Chapter 7

Tempo-Scaled Synthetic Speech

Introduction

In synthetic speech listeners may have different preferences with respect to 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 hearing proficiency
- density of information
- type of spoken text
- duration of synthetic speech
- individual tempo preference

It can be assumed that persons who are confronted with synthetic speech for the first time may well prefer slower synthetic speech than the default tempo. In contrast, people working with a speech synthesiser every day would probably require faster speech rates.

At present, if tempo in speech synthesisers is made adjustable, it is usually performed linearly: the segmental and prosodic structures are kept constant, just the segment durations are changed proportionally by the desired zooming factor. The result is similar to (but not the same as) a speech file being 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.

After a survey of existing approaches to non-linear tempo control in our own experiments 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 over linear methods. As a first step the speech tempo models applied here are 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.

7.1 Approaches to non-linear tempo control

In principle there are four ways to change the tempo of synthetic speech which are sketched in figure 7.1.

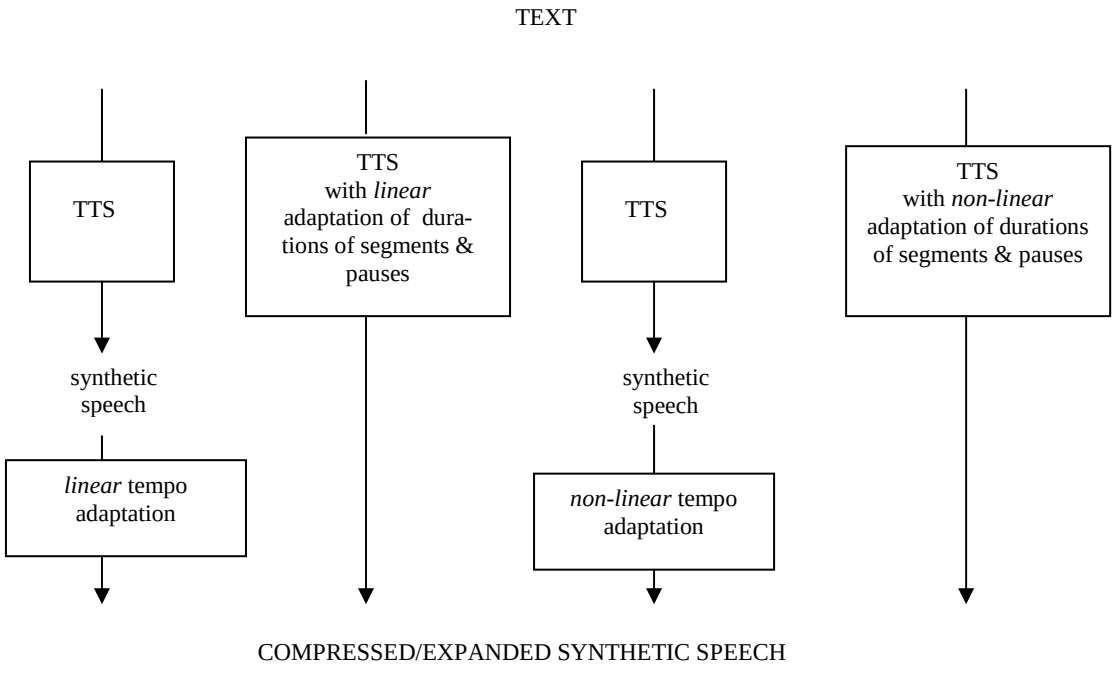

Figure 7.1 Four types of tempo adaptation for synthetic speech: 1) linear adaptation after synthesis, 2) linear adaptation during synthesis, 3) non-linear adaptation after synthesis, 4) non-linear adaptation during synthesis.

Either the adaptation of the synthesis output takes place *after* after the generation process (method 1 and 3 in figure 7.1) or the adaptation occurs *during* the generation of the synthetic speech (methods 2 and 4). Both methods can have *non-linear* or a *linear* time-scaling.

