journal of the Acoustical Society of America 90 (6), pp. 2956-2970.

Schröder, M. & Trouvain, J. (2001) The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching. Proc. 4th Speech Synthesis Workshop, Pitlochry, Scotland, pp. 131-136.

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

[i] See http://www.coli.uni-sb.de/~trouvain/tempo.html for test stimuli.