TEMPO VARIATION IN SPEECH PRODUCTION

IMPLICATIONS FOR SPEECH SYNTHESIS

Dissertation
zur Erlangung des Grades eines Doktors der Philosophie
der Philosophischen Fakultäten der Universität des Saarlandes

vorgelegt von

JÜRGEN TROUVAIN

Saarbrücken, im April 2003

Dekan: Prof. Dr. Klaus Martin Girardet

Berichterstatter: Prod. Dr. William J. Barry und Prof. Dr. Martine Grice

Tag der letzten Prüfungsleistung: 17.07. 2003

Zusammenfassung

Tempoveränderungen in gesprochener Sprache werden von vielen verschiedenen Variablen beeinflusst, die nicht genuin phonetischer oder phonologischer Natur sind. Wie in Kapitel 2 dargestellt, umfassen diese Variablen extra- und para-linguistische Faktoren wie Emotionen, Sprechereinstellungen, Stress, Alter, Sprachbeherrschung, Sprech- und Hörstörungen, die Rolle des Kommunikationspartners, oder die gewohnte Sprechgeschwindigkeit eines jeweiligen Sprechers. Auch sprachrelevante Faktoren, wie z.B. Textsorte (geschrieben oder gesprochen), Worthäufigkeit, Sprachplanung, Diskursorganisation und Informationsmanagement bestimmen das Sprechtempo.

Die rein phonetischen oder phonologischen Parameter werden in Kapitel 3 behandelt. Zunächst wird auf die zentrale Bedeutung von Pausen und prosodischer Phrasierung im Hinblick auf das Sprechtempo eingegangen, gefolgt von Abschnitten über den Einfluss von Intonation und Rhythmus. Der Abschnitt über phonologische Prozesse, die in zusammenhängender Rede auftreten, beschäftigt sich mit Veränderungen auf der Segment-Ebene, wie etwa Assimilationen, Tilgungen und phonemische Reduktionen, während der anschließende Abschnitt einen Überblick über die Faktoren liefert, die die Dauern von Segmenten und Silben beeinflussen. Das Kapitel schließt mit einer Einführung in Mechanismen auf der artikulatorischen Ebene als Einflussgröße bei Tempoveränderungen. Hierbei ist noch anzumerken, dass alle Prozesse der Tempovariation wie sie auf den genannten strukturellen Ebenen ablaufen als nicht-linear anzusehen sind.

Ein zentrales methodisches Problem besteht in der Messung von Sprechtempo. Kapitel 4 diskutiert die Eignung der linguistischen Einheiten Wort, Silbe und Lautsegment als angemessene Einheit für Sprechtempo-Messungen. Obwohl zeitliche Varianz am besten anhand von Lautsegmenten gemessen werden kann, erweist sich diese Einheit als schwierig zu definieren und zu zählen, im Gegensatz zur (phonologischen) Silbe und zum Wort. Die Einheit Wort hat allerdings Nachteile im Hinblick auf zeitliche Varianz und Vergleichbarkeit mit anderen Studien. Die Wahl der Einheit hängt letztlich von Ziel der jeweiligen Studie ab. Aus diesem Grunde ist es unmöglich, eine optimale Einheit für die Sprechtempomessung auszumachen.

Die zentrale Rolle der Pausen spiegelt sich in der wichtigen Unterscheidung zwischen Sprechgeschwindigkeit (mit Pausen) und Artikulationsgeschwindigkeit (ohne Pausen). Unabhängig vom globalen Tempo kann die Artikulationsgeschwindigkeit in einer Sprechphase zwischen zwei Pausen erheblich variieren, was sich zuweilen in einer Beschleunigung oder Verlangsamung innerhalb der Artikulationsphasen niederschlägt.

Der praktische Teil der Arbeit beginnt mit einer Analyse von natürlichen Daten in Kapitel 5. Anhand einer Fallstudie werden die prosodischen Merkmale des emotionalen Sprechstils von Pferderennkommentaren untersucht. Der auditive Eindruck eines hohen Sprechtempos im letzten Teil derartiger Kommentare schlägt sich nicht in einer Erhöhung der Artikulationsgeschwindigkeit nieder. Vielmehr bestätigen die Ergebnisse die wichtige Rolle der Pausen, wenngleich sie der Erwartung widersprechen, dass eine schnellere Sprechweise durch weniger Pausen markiert ist: Es treten mehr Pausen auf als in den als langsamer wahrgenommenen Abschnitten. Durch Interaktion mit den Faktoren Atmung und größere Tonhöhe entsteht der perzeptive Eindruck einer höheren Sprechgeschwindigkeit.

Die Sprachproduktionsexperimente, die in Kapitel 6 beschrieben werden, untersuchen die Tempowechselstrategien mehrerer Muttersprachler beim Vorlesen deutscher Texte. Die Ergebnisse zeigen viele idiosynkratische Unterschiede bei Pausierung, Artikulationsgeschwindigkeit, Segmentreduzierung, Phrasierung und Intonation.

Die Ergebnisse der Perzeptionstests mit tempo-skalierter synthetischer Sprache in Kapitel 7 können zur Verbesserung von Sprachsynthesesystemen genutzt werden. Die getesteten Modelle kontrollieren das Sprechtempo nur auf den Ebenen Pausierung und Phrasierung, mit vorausgesagter Position und Dauer von Pausen und phrasenfinal gedehnter Silben für verschiedene Sprechgeschwindigkeiten. Eine Sprachsynthese, die einige der nicht-linearen Aspekte berücksichtigt, die in Kapitel 3 vorgestellt wurden, wird gegenüber einer lediglich linear modifizierten Sprachsynthese bevorzugt, besonders bei niedriger Sprechgeschwindigkeit.

Acknowledgments

My initial idea of writing a PhD thesis was similar to the speaking style of news announcers which is marked by just few and short pauses, a clear and concise style, a high load of information, and a quick and fluent production.

During the process of writing this thesis, which took by far longer than expected, I made some observations which were more similar to a style of speech in which I have to formulate spontaneously on topics I have to think about. This resulted in a considerable amount of pausing time due to planning errors, and to a very dynamic production tempo (with too many phases of ritardando). But unlike speech where we can find a very strong effect of phrase-final lengthening, my writing was characterised by a thesis-final shortening. (It is of course just speculation whether this might be due to some higher-level constraints.)

I would like to thank all the people who discussed with me one aspect of the thesis or another in very different ways. First of all, I would like to say "Danke" to Bill Barry, my supervisor for many hours (and many pages) of discussion. Then I would like to thank all the people from the Institute of Phonetics at Saarland University who served as discussants or, additionally, as co-authors for various articles: Martine Grice, Jacques Koreman, Caren Brinckmann, Marc Schröder, Stefan Baumann, Bettina Braun, Attilio Erriquez, and Thomas Blug.

A special thank you goes to all those people who gave me the opportunity to present my work in several departmental colloquia in Edinburgh, Vienna, Bonn, Stuttgart, and the "obligatory" one in Saarbrücken. Thus, thanks to Hannes Pirker, Maria Wolters, and Michaela Atterer. A big [θæːːŋks] goes to Alice Turk from the University of Edinburgh who was a permanent motivator for my work during my stays in Edinburgh.

Last but not least, I am very grateful to Anja, Mattis and Finn who are probably glad that this piece of work has now found its end.

Saarbrücken, April 2003

Contents

Chapter 1

Introduction

1.1. Motivations - why study speech tempo?

Spoken language unfolds in time. Speaking - as every other form of motion - takes place by continuously moving parts of the body over time. Consequently there is always a given tempo of speech. Speech tempo is a characteristic of spoken language of which we can easily be made aware: speakers are able to change their rate of speech if they deliberately intend to do so. Similarly, on the speech perception side, listeners have an idea whether a given portion of speech was fast or slow relative to an expected normal tempo. But usually changing the rate of speech happens unconsciously, and there are a lot of instances where these changes can be observed. The dynamic nature of speech timing is one reason why we find so much variability in speech data in general.

Although tempo is a prosodic property, tempo is not a genuine linguistic property. Linguistic structures require units which can be described as invariant and distinctive. There is no *direct* linguistic property or contrast that can be attached to speech tempo. Tempo neither bears any meaning nor does it differentiate any meaning by itself. The sentence "John loves Mary." spoken slowly means exactly the same as the same sentence spoken a bit faster.

However, tempo can have a strong effect on the *realisation* of linguistic structures. The following four examples illustrate this effect.

- In German the difference between the two phonemes /a:/ and /a/ lies in its quantity, which is primarily based on vowel duration. A fast spoken /a:/, as in German "Staat" (Engl. 'state'), can show - everything else being equal - a duration which is not significantly different from a short /a/, as in German "Stadt" (Engl. 'town'), spoken at a slower pace.

- Pauses have a very important function for chunking information in speech. Prosodic phrase boundaries are very often marked by an acoustic pause. In fast speech, however, many of the pauses which might be observable at normal speed are temporally reduced or completely omitted. The consequence for the prosodic structure is that some prosodic phrase boundaries are realised differently or simply disappear. Compare the following sentence (taken from the German translation of "The North Wind and the Sun"). The indicated pauses can occur in a normally speeded version and there may be no pauses in a fast version: *Einst* [pause] *stritten sich Nordwind und Sonne,* [pause] *wer von ihnen* [pause] *der Stärkere wäre.*

- Speech rate can have a strong impact on the encoded sound and syllable structure. In the German sentence "Am Himmel ziehen die Wolken" (Engl. literally: 'In the sky move the clouds.') the underlying phonemic structure of the trisyllabic word sequence "ziehen die" would be /ts iː - ə n - d iː/. One possible fast realisation would be a disyllabic [tsini] where [n] changed its syllable position from coda to onset and the number of sounds and the number of syllables have been reduced.

- Speaking faster can also mean articulating the sound sequence faster. Three possible mechanisms in the above mentioned examples "Stadt/Staat" can be illustrated in the /tat/-/taːt/ sequences: 1) the lowering and the raising of the tongue can show a higher velocity; 2); the tongue can rest for a shorter period in the lowered target position for [a]; 3) the tongue does not reach this extreme target position.

The examples show that tempo affects many phonological and phonetic levels, prosodic as well as sound segmental properties. One aim of this study is to give an overview of all these levels. Most studies deal with only a small detail. We see it as essential for modelling speech tempo to consider all levels.

Changes in speaking rate happen all the time, all day long. There are numerous situations, conditions and circumstances in which these changes take place. Many disciplines dealing with spoken communication, other than phonetics and phonology, could benefit from a speech rate model: foreign language learning, language development studies, speech therapy, conversational analysis, psycholinguistics, social psychology, forensic phonetics, and last but not least speech technology.

An explicit aim of this dissertation is to develop a model for tempo control in speech synthesis. Listening to synthetic speech can be highly dependent on personal preference. A novice in this field or elderly, perhaps hard-of-hearing people might like it slower than a frequent synthesis user or some blind people who may desire a tempo faster than the fastest human speech. Users can determine the desired speed. In many current text-to-speech synthesis systems it is already possible to grade the speed without altering the pitch. However, this temporal adapation is achieved in a linear way, whereas the change of speech rate in *natural* speech can be characterised as non-linear. It therefore seems worthwhile investigating whether the effort of doing it in a non-linear way can substantially improve the acceptability of fast as well as slow synthetic speech.

1.2. Aims and structure of the thesis

The thesis is divided into two parts: first a theoretical part, and second an empirical part to illuminate some of the theoretical problems.

The first section of the theoretical part deals with the question *why* and *when* speakers differ in their speech tempo. In the past decades various sources have been identified which can be used to account for tempo variation. These sources range from linguistic ones such as text type and information structure, through paralinguistic ones such as emotion and stress, to extra-linguistic sources such as age and speech motor disorders.

Because tempo is manifested in the realised sound structure of a language, the phonetic aspects as well as the phonological aspects deserve a consideration of their own. Chapter 3 gives an overview of the phonetic and phonological details of when tempo changes occur. These include "higher level" phenomena such as the re-organisation of the prosodic phrase structure, as well as "lower level" phenomena such as the velocity of articulatory gestures. The considerations in this chapter will show that a change in tempo occurs at all levels in a non-linear rather than a linear way.

The problem of measuring speech tempo is addressed in chapter 4. As mentioned above, there are methodological problems in how to quantify and categorise speech rate. This complex issue encompasses subjective and objective tempo, local and global changes in articulation rate, changing tempo *between* different utterances, but also changing tempo *within* an utterance.

After the theoretical considerations of the first part, the analysis of real-world data (chapter 5) and the performance of an original production experiment (chapter 6) will be described. While the database analysis investigates tempo metrics, articulation rate characteristics of read and spontaneous speech, and segmental changes, the experiment focuses on the effect of tempo on the realisation of prosodic phrase boundaries.

The findings from both parts, the theoretical and the empirical, are used to build a tempo model for a speech synthesis system. The implications of the findings for such a task are presented in chapter 7, where a simple model for implementation is proposed. This model serves as a tool to perform perception experiments with tempo-scaled synthetic speech. The tests compare synthetic speech with standard *linear* time-scale modification and *non-linear* human speech-like tempo adaptation. The goal is to achieve a higher than usual acceptance of synthetic speech for different user groups and applications.

With these experiments it is shown that it is possible to alter the global tempo in a satisfactory way for text-to-speech-synthesis. This is particularly true for very slow speech which can be benificial for many applications, e.g. synthetic speech for those who are not familiar with this mode of speech (i.e. most potential users). The findings on very slow synthetic speech can also be transferred to natural speech that needs to be slowed down, which can be useful e.g. in language learning applications. The results also allow some interpretation of how fast or slow the default tempo of synthetic speech should be scaled. Moreover, more insight is gained about the impact of phrase boundaries and their realisations in fast synthetic speech.

Chapter 2

Sources of Tempo Variation

Introduction

Why and where do speakers change tempo? There are an infinite number of reasons why, and situations where we vary our speaking rate or can observe, more or less consciously, different tempos in other speakers.

In order to structure this discussion, the distinctions *extralinguistic - paralinguistic - linguistic* are chosen. Although there are no sharp boundaries between these three terms they help to illuminate different levels of the problem. Some aspects can be attributed to the individuality of a given speaker (*extralinguistic*). Other aspects can be explained solely by the situation and/or the inner state of the person who is speaking (*paralinguistic*). Third, other aspects can be unambiguously attached to how spoken language is performed in interaction and in order to convey verbal information (*linguistic*). The current chapter seeks to show the diversity of factors with observations from production as well as from the perceptual perspective. The subsequent chapter 3 is devoted to the encoding and the execution of the *phonetic plan*, i.e. how the phonological encoding is structured and how the resulting phonetic plan is realised in articulatory actions leading to actual speech.

Even though the present chapter cannot claim to represent a *complete* list of sources of tempo variation, it gives an idea of the range of sources of variability and it shows that attempts to explain variance in the linguistic-phonetic expression are poor if paralinguistic and extra-linguistic factors are not included. The "neutral" situation in recording laboratories simulating communication is not the same as communication in the real world. Lab-speech experiments can help to explain how speech communication may work, but only to a limited extent with respect to real speech.

2.1. Preliminary explanations

It is necessary to define some central terms and concepts at the very beginning to avoid a terminological confusion. These terms occur throughout the thesis and will be explained in detail with each chapter.

There are various terms used to denote the tempo of speaking such as *speech rate*, *rate-of-speech (ROS)*, *rate of speech production*, *speed of talking*, *talking rate*, *reading rate* (for read speech), *speaking tempo* or simply *tempo*. These terms are used here as synonyms and most of the time in this thesis the expression *tempo* is used.

The pause plays a central role when dealing with tempo and it frequently makes a big difference if pauses are taken into consideration or not. The usual distinction is that tempo can either be defined as articulation rate or as speaking rate. *Articulation rate* as a net rate refers to phases of articulation ***ex**cluding pauses*. *Speaking rate* as a gross rate refers to the entire speaking phase ***in**cluding pauses*.

Many expressions have been "invented" for *articulation phases* demarcated by two pauses: "chunk" (Fougeron & Jun, 1998), "run of pause-free speech" (Miller, Grosjean & Lomanto, 1984), "interpause stretch" (Dankovičová, 1997); "run" (Crystal & House, 1990), "articulatory run" (Tsao & Weismer, 1997), "interpausal speech run" (Koopmans-van Beinum & van Donzel, 1996), "phrase" (Fant, Kruckenberg, & Nord, 1992), "utterance" (Butcher, 1981), "T-phrase" (Eefting, 1991) and "speech chain" (Bartkova, 1991). The term used here is *inter-pause stretch*, because it seems the most informative.

Usually we want to categorise the tempo, i.e. whether a speech sample is considered fast, or slow, or slower than normal, or whatever the intended relational purpose might be. It must be emphasised that this categorisation strongly depends on the phonetic perspectives of speech production, speech acoustics, or speech perception. Although a speaker intends to speak "fast", the resulting speech will not necessarily be categorised as "fast" on the basis of a physical measurement (e.g. in syllables per second). Furthermore, this stretch of speech will not necessarily lead to the auditory impression of "fast" for listeners. Thus, tempo categories are only comparable and interchangeable under certain conditions. And it is with the same caution, that production studies have to be compared with perception studies in the review chapters.

The last point to be made in advance is that we can distinguish between the *subjective* tempo in production and perception, and a measured *objective* tempo. It is without doubt useful to have a standardised metric to quantify tempo. But bear in mind that there *is no* such standard, and that there *cannot* be such a standard as will be explained in chapter 4. The reasons are, among others, differences in the definition of the linguistic units, usage of pause, material and structure of language. Although there is no exact standard, similarly used measurements express similar things *more or less*. With this sensitivity in mind, the reader is referred to table 2.1. It contains a list of various studies dealing with different material in several languages where articulation and speaking rates were measured in syllables per second as the most popular tempo metric.

Table 2.1: Survey of studies investigating speaking rates in different languages and accents. Speaking rate (sr = including pauses) and articulation rate (ar = excluding pauses) is indicated in syllables per second for two different speaking modes: reading and spontaneous. For articulation rate the percentage of pause time of the whole speaking time is given. Numbers with an asterisk (*) are re-calculated as syll/sec for the number of speakers for the material on the basis of the data in the literature. Numbers with a double asterisk (**) indicate re-calculations either from the pause quotients or from the relationship between speaking rates and articulation rates in the original studies. For studies with various rates, only the data for the "normal" or "medium" rate are used here.

		subjects		**read**			**spontaneous**		
study	**text type**	**no.**	**language**	**sr**	**ar**	**pau**	**sr**	**ar**	**pau**
Dauer (1983)	prose	1	Engl. (UK)	5.9					
		1	English (US)	5.0					
Iivonen et al. (1995)	news	8	Engl. (UK)	5.3	5.4				
		9	English (US)	5.2	5.4				
Tsao & Weismer (1997)	prose	100	English (US)	4.39					
Hewlett & Rendall (1998)	neutral text & conversation	12	Engl.(Orkney)	4.50	5.49	18%	4.53	6.02	29%
		12	Engl. (Edinb.)	4.55	5.43	16%	4.34	5.52	24%
Tauroza & Allison (1990)	news / radio announcements		English (UK)	4.16					
	conversat.		English (UK)				4.39		
	interviews		English (UK)				4.18		
	lectures		English (UK)				3.24		
Grosjean & Deschamps (1975)	radio interviews	30	English					5.17	

Grosjean & Deschamps (1975)	radio interviews	30	French					5.29	
Fletcher (1987)	transribed interviews	6	French	4.49	5.53	19%			
Malécot et al. (1972)	conversations	60	French (Paris)				5.73		
Fougeron & Jun (1998)	prose	3x3	French (Paris)	4.32	5.65				
Slembek (1993)	news	2	French (FR)	5.32		11%			
	news	2	French (CH)	5.04		7%			
Slembek (1993)	news	2	German (CH)	4.24		15%			
	news	2	German (DE)	4.84		17%			
Meinhold (1967)	prose	14	German	5.40*		29%*			
	news	17	German	5.73*		18%*			
	poetry	8	German	3.63*		30%*			
Greisbach (1992)	texts	8	German	4.81*					
Iivonen et al. (1995)	news	5	German	5.80	5.90				
Künzel (1997)	news magazines	10	German	5.12	6.04	15%			
	monologues	10	German				4.28	5.89	29%
	monologues	10	German				4.18	5.83	29%
Trouvain (1999)	transcr. news	3	German	4.72	5.30				
	prose	3	German	4.67	5.52				
Wiese (1983)	cartoon retelling		German					4.57	
Strangert (1993)	transcr. news	1	Swedish	5.77	7.98	28%	4.17	7.80	46%
	news	1	Swedish	5.83	6.4	9%			
Koopmans-V.B. & Van Donzel (1996)	retold story	8	Dutch				5.85	3.79	35%
Iivonen et al. (1995)	news	18	Finnish	6.3	6.5				
Dauer (1983)	prose	3	Spanish	7.10					
Dauer (1983)	prose	3	Greek	7.47					
Dauer (1983)	prose	2	Italian	7.30					

2.2. Extra-linguistic sources of tempo variation

Habitual speech rate

Individual speakers can differ substantially in their typical speech rates. This can easily be shown by looking at a database containing the same text read by many different speakers. In the German "Kiel Corpus of Read Speech" (IPDS, 1994) 16 speakers read the IPA standard text "Nordwind und Sonne". The descriptive statistics of the speaking rate in table 2.2 show a considerable variation across the readers (half a syllable per second standard deviation) with the slowest reader more than two syllables per second slower than the fastest reader. Similar results were reported for read German speech (news magazine) by Künzel (1997), and have also been observed for other languages such as English (Goldman Eisler, 1968).

Tsao & Weismer (1997) investigated the habitual and the maximium articulation rates in extremely slow and extremely fast speakers (+/- 1 standard deviation from mean in a reading task). The results reveal that the slow readers at their maximum articulation rate could just articulate as fast as the fast readers at their habitual articulation rate. Since the speeding up magnitudes for both groups behave similarly, there are indications that the maximum rate for an individual can be predicted from his/her habitual rate.

Table 2.2. Mean speaking rate and mean articulation rate, with its standard deviations (sd), maximal (max) and minimal (min) values measured in underlying syllables per second of two German data collections: the "Nordwind und Sonne" recordings in the Kiel Corpus (excluding the pause between the two paragraphs) and readings of news magazine articles (Künzel, 1997).

	no.	**speaking rate**		**articulation rate**	
corpus	**speakers**	**mean (sd)**	**min-max**	**mean (sd)**	**min-max**
Kiel	16	4.27 (.51)	3.05 - 5.18	*to be added*	*to be adde*
Künzel	10	5.12 (.42)	4.52 - 5.82	6.04 (.50)	5.31 - 6.90

Age

Haselager et al. (1991) investigated articulation rate skills of Dutch-speaking boys and girls in four age groups between 5 and 11 years. They used a diadochokinetic task (repeating the same syllable as fast as possible) as well as spontaneous speech elicited in interviews. For both speaking modes the syllabic rate varied with the age group: the younger the group the slower the articulation rate. The same effect has been observed by Walker et al. (1992) for English speaking Canadian preschool children. Their results from a spontaneous speech and a speech imitation task show significant differences in the articulation rate between children at age 3 and at age 5. The age effect is also reported in a British study by Whiteside & Hodgson (2000). In a picture-naming experiment with children aged 6, 8, 10 years and an adult control group they found significant differences relating to articulation rate as a function of children's age. There is evidence that the increase in articulation rate during maturation proceeds in a non-linear way as pointed out by Hall, Amir & Yairi (1999).

An age effect seems to apply not only to the developmental phase of speakers. Malécot et al. (1972) report in their study on French spontaneous speech that older adults speak slower than younger adults: syllabic rate drops progressively by about half a syllable per second overall from 5.95 syll/sec to 5.52 syll/sec, for young (starting at 20) to older speakers (up to 69 years).

These findings are backed by an American English study by Sommers, Humes & Pisoni (1994) who investigated the effects of increased speaking rate and greater speaking-rate variability on spoken-word recognition in older and younger listeners. For younger subjects, neither increased speaking rate nor greater rate variability produced significant changes in perceptual identification scores. Older listeners, in contrast, exhibited significantly poorer identification scores for fast, compared to medium or slow speaking rates. In addition, trial-to-trial variations in speaking rate produced a significant decrease in identification scores for elderly subjects listening to fast-rate item.

Gender

In Whiteside & Hodgson (2000) a significant difference for gender was found (except for the six years old): females articulate slower than males. This confirms the findings in Whiteside (1996) for read sentence material. American variants of English also exhibit gender differences for tempo, as has been shown by Byrd (1992) for the TIMIT database. Considering only vowel duration, women exhibit longer values than men (Simpson, 1998, for German; Simpson, 2001, for American English). However for French, Malécot et al. (1972) found *no* significant differences between the sexes in terms of syllable rate, but there *were* differences in terms of utterance length. The findings of an English study (Deese, 1984) contradict the aforementioned studies. Here, women spoke faster (5.82 syll/sec) than men (5.48 syll/sec) but this difference was not statistically validated. Thus, whether gender influences tempo and articulation rate remains an open question.