The two types of non-linear time-scaling will be discussed here:

- attempts to integrate non-linear aspects in the TTS generation (method 4)
- attempts with a post-processing of the non-linear time-scaling (method 2), where it is irrelevant whether synthetic or natural speech has to be manipulated

There have been earlier attempts to scale the tempo of synthetic and non-synthetic speech non-linearly. These are described briefly below and summarised in table 7.1.

Attempts with synthetic speech

In the classic additive-multiplicative segment duration prediction by Dennis Klatt developed for American English, it is recommended that a short pause is inserted between a content and a following function word (Klatt, 1979) and that "individual segments are lengthened and shortened slightly depending on speaking rate" (Allen et al. 1987: 98).

Global speech rate in a German TTS system (Kohler, 1988) affects the segment durations through one of many factors in a modified Klatt rule set. The consequence is that segments are modified proportionally to their inherent durations.

For a French synthesiser (Bartkova, 1991) a mix of modelling pause and segment durations is also suggested. In her model, global speaking rate influences the segment durations independently of the additive-multiplicative duration model. Pauses are mapped directly onto syntactic breaks, which are classified as obligatory and optional. Optional break locations are used to insert new pauses for slow speech and to skip pauses for fast speech, respectively. This information on phrase breaks, mostly punctuation-based, determines the occurrence and duration of pauses.

For an English TTS system Monaghan (1991) focuses on altering the phonological structure of prosodic phrases and pitch accents to manipulate speech rate rather than on a pure phonetic change of segment duration. He argues that manipulations on the phonological level will more effectively alter the *perceived* speech rate rather than the objective one. For the latter he proposes to concentrate on segment durations.

Hertz (1991) modelled diphthongs for a formant synthesiser. She presented a procedure for modelling the target underhoot of the second formant according to Gay's (1968) results.

Higginbotham et al. (1994) performed text comprehension tests with two different American English TTS systems. The listener performance of read texts synthesised in two modes were compared: a) with the default speech rate, and b) versions where a pause of 10 seconds (!) had been inserted after each word. For each variable (text type, text length, TTS system) the slowed versions scored better than the standard settings in a summarising task. Thus, although many rather long pauses were inserted while the articulation rate was kept constant, the comprehension level increased.

Portele (1996) manipulated the temporal structure of segments such that particularly steady state phases were shortened or lengthened. The listening tests showed no significant difference between those signals with modified spectral dynamics and those without.

For a French TTS synthesiser Zellner-Keller (in press) applies re-syllabification and segmental rules as well as the addition of pauses and prosodic breaks. An important feature of the break assignment is that the breaks are not only determined by syntactic but also by rhythmical constraints. To calculate the actual segment durations, speech rate was taken into account as one of several factors.

Attempts with non-synthetic speech

The researcher team of Picheny et al. (1989) and Uchanski et al. (1996) published data where word intelligibility was tested with sentence material recorded in a conversational style and in a clear speech style. Both groups of material were manipulated such that the faster conversational-style sentences reached the duration of their clear-style counterparts, and vice versa. The first study performed a linear time-scale whereas the second study applied non-linear modifications. The word intelligibility scores of the test persons (with hearing loss) showed that the non-linear versions are superior to the linear ones for both manipulation methods (slowed down conversational style, and speeded up clear style).

In the more recent study it was shown that the manipulated versions were less intelligible than the original versions. This is true for persons with hearing deficiencies, for normal hearing persons under noise conditions, and for normal hearing people

with the speeded up clear speech (but not with the slowed down conversational speech). Thus, almost any less than ideal situation (e.g. noise, or synthetic rather than natural speech) as well as time-scale adjustment of both speaking styles have a negative effect on word intelligibility. Those factors have to be taken into account, particularly for speeding up synthetic speech, because material for speech synthesis is usually recorded in a clear style rather than in a conversational style.