Speech and hearing impairments

Apparently, most speech and hearing impairments also have an effect on speech tempo. In the area of motor speech disorders, tempo may be slowed down or shows great variations as has been described for the continuum of developmental dysarthria to developmental verbal dyspraxia (MorganBarry, 1995). These articulation disorders with neurological origin also show other prosodic symptoms such as erratic pausing, arhythmic structures, unfinished intonation units and disfluencies. The marked feature of various forms of stuttering (or stammering) is the abnormal number of disfluencies such as arhythmic pausing, blocking of articulatory air-flow, prolongations of sounds, restarts and repetitions of sounds, syllables and words. These disfluencies makes the overall tempo rather slow. Disfluencies such as repetitions, repairs, and filled pauses usually occur in spontaneous speech of non-stutterers. The moderate number and the type of disfluencies seem to count as criteria to classify them as "fluent". It is interesting to see here, that in fluent phases, persistent stuttering pre-school children show no significant difference with respect to articulation rate to a non-stuttering control group (Hall, Amir & Yairi, 1999). With regard to stuttering it is interesting to note that parental high speaking rate results in a degraded fluency in the children. On the other hand parental slowing often leads to an improved fluency, maybe because parents also show changes in behaviour in addition to slowing down speech rate. This can reflect more empathy with their children, as hypothesised by Guitar & Marchinkoski (2001).

Special forms of speech and hearing impairments are those caused by the use of drugs. As an example, alcoholic intoxication has an effect on speech rate as reported by Künzel et al. (1992) in a reading task: speakers under alcohol articulated more slowly and made more and longer pauses, including a greater number of hesitation pauses.

In a study investigating the intelligibility of sentences spoken in a conversational style and at a fast rate, and those spoken in a clear style and at a slow rate, Uchanski et al. (1996) found that listeners with a hearing loss show significantly better results for the clear speech condition (87% vs. 72% for conversational). Listeners with normal hearing show only a slight increase (98% vs. 92%) for the clear condition. This shows that speaking slower and more clearly to hard-of-hearing persons makes understanding easier for them.

Auditory conditions

In the same study by Uchanski et al. (1996) the same conditions (conversational/fast vs. clear/slow) were also tested with normal hearing listeners under noise conditions simulating a hearing loss. Under these adverse auditory condition the intelligibility effect of clear and slow speech is even more evident than for the hearing loss condition: 60% for clear vs. 44% for conversational style. These results give rise to speculations that under anything less than ideal listening situation clear and slow speech contribute significantly more to intelligibility than "natural" conversational and fast speech. That means that in unfavorable listening conditions – and here synthetic speech must be included – listeners would prefer slower speech.

Cultural and geographical background

Similar to dialectal and sociolectal phonetic differences one could imagine that differences in tempo occur between speakers of the same language but with a different cultural or geographical background. But Slembek (1993) found no tempo differences between broadcast news readings from stations of the different nations. French speaking news readers in France do not differ in syllabic rate from those in Switzerland, and

Swiss German radio news readers speak as fast as their colleagues in Germany. In a similar study Iivonen et al. (1995) found no tempo differences between American and British English radio news. However, Byrd (1992) reported in her analysis of the TIMIT database some statistically significant differences in tempo and pausing between speakers from different dialect regions in the United States.

Hewlett & Rendall (1998) investigated the question of whether lifestyle, in the form of urban vs. rural living, influences speech rate. Comparing Scottish English speakers from Edinburgh with those from the Hebrides, they rejected the claim that is sometimes made that city residents speak faster than those living in the countryside.

Language proficiency

Different languages may possibly differ in terms of rate of speech production units as can be seen in the studies for different languages in table 2.1. But there are certainly differences in terms of how speech rate is perceived across languages. Speech of the native language/s or those which are mastered with a higher level of proficiency is felt to be less fast than the speech of those languages with a lower level or no proficiency. Abercrombie (1967: 96) puts it as follows:

> "Everyone who starts learning a foreign language, incidentally, has the impression that its native speakers use an exceptionally rapid tempo."

In a study comparing spontaneous monologues of American English and Japanese speaking students Osser & Peng (1964) found no significant differences in terms of the phoneme production rate of the two groups. They explain the impression that an unknown language sounds faster than normal, i.e. than one's own language, with phonological differences such as different patterns of syllabic complexity: Japanese people tend to perceive unknown consonant clusters of English as syllables and therefore the number of perceived English syllables increases which results in a higher perceived syllabic rate. In contrast, English speaking people tend to interpret the many vowels and the many syllables but relatively few phonemes in Japanese as a higher syllabic rate compared to English.

Speaking a foreign language is usually linked with a higher cognitive activity. The problems include incompletely developed syntactic and morphological knowledge, slower lexical access, and articulatory difficulties in less well established segmental and prosodic patterns. The process of planning and executing speech is slowed down and this is normally mirrored by reduced fluency in the non-native talker. Evidence for this claim is given in e.g. Pürschel (1975) who examined German students of English. The students were first asked to read an English text and afterwards the same text in a German translation. There were more pauses in the reading of the foreign than in the native language text leading to a slowed down tempo. Wiese (1983) also investigated the temporal behaviour of German learners of English and native speakers of English with a cartoon retelling task. He found significant differences in terms of mean pause duration (slowing down the overall speaking rate) as well as in terms of mean articulation rate between the two groups. It seems that speaking in a foreign language means articulating more slowly and making more and also longer pauses than usual, i.e. than in the native language.

But *comprehension* of a foreign language is also affected by speech rate. As an example, Griffiths (1990) tested the comprehension of Japanese teachers listening to English texts delivered at three different rates. The moderately fast readings resulted in a significantly lower comprehension score than the slower readings. Anybody who has ever tried to learn a new language can probably confirm these findings. Therefore it is not surprising that learners sometimes explicitly ask for slowed down speech. This wish for special listening conditions is fulfilled e.g. by the Deutsche Welle, the German broadcast station abroad, that offers an additional version of slowed down broadcast news for foreigners (Deutsche Welle URL) where the news is spoken with a speaking rate of about 3 syllables per second compared to the usual speaking rate span of between 4.5 and 5.5 syllables per second for German broadcast news.

Apart from the fact that "normal" native speakers' tempo appears as "fast" for language learners (L2 judge L1 speech tempo), and that language learners produce the foreign language at a rather "slow" rate compared to their native language or to the tempo of native speakers (comparing L1 and L2 speech tempo), the tempo of language learners has an effect on the proficiency judgements by native-speaking listeners (L1 judge L2 speech tempo), in addition to segmental and prosodic errors. Munro & Derwing (2001) asked English native speakers to judge L2 speech (L1: Mandarin) on accentedness and comprehensibility. The articulation rate of the read sentences was manipulated with a speech compression-expansion editor by 10 % so that each sentence was presented in a slightly slower version, the actual version, and a slightly faster version. As expected the native-speaking listeners evaluate slightly faster versions as less accented and more comprehensible than the natural or even the slowed down tempo of foreigners' talk. However, talking too fast resulted in a downgrading.

Moreover, the study shows that tempo makes a small but significant contribution to both accent and comprehensibility ratings independent of segmental errors.

2.3. Paralinguistic sources of tempo variation

Emotions

Emotions can have a strong effect on speech tempo. Expressive speech is marked by global prosodic parameters such as F0 variation, voice quality and speech tempo. Several studies show evidence of general tendencies for some given emotional categories such as anger, joy or sadness (for an overview see van Bezooyen, 1984; Murray & Arnott, 1993; Banse & Scherer, 1996; Burkhardt, 2001). Although language, exploration methods, purpose, and speech material differ in these studies, the reported patterns look alike. Anger, rage, fear, but also happiness is generally marked by an increased tempo whereas boredom, sadness, sorrow, grief, and disgust is characterised by a slowed down tempo.

An alternative way of describing emotions is along three dimensions rather than with labels for "full-blown" emotions. The three dichotomies are dominant-subordinated (dominance), positive-negative (valence), and active-passive (activity). Speech tempo seems to correlate strongly with the activity dimension (Schröder, in prep.; Kehrein, 2002), i.e. the more active the speaker the faster s/he speaks, and the more passive the speaker the slower s/he speaks (Scherer, 1974; Apple, Streeter & Krauss, 1979).

Stress

Similar to emotional stress, cognitive stress can also result in high arousal. Lively et al. (1993) examined the effects of cognitive workload on speech production. Workload was manipulated by having speakers perform a compensatory visual tracking task while speaking short carrier sentences. In the workload condition, speakers produced utterances with increased amplitude and amplitude variability, decreased spectral tilt, increased F0 variability, and increased speaking rate.

Barber et al. (1996) also showed that a high cognitive workload leads to a faster tempo. In the two reported experiments (time-stress and dual-task performance) most of the English subjects doubled their speech rate. However, Dankovičová & Nolan (1999) were not able to show consistent speeding up when cognitive stress was present.

Competence and benevolence

Categories such as competence and benevolence have been related solely to perceptual impressions. Smith et al. (1975) synthesised utterances and manipulated them with respect to tempo. Then, subjects were asked to evaluate the stimuli on a list with attributes which could be summarised under the headings *competence* on the one hand and *benevolence* (cooperativeness, friendliness, politeness) on the other. The results show that the highest benevolence score correlates with the "normal" speaking rate, with a linear decrease to both sides, i.e. the more the speed increases or decreases the more the benevolence scores decrease. At the same time, the results show a tendency for faster rates to be given higher competence scores. In a similar study Apple, Streeter & Krauss (1979) summarise their results:

> "... slow-talking men are judged to be less truthful, fluent, emphatic, serious, and persuasive, and more passive, although they are also seen as more potent."

Ofuka et al. (2000) found in their comparisons of polite and casual Japanese utterances that speech tempo was used consistently by all six speakers: comparing polite with casual speech it appeared that polite speech was in general slower than the casual form. Apart from tempo and the F0 movement on the final vowel, the duration of the final vowel affected the politeness rating of Japanese judges. These ratings also revealed that the function of politeness related to speech rate was that of an inverted U-shape, i.e. the fastest and slowest versions get the lowest scores.

Communication partner

Everybody changes their speech tempo (among other prosodic parameters) more or less consciously when talking to people from whom one might assume less established information processing abilities, including infants, non-native speakers, elderly people, or persons with hearing difficulties. The study of Van de Weijer (1997) examining the speech of the mother, the father and the baby-sitter of a Dutch child between age 6 and 9 months confirms the general findings of infant-directed speech. So-called "motherese" compared to adult-directed speech features a higher F0 average and a larger F0 range along with a slower articulation rate as well as shorter utterance length (in syllables) between pauses. In an investigation of speech addressed to elderly people (American English) Kemper (1994) comes to similar results: "elderspeak" is marked by a slower articulation rate, shorter utterances, and more pauses.

Frequently speakers unconsciously adapt their speaking rate to their dialogue partner's or to the assumed speech rate ability of their partner. The accommodation theory (e.g. Street & Giles, 1982) distinguishes full and partial convergence, maintenance, and divergence, whereby speech convergence represents a move towards social integration. The different categories are examined with various non-content speech behaviour parameters such as response latency, utterance duration and speech rate of two interactants. While everyday interlocution normally reaches a low level of awareness, the speech behaviour "divergence" is brought into consciousness and perceived negatively (Street, 1982).

Adapting speech rate to the communication partner has also been observed for children as early as 3 years old. Guitar & Marchinkoski (2001) observed in their study with mother-child dyads that in five of six cases the children significantly slowed down their tempo when their mothers spoke slower (on average by 51 %). Based on the positive correlations they discovered, the authors hypothesise that children also speed up when their parents speak faster.

2.4. Language-relevant sources of variation

Speech planning

In her summary of investigations about cognitive activities during spontaneous speech Goldman-Eisler (1968: 31) claims that spontaneous and read speech do *not* differ basically in articulation rate:

> "Variations in the overall speed of talking were found to be variations in the amount of pausing. What is experienced as increase of speed in talking proved to be variation in amount of pausing. The rate of articulation based on vocal activity exclusively, on the other hand, was shown to be relatively invariant."

According to her the only temporal difference lies in pausing, namely that spontaneous speech compared to read speech is characterised by *more* pauses, *longer* pauses and the presence of *filled* pauses. This different pausing strategy can be expressed as an increase of pause time as the portion of the total speaking time, which also leads to a slower speaking rate in syllables per second. The figures in table 2.1 (p. 7-8) confirm the different pause time ratios for various languages and various studies: between 5% and 20 % for read speech, and 30% and 46 % for spontaneous speech.

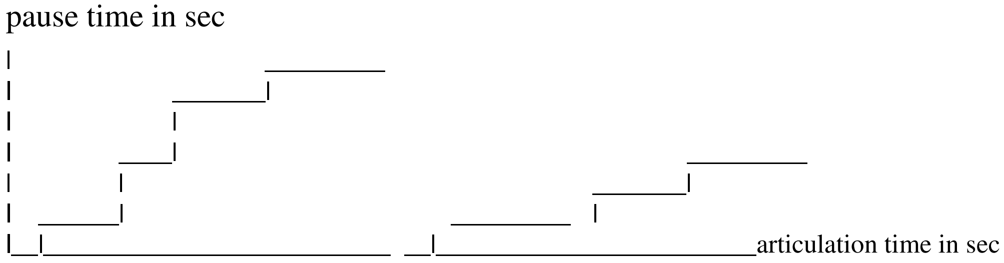
Read speech can be seen as a speech mode where the ideas to be expressed are completely prepared and formulated before speech production starts. In contrast, in many forms of spontaneous speech, the formulation process takes place "on-line" resulting in more pauses leading to a slower speaking rate. But the pauses seem to be unequally distributed over the utterance.

Levelt (1989:126) summarises the relevant studies:

> "There is some evidence that in longer monologues speakers slowly alternate between phases in which they spend much attention on information retrieval and inference (i.e. macroplanning) and phases in which they concentrate on finalizing messages for expression (i.e., on microplanning)."

The result is alternation of fluent phases (more articulation than pausing) and hesitant phases (more pausing than articulation) which reflects cognitive activity on the level of articulatory execution. The idealised scheme in figure 2.1 serves to illustrate these patterns of fluency and dysfluency in spontaneous and read speech styles.

Figure 2.1. Time course of spontaneous (left) and read speech (right) in articulating phases (x-axis) and pausing phases (y-axis). The degree of flatness mirrors the degree of fluency.



Let us ignore the pauses for now and consider only the articulation phases. Related to the quotation of Goldman-Eisler above, there seems no substantial differences in the *global* articulation rate between speakers or speaking styles. An articulation rate (e.g. in syllables per second) averaged over *all* articulation phases in the spontaneous mode of a speaker would not much differ from the average rate of *all* articulation phases in a read mode. However, the average rate does not tell anything about the variance *within* an articulation phase. In an investigation of interview responses by French speakers Miller et al. (1984) "discovered" this dynamic feature in the rate of articulation. They talk about "macro variables" which account for global tempo variation, in contrast to "micro variables" responsible for this local within-phrase tempo variation. As examples for these micro variables they suggest lexical access difficulties, syntactic construction delays and semantic planning problems.

All these "micro variables" point to some temporal delays during articulation typical for unscripted and more or less unplanned spontaneous speech, whereas read speech would lack these delays. In their Japanese study Hirose & Kawanami (2002) used dialogues simulated by actors which were also recorded as read isolated sentences. Although the above mentioned micro variables are not operating here, the dialogue speech samples show more dynamics than their read counterparts. These dynamics are expressed by the acceleration scheme between the two modes: compared to the readings, prosodic phrases in the dialogue turns were faster in the middle and slower at the end.

Types of spoken and written texts

Fónagy & Magdics (1960) measured tempo differences in text styles in their Hungarian study where sports news showed a higher tempo than read poetry. This is in line with Meinhold (1967) who found considerable tempo differences between readings of different text types, e.g. prose vs. poetry (see also table 2.1). Although the rate values for readings of prose and those of the news are close together, the results show only half as many pauses for the news (not in table 2.1) than for the prose text type whereas the mean pause duration is similar for both text modes. The data in table 2.1 give rise to the assumption that news reading has a faster speaking rate than other text types.

Abe (1997) investigated some prosodic characteristics of readings of different Japanese text types, such as a novel, advertisement phrases and paragraphs from an encyclopaedia. He found a much higher effect of pause and sentence boundary on the vowel duration of the preceding syllable for the novel compared to the other styles. The novel also showed the slowest speech rate, especially when pauses are included, whereas the encyclopaedia style is faster followed by the advertisement style which was the fastest.

In a study investigating recorded samples of ten different text types (such as prose, children's story, recipes, technical literature, dictionary) in three languages (Dutch, English, French) Fackrell et al. (2000) found that, in general, news readings are articulated faster than average and that dictionary entries, weather reports, and children's stories are read slower than normal for the three languages. However, the same tendency does not hold for each text type and language, e.g. advertising in English was faster than the average rate in contrast to a slower rate for the same text type in Dutch and French.

Similar to speaking styles based on texts, the styles of unscripted speech show variations of tempo. Kowal, Wiese & O'Connell (1983) performed a survey of various studies investigating spontaneous speech types such as descriptions of cartoons, pictures or films as well as speech in interviews (broadcast, television, medical patient). After a thorough recasting of all data available in five languages (German, English, French, Spanish, Finnish) they compared the monologic category of "storytelling" with the dialogic category of "taking part in interviews". It appeared that for both categories the articulation rate is comparable (5.17 syll/sec for storytelling vs. 5.26 syll/sec for interviews). However, in storytelling, pauses are made more often and are also longer, so that the percentage of pause time to total speaking time is greater (33%

vs. 17%) and the speaking rate is slower relative to interview speech (3.43 syll/sec vs. 4.31 syll/sec).

Tauroza & Allison (1990) showed in their British English data different speaking rates for various speech categories: turns in conversations are fastest (4.39 syll/sec), followed by radio announcements (4.16 syll/sec) and interview speech (4.18 syll/sec), and lectures being the slowest speech type investigated (3.24 syll/sec).

Dialogue management

A common phenomenon in a dialogue is that information is repeated, e.g. after a misunderstanding or to make something more explicit. One prosodic means is a slower articulation of the same word sequence and/or insertion of pauses, which also leads to a slowing down. This communication strategy is visible in an extreme form in man-machine communication: in German Wizard-of-Oz experiments the (human) operator of a dialogue system pretended to fail to understand (Fischer, 1999). Subjects showed a great repertoire of variation during the repetitions, depending mainly on the degree of cooperativeness. This variation in attitude is accompanied by various prosodic changes other than direct slowing down leading to a decreased rate, for example emphatic accentuation, more accents and hyper-articulation. These findings were confirmed in a similar experiment with synthetic speech for Viennese German (Pirker & Loderer, 1999).

In a dialogue, speakers are continuously sending and receiving signals on the status of the information exchanged. Confirmations and disconfirmations in the kind of an echoing response are usually marked by various prosodic means such as pause, duration, intonation contour and pitch range. As an example, Krahmer et al. (2002) found in their Dutch study that disconfirmative utterances were spoken more slowly than their positive counterparts, and that this feature was reliably used by listeners to classify those utterances as a negative response without context.

The study of Wells & Peppé (1996) can serve as an example of how tempo is used for turn organisation in dialogues. They found that in the Ulster variety as well as in the Tyneside variety of English a dialogue turn is delimitated by a markedly slowing down over the last two rhythmic feet (with a foot as a stretch of speech beginning with a stressed syllable). This turn-final lengthening is accompanyied by changes of other prosodic and non-prosodic phonetic properties such as loudness, voice quality, vowel quality and pitch contour.

Koiso, Shimojima & Katagiri (1998) claim for conversational Japanese that changes in tempo by the dialogue partners have a potential for cueing the structure of information collaboratively. In their data, openings of new information were marked by decelerations and the absence of information openings by accelerations.

Information management

Spoken language always features a certain degree of redundancy: it is not always necessary to understand all words to get the message. Greenberg (1999) speculates that

> "it is likely that frequently occurring words tend to be spoken faster and in more reduced fashion because of their inherent predictability."

This is well-known for the so-called function words (e.g. determiners, pronouns, auxiliary verbs, prepositions, interrogatives, conjunctions, degree adverbs) but also applies to frequently used lexical items. Normally, high-frequency words such as numbers are produced faster. In contrast to this, telephone numbers with their very low predictability are often spoken in a very slow way. If you miss a number or the correct order of the numbers you miss the whole message. Telephone numbers show *hardly any* or *no* redundancy. Thus, telephone numbers are optimally nested in a characteristic prosody which also features a slow speaking rate in terms of syllables per second as was shown by Baumann & Trouvain (2001) for German.

Uhmann (1989) found in her German data of everyday conversations that side comments such as parentheses and afterthoughts are marked by fewer pitch accents and a faster articulation whereas emphasised discourse segments are marked by more pitch accents and a slower articulation. This is in agreement with the analysis of Barden (1991) who showed for German dialogue speech that portions containing less central and less important information are spoken at a tempo faster than average, and inversely, that more central and more important portions are spoken at a slower speed than normal. This general notion was fleshed out in more detail in the Dutch study by Eefting (1991) where the well-established thema-rhema structure (or "given" - "new" information) was related to speech tempo. However, the effect of information value was only significant in those cases where additional accentedness was present, which is often, but not always the case in her data. A further link between information structure and tempo has been mentioned by Klatt (1976) by signalling contrastive information and/or emphasis by slowing down.

Tempo variation on the axis of hyper- and hypospeech

The communication situations listed above made it clear that tempo plays a crucial role in spoken communication which is always in a balance between the economic use of speech production and achieving the communicative goal of being understood by the listener. Using Lindblom's (1990) image that the speech production process changes continually on the hyper- and hypoarticulation axis. Related to synthetic speech production, the task is to find or to model an appropriate balance of the hyper- and hypo-continuum. Consider the two general goals of improving the acceptance of synthetic speech: intelligibility and naturalness. Intelligibility is usually increased by modelling clear speech, i.e. considering hyperarticulation. Improving naturalness is usually achieved by mapping features of conversational speech, i.e. by hypo-articulation. Both ends of this axis have their correlates to speech tempo. Thus, to improve the performance of synthetic speech, both ends of this axis must be considered, and this is dependends on the communicative situation in which a human listener is faced with synthetic speech.

Summary and discussion of chapter 2

This chapter presented a discussion of the most important factors underlying speech-tempo differences that have received attention in the literature (see figure 2.1). The amount of attention devoted to each factor varies considerably, and the factors addressed are presumably not the only ones that operate during speech production. The examples illustrate the great range of situations and conditions in which a change in speech tempo can take place. Since speech unfolds in time there can be no speech without speaking tempo. The tempo of speech is always changing, whether we are aware of it or not. This fact is rarely considered in speech analyses or speech applications (e.g. in speech synthesis where usually only *one* tempo is used, and, presumably, expected to fit all speakers, listeners, situations and text styles).

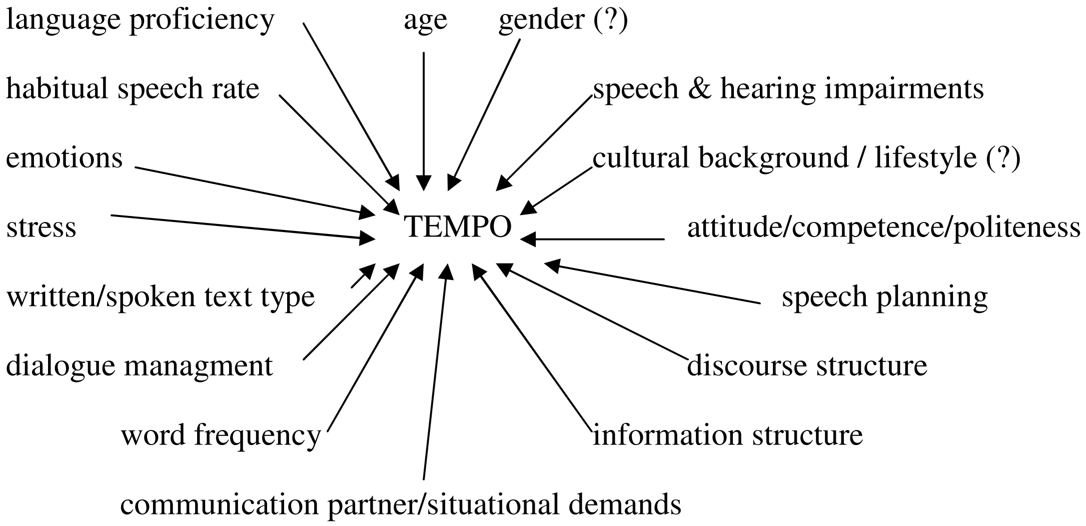


Figure 2.2. Sources of tempo variation.

Chapter 3

Phonetic and Phonological Aspects of Tempo Variation

Introduction

What happens when we talk faster than we normally do? And what happens when we talk slower than we normally do?

There are many phenomena observable in speeded up and slowed down speech, which can be assigned to different levels of speech production. It involves "high" levels of speech production such as the prosodic phrasing as well as "lower" levels such as the velocity with which articulator movements are executed during speech production.

This chapter aims to give a comprehensive view over the levels and the mechanisms on each level which might operate while speaking at tempos different to the "normal" one.

3.1. Phrasing and pausing

As pointed out in the previous chapter, it seems generally accepted since Goldman-Eisler (1968) that changes in tempo are mainly changes in pausing rather than changes in articulation rate. To vary the tempo, speakers change the number of pauses and change the duration of pauses. Consequently, slowing down is characterised by more pauses and longer pauses compared to speech at normal speed, and speeding up features fewer pauses and shorter pauses compared to normal speeded speech.

Indeed, Caspers & Van Heuven (1991) found deletions of phrase boundaries in fast Dutch speech. In addition to boundary deletions, demotions of phrases from major to minor phrases were reported for two of three French subjects in Fougeron & Jun (1998). A perception test with Dutch subjects revealed that sentences with an intonational phrase boundary in the contour were perceived ceteris paribus as slower (Rietveld & Gussenhoven, 1987).

In studies of French material read at different rates (Grosjean, 1979; Fletcher, 1987; Fougeron & Jun, 1998) the following observations have been made: slowing down is characterised by increasing, speeding up by decreasing the number of pauses. The higher the speed, the fewer the pauses. Evidence in (Bartkova, 1991). For German, however, Butcher (1981) observed different patterns for speeding up and slowing down: Increased tempo was characterised by reduced pause duration without a change in the number of pauses; slowing down was marked by a greater number of pauses without a change in the mean pause duration.

No clear picture about the reduction of pause duration arises from the French investigations. Fougeron & Jun (1998) found a reduction for their three speakers, only three out of six speakers applied pause shortening in Fletcher (1987), and Grosjean (1979) reported relatively constant pause durations. This is in strong contrast to Minifie (1963, cited in Lass, 1971) who, prior to Goldman Eisler, sees tempo variation more as a function of the compression and expansion of pause intervals than as a variation in articulatory intervals.

Table 3.1. increase (>), decrease (<), or no difference (=) to normal speeded tempo regarding number (no.) and duration (dur.) of pauses.

	slowing down		**speeding up**	
study	**number**	**duration**	**number**	**duration**
expectation	**>**	**>**	**<**	**<**
Grosjean (1979)	>	=	<	=
Fletcher (1987)	>	?	<	(<)
Fougeron & Jun (1998)			<	<
Butcher (1981)	>	=	=	<
Minifie (1963)	=	>	=	<

A naïve expectation might be that most speakers make maximal use of pausing mechanisms, namely reducing the number and duration of pauses for speeding up, and increasing the number and duration of pauses for slowing down. From this brief analysis of pausing strategies across different studies and different languages, it becomes clear that this is obviously not the case, and that different speakers make use of the various pausing mechanisms differently. Table 3.1 summarises all the main findings from the above mentioned studies. It would appear that there is *not* a tendency to use *all* possible pausing mechanisms to change tempo.

But despite these restrictions on the full use of pauses as a means of tempo variation, it is clear from table 2.1 (in the previous chapter) that the relative amount of pause time can be up to half the entire speaking time. The central role that pausing plays in the regulation of tempo requires a detailed answer to the question: what is a pause?

Definition of pauses

A first distinction must be made between a *perceived pause,* and an *acoustic pause* or a *silent pause*: listeners, when asked to react as they hear a pause, tend to miss some acoustic pauses at unexpected locations. Likewise they interpret pauses at grammatical locations where, actually, there was no acoustic silence but only other phrase boundary markers, such as phrase-final lengthened syllables and intonational markers.

However, an acoustic pause is by no means always silent. *Breath pauses,* characterised by an inhalation noise, belong to acoustic pauses. A further distinction has to be made between *filled pauses* and *unfilled pauses*. Filled pauses usually occur in spontaneous speech; e.g. hesitation phenomena caused by planning problems or as discourse markers in dialogues. Although there is a debate whether fillers are words or not (cf. Clark & Fox Tree, 2002), there seems to be agreement on the phonetic content of the fillers. Typically, fillers are either [ɛː] or [ɛmː] or [ə]. In contrast to filled pauses, unfilled pauses usually occur at grammatically motivated locations and usually consist of silence. However, some unfilled pauses are phonetically filled, namely with breathing. Of course, all combinations of breathing, fillers and silence can be found.

For an exact definition of a pause, it is necessary to determine a threshold where a pause starts to be a pause. While some studies work with a threshold below 50 ms (pause and segment detection with an automatic procedure) (e.g. Lee & Oh, 1999), in other (mainly psycholinguistic) studies only silent pauses longer than 500 ms are regarded (for a review see O'Connell & Kowal, 1983). Other examples are 200 ms (Grosjean & Collins, 1979); 150 ms (Tsao & Weismer, 1997); 130 ms (Dankovičová, 1997); 100 ms (de Pijper & Sandermann, 1994). Other silent intervals can complicate the pause defintion. The closure phase of plosives as articulatory activity should, of course, not be considered as a pause, although it is a silent interval. The closure phases typically ranges between 30 and 70 ms. But problems arise when an inter-pause stretch starts with a plosive. There, it is usually impossible to find the acoustic reflex of the beginning of the closure in the waveform. Then, either a default value must be set for all stops (or different classes of stops, e.g. fortis vs. lenis) or, as a less accurate method, these closure phases have to be ignored.

Realising prosodic boundaries

From a phonological point of view, pauses can be considered as phonetic realisations of prosodic boundaries (or prosodic breaks). Related to the claim that pausing plays the central role in speech tempo variation, it could be assumed that prosodic phrase structure, with pauses as the main markers of the prosodic boundaries, plays a central role in speech tempo variation. Therefore, further markers responsible for prosodic breaks should also be considered such as:

- the presence of an intonational boundary tone expressed as a F0 contour (e.g. de Pijper & Sanderman, 1994)
- phrase-final lengthening of specific parts of segments (e.g. Oller, 1973; Strangert, 1991; Turk, 1999)
- change of voice quality, e.g. creaky voice & whispering (e.g. Strangert, 1991; Kohler, 2000)
- declination in intensity and spectral tilt (e.g. Streeter, 1978; Strangert, 1991; Trouvain et al., 1998)
- phrase-initial strengthening (e.g. Fougeron & Keating, 1997)
- phrase-initial glottal reflexes, e.g. glottal stop and laryngealisation (e.g. Rodgers, 2000)

In the case of the wide-spread falling terminal intonation contours, all the above listed parameters fit into a picture where the acoustic silence is preceded by an utterance-final relaxation (Kohler, 2000) and followed by an utterance-initial tension.

However, for marking prosodic boundaries, all the above listed parameters are not necessarily fully realised and some of them may be missing. It seems generally accepted that silent intervals, F0 resets and finally lengthened sound segments are the primary cues which signal a boundary (Sanderman & Collier, 1996), whereas all the other parameters are considered secondary cues (Lehiste, 1970).

Levels of pauses and levels of prosodic phrase boundaries

Usually, there is more than just one level of prosodic phrase boundary. In many studies, a three-way distinction is made in major break, minor break, and no break (e.g. Crystal, 1969). The ToBI (Tone and Break Indices) scheme assigns six levels of breaks, with …

Intonation phrase normally marked by a pause …

There were also attempts to classify pauses into distinct categories. A division in *four* categories of pause length has been proposed by different authors such as Crystal (1969) (brief – unit – double – treble pauses). Butcher (1981) mentioned long pauses with corresponding 1400 ms, short pauses (500 ms) and unheard pauses (150 ms). In the study by Bartkova (1991) short pauses were those below 200 ms, medium ones were between 200 and 500 ms, and long ones were those exceeding 500 ms.

It can be implicitly assumed that a higher level break is marked by a longer pause than a lower level break. This is also what Strangert (1991) claimed for her Swedish news-reading data. She found a positive correlation between the acoustic signalling and the rank of the boundary for pauses as well as for other break markers such as F0 resetting and voice quality irregularities, but not for phrase-final lengthening. That means in relation to tempo, e.g. for fast speech, that the typically shorter durations are either a consequence of a general shortening of the pause durations, or that the breaks are stepped back to a lower level, or that both mechanisms operate.

Placement of pauses and prosodic boundaries

So far, we have learned that tempo variation can be reflected in a change of the number of pauses and a change of the duration of pauses. But an essential question is still open, notably which pause, i.e. at which location, should be skipped or inserted, or shortened or lengthened, respectively.

Lass & Deem (1971) found an increase in the number of within-sentence pauses for slow readings whereas fast readings were characterised by fewer between-sentence pauses.

Bartkova (1991) shows with a three-way division of pauses, that in her texts the number of pauses at „full stops“ are relatively constant over three rates; at „commas“, pauses change slightly (but more in the fast direction), and syntactic pauses show the greatest variability.

Strangert (1991) showed for her Swedish data that pauses at paragraph boundaries are by far the longest, followed by pauses at sentence boundaries, and with pauses at clause and syntactic phrase boundaries the shortest.

Caspers & van Heuven (1991) found that pauses at obligatory phrase boundaries do not tend to be skipped at fast rates, but those at voluntary phrase boundaries do.

In an investigation of German news reading, Mixdorff (2002) found that pauses at major phrase boundaries between sentences were considerably longer than major break pauses within sentences. Additionally, a noticeable number of the within-sentence pauses were detected at locations *not* marked by punctuation in the text. This, of course, has serious implications for pause prediction, and therefore tempo modelling in text-to-speech synthesis.

3.2. Intonation

There are several studies dealing with the effect of tempo on intonation, such as number of pitch accents, the choice of accent types and the truncation and compression of accentual F0 movements, pitch excursion size and overall pitch range, and the temporal alignment of pitch accents.

Caspers & van Heuven (1991) found no difference between normal and fast speech concerning the number of pitch accents.

Dutch listeners in Rietveld & Gussenhoven (1987) were asked to rate sentences that differed only in the complexity of the intonation contour. Examples with complex structures were judged ceteris paribus as slower than less complex structures. Correspondingly, in the Dutch production data in Caspers & Van Heuven (1991) pitch configurations marking the right boundary of intonation phrases tended to be simplified in fast speech. Similarly, there are simplifications of pre-nuclear pitch patterns. Fougeron & Jun (1998) e.g., report deletions of the initial high tones in French.

Ladd et al. (1999) found in their English data that F0 excursions are larger as rate slows down, while Kohler (1983a) found an increase in the average F0 level for fast speech in his German material.

In summary, these few consistent findings based on read speech do not allow as many generalisations as were possible e.g. for pausing. But there are clearly some interactions between tempo and intonation, with a general tendency for more accents at a slower tempo and fewer accents at a faster tempo. The number of pitch accents has, of course, implications for the ryhthmical structure sketched in the next section.

3.3. Rhythm

Since it reflects how (in our case) speech is temporally organised, rhythm clearly has implications for tempo variation. Problems arise when, for some authors, rhythm is more or less synonymous with tempo and the timing of speech. Another problem lies in the fact that, for the author, rhythm seems, paradoxically, to be one of the most obvious and at the same time one of the most controversial topics in phonetics and phonology. This seems particularly true when it comes to concrete temporal correlates of rhythmic units.

The classification of the languages of the world in stress-timed and syllable-timed languages by Pike (1945) led to the hypothesis that basic rhythmic units in a given language have approximately the same duration. However, this isochrony hypothesis has consistently failed when the actual durations of feet (stretches between two stressed syllables) in the stress-timed languages, or syllables in the other language rhythm-type were measured. Neverthless, there is a tendency to shorten syllable durations in feet with many syllables, and vice versa, evidence which can be seen as a sort of rhythmic compensation.

Several researchers see a tendency to isochrony in speech production in stress-timed languages such as English (Lehiste, 1975) and German (1982). In the latter case the isochrony tendency is manifested not only by durational compression but also in phonemic reductions (see next section) and modifications of the number of syllables and of the syllable complexity (see section 6.5) and the accent patterns (cf. last section). We therefore avoid the traditional complexities of the rhythm discussion and address these phenomena in the appropriate sections.

3.4. Segmental processes of connected speech

Fast speech is often characterised by processes on the sound segment level, such as deletion, assimilation and replacement of those sounds that are defined in the canonical or lexical structure of the words. These processes happen in connected speech and are normally treated as "post-lexical rules" by phonologists. The example in table 3.2. shows various segmental forms of the same sentence. Although the processes also occur in normal-tempo speech, the resulting segmental chain for fast speech differs from the one expected at a normal tempo. The corresponding speech signal shows a different stream of acoustic phonetic segments.

Table 3.2. The German example "Hast Du einen Moment Zeit?" in possible various forms starting with the abstract underlying form, the canonical realised form to more and more reduced forms (after Kohler, 1990).

full lexical (underlying) form	h a s t d u: aɪ n ə n m o m ɛ n t ts aɪ t
canonical form; assimilation voice; degemination	h a s **t** u: ʔ aɪ n ə n m o m ɛ n **t**s aɪ t
reduction of vowel quality, deletion of schwa; degemination	h a s t **ʊ** aɪ **n** m o m ɛ n ts aɪ t
reduction of vowel quality	h a s t ʊ **ə** n m o m ɛ n ts aɪ t
merging of vowels with reduction of vowel quality; assimilation of place of articulation with degemination	h a s t **ə** **m** o m ɛ n ts aɪ t
Deletion of schwa; assimilation of place of articulation	h a s **p** m o m ə n ts aɪ t

The canonical or lexical form of a word can be altered in different ways and at all levels of style and tempo. Deviations from the lexical form comprise basically assimilations of all types and complete deletions of segments. It is evident that in fast speech more deviations from the lexical form occur, so that sometimes the term "fast speech rules" or "allegro rules" is used.

These allegro rules are sometimes used to characterise certain speaking styles which are denoted as casual, informal or sloppy, as well as for diachronic phonological processes (Dressler, 1975). Of course, these processes have been observed and described for languages other than German, such as for Hebrew (Bolozky, 1977) to name just one example.

Greisbach (1992) tested the processes mentioned e.g. by Kohler (1990) for German speech read at maximal speed. He indeed found all the described processes and added some new rules to the ones already described. It is noteworthy that the realisations of the five speakers differed considerably in terms of rule selection.

Function words, which frequently occur in their weak forms deserve special consideration. The weak forms often show a tendency to reduce the vowel to a schwa, cf. "can" as [kæn] and [kən], "for" with [fɔː] and [fə]. Although there is a permanent propensity to use weak forms, the degree of "weakness" and the frequency of weak form occurrence seems to increase in styles marked as fast, or as Kohler (1995: 220) notes:

> "Je schneller das Redetempo ist, desto leichter stellen sich schwache Formen ein und desto weitreichender sind auch die Veränderungsprozesse."

3.5. Segment and syllable duration

A change in the articulation rate does, of course, imply a change in the temporal extension of speech stretches corresponding to linguistic elements such as syllables and sound segments. But as already pointed out with the change of pauses and the change of sound segmental processes, the change of the durations of sounds and syllables does not apply in a linear way.

It has been observed that sound segments reveal different behaviour in terms of compressing and expanding their durations according to their sound class. This phenomenon has been described as the *elasticity of sounds* (Gaitenby, 1965; Campbell & Isard, 1991). Roughly speaking, vowels expand and compress more than consonants, i.e. a slow articulation means primarily vowel lengthening, and fast articulation means primarily vowel shortening. However, there are many factors responsible for segment durations.

Extra-linguistic and para-linguistic factors influencing speech tempo have been presented in the previous chapter. There also, language-relevant factors were listed which are not genuinely of a phonetic or phonological nature. The phonetic or phonological factors which can influence the duration of sounds and syllables are presented in this section. Table 3.3 gives an overview of various factors influencing segment duration which have been reported in the literature for different languages. It must be noted that the studies and languages mentioned represent only a small selection out of a rich pool of research literature on the topic of speech segment duration.

The durations of the acoustic correlates of sound segments show great variability, even among the realisations of the same phoneme. Durational variability has also been found between speakers as well as within speakers (Klatt, 1976).

On the sound segmental level, an inherent duration of segments has been observed in the way that e.g. close vowels show shorter durations than open vowels (cf. the literature presented in Lehiste, 1970). The phonological quantity (sometimes expressed as degree of tenseness) can have an immense influence on the actual duration of segments, e.g. underlyingly long and tense vowels in German show longer average durations than short and lax vowels.

On the syllabic level, the number of segments (especially the number of consonants) have an influence on the duration of all segments in that syllable. In "strict" with five consonants, the durations of all sounds are expected to be shorter than in "tick" with only two consonants.

But also the position of the consonant in the syllable can make a difference for the consonant duration. The duration of [f] in syllable coda position (e.g. in "Schiff", engl. "ship") is expected to be longer than in syllable onset position as in "Fisch" (engl. "fish").

For vowel duration the (phonological) voicing status of the following consonant can be decisive. If the post-vocalic context is a lenis plosive like in "bag", the vowel is probably longer than in "back" with a fortis plosive as post-vocalic context. Despite word-final devoicing in German, this also applies where the fortis-lenis opposition is maintained (e.g. "leiten" vs. "leiden").

On the word level, the number of syllables in a word makes an essential contribution to the duration, especially for vowels. The [ɪ] in "stick" is longer than in "sticky" and the vowel in "stickily" is shorter than either of the others (Lehiste, 1972).

Similarly, the position of the syllable in the word can determine the segment duration. Word-final lengthening was observed e.g. for Dutch (Nooteboom, 1972), Swedish (Lindblom & Rapp, 1973) and American English (Oller, 1973; Beckman & Edwards, 1990).

One of the main factors determining duration is the lexical stress as the prominence on the word level. On the prosodic phrase level, the lengthening of pitch accented syllables, in addition to their inherent lexical stress, is a further factor. This illustrates the hierarchical and cumulative effects of duration-influencing factors.

A further prosodic condition leads to considerable durational changes, notably the lengthening effects at edges of prosodic constituents. The best known effect is phrase-final lengthening. But a phrase-initial strengthening with consequences for duration has also been reported (Fougeron & Keating, 1997).

Table 3.3. Factors which influence segment duration with evidence for different languages from a selection of studies. Studies that also looked for tempo are indicated with an asterisk (*).

factor	**study**
inherent durations	American English: Lehiste (1970); Klatt (1975); German: Neweklowsky (1975); Antoniadis & Strube (1984); French: O'Shaughnessy (1981)
phonological quantity	Amer. English: House (1961); German: Antoniadis & Strube (1984); Braunschweiler (1997)
consonant cluster	American English: Klatt (1975); German: Kohler (1988)
position in syllable	American English: Klatt (1975); Greek: Botinis et al. (1999)*; German: Kohler (1988)
pre-, post-vocalic context	Amer. English: Peterson & Lehiste (1960), House (1961), Lisker (1974); German: Antoniadis & Strube (1984); French: O'Shaughnessy (1981)
polysyllabic shortening / monosyllabic lengthening	Dutch: Nooteboom (1972); Swedish: Carlson & Granström (1986), Lindblom & Rapp (1973); German: Kohler (1986); American English: Klatt (1975)

position in word / word-final lengthening	Dutch: Nooteboom (1972); Swedish: Lindblom & Rapp (1973) ; Amer. English: Oller (1973); Beckman & Edwards (1990)*; French: O'Shaughnessy (1981)
word-initial consonant lengthening	Dutch: Nooteboom (1972); Swedish: Lindblom & Rapp (1973); Amer. English: Lehiste (1960); Oller (1973)
lexical stress	American English: Klatt (1975); Botinis et al. (1999)*; German: Jessen et al. (1995)
pitch accentual lengthening	German: Kohler (1988); English: Turk & White (1999); Dutch: Eefting (1991)
phrase-final lengthening	Amer. English: Gaitenby (1965); Klatt (1975); Lehiste, Olive & Streeter (1976); Beckman & Edwards (1990)*, Turk (1999); Hebrew: Berkovits (1991)*; Dutch: Gussenhoven & Rietveld (1992); French: Fougeron & Keating (1997)
phrase-initial strengthening	French: Fougeron & Keating (1997)
foot shortening	Dutch: Nooteboom (1991); Amer. English: Beckman & Edwards (1990)*

There is still much to discover about the domains in which these factors operate. The factors listed in table 3.3 are considered to result in local changes of articulation rate. A linguistic factor affecting global tempo is utterance length. A number of studies have shown that the length of an utterance has an influence on the tempo of the utterance. Fónagy & Magdics (1960) for Hungarian and Malécot et al. (1972) as well as Bartkova (1991) for French give evidence for the tendency that speech rate increases with length of utterance. Of course a long utterance would normally show fewer prosodic phrase boundaries than two or more short utterances, and that means fewer phrase-final lengthened syllables. Nevertheless, Haselager et al. (1991) found in their study with children, where they disregarded pauses, the vowel and the consonant(s) before the phrase break, that longer utterances are articulated at a faster rate than shorter utterances. Thus, it can be hypothesised that we have a shortening effect such as the shortening due to an increased number of segments in a syllable, or due to an increased number of syllables in a stress group. This shortening effect has been demonstrated e.g. by Lehiste (1972). In her material, not only utterance duration (in msec) and utterance length increase with the number of syllables and segments, but also the syllabic rate, with one word sentences the slowest and six word sentences the fastest utterance. Gaitenby (1965) makes the following observation in her sentence material: the longer the utterance in terms of number of segments, the shorter the absolute duration of any given segment, until an approximate minumum duration was reached beyond which segments could not be compressed further. This shortening effect has also been observed in natural recordings. In a longitudinal study of adult-child interaction, Van de Weijer (1997) offers evidence that articulation rate continously increases from utterance span of one syllable up to seven or more syllables, for both child-directed and for adult-directed speech.

All factors can interact with each other, and these interactions must be considered as well in a duration model. There are studies investigating some of the interactions of the mentioned factors with rate. Berkovits (1991) found a phrase-final lengthening effect which only operated with fast speech, not for slow speech in her Hebrew data. Thus, a duration model which attempts to model the effects of articulation rate as one of the main factors of durational variability must take into account the main potential interactions with the other factors.

3.6. Articulatory organisation

When humans articulate faster, different mechanisms of articulatory organisation may come into play. This can be achieved e,.g. by *shortening the duration* of the articulatory gesture, as discussed e.g. by Kröger (1996) and Gay (1981).

The best known mechanism applicable to fast speech is the *target undershoot*, i.e. a reduction in the magnitude of articulation. The theory of target undershoot (Lindblom, 1963) says that in a shortened sound segment the articulatory, and consequently the acoustic target has not been fully reached before the particular articulator starts the next gesture. The resulting spectral reduction of time-reduced vowels is expressed as a tendency for centralisation, i.e. that they are more central in the vowel space. Regarding different forms of shortening (induced by different tempo and different degree of stress, respectively), he found in his study support for the hypothesis that

> „it is immaterial whether a given length of the vowel is produced chiefly by the tempo or the degree of stress. Duration seems to be the main determinant of the reduction.‟

Contrary to Lindblom (1963), other reserachers were not able to find evidence for target undershoot in fast speech conditions (Engstrand, 1988; van Son & Pols, 1989; Nooteboom 1991).

A third mechanism for faster articulatory movement is the *increase of velocity.* The data investigated by Kuehn & Moll (1976) and Gay (1981) show that the velocity of articulators can increase in fast speech.

A further mechansim to speed up articulation is to *increase the gestural overlap*. Adjacent sound segments that use different articulators can also overlap in production. For example, [t] does not require the use of the lips, so lip rounding in an adjacent segment (like [u:]) can begin during the [t], confer the first [t] in "tourist" with the one in "tick". As an example study, Engstrand (1988) varied stress and tempo in vowel-consonant-vowel sequences with Swedish speakers. He found that vowel- and consonant-related gestures were coproduced to a greater extent at fast tempo compared to slow tempo.

Also the *degree of coarticulation* can vary as a function of speech rate. In coarticulatory assimilation, neighbouring sound segments that require the same articulator use a single articulatory gesture for both sounds. E.g. in the phrase "Er hat ja gelogen.", [t] and [j] both require particular articulations of the tongue tip and blade. In this case, [t] is often produced with the palatal gesture of the [j], resulting in a [ç]-like release of the plosive.

Although the mechanisms are mentioned separately they do not necessarily apply separately. In an American English study investigating different consonant clusters under different speech rate conditions eliciting electropalatographic data, Byrd & Cheng Tan (1996) report individual consonant shortening in duration and a relatively increase in the overlap of the articulations.