The study by Covell, Withgott & Slaney (1998) also provides evidence for the superiority of a non-linear over a linear approach for speeding up. To compress prerecorded speech they cut down the durations of

- pauses (but not below a threshold of 100 ms)
- unstressed vowels (by an intermediate amount)
- stressed vowels (to a lesser degree)
- consonants (based on the stress level of the neighbouring vowel)

They paid special attention to spectrally changing transitions and to already short segments so that these portions were not affected too much. In listening tests comparing linearly vs. non-linearly compressed speech, the non-linear versions scored significantly better in comprehension tasks for short dialogues and monologues as well as for A-B preference tests. Interestingly, there was no significant difference between the two compression methods for longer dialogues. A possible explanation for this is that there is a perceptual adjustment to all sorts of speaking styles, and that a perceptual adjustment to the unnatural speaking style takes longer for the linear compressed speech, with consequences for shorter utterances rather than for longer utterances.

In contrast to the expectation that non-linear methods of compressing speech yield better results than linear methods, the work of Janse (2003) revealed that word intelligibility in Dutch performs better when linearly adapted. The results of her experiments with a high and a very high compression rate (40% and 60%, respectively) are interpreted under assumption that segmental information of the more temporally reduced unstressed syllables are lost for the listener.

He & Gupta (2001) tested three time-compression techniques in terms of intelligibility and preference: a) linear time-compression, b) pause removal with following linear compression, c) a non-linear compresssion method similar to the MACH1 algorithm described Covell, Withgott & Slaney (1998). Their results show that there was

no significant difference neither in preference nor in intellegibility between non-linear and linear compression algorithms at moderate compression rates which correspond to about 60% normal rate duration. However, for the high speedup factor (2.5 faster than normal) the non-linear compression methods show significantly better results than a linear adaptation in the comprehension as well as in the preference tests.

To summarise the presented approaches to non-linear tempo change of recorded speech: For very extreme changes of articulation rate it seems insufficient to regulate only one property, e.g. segment duration, as was done for extremely fast articulation (Janse, 2003) as well as for extremely slow articulation (Neijme & Moore, 1998). It seems more promising if a number of phonetic and phonological mechanisms are taken into account as was done in Covell, Withgott & Slaney (1998) where the markers of prosodic re-structuring such as pauses, stress conditions as well as segment class and sub-segmental structure were considered.

Conclusions

The attempts discussed above to scale the tempo of synthetic speech in some non-linear way are summarised in table 7.1. Two points about them are remarkable.

First, very few of the models scaling the tempo of synthetic speech were actually tested with listeners such as Higginbotham et al. (1994), Portele (1996) and Janse (2000). The others are either grounded in formal assumptions based on observations of natural speech (Klatt, 1979; Kohler, 1988; Monaghan, 1991; Hertz, 1991), or they depend on speech production data with an evaluation of the model against these production data (Bartkova, 1991; Zellner-Keller, in press).

Second, none of the above mentioned models considered *all* structural levels presented in the chapter on the phonetic and phonological aspects of tempo change.

For an efficient tempo modelling it would seem obvious a) to consider *all* levels in the model, and b) to perform perception tests. However, there are arguments against such all-or-none model tests. Even if the results are in favour of our hypothesis that a "full" non-linear tempo model is preferred over a linear modification it cannot explain *which* aspect of modelling accounts for the better performance. Additionally, it cannot be assured that all aspects presented can be modelled in a comparable and appropriate way. And last but not least, there are reasons to doubt that simply copying observations from natural speech to synthetic speech are appreciated by listeners, as the examples for segmental reductions (Portele, 1997) and spectral tilt (Barry et al., in press)

show. Thus, it was decided to start with a non-linear tempo model which seems rather simple at the first glance.

Table 7.1. Approaches of non-linear tempo control in speech synthesis (except * for recorded speech). Language (AmE=American English; BrE=British English; Fre=French, Dut=Dutch, Ger=German), tempo (sl=slower; fa=faster), evaluation method (production data or perception test), and considered levels of observed phenomena: prosodic breaks, pitch accents, segmental and syllabic structure, pause duration, segmental duration, sub-segmental timing.

study	lang.	tempo	eval.	pros. breaks	pitch acc.	segm & syll	pause dur.	segm. dur.	sub-segm. timing
Klatt (1979)	AmE	sl/fa	-	x			x	x	
Kohler (1988)	Ger	sl/fa	-					x	
Bartkova (1991)	Fre	sl/fa	prod	x			x	x	
Hertz (1991)	AmE	fa	-						x
Monaghan (1991)	BrE	sl/fa	-	x	x				
Higginbotham et al. (1994)	AmE	sl	perc	x			x		
Covell, Withgott, Slaney (1998)*	AmE	fa	perc				x	x	x
Portele (1996)	Ger	sl/fa	perc						x
Uchanski et al. (1996)*	AmE	sl/fa	perc				x		x
He & Gupta (2001)*	AmE	fa	perc				x	x	x
Janse (2003)*	Dut	fa	perc					x	
Zellner-Keller (in press)	Fre	sl/fa	prod	x		x	x	x	

7.2. Prosodic phrasing in the MARY text-to-speech synthesiser

The hypothesis is that tempo-scaled synthetic speech with non-linear changes found in human speech would be preferred by listeners over linear methods. In this section a model is described which takes the non-linear changes found in human speech into consideration.

As already pointed out in chapter 2, Goldman Eisler (1968) claims that changes in speech rate are predominantly changes in pausing with a more or less constant articulation rate, an observation confirmed for perceptually extreme changes by the study presented in chapter 5. Based on this assumption the model presented here focuses on pausing as phonetic marker and phrasing as determination of pausing structure. This should include more than just changes in the *duration* of predicted pauses. It should also consider changes in the *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 synthesis systems is primarily based on punctuation and/or syntactic analysis. Thus, a prediction of inserted breaks/pauses and of skipped breaks/pauses must be handled at this stage of linguistic analysis.

The strength of the prosodic breaks influences their realisations. A higher-level break may be marked by a longer pause, increased phrase-final lengthening and a more distinct F0 movement. For slowing down, our first model proposes to insert minor prosodic breaks in addition to the default breaks. Additional breaks will result in more pauses and more phrase-final lengthened syllables. For reasons of simplicity, a new break will occur after each syntactic noun phrase and after each syntactic adjective phrase. Moreover, the duration of pauses will be considerably lengthened. This procedure is slightly different to those in Bartkova (1991) and Klatt (1979), where a pause is inserted between *each* content and function word, and very different to Higginbotham et al. (1994), where a pause is inserted after each word. The duration of pauses will be changed considerably according to the desired tempo.

Conversely, for speeding up, predicted breaks will be skipped, resulting in fewer pauses and fewer cases of phrase-final lengthening. Pause durations shall be shortened.

Default phrasing in MARY

Before going into the details of the model, which alters the prosodic structure and prosodic events for changing speech tempo, it is necessary to present the default mechanism of the synthesiser used for the experiments. The default output of the German TTS system MARY (Schröder & Trouvain, 2001) serves as the baseline for the model that is summarised in table 7.2. There are four types of breaks to be predicted, which are all based on the German ToBI conventions (Baumann, Grice & Benzmüller, 2001). These, in agreement with the original ToBI model for American English (Beckman & Ayers, 1994) define six levels of break indices.

A break "2" occurs before a prepositional phrase (PP) and before a conjunction in coordinated noun phrases (NP) or coordinated adjective phrase (AP), e.g. in "Er sprach [break 2] mit belegter Stimme." The default realisation does not currently manifest a pause in the temporal segmental structure, nor does it trigger a boundary tone.