To summarise, the mechanisms mentioned above - as well as those mechanisms on the other structural levels – are non-linear in nature. This phenomenon has also been recognised as a general principle of articulatory tempo variation by Gay (1977):

> "The reduction in duration of all segments coupled with the relative constancy of acoustic (vowel) targets, suggests that this adjustment [of articulatory movement] involves primarily a horizontal compression. This [horizontal] compression [...] is a non-linear one, and one that causes both a decrease of duration and an increase in coarticulation."

Summary and discussion of chapter 3

A change of tempo results in changes at *many* levels of phonetic and phonological characterisation, and not only at one level as one might think if just pausing or articulatory velocity of a specific articulator is the subject of study. These levels can be more or less closely linked to each other, e.g. phrase break and segment duration vs. phrase break and pitch accent. Thus, tempo variation is very complex seen from the articulatory point of view.

It seems a general principle at all levels that changes in acoustic duration resulting from changes of speech tempo are *non-linear*. This fact makes the modelling of speech tempo much more complicated than a simple model based on a combination of linear changes.

If speech tempo is to be modelled, then it is necessary to develop the model at the different levels presented above. If speech tempo is to be modelled in a non-linear way, as observed in natural speech, then knowledge must be acquired about the non-linearity, and the magnitude of change at each level must go into the model. And finally, if speech tempo is to be modelled non-linearly for a speech synthesis application, then the general model based on speech production must be adapted for a given artificial speech generation architecture and the performance of the implemented model should be tested for speech perception, i.e. with actual listeners.

Chapter 4

Measuring Tempo

Introduction

4.1. Categorisation of tempo

Before talking about measuring speech tempo let us make clear with some examples what kind of tempo we deal with when we want to measure and to categorise tempo in speech. If we instruct two speakers to read a given text at three different speeds, first at a pace that is normal for them, then at a slow pace, and finally at a fast pace, then we have speech with three different tempi. If we measure the durations of each of these text recordings it can be assumed that the slow versions take longer than the normal ones, and the normal versions take longer than the fast ones. However, it might be that the slow version is shorter than the normal one, as happened with one speaker in Trouvain (1999). In other words, the durations of the various productions as *objective* measurements do not necessarily mirror the intended tempo as a *subjectively* produced speech tempo.

Moreover, if we ask listeners to judge which of two recordings of the same text they think is faster, the choice does not necessarily fall onto the production with the shorter duration. For this judgement, other factors could play a role such as dysfluencies, deletions and assimilations, number and duration of pauses (Goldman Eisler 1968) but there is also an influence of fundamental frequency on perceived rate (Kohler, 1986; Rietveld & Gussenhoven, 1987). That means that the *subjective* impression of speech tempo does not exactly match the *objective* measurement.

Thus, the following three types of speech tempo must be distinguished from each other:

- the subjective, intended tempo of speech production
- the objective, measured tempo reflected by durational correlates of linguistic units
- the subjective, perceived tempo

Tempo relates a distance to a duration, both measured with objective criteria. A tempo which is based on the relation of a "distance" in speech to a duration is able to quantify a given piece of speech *quantitatively on a continous scale*, e.g. in number of syllables per second. In contrast, the intended as well as the perceived tempo can be assigned to *categories*, e.g. slow or rapid. Each listener/speaker has an idea what slow, fast, normal (or however the category is named) means, but obviously everybody has her/his own interpretation of these categories, otherwise renditions of the same text at the same intended tempo would not diverge in their durations.

The intended as well as the perceived tempo compare speech tokens *relative* to one another. One can ask people to produce an utterance slower than normal, and people can judge whether a given utterance in one recording is faster than in another recording.

In order to make the tempo of instances of speech comparable, an objective metric seems the most promising method to do the job. The following sections deal with the problem of how to measure speech tempo quantitatively.

4.2. Units of tempo measurement

Measuring speed means relating a distance covered by a body to the time used. In speech, the articulators are our bodies moving in time and space. However, with respect to the articulators there is a lack of homogeneity: some articulators move inherently faster than others. The tip of the tongue e.g. is able to execute many more movements in a given time compared to the velum (Hudgins & Stetson, 1937, cited in Lehiste, 1970). Moreover, the articulators neither move all the time nor do the they move to the same extent. Measuring distances (here in millimetres) could of course be done for each of the different articulators. Although the generation of speech and its sound segments can be seen as the result of the execution of articulatory gestures (whose distance can be measured), it is crucial to understand that speech is the result of the temporally coordinated execution of articulatory gestures that lead to speech *events*. That means, there is no distance that we can measure, we must seek a unit which describes speech events.

A number of different linguistic units have been proposed to serve as the substitute for a distance measurement unit in speech. In the literature we encounter a great variety of tempo denotations such as words per minute (wpm), syllables per minute (s/min), syllables per second (syll/sec or s/s), average syllable duration (ASD in msec), phones per second, or average phone duration (in ms). That means that units in use for measuring speech rate are, among others, the word, the syllable and the phone. Although commonly used, the definition of these linguistic units is not always straightforward. The advantages and disadvantages of these units shall be presented and discussed in the following sub-sections.

The word

Superficially, the word is easy to define and to count, and therefore apparently a useful unit for tempo measurement. A word can be defined as a sequence of letters that is not interrupted by a space or by an additional punctuation sign in a written text. However, the length of the units vary so much that the word is useless as a basis except for extremely long texts.

In some regards the graphical word is in conflict to other definitions of the word. The writing of the same words can differ within one language, depending on current orthography standards in a given language, e.g. in German "zusammenschreiben" vs. "zusammen schreiben"; English "shop-assistant" vs. "shop lifter", or "infra-red" vs. "infra red". Not every lexical word is expressed as one graphical word, e.g. French (and also German) "à la carte" or "San Francisco". Clitic groups such as German "ich hab's" or English "I don't" can be seen as two or three words. Also, the word can be seen as a morphological word, e.g. German "Berlin-Tegel" are two morpheme-based words but only one graphical word. Further, the number of words is unclear for many numerical expressions, e.g. German "17,50 €" are two graphical words, but in the spoken form three lexical words.

It would be reasonable for cross-linguistic studies to have a linguistic unit which allows comparisons across languages. Here, the length of words can vary to a very high degree, if we think of morphologically rich languages such as Finnish or German (e.g. German "Donaudampfschifffahrtskapitän") or agglutinating languages such as

Turkish. Similar lexical concepts should show a comparable length in compared languages. The three graphical words in the American English "Federal Supreme Court" are opposed to only one German "Bundesverfassungsgericht".

The syllable

A syllable can mean the underlying syllable derived from the lexical form of the word, or a syllable can mean the realised syllable. For the underlying syllable the number of syllables seems always clear. An exception to this clarity of syllable count is in German non-syllabic vowels as in *Piano*, *Lineal* or *genial*. In contrast to phonemic syllables the presence of a realised syllable is sometimes hard to detect. Syllables can be skipped (even words can be skipped or completely blended), and in many cases it is hard to decide when a syllable is skipped or still there. As a frequently occuring example in German, the phoneme sequence vowel-schwa-/n/ as in “ziehen” (Engl. "to pull") the syllable /ən/ can be realised as a syllabic [n̩] or as an non-syllabic [n], leading to different syllable counts.

The sound segment

One interpretation of a phone is the phonemic segment of a lexical word, whereby the phonemic status of certain sound segments are still a matter of debate. Another issue is whether affricates and diphthongs should be seen as mono- or bi-phonematic, i.e. one or two segments. Usually, a glottal stop is not given a phonemic status. Further delicate aspects include the results of phonological processes such as diphthongisations (e.g. in German homosyllabic vowel+/r/-sequences like in "Start" (Engl. "start"); or the phonemics of certain affixes, e.g. the "er" in "ersetzen" (Engl. "to replace"), which could be /ɛr/, /ər/, /ɛɐ/ or simply /ɐ/; or the degemination of homorganic consonants as in "kann nicht" (Engl. "cannot"). Even if the listed problems do not contain central concerns about the sound segment as the optimal unit for tempo measurement it should be clear that the segment is not unproblematical.

Intended vs. realised forms

A distinction which is infrequently made is between the intended forms (corresponding to the canonical or lexical or underlying form) and the realised forms, the latter is called "effektive Lautzahl" (Engl. "effective number of sounds") by Hildebrand (1963). *Intended* forms have the advantage that they can be easily derived from the lexical representation of the uttered words, whereas their actual *realisation* can vary strongly. This fact has already been pointed out by von Essen (1979) and can be illustrated with the German sentence "Am blauen Himmel ziehen die Wolken." (Engl. lit. "In the blue sky wander the clouds."). The transcription of this sentence consists of 26 phonemes and 10 syllables:

/ʔam blaʊən hɪməl tsiːən diː vɔlkən/

However, a typical reduced realisation of this sentence, is shrunk to 20 phones in 8 syllables:

[am blaʊn̩ hɪml̩ tsini vɔlkŋ̩]

If we assume a duration of 2 seconds for a realisation of this sentence, the measured or "objective" tempo in phones/sec would either be 13 phones/sec (intended) or 10 phones/sec (realised); the "objective" tempo in syllables/sec would either be 5 syll/sec (intended) or 4 syll/sec (realised). Ironically speaking, a speaker can speed up or slow down the speech tempo by a quarter just by defining the unit of measurement. This example shows that just one criterion of the definition of the unit of tempo measurement, here the sound segment, can be decisive on the meaning of what has been measured. This, of course, has serious implications for comparing data of different studies.

Other units

In music, tempo is measured by a metronome in beats per minute. The composer can either indicate the metronome value or can use a tempo term such as *adagio*, *lento, largo, grave* for slow tempi and *moderato, allegro, vivace* for faster tempi. These terms correspond to metronome values where the normal range is considered to lie between 75 and 80 beats per minute, i.e. values which are slightly higher than the 72 heart beats per minute of a middle-aged healthy adult person.

The idea of also using beats per minute in speech has been applied by a few researchers, e.g. by Uhmann (1989). In addition to syllables per second, she proposes *accents per second* as an additional measurement of tempo or "density". In her analysis of German conversational data she has shown examples of what she calls "contextualisation cues" in which speakers make utterances interpretable in dialogues. For example, a low number of accents per second combined with a high number of syllables per second serves to contextualise parenthetical utterances, sidesequences and afterthoughts. In contrast to these passages of low relevance, portions of high relevance such as emphatic utterances are contextualised by a high number of acc/sec and a low number of syll/sec. A combination of a high number of acc/sec and a high number of syll/sec can be found in repair sequences. The problem with accents is of course to define this unit with the aim of a consistent and reliable use across researchers. Uhmann (1989) transcribed primary, secondary and emphatic accents, but the transcription of these can differ between labellers, which is counter-intuitive to the idea of having a quasi-exact quantification.

Last but not least, non-linguistic units that were derived from the acoustic signal have been applied to quantify speech tempo. In studies aiming at detecting articulation rate automatically, e.g. for use in automatic speech recognition. This is done to improve the modelling of fast speech with a high number of segment deletions and replacements (cf. chapter 3) usually featuring an disproportionally high word error rate. Morgan, Fosler & Mirghafori (1997) calculated energy fluctuations to determine articulation rate whereas Samudravijaya, Singh & Rao (1998) enhanced the parameter set and also tested measures of non-stationarity and voicing switch rate.

Selecting a unit of tempo measurement

The previous sub-sections make it clear why there cannot be an objective "metre" for speech tempo measurement. Nevertheless, one linguistic unit must be selected if speech tempo is to be quantified. The following criteria may give an orientation for selection:

- degree of popularity
- comparability across studies
- ease of counting
- ease of definition

- reflection of temporal variance

The word (as words per minute) and the syllable (as syllables per second or as average syllable duration) seem to be widely used as tempo metrics, whereas the sound segment (usually as phones per second) seem to be less frequently used. Regarding the comparability to data of other studies of the same as well as of a different language, the syllable and the sound segment seem to be preferred rather than the word. Counting tokens is not a problem for the word, and counting does not cause greater problems for the syllable. However, counting sound segments requires a transcription of all recordings, and that is often not available. The easiest definition can be given to the phonemic syllable followed by the word, and here again the sound segment seems to be the most problematical case. Nevertheless, the essential characteristic of a unit expressing tempo is the reflection of temporal variance. Here, the word seems to score worst, and the sound segment best, followed by the syllable, i.e. the smaller the better.

In order to check the tempo fluctuations due to the choice of the unit, Carroll (1966, cited in Kowal, 1991) investigated which differed in the number of syllables per word in a reading aloud experiment texts. The measurement of words, syllables and phonemes per minute showed that the variation coefficient of the mean values per text was highest for the word and lowest for the phoneme. The most reliable results for the different texts were found for phonemes per minute.

This finding is in agreement with the results in Trouvain et al. (2001) with German data where the number different linguistic units were correlated with articulation time. It was shown that realised phones correlated best, followed by intended phones, realised syllables, intended syllables, and words (in this order).

In a study recommending standard speech rates for foreign language training, Tauroza & Allison (1990) compared the word rates and the syllable rates of four different speech styles. For reasons of different word-to-syllable relation for each style (news texts having more syllables per word than interview speech), the two rates were not at all in agreement with each other. Syllable rate was found to be better as an expression of one standard tempo for various styles than word rate.

An argument against the syllable as a quasi-universal unit is that in mora-timed languages such as Japanese, speech tempo is frequently measured in mora per second (e.g. Kuwabara, 1996; Koiso, Shimojima & Katagiri, 1998).

Also in testing the sensitivity of different tempo measurements for the classification of (English) speech according to their tempo (for use in automatic speech recognition), the recognition rate is more sensitive to phone rate than to word rate (Siegler & Stern, 1995).

Assessing the contributions of words, syllables and segments to utterance duration with reference to articulation rate measurement, Faulkner (1997) identified for English texts the phoneme as the single most significant variable to explain durational variance.

Den Os (1985) gives evidence that phonological syllables per second and phonetic segments per second best fits the perceived speech rate for Dutch and Italian short utterances. Phonetic syllables were worst.

In perception tests investigating the estimation of local speech rate, Pfitzinger (1999) found out that a combination of phone rate and syllable rate matches the subjective evaluations best when listening to short windows of speech (625 ms).

Referring to differences between languages, where the rhythm type of the language play an important role, Roach (1998) favours the sound segment as unit to be preferred:

> "Dauer (personal communication) has found that Greek and Italian are spoken more rapidly than English in terms of syll/sec, but this difference disappears when sounds/sec are counted. […] It seems that on the evidence available at present, there is no real difference between different languages in terms of sounds per second in normal speaking cycles."

To summarise, among the existing units there is obvioulsy not *the* optimal unit for tempo measurement. The selection of the unit depends on the purpose of the study. However, although word per minutes seem a rather widely used metrics it is obviously less favourable for most purposes. An exception may be when more abstract units are compared, as done in the study by Grosjean (1979) who investigated the articulation rate and the pause rate of signers (American Sign Language) and speakers (American English).

The criteria listed have not been weighted so far, but it seems clear that the unit that mirrors tempo best is the one that is most sensitive to temporal variance. By nature this is the smallest unit, i.e. from the units presented here the (realised) sound segment. However, there are other factors worth consideration. One usual way to economise articulatory effort, with the consequence of speaking faster, is to reduce the number of realised segments. That means that the intended phone would be the appropriate candidate because it additionally accounts for an important tempo variation factor, anemely degree of segmental reduction. And the last note on the relative importance of the listed criteria refers to the ease of definition and the ease of counting, which speaks for the intended syllable (ignoring the word). These two criteria will be the decisive ones for many studies and many applications, simply for practical reasons.

The role of pauses in tempo measurement

With reference to the beginning of chapter 2, articulation rate was defined as the net speech rate, and speaking rate including the pauses was defined as the gross speech rate, in line with many other researchers (e.g. Goldman Eisler, 1968; Wood, 1973). A look at table 2.1 (p. 7-8) makes it clear that the differences can be substantial between these two measurements, ranging up to several syllables per second difference for the same recordings, especially in spontaneous speech with a high percentage of pausing time.

If the differences can be so dramatic, then the defintion of a pause is crucial to determine speech tempo. In chapter 3, several thresholds in different studies were listed, ranging from 50 ms up to several hundred ms. It goes without saying that an articulation rate measured with a pause threshold of 100 ms can differ considerably from the articulation rate measurement of the same recording with a pause threshold of 500 ms (cf. Kowal, Wiese & O'Connell, 1983).

Besides pause thresholds, unintentionally articulated speech also causes problems, e.g. in a filled pause ("die äh meiner Meinung nach") or in corrections of slips of the tongue ("die deiner Mei, nein meiner Meinung nach") or in word repetitions in spontaneous speech ("also die die die nicht das nötige Kleingeld haben"). There is no common standard whether to consider these dysfluencies as ordinary speech articulation, or as a pause, or simply to ignore these instances of badly formed articulatory performance.

4.3. Dynamics of global and local tempo

Global and local levels of articulation rate

Another uncertainty when dealing with speech rate concerns the stretch of speech taken into consideration. Speech rate changes continuously while speaking (cf. Wood, 1973; Miller, Grosjean & Lomanto, 1984), so that the first part of an utterance can be spoken fast, while the second part can be rather slow, or vice versa. An average rate calculated for an utterance does not necessarily reflect the tempo characteristics of different parts. When the domain is not specified, it is not clear whether the speech rate quantifications are related to a more global or to a more local level. Most of the time, when people talk about speech rate, they use the term globally, referring to an entire text, sentence or whatever the utterance might be. The problem of local variations has long been neglected. The main question to be answered is: How "locally" should speech tempo be considered?

No matter what the local unit will be, despite one global rate that can be determined, there are tempo differences between the individual phrases. Spontaneous speech can be expected to be marked by more changes in articulation rate than we find in read speech: Planning problems are likely to cause hesitations (e.g. syllable drawls) leading to slow stretches followed by fluent, fast stretches. These planning problems in spontaneous speech also increase the number of filled and unfilled pauses which lead to shorter inter-pause stretches. Especially utterances consisting of only one or two discourse particles such as "ja" contribute to a high number of short but very slow inter-pause stretches. The last points would support the reported tendency that "the longer the utterance the faster its rate" (cf. Fónagy & Magdics, 1960; Malécot, Johnston & Kizziar, 1972, Martínez et al., 1997, but see also Koopmans-Van Beinum & Van Donzel (1996) for different results). Emphasis, which occurs more often in spontaneous speech, represents another factor which results in a slower tempo.

In an inspection of the German "Kiel Corpus of Read and Spontaneous Speech" (IPDS, 1994) we compared the rate characteristics of read versus spontaneous speech (Trouvain et al., 2001). The results of this study (replicated in table 4.1) show that in spontaneous speech inter-pause stretches (ips) as well as intonation phrases (IP) are shorter on average and show a greater variance than in read speech.

Table 4.1: Mean duration (in sec) and mean articulation rate (real. phones/sec) of inter-pause stretches (ips) and intonation phrases (IP) for spontaneous and read speech with standard deviations.

		duration mean (sd)	**articulation rate mean (sd)**
spontaneous	**ips**	1.81 (1.29)	13.24 (3.29)
	IP	1.17 (0.73)	13.18 (3.75)
read	**ips**	1.98 (1.03)	13.06 (2.03)
	IP	1.49 (0.67)	13.01 (2.23)

With respect to articulation rate, spontaneous speech is slightly faster and shows a greater variance (see also figure 4.1). Although faster on average, spontaneous speech features a high number of slow utterances. One reason lies in the large number of very short inter-pause stretches (<1 sec) in this speaking mode. Indeed, one and two word utterances are slower than the mean. Intonation phrases are generally shorter than inter-pause stretches, but there is basically no difference in articulation rate.

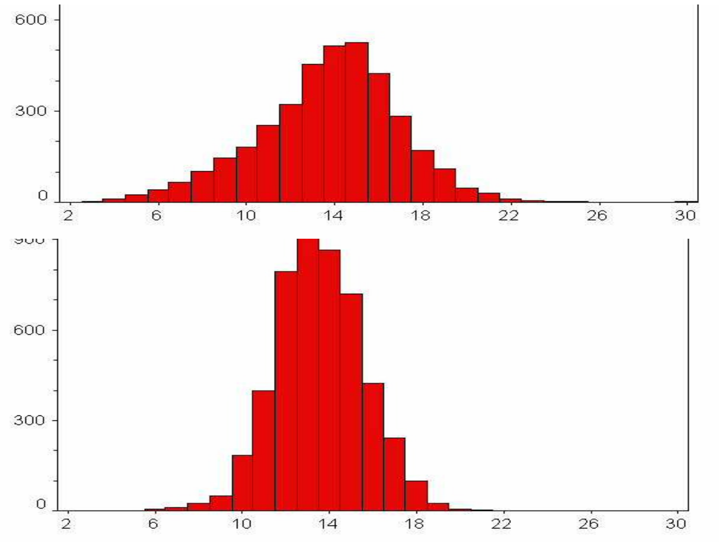


Figure 4.2: Histograms of articulation rate (realised phones/sec per inter-pause stretch) in spontaneous (top) vs. read speech (bottom) in the "Kiel Corpus of Read and Spontaneous Speech" (data from Trouvain et al., 2001).

Domains of articulation rate

When we looked for the "optimal" unit to describe tempo, the main criterion was that this unit expresses the temporal variability best. Now, searching for the "optimal" domain, we look for a stretch of speech in which the tempo variation is smallest, or, in other words, where articulation rate shows the highest degree of constancy.

Whatever the optimal utterance domain may be, tempo changes can occur not only between adjacent phrases but also within a phrase. The problem lies in the acceleration and deceleration within the local section. Each syllable lengthened due to accentedness or phrase finality is decelerated. We can focus domains as small as the syllable or even the syllable rhyme (phrase-final lengthening). All these very local phenomena can be seen as *accelerando* and *rallentando* (or *ritardando)* as labelled by Crystal (1969) in his list of prosodic systems under the heading *complex tempo system* in addition to the *simple tempo system* of global rates such as allegro and lento.

The previous paragraphs showed that the global tempo of a longer utterance can be distinguished from the local tempo of a single phrase within this utterance (confer left and mid pattern in figure 4.1). Moreover, there can be tempo variations within this single utterance showing e.g. a rallentando pattern, as illustrated in figure 4.2 (right side).

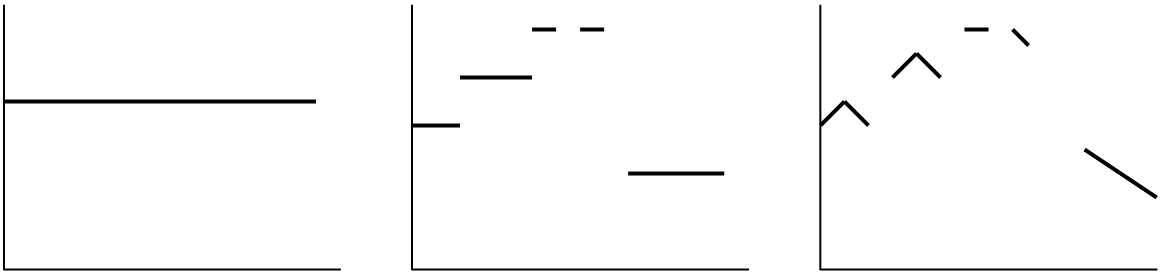


Figure 4.1: Global and local levels of tempo in idealised schemata of time course (x-axis) and tempo (y-axis); left: global rate for the entire utterance; mid: local rates for single phrases (e.g. inter-pause stretches); right: local rate shapes within the single phrases.

In her studies of Czech and British English, Dankovičová (1997) investigated the following spans of speech production as domains to measure articulation rate: the inter-pause stretch, the intonation phrase and the syntactic clause. She showed that the duration of phonological words is best mirrored by the intonation phrase.

Normalisation of rate dynamics

Crystal & House (1990) showed with articulation rates in a reading task with American English speakers that slow talkers and fast talkers differ in their global tempo (as illustrated in figure 4.2 left side) and they differ in the rates of the inter-pause stretches (cf. figure 4.2 mid). However, slow and fast readers have very similar patterns of local rate changes, i.e. the pattern in figure 4.2 (mid) is shifted upwards for fast speakers and shifted downwards for slow speakers.

The aim is to determine and to weight the factors responsible for the variation. Such a normalised value makes it easier to compare utterances differing in rate. But how can we relate given (phonological) information about syllable structure, number of segments, phrasal stresses, phrase boundaries and so on to a "normalised" rate? Koopmans-Van Beinum & Van Donzel (1996) tried to do so by assigning different weights to various phonological factors such as vowel quantity and schwa syllable. Although they consider their attempt to normalise rate dynamics in inter-pause stretches as preliminary they were able to show that the normalised rates of these phrases point to the discourse structure of the text. Phrases which are used to introduce something new are marked by a slow normalised rate. This picture was not so clear without the normalisation. It might be that such a normalisation of speech tempo could be a helpful instrument in order to improve the detection of temporally marked elements of information structure. This was also done by Uhmann (1989) who identified in her data less relevant passages such as parentheses with a high syllabic rate and a low rate of pitch accents and, in contrast to that, highly relevant passages with a low syllabic rate and a high density of pitch accents (cf. also Barden, 1991). However, looking at the experiences of Koopmans-Van Beinum & Van Donzel (1996) there is a big need for research:

> "The main conclusion of our study must be that accounting for variations in speaking rate of what may be considered as 'spontaneous speech', is a very complicated task."