A break "3" which corresponds to a "minor prosodic break" or a boundary of an "intermediate phrase (ip)" is assigned in two cases: 1) before the finite verb, i.e. after the German "Vorfeld" if this stretch of the sentence exceeds two syllables; example: "Der amerikanische Präsident [break 3] sagte gestern …". 2) before the conjunctions "und" (English "and") and "oder" (English "or") example: "Er fuhr nach Köln [break 3] und besuchte eine Freundin.". A break "3" is marked by a 120 ms pause, a final lengthening factor for parts of the last syllable in the duration model (see table 7.2), and a minor boundary tone (H-) which changes the F0 excursion size to a small degree.

A break "4" is linked with a comma in the text which in most cases represent the division of clauses, tokens of an enumeration, or tags. An example is "Er trank das Bier, [break 4] obwohl er keinen Alkohol mochte." The realisation of a break "4" consists of a 200 ms pause, the same final lengthening factor as with "3", but major boundary tones (e.g. H-% and L-%) leading to bigger changes of the F0 excursion size.

A break "6" is assigned at the end of a sentence and is marked by a longer pause than "4" (410 ms). Roughly speaking, a break "4" as well as a break "6" can be seen as an "intonation phrase" boundary. The difference between "4" and "6" in MARY lies in the syntactic embeddedness expressed by punctuation.

Neither a break "5" nor a break "1" is currently used in the synthesiser.

The default states described here and summarised in table 7.2 will not just modified in terms of existing pause durations. Pauses will also be inserted, e.g. break "2" becomes a pause for slowing down.

Table 7.2. Default mechanism for predicting the position, break stength and realisation of a prosodic break (pause duration in ms; final lengthening factor in duration model; boundary tone triggering F0 excursion size).

break	predicted position	pause duration	factor final lengthening	boundary tone
"2"	PP; Conjunction in coordinated NP or AP	-	-	-
"3"	finite verb > 2 tokens; "und"/"oder"	120	1,4 (nucleus) 1,1 (coda)	H-
"4"	comma	200	0,6 (elsewhere)	H-%,
"6"	end of sentence	410		H-^H%, L-%

Like all TTS systems, this default model shows potential caveats such as an unclear correlation between punctuation signs especially commas and break strength, and a missing theory-bound classification of the various break strengths. It would certainly be helpful to have a more distinct modelling of phrase-final lengthening and production based pause duration. Further missing aspects are rhythmical balance (as considered e.g. by Zellner-Keller, in press), as well as semantic and pragmatic contexts. Although it is clear that this default model does not fully reflect speech production data, it produces acceptable prosodic phrases for German texts, as perception tests confirmed.

7.3. Perception experiment 1

Methods

In order to compare different tempo adaptation methods all versions to be compared need to show the same total duration. It was decided to test the preference of two consecutively played speech stimuli (paragraph-length) that differ just in the way the tempo was adjusted. Stimuli were generated for four tempo categories with the German text-to-speech synthesiser "Mary" using diphones (Schröder & Trouvain, 2001). Each of the tempo categories has a certain compression or expansion factor relative to the default duration assigned in "Mary". That means, that an expansion of the duration of the entire speech stimulus by 20% would result in a 120%-version (relative to the default), and a compression by 40% would lead to 60%-version. The tempo categories and their stretching values are as follows:

- very slow (140%)
- rather slow (120%)
- rather fast (80%)
- very fast (60%)

For each of the four tempo categories, versions were generated according to two methods:

- a purely *linear* time-scaled version with preserved pitch characteristics
- a hybrid version with *adjusted* break prediction

In total, there were eight versions (4 tempo x 2 methods) to be used in four pairs for the preference test. In order to minimise a list effect, each stimulus appeared once in the first position of a stimulus pair, and in the second position in a further stimulus pair. This resulted in eight stimuli containing *linear–adjusted* pairs.

The versions with the adjusted break prediction were generated in three steps:

- step 1: adjusting prosodic breaks
- step 2: adjusting pause duration according to break level and tempo category
- step 3: linear time-scaling of the remaining signal

Step 1 and 2 were considered by the first model that features the following modifications of the default set-up: for both slow rates, breaks of strength "2" are inserted after *each* noun phrase (NP) and *each* adjective phrase (AP). For both fast rates the breaks of strength "3" are demoted to "2". The envisaged effect is to insert more pauses with their accompanying final lengthening for slow rates, and that pauses are skipped with their accompanying final lengthenied syllables for fast rates. As can be seen in table 7.3, the duration of pauses are dependent on two factors: the break strength and the envisaged tempo.

Table 7.3: Pause durations of model 1 according to prosodic break strength and tempo.

break	very fast (60%)	rather fast (80%)	default (100%)	rather slow (120%)	very slow (140%)
"2"	-	-	-	120	200
"3"	20	80	120	200	410
"4"	50	100	200	410	700
"6"	100	200	410	700	1000

An example for both versions can be seen in a sentence of the text of the first experiment in table 7.4. Note that in the non-linearly adapted versions, pauses are longer and more frequent, and the articulation phases are shorter compared to the linearly adapted versions.

15 students of phonetics and computational linguistics, all German native-speakers served as subjects. Their experience with synthetic speech ranged from none to some. Subjects were told that a newsreader with an artificial voice would be tested and that this voice can speak at various speeds. They were asked to select the version they preferred from each pair of news paragraphs (for texts see appendix). All pairs occurred in both orders, and all stimuli pairs were randomised. They were presented via loudspeakers in a quiet office with one warm-up stimulus at the default tempo. The test took about 10 minutes per subject.

Table 7.4. The second sentence extracted from the two *very slow* versions (A = linear; B = hybrid). For each stretch of text (top line) and prosodic breaks (upper line for A & B) the duration of pause and articulation phases in ms are given (bottom lines). In cases where a break "2" is indicated for the *hybrid* 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

Results

The first hypothesis was that the hybrid versions would always be preferred over the linear versions. In addition, it was expected that the break/pause effect would be more distinct at slower rates since slower readings usually show more pauses. The results presented in table 7.5 confirm both hypotheses for three of the four speech rates, with the exception of rather slow (120%): listeners preferred the adjusted versions, especially for "very slow" reflected by the high number of consistent answers.

Table 7.5. The preferences (15 listeners) in percent for the first perception experiment comparing the linear versions and the adjusted versions (model 1). The percentage of inconsistent judges are in parentheses.

tempo	linear – adjusted 1
very slow	17% – 83% (33)
rather slow	83% – 17% (33)
rather fast	23% – 77% (46)
very fast	40% – 60% (80)

Subjects differed with regard to the consistency of their answers reflected in different preferences in the two pairs containing the same versions. The number of inconsistent answers increased from 33% for very slow and rather slow rates up to 80% for the very fast rate.

Discussion

One possible explanation for the exception at "rather slow" is that in both slow versions the number of pauses was more than doubled. It might be that for the adjusted 120% version the "interruption" of normal-tempo speech by so many pauses left a "choppy" impression and for this reason the word sequence was not amenable to a reasonable information chunking. Obviously, what seems good for *very* slow rates need not necessarily be good for *rather* slow rates. A more moderate increase in the number of pauses seems advisable. Some subjects reported that pauses at some locations were perceived as a disturbance. This implies that - for slower speech rates - not every syntactic break can be treated in the same way to predict prosodic breaks. Here, a refined syntax-prosody mapping as well as the consideration of the rhythmical balance across prosodic phrases is needed.

In contrast to speeding up, slowing down seems to be sufficiently well modelled by a longer *relative* pause duration (reflected in pause-to-articulation ratio) at more pause locations with a moderately slower articulation rate. Too slow an articulation can strengthen the effect of boredom that is sometimes reported. Although the described mechanism was shown to work for "very slow", the "rather slow" tempo clearly needs a refined break/pause prediction model. But also the "very slow" version deserves a refinement, because the "very slow" versions left the impression of rather fast articulation phases with a very high number of pauses with some overlong pauses.