It is one thing to normalise objective tempo by calculation, it is another thing to test how actual listeners normalise for, or indeed whether or how they perceive tempo variation found in objective tempo values. A listener appears to compensate for the numerical variation in rate, a fact that can be explained by the linguistic and phonetic (rhythmic) restrictions. There are comparably few instances of noticeable tempo changes in spontaneous speech (Batliner et al., 1997) and there are expected to be no noticeable tempo changes in neutral read speech, e.g. news reading.

Summary and discussion of chapter 4

Measuring speech tempo contains various sources of confusion. In this chapter we attempted to make clear that we must distinguish whether tempo means the intended tempo category in speech production, or a perceived tempo category, or a quantified objective tempo, where acoustic correlates of linguistic units are related to physical time.

The latter consideration is often expressed as word rate, syllable rate or phone rate, leading to the central question of speech tempo measurement: what is an optimal unit to quanitify speech tempo? What are criteria to determine an appropriate unit? The pros and cons of the syllable on the one hand, and the sound segment on the other were discussed. The word was considered to be the least optimal tempo unit – despite its frequent use, e.g. in speech synthesis markup languages such as SABLE (Sable URL). Although there are many arguments for the sound segment as the preferrable unit, the arguments from a practical perspective favour the phonological syllable as standard unit for tempo measurement.

But apart from the unit itself, further perspectives should be taken into consideration when speech tempo is quantified. These include the vital role of pauses (leading to a net rate excluding pauses or a gross rate including pauses), consequently the definition of a pause (there are great variety of thresholds), and also the domain in which articulation rate is measured.

The last sub-section was dedicated to the question of tempo variability found across and within phrases, with supporting data from our own corpus analyses. This led to a further distinction to bear in mind, namely the necessity to keep apart the global tempo of a longer utterance from the local tempo of single phrases within this utterance.

Chapter 5

Real-World Data Analysis: A Case Study

Introduction

The discussion in the previous chapters have shown that there are:

- many situations and conditions in which tempo variation can be observed, e.g. text sort or emotive speech
- many mechanisms with which tempo variation can be achieved, e.g. by varying pauses or reducing the duration of vowels and consonants
- methodological problems in measuring tempo, e.g. how to indicate tempo values for various inter-pause stretches for spontaneous speech.

In this chapter a case study with real-world data is given, namely a prosodic analysis of horse race commentaries. The advantages of this somewhat special text sort are the following ones:

- perceptually, the impression is of continously increasing tempo
- it is emotive speech, and the integration of emotions in synthetic speech is one type of application of tempo-scaling in synthetic speech
- it is real-world data, not laboratory speech or a meta-form of speech

The intriguing question is how the tempo increase that iscontinously manifest over the entire commentary is achieved. Moreover, other prosodic effects can also be observed, not just those realised on a durational level. Thus, the question can be extended to "how do prosodic systems interact?"

5.1. The prosody of excitement in horse race commentaries

As mentioned in chapter 2, high emotional activity in speech is characterised by a higher pitch average, a wider pitch range, a higher intensity, and a faster speech rate compared to some neutral or default way of speaking (Banse & Scherer, 1996; van Beezooyen, 1984; Murray & Arnott, 1993). Terms used to describe excitement or arousal are e.g. "anger", (especially "hot anger" or "rage") on the negative side, and "elation", "joy" and "happiness" on the positive side.

A further form of arousal, which is investigated in this study, could be named "suspense" and is exemplified in horseracing commentaries (henceforth HRC).

Prosody of horse race commentaries

In HRC, the course of emotional arousal is dependent on the fixed framework of the race, progressing from relative calm, through increasing excitement to the climax at the finish, then returning to a post-race calm. The form of arousal is presumably specific to sports commentaries, being neither negative (as with "fear" or "hot anger") nor positive (as with "elation"), but expressing the commentator's sense of excitement and suspense.

Barry (1995) gives an auditorily based description of a typical HRC pattern:

> "In British English there's a clear monotonisation rule, with definite, race-stage oriented resets (to a high pitch) with tempo and volume increases from one series of 'intonation units' to the next, and with sudden rallentando and decrescendo combined with a short series of resets to a lower pitch and a final low falling contour from the moment the winning horse finishes."

A stylised HRC pattern features an increase of pitch level, tempo and intensity, and a decrease of pitch range (figure 5.1).

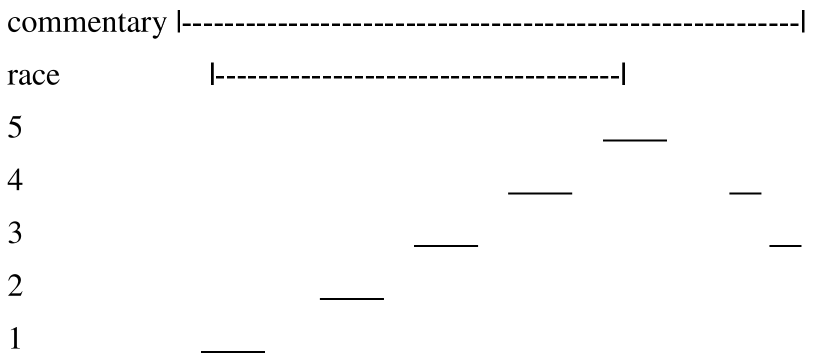


Figure 5.1: Stylised time course of horse race commentary according to Barry (1995). Parameters: pitch level (1 = normal, 5 = high); pitch range (1 = normal, 5 = narrow); tempo (1 = normal; 5 = very fast); intensity (1 = normal; 5 = high).

Spontaneity of horse race commentaries

The advantage of sports commentaries is that they offer examples of non-scripted speech with genuine informational and emotional expression. Although one cannot claim that HRC is completely "spontaneous" since it has a conventionalised form, it is taken from real situations compared to emotive speech acquired in laboratory situations from actors, which is the usual way of eliciting emotive speech production.

A further advantage of sports commentaries is the quality of the speaker, which guarantees a certain level of fluency, allowing an analysis of prosodic structuring which is less disturbed by dysfluencies (e.g. filled pauses, false starts, mispronunciations, syllable drawls, repeats, ungrammatically located pauses) than is the case with non-professional speakers.

Purpose of this study

The present study pursues a number of aims: Firstly, it seeks a verification of the auditory description of HRC given in Barry (1995). Secondly, it offers quantitative data on a number of prosodic parameters, among them several tempo-related parameters, which vary systematically with degree of excitement/arousal within a situation. This provides independent evidence for the pattern of increasing and subsiding excitement reflected by those parameters. Thirdly, it thus provides an empirical basis for comparison with other forms of excitement.

We therefore attempt to identify and discuss (1) what is specific to the HRC speaking style, (2) what is specific to the excitement component of the HRC.

5.2. Methods

Material

Three horse-race commentaries were chosen at random from a selection of HRCs recorded during BBC Grandstand (television) transmissions. The durations (time period from start of the race to the moment the winning horse finishes) of the three HRCs range from 79 seconds for race 1 to 145 seconds for race 3. Each race was commentated by a different (male) speaker. Although their identities are not known to the author, the speakers in race 2 and 3 are recognisable as speakers of New Zealand and Australian English, respectively. The commentator in race 1 speaks a standard southern English accent.

Analysis

In the first analysis step, three phases were selected to encompass the build-up and decline of tension expressed in the commentary. These are located

1) near the beginning of the race
2) in the middle
3) spanning the finish until the moment the winning horse passes the line.

These phases are sandwiched between breath pauses and they comprise around 25 seconds in the first race, 26 seconds in race 2, and 32 seconds in race 3.

For each phase an orthographic transliteration with special symbols for silent pauses, breath pauses, and filled pauses such as [ɛː] was carried out. Filled pauses were regarded as articulation, so that only silent and breath pauses are considered as ("unfilled") pauses below.

Each pause was marked in the time course, providing data on the number of pauses, the pause durations, and the durations of inter-pause stretches and inter-breath stretches, respectively.

In order to determine speech rate, the number of phonological syllables for each inter-pause stretch was counted. Filled pauses were regarded as one articulated syllable. Thus, "speaking rate" is defined as the number of syllables divided by total speaking time including pauses - in contrast to "articulation rate", which is based on total speaking time excluding pauses.

Prosodic phrases demarcated by intonational and rhythmical means other than pauses were also transcribed. However, these phrase markers were used for further instrumental investigation due to the lack of reliable labelling criteria.

In the second analysis step, instances of prominent horse names were excised. For race 1 "Two Clubs" occurred six times, in race 2 "Pentland's Flyer" 11 times, and race 3 "Hot 'n' Saucy" 15 times. These occurrences, which are more or less equally distributed over the entire race, were used to determine the pattern of pitch and intensity characteristics.

F0 range, defined as the difference between the highest and lowest value in comparable syllables, was measured in the second stressed syllable of each horse name.

Spectral tilt as an acoustic correlate of glottal excitation, was calculated as the difference in dB between the amplitude of the first harmonic (H1) and the amplitude of the second formant (A2) (cf. van Sluijter & van Heuven, 1996). Only [a]-like sounds were considered: [ʌ] in "Two Clubs", the [a]-portion of the diphthong in "Pentland's Flyer", and [ɑ] in "Hot 'n' Saucy" (Australian accent).

The same vocalic portions were used to determine intensity in dB.

5.3. Results

Filled pauses

One of the most typical phenomena of spontaneous speech is the filled pause. For all three speakers filled pauses were observed, though with differences in the number of instances (see table 5.1). Apparently, commentator 1 has a lower level of fluency compared to the commentator of race 3. The fact that all speakers show some form of dysfluency (despite their being professional speakers) can be seen as an indicator of spontaneous communication, and therefore as evidence for the naturalness of the speaking situation.

Table 5.1: Number of filled pauses and their locations with reference to the inter-pause stretch (ips). Fillers within ips ("ips-mid") usually occurred at syntactic phrase boundaries.

speaker	**ips-onset**	**ips-mid**	**ips-offset**	**total**
1	6	5	-	= 11
2	2	3	1	= 6
3	1	1	-	= 2

Duration of pauses and phrases

In general, the data (table 5.2) show that, as the race progresses, the average pause length does not get shorter in all cases. However, the inter-pause stretches do get shorter, and a parallel reduction in the inter-breath stretches reflects the increased breath rate towards the end. The corollary of this trend is, of course, more pauses per time unit as the race progresses.

Exceptions to the trend are, however, the central portion of race 1, which has longer pauses than in the beginning and end phases, and of race 2, which has both longer pauses and longer inter-pause stretches.

From table 5.2 it is evident that the shortest pause durations occur in the final part. This is in agreement with the expected tempo increase towards the end, because one strategy for achieving a higher tempo is to shorten the pauses. From the overall HRC pattern we might propose a general rule like "the later in the race, the shorter the pauses". This idea, however, is spoiled by the pause durations in the middle part of race 1 and 2 which are longer than those for the initial part.

Table 5.2: Average duration in seconds of unfilled pauses, inter-pause stretches (articulation phase between two unfilled pauses), inter-breath stretches (articulation phase between two breath pauses). Phases of approx. 25-32 seconds at the beginning, middle, and end of the race were investigated.

	race	**beg**	**mid**	**end**
pause	1	0.348	0.408	0.279
	2	0.356	0.396	0.226
	3	0.635	0.495	0.263
inter-pause stretch	1	4.303	2.349	2.063
	2	2.587	4.115	2.231
	3	3.763	3.127	2.794
inter-breath stretch	1	4.303	3.523	3.438
	2	3.880	6.173	3.069
	3	4.390	4.020	3.841

Another way of speeding up is to reduce the number of pauses, which results in longer inter-pause stretches. The expectation of "the later in the race, the fewer the pauses, the longer the inter-pause stretches" is supported by no speaker, except in race 2, again in the middle part.

The shortened inter-breath stretches in the final phase of each race points to an increase in air-flow during speech towards the end.

Global tempo

Neither for speaking rate nor for articulation rate can the expected pattern of speeding up during the commentary be confirmed (see table 5.3). For the three parts of each of the race commentaries the middle part deviates from the other two, either as the slowest phase (race 1), or as fastest phase (race 2 and 3). Even if we compare only the beginning and the final parts, only two out of six measurements show a slight increase in the syllabic rate. Obviously the global tempo measured in syllables per second does not reflect the pattern described in Barry (1995). Other factors must be responsible for the impression of gradual acceleration.

Table 5.3: Speaking rate (SR, including pauses) in syllables per second for three different phrases. Articulation rate (AR, excluding breath and silent pauses) in syllables per second for three different phrases.

	race 1		race 2		race 3	
part	**SR**	**AR**	**SR**	**AR**	**SR**	**AR**
beg	4.61	4.92	4.28	4.81	3.91	4.48
mid	3.98	4.59	4.54	4.90	4.08	4.66
end	4.58	5.14	4.10	4.48	4.08	4.43

Fundamental frequency level

Figure 5.2 demonstrates the successive build-up of tension reflected in the fundamental frequency in all three races investigated.

The most extreme example in race 2 reveals a difference of 15.3 semitones between the lowest token at the beginning and the highest token at the finish, i.e. one and a quarter octaves. Almost one octave difference (11.5 semitones) can be observed for race 3, and over half an octave (7.5 semitones) for race 1.

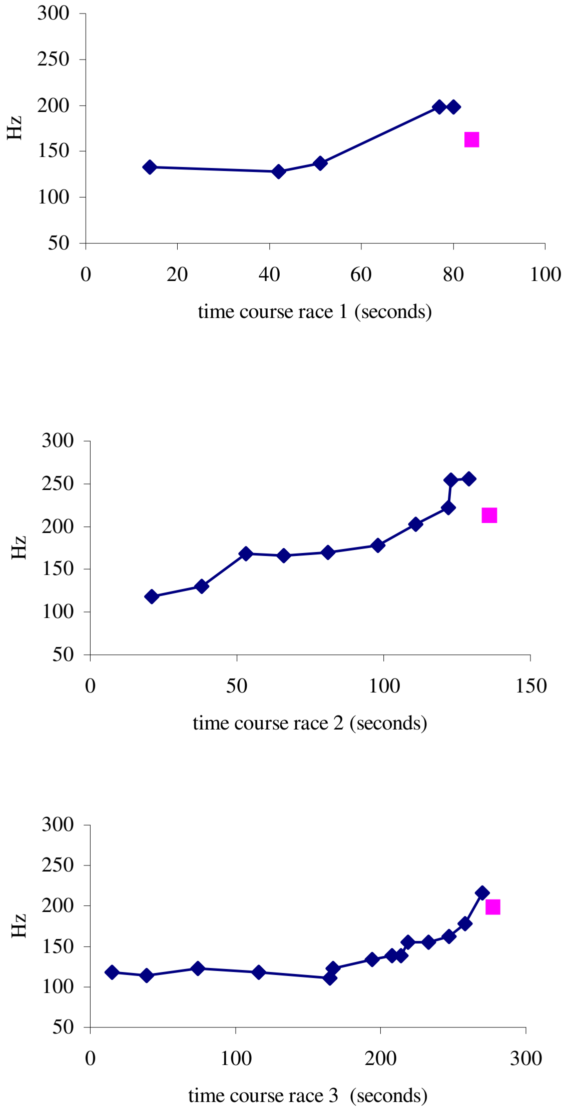


Figure 5.2: The F0 course (in Hz) at the occurrences of same words (horse name) for the three races. Values taken at the second stressed syllable are marked on the time scale (x-axis). In each graph the rightmost value is the one taken directly after the finish.

Fundamental frequency range

The F0 range, i.e. the difference between the highest and the lowest F0 value found in the examined vowels, is expected to be narrowed. We see here a variation as a function of speakers, position in the race, position in the intonational phrase, and length of the investigated vowel. Speaker 1 gradually reduces the range from 9 to 0 Hz in the five examples of [ʌ] in "Clubs". Speaker 3 also shows a gradual reduction of the F0 range from 25 Hz to 5 Hz in a phonemically long [ɔː] in "Saucy". However, in the middle part small ranges (< 10 Hz) alternate with larger ranges (> 10 Hz). This pattern of alternation in the middle phase is similar for speaker 2, who shows more extreme ranges (up to 57 Hz) in the phonemically long diphthong [αI], or [αI↔], in "Flyer". Additionally, speaker 2 widens the range again towards the very end.

As far as it is possible to interpret these numbers, one can say that F0 range is not homogeneous across the race nor across the speakers. In this sense the narrowing of the F0 range described by Barry (1995) can be confirmed for speaker 1, but only partially confirmed for speakers 2 and 3.

Fundamental frequency movements

Regarding F0 movements on "nuclear" tone occurrences of the horses' names, race 1 maintains a fall in the first two occurrences, but shows quite a level tone in the remainder. This pattern of starting with a fall, and ending with a level tone is repeated in race 3. Race 2, however, starts with a fall on "Flyer" and changes to a rise pattern or, alternatively, a rise-fall pattern for the finishing part. We see here, that the default falling contour (a high-low pitch accent) alternates either with a level tone (monotonal pitch accent) or a rising contour (a low-high tone). It is debatable whether this alternation of pitch accent patterns is a change in the phonological structure (choice of tonal accent with a linguistic function) or a phonetic realisational phenomenon.

Overall intensity

Since F0 and Sound Pressure Level (SPL) are known to correlate with sub-glottal pressure (Ladefoged, 1967), an increase in intensity may be expected to accompany the observed rise in F0 over the course of the HRC. The auditory observation of increased loudness is of course a further reason to expect this. Table 5.4 confirms these expectations, if we ignore one outlier. All three speakers have in common that the intensity is lowered after the finish. This can be seen as a reduction of sub-glottal pressure, an indication of relaxing tension.

Table 5.4: Overall intensity and spectral tilt in dB at occurrences of same words (horse name) for the three races. Mean values from the first third (beg), second third (mid), last third (end) of each race are given. Numbers of occurrences (#) for each phase are indicated in the first rows. The value taken directly before, and the one directly after the finish are listed separately.

	race	**beg**	**mid**	**end**	**last before**	**first after**
no. of occur-rences	1	1	2	2		
	2	3	3	4		
	3	3	3	8		
overall intensity	1	75.6	74.9	76.4	77.1	76.7
	2	73.5	75.7	76.5	76.9	76.0
	3	72.7	74.0	75.0	74.8	73.7
spectral tilt	1	+9.8	+6.1	+0.5	-2.6	+1.9
	2	-0.3	-10.1	-13.0	-16.0	-22.0
	3	-9.6	-10.7	-11.7s	-14.6	-9.6

Spectral tilt

If we think of a gradual inclination of tension during the race, this tension should also be transmitted to the vocal folds. As a consequence, the excitation of the glottal pulse should be strengthened. Decreased spectral tilt is accepted as a correlate of the increased effort associated with accented vs. unaccented syllables (van Sluijter & van Heuven, 1996). Since the increased effort apparent in the raised F0 and increased intensity during the HRC can be seen as the result of a similar - though longer term – physiological adjustment of e.g. sub-glottal pressure, a decrease in spectral tilt can be expected as the race progresses.

In table 5.4 the tendency for the spectral tilt to decrease is clearly visible for all speakers. For speakers 1 and 3 the increasing tilt directly after the finish can be interpreted as a relaxation of tension. In contrast, speaker 2, however, seems to remain and even to strengthen this tension.

5.3. Discussion of the analysis

HRC characteristics

The general pattern of increasing excitement during the course of a horse race, with the climax at the finish, followed by a rapid relaxation after the finish is clearly reflected in the following prosodic properties (compare table 5.5): more pauses, shorter pauses, higher breathing rate, much higher F0, higher intensity, flatter spectrum.

The direct dependency of these parameter trends on the race development (and the development of tension) is clearly reflected in the occasional abrupt up-step of F0 and other values when there is a sudden and unexpected event (e.g. race 2, 3rd and 9th mention of "Pentland's Flyer").

However, despite the overall similarity of pattern in F0 and effort (intensity level), the individual levels and ranges remain a characteristic of the individual commentator.

Prosody of excitement

In some respects a HRC shares the features of other forms of excitement. Table 5.5 lists the general trends of general emotional classes which can be subsumed under the label "excitement". The values in the table are taken from similar tables in Banse & Scherer (1996), van Beezoyen (1984), Murray & Arnott (1993). There is clear agreement among the studies regarding the general tendencies of the prosodic realisations.

The differences in the temporal parameters must be seen against the length of the investigated speech samples. It seems to be a popular paradigm in emotional speech research to study one-sentence utterances. This explains why pausing behaviour is not taken into account in the tables in Cowan (1936), van Beezooyen (1984) and Banse & Scherer (1996). Cowan (1936), however, remarks that fewer and shorter pauses occur in emotional speech, whereas we found *more* shorter pauses. This difference in results show that pausing strategies, and especially breathing, might be aspects worth studying in an emotional context, as well as tempo.

Table 5.5: Changes in prosodic parameters compared to neutral speech for different forms of excited speech: last phase of horse race commentaries (HRC), based on this study; anger, joy, surprise, fear, summarised from similar tables in Banse & Scherer (1996), van Beezoyen (1984), Murray & Arnott (1993). '+' = increase; '++' = large increase; '-' = decrease; '±' = unclear; empty cell = unknown.

prosodic parameter	**HRC**	**Anger**	**Joy**	**Surprise**	**Fear**
pausing rate	+				
breathing rate	+				
pause duration	-				
tempo	±	+	±	±	+
F0 level	++	++	+	+	+
F0 range	±	++	++	+	±
intensity	+	+	+	+	±
spectral tilt	-				

Unanimous agreement exists in terms of the increased average F0, whereas the F0 range seems greatly increased for anger and joy, less increased for surprise, but unclear for fear and HRC.

Probably due to the increased F0 level one can find the increase in overall intensity in all classes of emotions, except for the unclear case of fear.

In the study of Banse & Scherer (1996), a number of other parameters relatable to spectral tilt were investigated, but the results are not directly interpretable in the framework of the present study. Banse & Scherer (1996) show that the relative amount of high to low frequency energy varies with the expression of *different* emotions. However, they do not address the question of the degree to which a single emotion category is reflected in the strength of the parameter.

In the present study we can maintain that the activity level is clearly reflected in the prosodic patterns discussed. However, as Murray & Arnott (1993) noted, active forms of emotional speech show strong tendencies to be confused with others. Thus, it is no surprise that HRC fits well with the prosodic characteristics of high activity level.

It must be conceded that a differentiated view on an emotion such as "anger", which can be further subdivided in hot anger or rage, cold anger, threat, frustration or further nuances, would show differences in the extent of the use of prosodic parameters.

Learned vs. natural excitement

HRC as a product of an individually developed, public oriented, professional speaking style leads to the important question: although the data are from the "real" world, is the excitement in a HRC "real" in the sense of spontaneous and natural?

Although HRC recordings are considered as "spontaneous" speech, because it is unscripted, there is definitely a certain degree of routine, and therefore a certain lack of spontaneity:

- As shown above, the frame and the period of "getting prosodically excited" is given.
- The presumed expectations of the listeners are present. A non-excited commentator would probably be classified as "boring" and hence "non-professional".
- The usual acceptance of being excited is shifted: the commentator is allowed to show a degree of excitement which would be seen as exaggerated in a "default" situation.

The higher breathing rate indicates a higher level of physiological activity. However, a higher breathing rate can be explained by an increased phonatory and pulmonic effort which is reflected in the higher values for spectral tilt and intensity, respectively. It remains speculation whether this special "shouting"-style is a consciously controlled and/or a trained behaviour, or whether it represents a natural verbal manifestation of the excitement which can be observed in other people towards the finish of the race as well.

Multiple functions of prosody

It is a truism that prosody has functions on various levels of speech communication: linguistically, paralinguistically, and extra-linguistically.

We are still a long way from being able to exploit this knowledge. In the area of speech synthesis, attention has been shifted from "intelligibility" to "naturalness". In the context of emotions this means e.g. overcoming the boredom of artificial voices (Trouvain, 2000), or an explicit modelling of certain emotion types (e.g. Cahn, 1989; Murray, Arnott & Rohwer, 1996; Schröder, 1999).

In our opinion, it is important to have a theoretical framework such as the linguistic - paralinguistic - extralinguistic distinction (Crystal, 1969), and to fill it with data-supported life, in order to identify what kind of information a speaker transfers to the listener by prosodic means. An example from the HRCs shows the semantic/pragmatic content located at the paralinguistic level. A HRC tends to have a rather low sound segment or word based information value, but a relatively high prosody based information value. E.g. we assume that the placement of a horse (as it appears in the example in section 3 can be (partially) decoded by prosody, if the listeners know the prosodic context of the commentary before the climax. A further indication supporting the prime importance of prosody is the fact that some words are not comprehensible (even to phonetically trained native listeners).