7.4. Perception experiment 2

Based on the outcome of the first listening experiment the first model of break and pause prediction has to be refined (henceforth model 2) and tested again with listeners. Thus, the goal of the second perception experiment is to find answers to the following research questions:

- Can we replicate the good result for *very slow* in experiment 1, either with model 1 or with model 2?
- Does model 2 perform better than model 1 for *rather slow*?
- Does model 2 perform better than the linear model for *rather slow*?
- Does model 1 or model 2 *generally* perform better?

Methods

Model 2 aims to avoid the deficits of model 1 that appeared in the first experiment and to deliver some refinements. The rather fast articulation phases for the slow versions should be slowed down, the excessive number of pauses should be avoided, and the overlong pauses should be shortened. Furthermore, the very slow version should be improved by a higher degree of phrase-final lengthening. Therefore the following changes apply to model 2:

- insert break "2" after a NP or VP just in those cases where the new minor phrases (after insertion of a break "2") also show a predicted pitch accent
- apply additional factor 1.5 for each syllable rhyme (nucleus plus coda) in each pitch accented word (all speech rates)
- apply shorter pause durations according to the values in table 7.6
- maintain the break "3" for fast rates (in contrast to model 1 where it has been skipped)

Table 7.6: Pause durations of model 2 according to prosodic break strength and envisaged tempo. If different, pause durations of model 1 in parentheses.

break	very fast (60%)	rather fast (80%)	default (100%)	rather slow (120%)	very slow (140%)
"2"	-	-	-	100 (120)	120 (200)
"3"	40 (20)	80	120	180 (200)	300 (410)
"4"	50	100	200	300 (410)	700
"6"	100	200	410	620 (700)	1000

The same test paradigm is applied as in experiment 1, but with a different news text (2 sentences, 36 words, 74 syllables; see appendix). In total 10 German native listeners took part.

Results

At the "very slow" rate, the second model performs slightly better than the first model in the first experiment with 80% of the preferences. However, the repeated test of the first model in this experiment scored only 30% preference, in contrast to 83% in the previous experiment. A direct comparison of the two models at this tempo showed a very clear preference for model 2.

At the "rather slow" rate, the second model improves compared to model 1 in the first experiment. Nevertheless, the listeners still preferred the linear version at this rate. Since model 1 was considered inferior for "rather slow" in experiment 1, no direct comparisons between model 2 and model 1 were performed for that specific tempo category in experiment 2.

At both fast rates, the results show a preference for model 1 compared to the linear versions, slightly weaker than in experiment 1 for "rather fast" and slightly stronger for "very fast". There is no preference for model 2. In the direct comparison of the two models, the model 1 is clearly preferred at the "very fast" rate.

Table 7.7. The preferences in percent for the comparison from the first experiment (replicated from table 7.5) and the three comparisons of the second perception experiment. Percentages of inconsistet judges are in parentheses.

	test 1	test 2		
	linear – adj. 1	linear – adj. 1	linear – adj. 2	adj. 1 – adj. 2
very slow	17% – 83% (33)	70% – 30% (40)	20% – 80% (40)	10% – 90% (40)
rather slow	83% – 17% (33)	-	60% – 40% (40)	-
rather fast	23% – 77% (46)	40% – 60% (80)	45% – 55% (50)	55% – 45% (50)
very fast	40% – 60% (80)	30% – 70% (60)	55% – 45% (50)	70% – 30% (40)

Discussion

The following discussion is oriented along above mentioned research questions.

- Can we replicate the good result for *very slow* in experiment 1, either with model 1 or with model 2?