We conclude that a more detailed investigation in prosody can help to model specific speaking styles such as HRC, and to model specific emotions such as excitement and its derivatives.

Summary and discussion of chapter 5

Tempo variation in horse race commentaries is clearly reflected in changes of pausing behaviour. This was expected on the basis of the general statement that tempo variation is primarily variation in pausing (Goldman Eisler, 1968). However, this case study has provided evidence against the generalisation that speeding up is characterised by fewer and shorter pauses. There are more pauses in the auditorily faster last bit of those commentaries, and an important influence of breathing and average pitch level. This study allows a differentiated view on tempo stressing the role of breathing, a parameter not frequently taken into account in emotional speech research.

In view of the sub-title of this thesis, the following question is allowed: what is the use of an analysis of a special emotive and spontaneous speaking style for *text*-to-speech synthesis? First, horse race commentaries are generated with speech synthesis in some cultures, e.g. in Japan (Campbell, personal communication). Second, the results of this case study confirm the importance of pauses for analysing tempo in particular, and prosody in general. This means for speech synthesis that tempo modelling is first of all pause modelling. Third, one of the future applications of synthetic speech is to model emotive speech. Here, tempo and pausing are essential prosodic parameters along the activation axis.

Chapter 6

Production Experiments: Effects of Tempo on Phonological Structure

Introduction

Usually, natural speech occurs as spontaneous face-to-face dialogue. Reading texts aloud is normally tied to specific sorts of speech, e.g. reading newpapers articles or letters to somebody in the room, reading books aloud to children, many types of broadcasted speech (radio and television news, teleprompted speech, weather forecast), many forms of aesthetic communication (radio play, read novels, recitation of poems), or other forms of public-oriented speeches (press conferences, official speeches, judges' verdicts, prayers, sermons, university lectures). For experimental studies performed under laboratory conditions, read speech has the advantage that different versions of one text are more comparable in contrast to different examples of spontaneous speech. Since one of the aims of the study is to explore implications for *text*-to-speech synthesis systems it seems legitimate to examine the production of read rather than spontaneous speech.

We report two production experiments here in which we asked people to read texts of paragraph length at three different rates "medium", "slow", and "fast". In both experiments we ascertained speaking and articulation rate, mean pause duration as well as the number of pauses and the number of prosodic phrases. In experiment 1 we focus on pausing structure, phrase structure and segmental reductions, whereas in the second experiment (which has also been reported in Trouvain & Grice, 1999) the focus lies on phrasing and tonal structure.

Apart from the question whether speakers make use of the various possibilities for phonological rearrangement, the analysis and the interpretation of the results are discussed against the background of homogeneity and symmetry assumptions versus individual strategies.

We summarise the questions as follows:

- Are the strategies observed at the various levels of analysis used by all speakers?
- Are the strategies that are used for speeding up also used to a similar extent in reverse for slowing down?
- Is it possible to sketch an individual tempo profile at the phonological level?

6.1. Production Experiment 1

Methods

In the first experiment three female native speakers of German recorded three readings of a five-sentence newspaper article (see Appendix for text) at three self-selected rates, "normal", "fast", and "slow". This resulted in 27 versions for analysis (3 speakers x 3 rates x 3 readings). Each cycle of readings started with "normal" followed by "slow", and finished with "fast". The speakers (labelled *speaker 1, 2, 3*) come from different dialect regions (Moselfränkisch, Badisch, Schwäbisch). Although no one showed obvious dialectal peculiarities at the segmental level, regional influences cannot be completely excluded.

For each version the total duration (in ms) was measured, as well as the durations of pauses longer than 100 ms (cf. Butcher, 1981), a threshold that was extended to 150 ms when the pause was followed by a stop consonant. Closure durations in post-pausal positions were counted as part of the total pausing time, however. The speaking rate (*including* pauses) and the articulation rate (*excluding* pauses) are calculated as a function of the number of phonological syllables (which is the same for all versions). For several reasons it was decided to measure the tempo in phonological syllables rather than in sounds although the discussion in chapter 4 has shown that the sound segment, either as realised phone or as phoneme, seems to be the most favourable unit in tempo measurement. First, there are the advantages of the syllable compared to the sound segment: easier counting, easier definition and higher degree of popularity and therefore a higher degree of comparability across studies. Second, there are the advantages of counting intended forms rather than realised ones in terms of reliable identification and, for the sound segment, acceptable definition.

For the second of the five sentences in the text (arbitrarily chosen), phrase breaks were transcribed impressionistically by the author, allowing for a three-level distinction (0 = no break, 1 = minor break, 2 = major break).

As an illustration of segmental processes, a transcription of a short excerpt from the first sentence (consisting of 4 words) is given.

Results

Speaking rate and articulation rate

In neither representation of the global rate (see table 6.1) is there any overlap of the three tempo categories between speakers. That means that, across our subjects, the realised rate categories correspond to the intended ones.

Table 6.1. Mean values (in phonological syllables per second) for speaking and articulation rate for the three speakers for each of the intended rate.

	speaking rate			**articulation rate**		
speaker	**S1**	**S2**	**S3**	**S1**	**S2**	**S3**
fast	5.54	6.12	6.75	6.06	6.49	7.43
normal	4.33	4.84	4.98	4.81	5.42	5.68
slow	3.44	3.80	3.55	3.90	4.49	4.13

The differences between the speakers are seen in the mean values for their normal tempo and in the values expressing their tempo range. Speaker 3 prefers a higher speed than the others, and she also shows a wider range.

Pausing

All three subjects have a greater increase in the number of pauses from "normal" to "slow" than from "fast" to "normal". The greatest increase is found for speaker 2, followed by speaker 3, and then by speaker 1 (see table 6.2).

All speakers also show a homogeneous picture in terms of mean pause duration: the slower they speak, the longer the pauses. Nonetheless, the differences for slowing down are smaller than those for speeding up.

Table 6.2. Distribution of all realised pauses in 100 ms bins for each of the three speakers for each rate. Maxima are in bold (see text). Mean number of pauses per version (per rate and speaker) and the mean pause durations per version are given in the last two rows.

speaker	**S1**			**S2**			**S3**		
pause dur.	**slow**	**normal**	**fast**	**slow**	**normal**	**fast**	**slow**	**normal**	**fast**
100-200	1	-	-	**8**	1	2	2	-	-
200-300	-	-	-	2	-	1	-	-	1
300-400	2	-	**6**	**4**	1	**5**	3	-	**5**
400-500	1	**4**	2	2	**5**	1	2	1	**4**
500-600	1	2	3	**4**	**6**	-	-	**4**	-
600-700	**6**	**4**	-	**5**	2	-	2	**5**	-
700-800	**5**	1	-	2	-	-	2	1	-
800-900	-	1	-	2	-	-	**5**	1	-
> 900	1	-	-	-	-	-	3	-	-
mean no.	5.3	4.0	3.7	9.7	4.7	4.0	6.3	4.0	3.7
mean dur.	666	594	428	438	424	258	664	635	379

Pauses show a great diversity in temporal extension. The distributions of pauses in bins differing in steps of 100 ms duration can illustrate some of the regularities in pause timing. In the "fast" condition all speakers produce most of their pauses in the duration group between 300 and 400 ms, with a slight tendency to shorter durations for speaker 2, and a tendency to longer durations for the other two speakers.

This tendency is continued in the "normal" pauses where speaker 3 has her maxima between 500 and 700 ms, speaker 2 between 400 and 600 ms, and there is a "bimodal" distribution for speaker 1 (400-500 ms, and 600-700 ms). Such a division is also present in the "slow" data for speaker 2: although like speakers 1 and 3 she also uses the whole range of durations, she structures her pause durations in long (500-700 ms), medium (300-400 ms) and short (100-200 ms) pauses. Regarding the "short" pause, it was striking that some silent intervals below the 100 ms threshold were observed for this speaker, though they were not taken into account.

The other two speakers also make use of the entire durational spectrum with higher value maxima, speaker 1 between 600 and 800 ms, and speaker 3 between 800 and 900 ms.

Phrasing

Changes in phrasing are illustrated with one example sentence. Table 6.3 shows the mean break strength for each potential prosodic boundary of the second sentence: *Nach Auskunft* (A) *der Polizei* (B) *war der Junge* (C) *bei einer Klettertour* (D) *an einem Steilhang* (E) *ausgerutscht.* (F) *Im Fall ...*

It can easily be seen that the principle generally holds: the faster the rate, the lower the break level. Exceptions are two cases where "normal" has a slightly higher mean level than "slow" (S1, B, and S2, C), and the end-of-sentence break (F) which remains constant (except for S2 "fast").

As expected, the pause duration for breaks of the same level decreases with increasing rate (break F). However, we can see different pause durations for comparable boundaries not only across rates, but also within a rate category. The pause durations associated with the two transcribed major breaks of the "slow" versions of speaker 2 differ considerably (176 ms vs. 687 ms). The same is true for speaker 1. Her major breaks for "normal" speaking rate are realised at location F with a rather long pause, and at location B with no pause at all. Further examples can also be found for minor breaks.

Table 6.3. Means within versions of transcribed break strength (0 = no break, 1 = minor break, 2 = major break) for each potential break location. Breakdown per rate (slow, normal, fast) and for each speaker (S1, S2, S3). Numbers after the slash give the mean pause duration (in ms). If no mean pause duration is indicated, no pause had been produced.

	S1			**S2**			**S3**		
	slow	**normal**	**fast**	**slow**	**normal**	**fast**	**slow**	**normal**	**fast**
A	0.7	0.3	-	0.7	0.7	-	0.7	-	-
B	1.7/143	2.0	0.7	2.0/176	1.7	0.3	1.7/124	1.3	0.3
C	1.0	0.7	-	1.0	1.3	-	1.0	-	-
D	1.0	1.0	-	1.0/171	0.3	-	1.7/114	0.3	0.3
E	1.0/038	0.3	-	0.3/071	0.3	-	1.3	-	-
F	2.0/740	2.0/609	2.0/448	2.0/687	2.0/531	1.7/140	2.0/910	2.0/665	2.0/346

If the values for mean break strength and mean pause durations (table 6.3) are averaged across the three speakers (figure 6.1), the following features of the non-linear nature of changes of speech rate in terms of phrasing and pausing can be noted:

- pause duration of breaks of the same strength and of the same location decreases non-linearly from slow to fast (cf. break F)
- some breaks reduce in strength whereas others keep the same strength from slow to fast (cf. breaks A-E vs. F)
- higher level breaks are not necessarily marked by a pause, whereas lower level breaks can be marked by a pause at the same rate (cf. break B vs. E at slow)
- mean break strength correlates non-lineraly with pause duration (cf. break F vs. B vs. D)

The non-uniform changes of the break strength at the same break location across the rates as well as the considerably different pause duration for the breaks of the same level across the rates demonstrate the non-linearity of how tempo change is achieved with respect to phrasing.

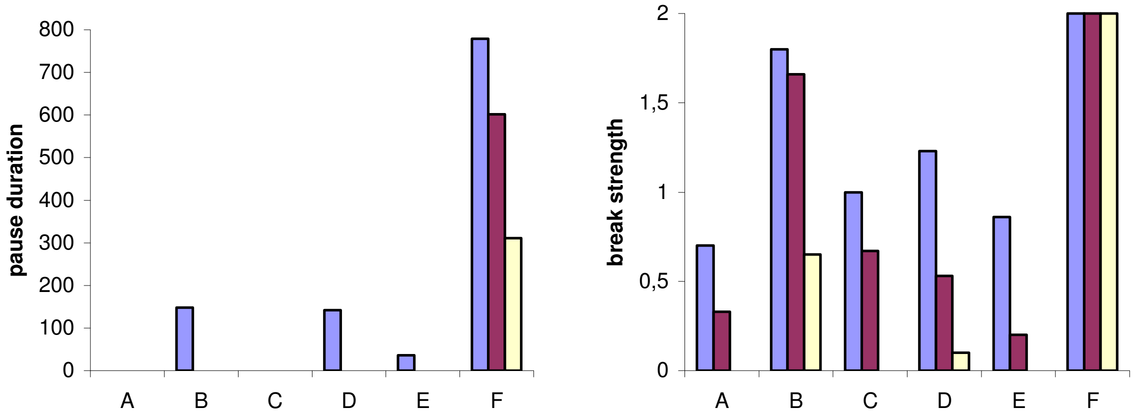


Figure 6.1. Mean pause duration and mean break strength for each break location averaged across the three speakers for each rate category (slow=bright, left; normal=dark, central; fast=white, right).

Segmental reductions

Many reduction processes can be observed in connected speech and they are well described for German (e.g. Kohler, 1990) as well as for other languages as presented in chapter 3. The aim of the analysis in this experiment is to check whether processes apply in a consistent way along speech rates and across speakers. Thus, segmental reduction processes are shown in the following example. The word sequence *hat am Morgen einen* from the first sentence (word-by-word translation: a 16 year old *has in the morning an* 80 meter fall ... survived) has as its phonemic form:

/ h a t a m m ɔ r - g ə n aɪ - n ə n /

After the application of several realisation rules such as aspiration of fortis stops, glottal stop insertion before vowels, degemination, and r-vocalisation, one could predict the following phonetic form for clear and slow speech:

[h a th ʔ a m ɔɐ g ə n ʔ aɪ n ə n]

One location within this word sequence was selected to describe various phonological processes of connected speech that can apply as reductions from the predicted form. In this example we consider the phoneme sequence at the boundary between the words *ha*_**t a**_*m*. Three different processes can occur:

- Omission (or non-realisation) of aspiration [h a th] -> [h a t]
- Omission (or non-insertion) of glottal stop [ʔ a m] -> [a m]
- Lenition of fortis plosive (with omission/non-insertion of glottal stop) [h a th ʔ a m] -> [h a d a m]

Four different versions of the realisation of this bi-phonemic combination are possible. The level of reduction can be expressed by the number of missing or changed phonetic segments compared to the predicted slow/clear form:

- level 0: [th ʔ a]
- level 1: [t ʔ a] or [th a]
- level 2: [t a]
- level 3: [d a].

In table 6.4 one can see that all possibilities are indeed used and that the speakers reduce more at higher rates. An exception in this respect is speaker 1 who always uses the same forms for "normal" and "fast". Apart from this exception, there is no case where a speaker always uses one form for one rate category. This is particularly true for the slow versions, where we would expect a careful and precise articulation. But only three out of nine realisations correspond to the predicted slow/clear form.

Table 6.4. Frequencies of realisations of "ha**t** **a**m Morgen". The level of reduction (from 0 to 3) depends on the number of the phonological processes applied to a quasi-canonical form. Each speaker (S1, S2, S3) produced 3 versions at each rate (slow, normal, fast).

		S1			**S2**			**S3**		
reduction		**S**	**N**	**F**	**S**	**N**	**F**	**S**	**N**	**F**
0	t h ʔ a	1	-	-	2	-	-	-	-	-
1	t ʔ a	2	-	-	-	2	-	2	1	-
1	t h a	-	-	-	1	-	1	-	-	-
2	t a	-	3	3	-	1	2	1	1	1
3	d a	-	-	-	-	-	-	-	1	2

Several reduced forms are chosen for more than one rate, e.g. [t a] for all speakers, and in case of speaker 3 for all rates. So, a specific reduced form seems not to be bound to a specific rate, and conversely a specific rate seems not to be bound to a specific reduced form and the processes which generate it. This is shown in the fact that speaker 1 is the only one not to use different forms for any given rate.

From the individuality point of view it can be said that speaker 3 has the strongest propensity for reduction in this example, followed by speaker 1, and speaker 2 with the least reduced productions.

A similar pattern can be reported from another example in the phrase: *Mor**gen** **ei**nen*. Here, four processes can be employed: omission (or non-realisation) of glottal stop, schwa elision, place assimilation of nasal, and deletion of homorganic plosive. These processes result in six possible forms ranging from [gənʔaɪ] > [gənaɪ] / [gŋʔaɪ] > [gŋaɪ] / [ŋʔaɪ] > to [ŋaɪ]. Each speaker used the most careful pronunciation only once and the degree of reduction increases gradually across the rates, but the type of reduction is not confined to one tempo category.

Summary Experiment 1

In general, the assumptions about the phonological mechanisms of speeding up and slowing down presented in chapter 3 were confirmed in the experiment. After making sure that the speakers were able to match the intended rate categories "fast", "normal" and "slow" temporally, it has been shown that pause timing works as expected: the slower the speed, the longer and the more frequent the pauses, and vice versa. Both pausing features become evident in the mean values as well as in the overview with the temporal distribution of pauses (table 6.2).

Similarly in the case of phrasing, which is illustrated on the basis of one example sentence: the slower (or faster) the speaking rate, the more (or less) prosodic breaks and the higher (or lower) the break level.

For the segmental reduction phenomena, too, the expectations were fulfilled on a general level: the faster the speech, the more reduced forms are selected.

But these observations can neither express the degree of generalisation nor individual tendencies. Even if we can say something general about rate and reduced forms, that does not mean that at a slow rate *in general*, i.e. in the majority of the cases, the most precise form is produced, and for fast rate the most reduced one. There seems to be a scope for variation, which sometimes results in individual patterns such as the three-fold distinction of the pauses for slow speech of speaker 2.

Another example of the restricted value of a general statement is the assumption that higher level boundaries are associated with longer pauses, and lower level boundaries with shorter or even absence of pauses. This general hypothesis was confirmed here. Nevertheless, pause duration is not tied to a certain break level: speakers select long pause durations as well as short pause durations for the same break level within a certain rate, and they make differences in pause duration for the same break level across rates. A more differentiated analysis of the break levels, as in De Pijper & Sanderman (1994), might lead to a better correlation between break level and pause duration.

6.2. Production Experiment 2

Methods

For the second experiment, three female native speakers of German recorded three readings of the German version of "The North Wind and the Sun" (see Appendix for text) at three self-selected rates, "normal", "fast", and "slow". This resulted in 27 versions for analysis (3 speakers x 3 rates x 3 readings). The experiment is described in detail in Trouvain & Grice (1999).

The procedure is the same as for experiment 1 with the following exceptions: the phrase breaks are transcribed by two labellers (one was the author). To illustrate the change of boundary strength an index with three levels was defined for each reading: a shift from major to minor boundary would involve a -1 step, a shift from no boundary to a major boundary is +2, and so on. All steps are summed to give a cumulative shift value. Furthermore, transcribed pitch accents are divided into two groups, bitonal and monotonal ones.

All speakers (labelled *speaker 1, 2, 3*) stem from southwest German dialect regions (Saarbrücken (=Rheinfränkisch) & Badisch). Again, no one showed obvious dialectal peculiarities at the segmental level. Regional influences, especially in intonation, cannot be excluded. Speaker 2 also participated in the previous experiment.

Results

Speaking and articulation rate

Table 6.5 shows the results for the rate characteristics which are similar for both measurements, speaking rate and articulation rate: speakers 1 and 2 make clear differences between the three rates whereas for speaker 3 the difference between "fast" and "normal" is only small.

Table 6.5. Mean values (in phonological syllables per second) for speaking and articulation rate for the three speakers for each of the intended rate.

	speaking rate			**articulation rate**		
speaker	**S1**	**S2**	**S3**	**S1**	**S2**	**S3**
fast	5.51	6.27	4.39	6.33	6.93	5.19
normal	4.84	4.91	4.26	5.60	5.85	5.11
slow	4.30	4.14	3.48	5.14	5.08	4.58

Unlike experiment 1, the speakers in this experiment do not form a homogenous group with regard to their speech rate categories. Speaker 3 is generally slower for all three categories and also has a smaller range between "slow" and "fast". Her "fast" category is almost as slow as the "slow" versions of the other two speakers.

Pausing

No difference is observed in the number of pauses for speaker 3 between "fast" and "normal" (see table 6.6). However, with pause duration a different picture emerges (also in table 6.6). All three speakers distinguish the three rates in terms of pause duration, though not in the same way. Speakers 3 and 2 both increase pause duration as rate decreases while speaker 1 does the opposite, lengthening the average pause duration as she increases the rate, though she reduces the number.

Table 6.6. Mean number of pauses and mean pause duration in ms.

	number of pauses			**pause duration**		
speaker	**S1**	**S2**	**S3**	**S1**	**S2**	**S3**
fast	6.7	6.0	13.3	646	465	475
normal	8.7	11.3	13.0	592	533	548
slow	11.7	13.3	17.0	583	608	772

Phrasing

Regarding the number of transcribed breaks (see table 6.7) speaker 1 & 2 make distinctions between the three rates, although they do this to a different extent. Speaker 3 again makes no distinction between "fast" and "normal", but we can see a clear difference between "slow" and "normal".

Table 6.7. Mean number of transcribed prosodic breaks.

	number of breaks		
speaker	**S1**	**S2**	**S3**
fast	18.0	15.3	19.7
normal	19.7	18.7	20.3
slow	20.7	21.3	26.4

The summing of all boundary strength steps shows that speaker 2 demotes phrases for speeding up and promotes phrases for slowing down (figure 6.2). Speaker 1 only applies demotion for speeding up, and speaker 3 only uses promotion for slowing down (the sum of the break indices are equal for "fast" and "normal").

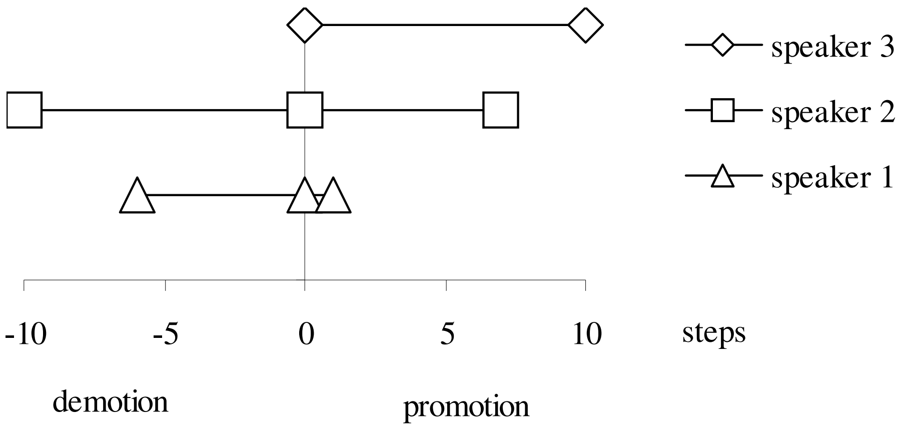


Figure 6.2. "Promotion" and "demotion" of prosodic boundaries taken for each speaker separately, comparing normal rate to fast and normal to slow. Break index score changes are calculated in steps.

Pitch accents

As stated in chapter 3, pitch accents can be expected to be reduced in number and complexity. The results of the pitch accent analysis show considerable variation (see figure 6.3). Speaker 1 has almost no change from "normal" to "slow" and no change in the total number of accents from "normal" to "fast", but a considerable reduction in the number of bitonal accents. Speaker 2 increases and decreases, respectively, both the total number of accents and the number and proportions of bitonal accents from "normal" to "slow" and from "normal" to "fast". Speaker 3 shows the same pattern of increase in total accent number, but there is no change in the number of bitonal accents with rate, resulting in a reverse pattern in the proportion of bitonal accents fast > normal > slow.

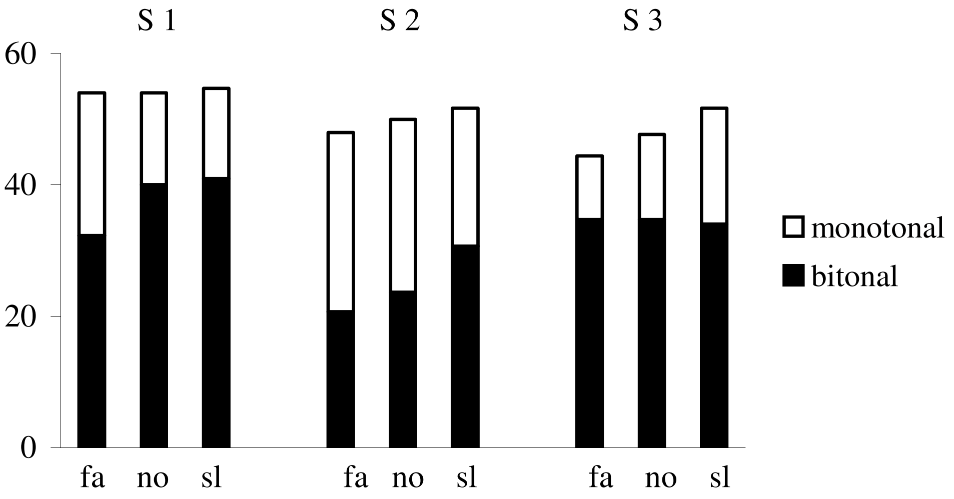


Figure 6.3. Distribution of pitch accents for all realisations. Pitch accents are classified either as monotonal or bitonal.

Summary Production Experiment 2

In the second experiment we checked the consistency of the results from the first experiment with respect to articulation and speaking rate as well as mean number and mean duration of pauses. Although the same general patterns were found in experiment 2, there are some interesting differences. Speakers differed more strongly in their choice of strategies, and also in the implementation of these strategies. Even within one speaker (speaker 2 participated in both experiments) we find a change in magnitudes for the examined parameters between experiments.

In experiment 2 phrasing was investigated in terms of number of breaks and the change of break level, whereas in experiment 1 the break locations and their reflection in pause duration was explored. Additionally, the tonal pattern was analysed in the second experiment.

The second experiment reveals individual patterns of change rather than general tendencies such as those in the first experiment. Speaker 2 conforms to the patterns of change regarding the number of pauses, the number of phrases, the number of pitch accents (especially the number of bitonal ones), the pause duration, and the promotion and demotion of phrase boundaries, that are expected on theoretical grounds. The other two are inconsistent in several of the analysed properties: Speaker 1 with respect to a) mean pause duration, b) promotion of phrases, and c) number of pitch accents.

Speaker 3 concerning a) articulation and speaking rate, b) mean pause duration, and c) demotion of phrases.

6.3. Discussion of experiments 1 + 2

Pausing strategies

An expectation from many studies of speech rate and pause behaviour would be that most speakers would make maximal use of pausing mechanisms, namely reducing number and duration of pauses for speeding up, and increasing number and duration of pauses for slowing down. From a brief analysis of pausing strategies across different studies and different languages it becomes clear that this is obviously not the case and that speakers use pausing mechanisms only partially as illustrated in table 3.1 (chapter 3, p. 26). Table 6.8 summarises all the main findings from the above mentioned studies with the result that there is virtually *no* tendency to use *all* possible pausing mechanisms to change tempo. This picture is also mirrored in the strategies of the six speakers described in Trouvain (1999) where only one speaker comes to the maximum exploitation of pausing changes.

Table 6.8: increase (>), decrease (<), or no difference (=) to normal speeded tempo regarding number and duration of pauses.

	slow		fast	
	number	duration	number	duration.
Exp. 1 S1	>	>	=	<
S2	>	=	<	<
S3	>	≥	=	<
Exp. 2 S1	>	=	<	> (sic!)
S2	>	>	<	<
S3	>	>	=	<

Phonological reorganisation

In Ladd et al. (1999) the rather general observations in investigations of the effects of speech rate on intonation were criticised:

> "Relative number of prosodic boundaries and relative F0 level are global properties of contours, and it is therefore difficult to incorporate the findings (...) into a quantitative model. In particular, knowing about such global effects is of little or no use for predicting or modeling the effect of changes in speech rate on the detailed course of F0 in individual pitch accents."

The same criticism can be applied to the more global analysis of the experiments presented here. But it is necessary to know the overall patterns of change before details can make sense. Even though global statements about changing phonological properties cannot predict the final shape of phonetic parameters, they can help to model the changed phonological frame which forms the basis for predicting phonetic parameter values.

It is evident that speech rate affects the phonological structure in such a way that, on this basis alone, segment and syllable durations are changed: de-accentuation results in a lack of accentual lengthening; the promotion (or demotion) of a phrase boundary results in increased (or reduced) phrase-final lengthening; vowel reduction can lead to a vowel type which is reduced in its inherent duration; a degeminated consonant is shorter than two consonants, and a deleted consonant means zero duration. So, the starting-point for predicting or explaining segment durations is highly dependent on the reorganised phonological facts.

In the case of spectral reduction we can see that a (quite complex) re-structuring of the phonological frame, ultimately linked with tempo change, needs to be considered when purely phonetic properties such as spectral quality are investigated. Fast speech alone need not trigger spectral reduction, but it usually occurs together with other prosodic conditions like accentedness, or position in a prosodic phrase. These conditions are affected by tempo, however: the degree of accent can be reduced, or the length (as well as the duration) of a prosodic phrase can change.

Knowing about the re-structuring of pitch accentuation is important, whether or not we assume that different underlying tones were realised, or think that monotonal pitch accents are reduced bitonal ones.

The problem with global statements of the kind cited above is that a) they only show general tendencies, and do not capture the strategies applied by a real speaker, and b) they say something of *which* variables change, sometimes a little bit of *how* variables change, but they usually say nothing about *when and where* variables change.

This study attempts to generalise in various respects. It looks at slow *and* normal *and* fast speech, not only the fast-normal distinction, though the author is aware of the fact that these tempo categories are artificial. It looks at several segmental and suprasegmental phenomena, not only one aspect of phonology. But it also looks behind the scene of the general tendencies and tries to shed some light onto the mechanisms for achieving tempo change.

What the study does not do is to make clear the *when and where*, i.e. under which conditions exactly a modification rule is applied and to what degree. It is good to know, that, let us say, 20% of minor prosodic boundaries should be "demoted" in fast speech, but this says nothing about the exact conditions nor about which breaks are concerned. It is of course necessary to have more insight into the location of boundaries. It is a truism that "more important" boundaries are realised more elaboratedly (longer pauses, more final lengthening, boundary tones, creaky voice, ...)

Individual strategy profiles

What is clear from this study is that speakers differ in their strategies for achieving another tempo, and that these differences can be quite considerable. It also becomes evident that strategies for slowing down are not reversed speeding up strategies (see figure 6.3). This lack of homogeneity among speakers and the lack of symmetry within speakers are important features for modelling speech rate, both for a general tempo model, and for an individual model.

Of course, individual strategies have frequently been observed on various phonetic levels. E.g., in the study by Ladd et al. (1999) only certain speakers enlarge the pitch excursion size as rate slows down, and Kuehn & Moll (1976) report different preferences in terms of velocity and displacement of articulators.

Personality markers are also apparent in spontaneous speech. The type of pause fillings, the way syllables are drawled, the locations of interruptions in the speech flow, and the frequency of all kinds of dysfluencies are substantial features of an idiolect.

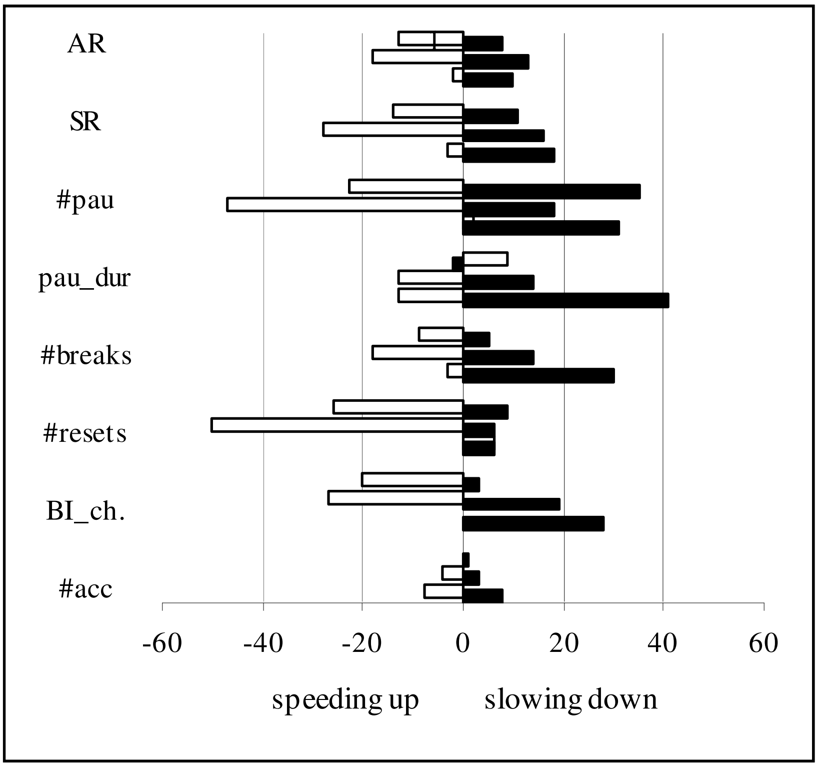


Figure 6.3. Summary of strategies for speeding up and slowing down, expressed as percentages of the normal tempo value for eight parameters. For each parameter, the values for each speaker 1, 2, 3 are given separately: 1. Articulation rate (AR), positive values here indicate fewer syll/s leading to a slower rate, negative values more syll/s; 2. Speech rate (SR), calculated as for AR; 3. Number of pauses (#pau); 4. Average pause duration (pau_dur); 5. Number of transcribed breaks at level 3 or 4 (#breaks); 6. Number of F0 topline resets (#resets); 7. Promotion or demotion of transcribed prosodic breaks (BI_ch.) calculated in steps as for table 6.7; 8. Number of pitch accents (#acc).

These are only three aspects which contribute to an individual tempo profile. Idiosyncrasies need to be considered on many levels. It is likely that strategies of phonological restructuring are important. If we think of speech synthesis applications, it is a necessary start. No matter whether one wishes to develop an individual synthetic voice, e.g. for an animated character, or to enhance the variety of speaking styles for different situations or text styles, symbolic (= phonological) input is always required.

Summary and conclusion chapter 6

The production experiments described here confirmed many points which were expected from the knowledge summarised in chapter 3, but they also revealed some details not expected in this form. Some general tendencies for speeding up and slowing down were found to apply, like segmental reduction processes, changing the number of pauses and prosodic breaks, altering the pause durations, and changing number and type of pitch accents. A closer look into individual strategies achieving tempo change shows that this general tendency does not take place for all speakers nor does it occur in the same magnitude nor in each realisation.

With regard to pause durations speakers obviously apply different classes for the three rates. But there is no cue that a general pattern of short, medium and long pauses (cf. Crystal, 1969; Butcher, 1981) holds for one of the speakers in the first experiment. Similarly, the general correlation between break strength and pause duration is questioned. Prosodic breaks of the same strength were marked by considerably higher pause durations when located between sentences compared to breaks located within sentences. Here, syntactic embeddedness plays an important role. Rules that map syntactic breaks to prosodic breaks remain only tendential, while we have to observe that, despite the same syntactic structure, the location of prosodic phrase breaks differ in *all* of the 54 versions in the two experiments. Optionality seems also to play a role in the way segmental structure is re-organised due to tempo change. Rules can describe various processes which are likely to occur but it remains unclear whether, and how systematically these many processes apply. A similar picture emerges with the prosodic processes investigated in the second experiment. Although one of the three speakers could be considered as prototypical, two of them follow their own and sometimes not very consistent ways to vary speech tempo.

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.

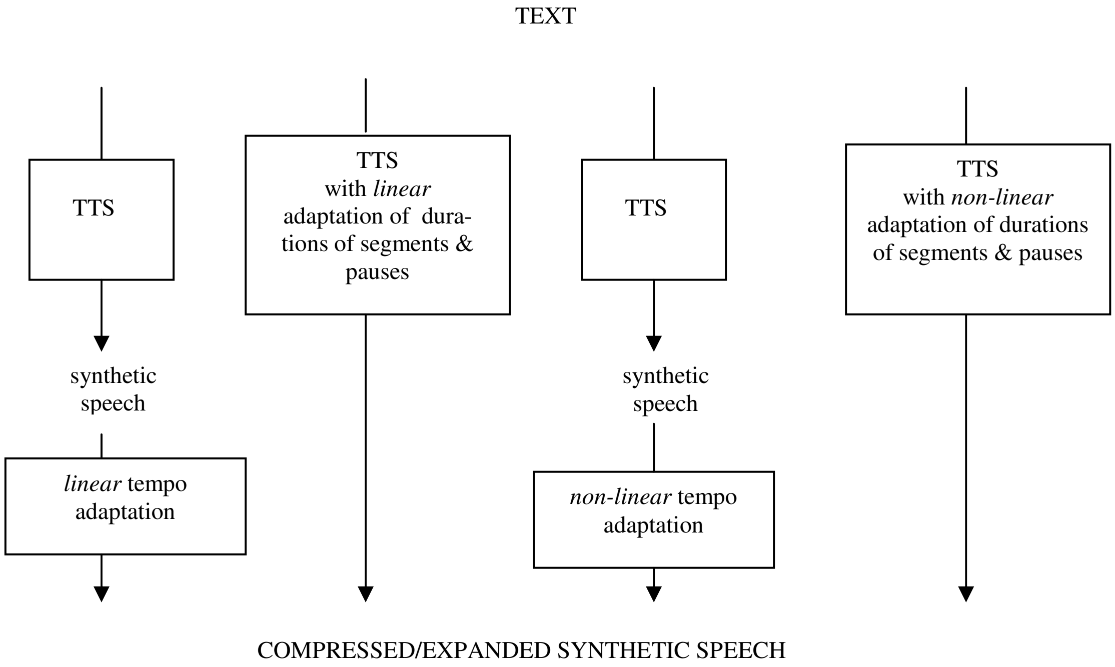

TEXT
TTS
synthetic speech
linear tempo adaptation
TTS with *linear* adaptation of durations of segments & pauses
TTS
synthetic speech
non-linear tempo adaptation
TTS with *non-linear* adaptation of durations of segments & pauses
COMPRESSED/EXPANDED SYNTHETIC SPEECH

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) 0,6 (elsewhere)	H-
"4"	comma	200		H-%, H-^H%, L-%
"6"	end of sentence	410		

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.

Chapter 8

Summary and Conclusions

8.1. Summary

Tempo variation in speech production is dependent on many variables which are not genuinely phonetic or phonological in nature. As discussed in chapter 2, these variables include extra-linguistic and para-linguistic factors like emotions, attitude, stress, age, language proficiency, speech and hearing impairments, the role of the communication partner, and habitual speech rate. There are also language-relevant factors that determine speech tempo, such as text type (written or spoken), word frequency, speech planning, discourse organisation and information management.

The purely phonetic and phonological parameters are presented on structural levels in chapter 3. It starts with the central role of pauses and prosodic phrasing for tempo, followed by sections on intonation and rhythm, in which further aspects of prosody are discussed. The section on connected speech processes deals with changes on the sound segmental level in terms of assimilations, deletions and phonemic reductions, while the section on duration attempts to give an overview of the factors which influence the durational correlates of sound segments and syllables. The chapter closes by introducing some mechanisms on the articulatory level as variables for tempo variation. It is important to note that all processes of tempo variation applying on the structural levels mentioned are non-linear in nature.

A central methodological issue is how to measure speech tempo. Chapter 4 discusses the pros and cons of the linguistic units word, syllable, and sound segment for the selection as the appropriate tempo measurement unit. Although temporal variations are best captured by measuring sound segments, there are disadvantages with this unit with respect to a clear-cut definition and ease of counting, in contrast to the (phonological) syllable and the word. However, the word shows shortcomings in terms of temporal variance and comparability across studies. The choice of the unit heavily depends on the purpose of the study. This is why there is no unambigously optimal unit for measuring speech tempo. The central role of the pause is also mirrored in the important distinction between speaking rate (with pauses) and articulation rate (without pauses). Independent of a global tempo, articulation rate in inter-pause stretches can vary considerably, which is sometimes observed as acceleration and deceleration within the phases of articulation.

The empirical part begins with an analysis of real-world data in chapter 5. In this case study, the prosodic characteristics of the emotive speaking style of horse race commentaries were investigated. The auditory impression of a high speech tempo during the last part of a commentary is not reflected by an increase of articulation rate. Thus, the results confirm the important role of pausing. However, they contradict the expectation that speeding up is marked by fewer pauses: in our data, pauses occur more often compared to perceptually slower parts. This finding must be seen in interaction with breathing and the use of a very high average pitch level, which, together, lead to the speech tempo being perceived as higher.

The speech production experiments described in chapter 6 investigated (German) speakers' strategies for achieving tempo variation when reading aloud. The results of pausing behaviour, articulation rate, segmental reductions, phrasing and intonation revealed many idiosyncratic differences on these levels.

The perception tests with tempo-scaled synthetic speech in chapter 7 suggest improvements for the development of speech synthesis. The models presented control tempo just on the level of pausing and phrasing, with predictions of locations and durations of pauses and phrase-final lengthened syllables for different rates. Speech synthesis which included some of the non-linear aspects presented in chapter 3 was preferred over linearly modified speech synthesis, especially for a very slow tempo.

8.2. Conclusions

It has been shown that tempo modelling is, first and foremost, pause modelling. This was expected on the basis of the general statement that tempo variation is primarily variation in pausing (Goldman Eisler, 1968). However, the case study looking at horse race commentaries (chapter 5) has provided evidence against a generalisation that speeding up is characterised by fewer and shorter pauses. We found more pauses in the auditorily faster last bit of those commentaries, and an important influence of breathing and average pitch level.

The reading rate experiment described in chapter 6 has shown that there are various strategies - and not just one general strategy - how to vary tempo in speech production. This concerns the area of pauses and prosodic breaks, the use of segmental reductions, and F0 characteristics.

A further critical view on the general statement on pausing as the main factor of tempo variation aims at the prediction of pause location on the one hand, and actual duration of pauses on the other. With regard to speech synthesis it can be stated that these two points are not very well modelled in most of today's text-to-speech synthesisers. Especially sentences with long portions without punctuation marks make these shortcomings obvious. Going beyond a break and pause prediction that is solely based on punctuation seems to be essential, as demonstrated in chapter 7 by the perception experiment using tempo-scaled synthetic speech.

Further implications for synthetic speech concern the paradigm of generating artificial speech with natural speech as the ideal model. The copying of what researchers found in natural speech does not necessarily lead to improvements in synthetic speech. This has been shown in perception experiments by Portele (1997) for schwa deletion in German, for energy modelling by Barry et al. (in press), and for very fast playback rates by Janse (2003). One of the conclusions drawn from the outcome of the perception tests presented here is that most listeners prefer - as default speech tempo in synthesis - a tempo which is slower than the one they prefer under natural speech conditions. Taken together, these findings mean that e.g. news reading speech, which can be considered as one of the faster speaking styles (see chapter 2), provides less optimal data for modelling speech synthesis, either by statistical methods or by rule. This also means for the evaluation of a synthesiser's performance that a good match with the natural test material does not guarantee similar results for perceptual scores of intelligibilty and pleasantness.

Doing without perception tests in evaluating synthetic speech ignores the crucial point that synthetic speech is made for *listeners* rather than for speakers. Looking into details of human speech production may provide helpful information but is not necessarily the most effective way. At the end of the synthetic speech chain there are listeners, and listeners can have very different needs. These needs can be illustrated by the effect of tempo in speech synthesis: a blind user of synthetic speech or fast playback speech in everyday life wishes to have a very fast rendering of synthetic speech, without any special interest in a "nice-and-natural sound". What counts here is intelligibility at high speed. In contrast, a first- or second-time user of synthetic speech, i.e. nearly everyone, probably requires a slower tempo to achieve the highest degree of intelligibility and acceptability. This is even more relevant for users who also prefer a rather slow tempo in natural speech, such as elderly people, hard-of-hearing persons, and language learners (cf. chapter 2).

However, most present-day synthesisers bear the risk of sounding "bored" when speech tempo is reduced. Here, counter-initiatives like a more elaborated use of pitch range and pitch contour could help, to name just two other prosodic parameters. As several examples - both in the theoretical part and in the practical part of the thesis - showed, a change of speech tempo is frequently accompanied by other prosodic properties.

Last but not least, the thesis has implications for recorded natural speech. The advantages of a non-linear speech compression which focuses mainly on pauses has been demonstrated by Covell, Withgott & Slaney (1998) and He & Gupta (2001). Applications of this technique are e.g. audio and audio-video browsing, or fast playback for blind people. The advantages of a non-linear speech expansion with a main focus on pauses and an appropriately slow articulation rate, as shown for synthetic speech in chapter 7, could also apply to recorded speech. Slow playback rates are required e.g. in language learning.

The thesis attempts to provide an overview of tempo variation in speech production. Doing basic research by inspecting real-world data as well as laboratory speech, new insights in the field of timing and tempo of speech have been gained. The tempo model, which is based on these and other findings from basic research, has been implemented in a text-to-speech synthesiser and tested in perception experiments. I hope this thesis is an example of how basic research and technology-oriented research in phonetics and phonology can be combined and used for various applications based on synthetic speech.

References

The following abbreviations are used:

J Phonetics   Journal of Phonetics

J Acoust. Soc. Am.   Journal of the Acoustical Society of America

J Sp Lang Hear Res Journal of Speech Language and Hearing Research

JIPA Journal of the International Phonetic Association

ICSLP International Conference on Spoken Language Processing

ICPhS International Congress of Phonetic Sciences

Abe, M. (1997): Speaking styles: statistical analysis and synthesis by a text-to-speech system. In: van Santen et al. (eds) *Progress in Speech Synthesis.* Springer: New York etc. pp. 495-510.

Abercrombie, D. (1967). *Elements of General Phonetics*. Edinburgh University Press.

Abramson, A.S. & Lisker, L. (1972). Voice-timing perception in Spanish word-initial stops. *J Phonetics* **1**, pp. 1-8.

Antoniadis, Z. & Strube, H.W. (1984). Untersuchungen zur spezifischen Dauer deutscher Vokale. *Phonetica* **41**, pp. 72-87.

Apple, W., Streeter, L.A. & Krauss, R.M. (1979): Effects of pitch and speech rate on personal attributions. *Journal of Personality and Social Psychology 37* (5), pp. 715-727.

Banse, R. & Scherer, K.R. (1996) Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology* 70 (3), pp. 614-636.

Barber, C., Mellor, B., Graham, R., Noyes, J.M., Tunley, C. (1996). Workload and the use of automatic speech recognition: The effects of time and resource demands. *Speech Communication* **20**, pp. 37-53.

Barden, B. (1991). Sprechgeschwindigkeit und thematische Struktur. Arbeitspapier **15** *"Kontextualisierung durch Rhythmus und Intonation"*, University of Konstanz.

Barik, H.C. (1977). Cross-linguistic study of temporal characteristics of different types of speech materials. *Language & Speech* **20**, pp. 116-135.

Barry, W.J. (1995). Phonetics and phonology of speaking styles. Proc. ICPhS (2), Stockholm, pp. 4-10.

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

Batliner, A., Kießling, A., Kompe, R., Niemann, H., Nöth, E. (1997). Tempo and its change in spontaneous speech. *Proc. Eurospeech Rhodes*, pp. 763-766.

Baumann, S. & Trouvain, J. (2001). On the Prosody of German Telephone Numbers. *Proc. Eurospeech 2001* Scandinavia, pp. 557-560.

Beckman, M.E. & Ayers, G.M. (1994). Guidelines for ToBI labeling. http://ling.ohiostate.edu/Phonetics/E_ToBI/etobi_homepage.html

Beckman, M.E. & Edwards, J. (1990). Lengthenings and shortenings and the nature of prosodic constituency. *Laboratory Phonology* **I**, pp. 152-178.

van Bezooyen, Renee (1984): *Characteristics and Recognizability of Vocal Expressions and Emotions*. (Netherland Phonetic Archives 5). Foris: Dordrecht.

Berkovits, R. (1991). The effect of speaking rate on evidence for utterance-final lengthening. *Phonetica* **48**, pp. 57-66.

Bolozky, Shmuel (1977). Fast speech as a function of tempo in natural generative phonology. *Journa of Linguistics* **13**, pp. 217-238.

Botinis, A., Fourakis, M., Priniou, I. (1999). Prosodic effects on segmental durations in Greek. *Proc. Eurospeech* Budapest, pp. 2475-2478.

Braunschweiler, N. (1997). Integrated cues of voicing and vowel length in German: a production study. *Language & Speech* **40** (4), pp. 353-376.

Brubaker, R.S. (1972). Rate and pause characteristics of oral reading. *J Psycholinguistic Research* **1** (2), pp. 141-147.

Burkhardt, F. (2001). *Simulation emotionaler Sprechweise mit Sprachsynthese-verfahren.* Diss. TU Berlin. Shaker-Verlag.

Butcher, A. (1981). Aspects of the speech pause: phonetic correlates and communicative functions. *Aipuk* **15** (Arbeitsberichte Institut für Phonetik Kiel).

Byrd, D. (1992). Sex, dialects, and reduction. *Proc. ICSLP* Banff (1), pp. 827-830.

Byrd, D. & Tan, C.C. (1996). Saying consonant clusters quickly. *J Phonetics* **24**, pp. 263-282.

Campbell, W.N. & Isard, S.D. (1991). Segment durations in a syllable frame. *J. Phonetics* **19**, pp. 37-47.

Carlson, R. & Granstrom, B. (1986). A search for durational rules in a real-speech data base. *Phonetica* **43**, pp. 140-154.

Carlson, R., Granstrom, B & Klatt, D.H. (1979). Some notes on the perecption of temporal patterns in speech. In: Lindblom & Öhmann (eds): *Frontiers of Speech Communication Research.* pp. 233-243.