Regarding model 2, the answer is yes, regarding model 1 the answer is no. On the one hand it is satisfying to know that model 2 in experiment 2 performs as well as model 1 in experiment 1. On the other hand it is surprising that the same model which gave a very good performance in one experiment, fails in a second experiment. The essential difference between the two experiments was the text. This means that break predictions are unreliable, in turn, implies that too few of the relations between syntactic and prosodic breaks were considered, and possibly that the rhythmic balance of prosodic phrase length play a greater role than expected. Future modelling of prosodic phrasing needs to take these two aspects into consideration. A particular feature of the linear versions at a very slow rate is the highly unnatural slow articulation rate. This was avoided in the adapted versions by inserting more pauses and lengthening of them. This finding can play an important role for many types of users, e.g. older people, or those unexperienced with synthetic speech.

- Does model 2 perform better than model 1 for *rather slow*, and does model 2 perform better than the linear model for *rather slow*?

Model 2 indeed performs better for *rather slow* but is still inferior to the linear model. One possible explanation for this unexpected result is that those listeners generally prefer slower rates when speech is distorted in any way, and this is the case for synthetic speech. That means that the rate we declared here as *rather slow* - seen from a speech production perspective - is in fact for most listeners the *normal* rate - for perceiving synthetic speech. Obviously, normal articulation rate with as many breaks as in slow speech is not appreciated by the listeners. The implication from this interpretation is that the default tempo of synthetic speech should be slower than the default tempo of natural speech. However, this recommendation should not be generalised for all types of users of synthetic speech: a blind person who uses speech synthesis every day will express tempo needs which are completely different from those just described.

- Does model 1 or model 2 *generally* perform better?

Here, it is impossible to give a clear answer. For "very slow" it cannot be definitely decided which model is better. Model 1 performed well in one experiment but failed in the other. Model 2 (in experiment 2) was equally as good as model 1 (in experiment 1). Model 2 showed improvements for *rather slow*, but not with the envisaged result that it outperforms the linear method. For both *fast* categories, the first model generally performed better than the second one. This means that the first model seems to show a possible direction for altering fast synthetic speech by means of prosodic phrasing.

Summary and discussion of chapter 7

With these experiments it has been shown that it is possible to alter the tempo in a satisfactory way for text-to-speech synthesis. Compared to the use of changing tempo in natural speech, the modelling demonstrated here is restricted to changes of the *global* tempo, for *read* speech, and in *monologues*. This is in contrast to local tempo changes in spontaneous dialogues presented in chapter 4.

In contrast to most other studies dealing with tempo control, we performed perception experiments. We were able to show that just modelling prosodic phrasing

leads to partial improvements. However, modelling just phrases seems more complicated than expected, and is not as easy as e.g. the Klatt rules "predict". Not only two categories such as slow and fast, which are rather abstract and therefore vague, were tested; there were four categories in total, with an exact reference to a default speed.

The results for "very slow" are evidence that improvements are possible for this tempo category, at least for German speech synthesis. The findings can be integrated in several speech synthesis applications such as general information systems where users are confronted with synthetic speech for the first time or in user-adaptive systems aiming at non-native speakers or those with hearing deficiencies (see introduction of this chapter and also chapter 2). But the findings can also be used to slow down pre-recorded natural speech in the area of language learning.

Despite a 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 fewer cases of accentual lengthening. De-accenting could also counteract the impression of over-accenting whereas phonemic reductions as well as spectral reductions could oppose the impression of segmental hyper-articulation which is often felt. Further benefits can be expected from modelling the segment durations considering the different degrees of sound segment elasticity.

The results for *rather slow* suggest that the determination of the default speed is the first problem when controlling tempo. On the basis of these results and the study of Uchanski et al. (1996) it can be assumed that listeners prefer a slower tempo for synthetic speech than they do for natural speech. This has consequences for defining the default tempo of synthesisers, but also for the test and training material used for timing prediction in TTS systems, especially the modelling of segment duration. Here, fast reading styles such as news readings do not seem very appropriate (see chapter 2). Finally, it must be said that any improvement of the timing for the default tempo also improves the quality of speech rates other than default.