Caspers, J. & Van Heuven, V.J. (1991). Phonetic and linguistic aspects of pitch movements in fast speech in Dutch. *Proc. ICPhS* Aix-en-Provence (5), pp. 174-177.

Clark, H. H. & Fox Tree, J.E. (2002). Using *uh* and *um* in spontaneous speaking. *Cognition* **84**, pp. 73-111.

Covell, M., Withgott, M. & Slaney, M. (1998). MACH1: Nonuniform time-scale modification of speech. *Proc. IEEE Confer. on Acoustics, Speech and Signal Processing*, Seattle.

Cowan, M. (1936). Pitch and intensity characteristics of stage speech. *Arch Speech*, Suppl. to December issue.

Crown, C.L. & Feldstein, S. (1991). The perception of speech rate from sound-silence patterns of monologues. *J Psycholinguistic Research* **20** (1), pp. 47-63.

Crystal, D. (1969). *Prosodic Systems and Intonation in English.* Cambridge University Press: Cambridge.

Crystal, T.H. & House, A.S. (1990). Articulation rate and the duration of syllables and stress groups in connected speech. *J Acoust. Soc. Am.* **88** (1), pp. 101-112.

Dankovičová, J. (1997). The domain of articulation rate in Czech. *J Phonetics* **25**, pp. 287-312.

Dankovičová, J. (1999). Articulation rate variation within the intonation phrase in Czech and English. *Proc. ICPhS* San Francisco, pp. 269-272.

Dankovičová, J. & Nolan, F. (1999). Some acoustic effects of speaking style on utterances for automatic speaker verification. *JIPA* **29** (2), pp. 115-128.

Dauer, R.M. (1983). Stress-timing and syllable-timing reanalyzed. *J Phonetics* **11**, pp. 51-62.

Deese, J. (1984). *Thought into Speech: The Psychology of Language.* Prentice-Hall: Englewood Cliffs.

Deutsche Welle, URL "Langsam gesprochene Nachrichten": http://www.dw-world.de/english/0,3367,2146-0-0-B__,00.html, retrieved 30/04/2003 van Dommelen, W. (1982). A contrastive investigation of vowel duration in German and Dutch. *Phonetica* **39**, pp. 23-35.

van Donzel, M.E. & Koopmans-van Beinum, F.J. (1996). Pausing strategies in discourse in Dutch. *Proc. ICSLP* Philadelphia, pp. 1029-1032.

Dressler, W. (1975). Methodisches zu Allegro-Regeln. In: Dressler, W. & Mareš, F.V. (eds): *Phonologica 1972*. Wilhelm Fink: München & Salzburg.

Edwards, J., Beckman, M.E. & Fletcher, J. (1991). The articulatory kinematics of final lengthening. *J Acoust. Soc. Am.* **89** (1), pp. 369-382.

Eefting, Wieke (1991): *Timing in Talking. Tempo Variation in Production and Its Role in Perception*. Diss. Utrecht.

Fackrell, J., Vereecken, H., Martens, J.-P. & van Coile, B. (2000): The variation of prosody with text type. *Proc. ESCA/IEEE Workshop on State-of-the-Art in Speech Synthesis*, London, 5/1-5/8.

Fant, G., Kruckenberg, A. & Nord, L. (1992). Prediction of syllable duration, speech rate and tempo. *Proc. ICSLP* Banff, Canada, pp. 667-670.

Faulkner, D.S. (1997). *Assessment of the independent contributions of words, syllables and segments to the durations of utterances with reference to the measurement of speech rate.* Ms. Edinburgh University.

Fischer, K. (1999): Discourse effects on the prosodic properties of repetitions in human-computer interaction", Proc. *ESCA Workshop on Dialogue and Prosody*, Veldhoven (NL), pp. 123-128.

Fletcher, J. (1987). Some micro and macro effects of tempo change on timing in French. *Linguistics* **25**, pp. 951-967.

Fónagy, I. & Magdics, K. (1960). Speed of utterance in phrases of different lengths. *Language & Speech* **3**, pp. 179-192.

Fougeron, C. & Keating, P.A. (1997). Articulatory strengthening at edges of prosodic domains. *J Acoust. Soc. Am.* **101** (6), pp. 3728-3740.

Fougeron, C. & Jun, S.-A. (1998). Rate effects on French intonation: prosodic organization and phonetic realization. *J Phonetics* **26**, pp. 45-69.

Fourakis, M. (1991). Tempo, stress, and vowel reduction in American English. *J Acoust. Soc. Am.* **90** (4), pp. 1816-1827.

Gaitenby, J.H., (1965). The elastic word. *Haskins Report* **SR-2**, pp. 3.1-3.12.

Gay, T. (1978). Effect of speaking rate on vowel formant structures. *J Acoust. Soc. Am.* **63**, pp. 223-230.

Gay, T. (1981). Mechanisms in the control of speech rate. *Phonetica* **38**, pp. 148-158.

Gee, J.P. & Grosjean, F., (1983). Performance structures: a psycholinguistic and linguistic appraisal. *Cognitive Psychology* **15**, pp. 411-458.

Gilbert, J.H. & Burk, K.W. (1969). Rate alterations in oral reading. *Language & Speech* **12**, pp. 192-201.

Goldman-Eisler, F. (1961). The significance of changes in the rate of articulation. *Language & Speech* **4**, pp. 171-174.

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

Greenberg, S. (1999). Speaking in shorthand - A syllable-centric perspective for understanding pronunciation variation. *Speech Communication* **29**, pp. 159-176.

Greisbach, R. (1992). Reading aloud at maximal speed. *Speech Communication* **11**, pp. 469-473.

Griffiths, R. (1990). Speech rate and NNS comprehension: a preliminary study in time-benefit analysis. *Language Learning* **40** (3), pp. 311-336.

Grosjean, F. (1979). A study of timing in a manual and a spoken language: American sign language and English. *J Psycholinguistic Research* **8**, pp. 379-405.

Grosjean, F. & Collins, M. (1979). Breathing, pausing and reading. *Phonetica* **36**, pp. 98-114.

Grosjean, F. & Deschamps, A. (1975). Analyse contrastive des variables temporelles de l'anglais et du français: vitesse de parole et variables composantes, phénomènes d'hésitation. *Phonetica* **31**, pp. 144-184.

Grunwald, T. (1983). *Reduktion und Kompensation als Funktion der Sprechgeschwindigkeit im Deutschen.* (Forum Phoneticum 28). Hamburg: Buske.

Guitar, B. & Marchinkoski, L. (2001). Influence of mother's slower speech on their children's speech rate. *J Sp Lang Hear Res* **44**, pp. 853-861.

Gussenhoven, C. & Rietveld, A.C.M. (1992). Intonation contours, prosodic structure and preboundary lengthening. *J Phonetics* **20**, pp. 283-303.

Hall, K.D., Amir, O. & Yairi, E. (1999): A longitudinal investigation of speaking rate in preschool children who stutter. *J Sp Lang Hear Res* **42**, pp. 1367-1377.

Harris, M.S. & Umeda, N. (1974) Effect of speaking mode on temporal factors in speech: vowel duration. *J Acoust. Soc. Am.* **56** (3), pp. 1016-1018.

Haselager, G.J.T., Slis, I.H. & Rietveld, A.C.M. (1991). An alternative method of studying the development of speech rate. *Clinical Linguistics & Phonetics* **5** (1), pp. 53-63.

He, L. & Gupta, A. (2001). Exploring benefits of non-linear time compression. Proc. *Conference on Multimedia*, Ottawa, pp. 328-391.

Hertz, S.R. (1991). Streams, phones and transitions: toward a new phonological and phonetic model of formant timing. *J Phonetics* **19**, pp. 91-109.

Hewlett, N. & Rendall, M. (1998): Rural versus urban accent as an influence on the rate of speech. *JIPA* **28**, pp. 63-71.

Hieke, A.E., Kowal, S. & O'Connel, D.C. (1983). The trouble with "articulatory" pauses. *Language & Speech* **26** (3), pp. 203-214.

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.

House, A.S. (1961). On vowel duration in English. *J Acoust. Soc. Am.* **33** (9), pp. 1174-1178.

IPDS, Institut für Phonetik und Digitale Sprachverarbeitung (1994). *The Kiel Corpus of Read Speech*, Vol. 1 (CD-ROM). University of Kiel, Germany.

Iivonen A., Niemi. T. & Paananen, M., (1995). Comparison of prosodic characteristics in English, Finnish and German radio and TV newscasts. *Proc. ICPhS* Stockholm (2), pp. 382-385.

Janse, E. (2000). Intelligibility of time-compressed speech: three ways of time-compression. *Proc. ICSLP* Beijing.

Janse, E. (2003). *Production and Perception of Fast Speech*. PhD Thesis Utrecht.

Jessen, M., Marasek, K., Schneider,K. & Claßen, K. (1995). Acoustic correlates of word stress and the tense/lax opposition in the vowel system of German. Proc. *ICPhS* (4) Stockholm, pp. 428-431.

Kehrein, R. (2002). *Prosodie und Emotionen*. (Reihe Germanistische Linguistik). Niemeyer: Tübingen.

Keller, E. & Zellner, B. (1996). A timing model for fast French. *York Papers in Linguistics* **17** (University of York), pp. 53-75.

Kemper, S. (1994). Elderspeak: Speech accommodations to older adults. *Aging and Cognition* **1**, pp. 17-28.

Klatt, D.H. (1973) Interaction between two factors that influence vowel duration. *J Acoust. Soc. Am.* **54** (4), pp. 1102-1104.

Klatt, D.H. (1975). Vowel lengthening is syntactically determined in a connected discourse. *J Phonetics* **3**, pp. 129-140.

Klatt, D.H. (1976). Linguistic uses of segmental duration in English: acoustic and perecptual evidence. *J Acoust. Soc. Am* **59** (5), pp. 1208-

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

Kohler, K.J. (1982). Rhythmus im Deutschen. *Aipuk* **19** (Arbeitsberichte Institut für Phonetik Kiel), pp. 89-105.

Kohler, K.J. (1983a). F0 in speech timing. *Aipuk* **20** (Arbeitsberichte Institut für Phonetik Kiel), pp. 55-97.

Kohler, K.J. (1983b). Prosodic boundary signals in German. *Phonetica* **40**, pp. 89-134.

Kohler, K.J. (1986). Invariance and variability in speech timing: from utterance to segment in German. Perkell, J.S. & Klatt, D.H. (eds): *Invariance and Variability in Speech Processes*. Hillsdale, NJ & London. pp. 268-289.

Kohler, K.J. (1988). Zeitstrukturierung in der Sprachsynthese. *ITG-Fachberichte* **105**, pp. 165-170.

Kohler, K.J. (1990). Segmental reduction in connected speech in German: phonological facts and phonetic explanations. In: Hardcastle & Marchal (eds): *Speech Production and Speech Modelling*, pp. 69-92.

Kohler, K.J. (1995). *Einführung in die Phonetik des Deutschen*. Berlin: Erich Schmidt Verlag. 2. Auflage.

Kohler, K. J. (2000). Linguistic and paralinguistic functions of non-modal voice in connected speech. *Proc. 5th Seminar on Speech Production: Models and Data*. Kloster Seeon, Bavaria, pp. 121-124.

Koiso, H., Shimojima, A. & Katagiri, Y. (1998). Collaborative signaling of informational structures by dynamic speech rate. *Language & Speech* **41** (3-4), pp. 323-350.

Koopmans-Van Beinum, F.J. & Van Donzel, M.E. (1996). Discourse structure and its influence on local speech rate. *Proceedings* **20** (Institute of Phonetic Sciences, Amsterdam University), pp. 1-11.

Kowal, S. (1991). *Über die zeitliche Organisation des Sprechens in der Öffentlichkeit.* Bern: Huber.

Kowal, S., Wiese, R. & O'Connell, D.C. (1983). The use of time in storytelling. *Language & Speech* **26** (4), pp. 377-392.

Kröger, B.J. (1996). Zur phonetischen Realisierung von Sprechtempoänderungen unter Einbeziehung von artikulatorischer Reorganisation: Artikulatorische und perzeptive Untersuchungen. In: Gibbon, D. (ed): *Natural Language Processing & Speech Technology, 3rd KONVENS Conference.* Mouton de Gruyter: Berlin & New York, pp. 171-185.

Kuehn, D.P. & Moll, K.L. (1976). A cineradiographic study of VC and CV articulatory velocities. *J Phonetics* **4**, pp. 303-320.

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

Künzel, H.J., Braun, A. & Eysholdt, U. (1992). *Einfluss von Alkohol auf Sprache und Stimme.* Heidelberg: Kriminalistik-Verlag.

Kuwabara, H. (1996): Acoustic properties of phonemes in continous speech for different speaking rate. *Proc. ICSLP* Philadelphia, pp. 2435-2438.

Ladd, D.R., Faulkner, D., Faulkner, H., Schepman, A. (1999): Constant "segmental anchoring" of F0 movements under changes in speech rate. *J Acoust. Soc. Am.* **106** (3, Pt. 1), pp. 1543-1555.

Lane, H. & Grosjean, F. (1973). Perception of reading rate by speakers and listeners. *J Experimental Psychology* **97** (2), pp. 141-147.

Lehiste, I. (1970). *Suprasegmentals*. Cambridge, MA: MIT Press.

Lehiste, I. (1972). The timing of utterances and linguistic boundaries. *J Acoust. Soc. Am.* **51** (6 Pt. 2), pp. 2018-2024.

Lehiste, I. (1973). Rhythmic units and syntactic units in production and perception. *J Acoust. Soc. Am.* **54** (5), pp. 1228-1234.

Levelt, W.J.M. (1989). *Speaking. From Intention to Articulation.* Cambridge, MA & London: The MIT Press.

Lindblom, B. (1963). Spectrographic Study of Vowel Reduction. *J Acoust. Soc. Am.* **35** (11), pp. 1773-1779.

Lindblom, B. (1990). Explaining phonetic variation: a sketch of the H&H theory. In: Hardcastle, W. & Marchal, A. (eds): *Speech Production and Speech Modelling*, pp. 403-439.

Lindblom, B. & Rapp, K. (1973). *Some temporal regularities of spoken Swedish.* PILUS **21, (**Linguistics Stockholm).

Lisker, L. (1974). On "explaining" vowel duration variation. *Glossa* **8**, pp. 233-246.

Lively, S., Pisoni, D., van Summers, W., and Bernacki, R. (1993). Effects of cognitive workload on speech production: Acoustic analyses and perceptual consequences. *J Acoust. Soc. Am.* **93** (5), pp. 2962-2973.

Luce, P.A. & Charles-Luce, J. (1985). Contextual effects on vowel duration, closure duration, and the consonant/vowel ratio in speech production. *J Acoust. Soc. Am.* **78** (6), pp. 1949-1957.

Malécot, A., Johnston, R. & Kizziar, P.-A. (1972): Syllabic rate and utterance length in French. *Phonetica* **26**, pp. 235-251.

Martin, E. (1970): Toward an analysis of subjective phrase structure. *Psychological Bulletin* **74** (3), pp. 153-166.

Meinhold, G. (1967). Quantität und Häufigkeit von Pausen in gelesenen deutschen Texten im Zusammenhang mit dem Sprechtempo. *Gesellschafts- u. Sprachwissensch. Reihe*, Jg. **16** (1) (Wiss. Zeitschrift Universität Jena), pp. 107-111.

Miller, J.L., Grosjean, F. & Lomanto, C. (1984). Articulation rate and its variability in spontaneous speech: a reanalysis and some implications. *Phonetica* **41**, pp. 215-225.

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. 41/1–41/5.

Monaghan, A.I.C. (2001). An auditory analysis of the prosody of fast and slow speech styles in English, Dutch and German. In: E. Keller, G. Bailly, A. Monaghan, J. Terken and M. Huckvale (eds.) *Improvements in Speech Synthesis*. Wiley & Sons, pp. 204-217.

Morgan, N., Fosler, E. & Mirghafori, N. (1997). Speech Recognition using On-line Estimation of Speaking Rate. Proc. *Eurospeech* '97, Rhodes, Greece.

Morgan Barry, R. (1995). The relationship between dysarthria and verbal dyspraxia in children: a comparative study using profiling and instrumental analyses. *Clinical Linguistics & Phonetics* **9** (4), pp. 277-309.

Munro, M.J. & Derwing. T.M. (2001). Modeling perceptions of the accentedness and comprehensibility of L2 speech. *Studies in Second Language Acquisition* (SSLA) **23**, pp. 451-468.

Murray, I.R. & Arnott, J.L. (1993). Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion. *J Acoust. Soc. Am.* **93** (2), pp. 1097-1108.

Neweklowsky, G. (1975). Spezifische Dauer und spezifische Tonhöhe der Vokale. *Phonetica* **32**, pp. 38-60.

Nooteboom, S.G. (1991). Some observations on the temporal organisation and rhythm of speech. *Proc. ICPhS* Aix-en-Provence, pp. 228-237.

O'Connell, D.C. & Kowal, S. (1972). Cross-linguistic pause and rate phenomena in adults and adolescents. *Journal of Psycholinguistic Research* **1** (2), pp. 155-164.

O'Connell, D.C. & Kowal, S. (1983). Pausology. In: Sedelow, S.Y. & Sedelow, W.A. (eds) *Computers in Language Research 2*. Berlin, New York, Amsterdam: De Gruyter.

Ofuka, E., McKeown, J.D., Waterman, M.G. & Roach, P.J. (2000): Prosodic cues for rated politeness in Japanese speech. *Speech Communication* **32** (3), pp. 199-217.

Oller, D.K. (1973). The effect of position in utterance on speech segment duration in English. *J Acoust. Soc. Am.*, pp. 1235-1247.

den Os, E. (1985). Perception of speech rate of Dutch and Italian utterances. *Phonetica* **42**, pp. 124-134.

O'Shaughnessy, D. (1981). A study of French vowel and consonant durations. *J Phonetics* **9**, pp. 385-406.

Osser, H. & Peng, F. (1964). A cross cultural study of speech rate. *Language & Speech* **7**, pp. 120-125.

Peterson, G.E. & Lehiste, I. (1960). Duration of syllable nuclei in English. *J Acoust. Soc. Am.* **32** (6), pp. 693-703.

Pfitzinger, H.R. (1999). Local speech rate perception in German speech. *Proc. ICPhS* San Francisco, pp. 893-896.

de Pijper, J.R. & Sanderman, A.A. (1994): On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues. *J Acoust. Soc. Am.* **96** (4), pp. 2037-2047.

Pirker, H. & Loderer, G (1999). "'I said two ti-ckets': how to talk to a deaf wizard", Proc. *ESCA Workshop on Dialogue and Prosody*, Veldhoven (NL), pp. 181-185.

Pisoni, D.P. (1993). Long-term memory in speech perception: some new findings on talker variability, speaking rate and perceptual learning. *Speech Communication* **13**, pp. 109-125.

Pisoni, D. (1997). Perception of Synthetic Speech. In: van Santen et al. (eds), Progress in Speech Synthesis, pp. 541-560.

Pols, L.C.W., van Santen, J.P.H., Abe, M., Kahn, D. & Keller, E. (1998). The use of large text corpora for evaluating text-to-speech systems. *Proc. LREC* Granada, pp. 637-640.

Port, R.F. (1981). Linguistic timing factors in combination. *J. Acoust. Soc. Am.* **69** (1), pp. 262-274.

Portele, T. (1996). Dynamische Anpassung der Sprechgeschwindigkeit. Mehnert, D. (ed.): *7. Konferenz Elektronische Sprachsignalverarbeitung*, Berlin, pp. 238-243.

Portele, T. (1997). Reduktionen in der einheitenbasierten Sprachsynthese. Proc. *Fortschritte der Akustik - DAGA 97* Kiel, Germany, pp. 386-387.

Pürschel, H. (1975). *Pause und Kadenz. Interferenzerscheinungen bei der englischen Intonation deutscher Sprecher.* (Linguistische Arbeiten 27). Tübingen: Max Niemeyer Verlag.

Rieber, R.W., Breskin, S. & Jaffe, J. (1972). Pause time and phonation time in stuttering and cluttering. *J Psycholinguistic Research* **1** (2), pp. 149-154.

Rietveld, A.C.M. (1975). Untersuchungen zur Vokaldauer im Deutschen. *Phonetica* **31**, pp. 248-258.

Rietveld, A.C.M. & Gussenhoven, C. (1987). Perceived speech rate and intonation. *J Phonetics* **15**, pp. 273-285.

Roach, P. (1998). Some languages are spoken more quickly than others. In: Bauer, L. & Trudgill, P. (eds): *Language Myths*. Penguin, pp. 150-158.

Rodgers, J. (2000). The phonatory correlates of juncture in German. Proc. *5th Seminar on Speech Production: Models and Data*. Kloster Seeon, Bavaria, pp. 289-292.

de Rooij, J.J. (1979). *Speech punctuation*. Diss. Utrecht.

Sable, URL "Speech synthesis markup language": http://www.cstr.ed.ac.uk/projects/sable/sable-spec2.html, retrieved 30/04/2003

Samudravijaya, K., Singh, S.K., & Rao, P.V.S. (1998). Pre-recognition measures of speaking rate. *Speech Communication* **24**, pp. 73-84.

Sanderman, A.A. & Collier, R. (1996). Prosodic rules for the implementation of phrase boundaries in synthetic speech. *J Acoust. Soc. Am.* **100** (5), pp. 3390-3397.

Scherer, K.R. (1974). Acoustic concomitants of emotional dimensions: judging affect from synthesized tone sequences. In: Shirley Weitz (ed). *Nonverbal Communication*. New York etc.: Oxford University Press.

Schröder, M. (in preparation). *Emotion and Speech*. Dissertation, University of the Saarland.

Schröder, M. & Trouvain, J. (2001): The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching. To appear in the Proceedings 4th Speech Synthesis Workshop Perthshire, Scotland.

Siegler, M.A. & Stern, R.M. (1995). On the effects of speech rate in large vocabulary speech recognition systems. Proc. *International Conference on Acoustics and Speech Signal Processing*.

Simpson, A.P. (1998). *Phonetische Datenbanken des Deutschen in der empirischen Sprachforschung und der phonologischen Theoriebildung.* (Aipuk 33), Arbeitsberichte Universität Kiel.

Simpson, A.P. (2001). Dynamic consequences of differences in male and female vocal tract dimensions. *J Acoust. Soc. Am.* **109**(5), pp. 2153-2164.

Slembek, E. (1993). Vorüberlegungen zu Sprechtempo und Pausierung in verschiedenen Kulturen. In: Bonner, M., Braun, E. & Fix, H. (eds): *Nachbarschaften. Festschrift für Max Mangold.* (Beiträge zur Sprache im Saarland. Bd. 11). Saarbrücken, pp. 381-394.

van Sluijter, A.A.M. & van Heuven, V. (1996). Spectral balance as an acoustic correlate of linguistic stress. *.J Acoust. Soc. Am.* **100**(4), pp. 2471-2485.

Smith, B.L., Brown, B.L., Strong, W.J. & Rencher, A.C. (1975). Effects of speech rate on personality perception. *Language & Speech* **18**, pp. 145-152.

Sommers, M.S., Humes, L. & Pisoni, D.B. (1994): The effects of speaking rate and stimulus variability on spoken word recognition by young and elderly listeners. *Progress Report* **19**, Speech Research Lab, Indiana University.

van Son, R.J.J.H. & Pols, L.C.W. (1989): Comparing formant movements in fast and normal rate speech. *Proc. Eurospeech* Paris (2), pp. 665-668.

Strangert, E. (1991). Pausing in texts read aloud. *Proc. ICPhS*, Aix-en-Provence, pp. 238-241.

Street, R.L. (1982). Evaluation of noncontent speech accommodation. *Language & Communication* **2** (1), pp. 13-31.

Street, R.L. & Giles, H. (1982). Speech accommodation theory. In: Roloff, M.E. & Berger, C.R. (eds) *Social Cognition and Communication*. pp. 193-226.

Streeter, L. (1978). Acoustic determinants of phrase boundary perception. *J Acoust. Soc. Am.* **64** (6), pp. 1582-1592.

Summerfield, Q. (1981). Articulatory rate and perceptual constancy in phonetic perception. *J Experimental Psychology: Human Perception & Performance* **7** (5), pp. 1074-1095.

Swerts, M. (1997). Prosodic features at discourse boundaries of different strength. *J Acoust. Soc. Am.* **101**, pp. 514-521.

Tauroza, S. & Allison, D. (1990): Speech rates in British English. *Applied Linguistics* **11**, pp. 90-115.

Trouvain, J. (1999): Phonological aspects of reading rate strategies. *Phonus* **4** (Phonetics Saarbrücken), pp. 15-35.

Trouvain, J. (2000): Zur Langweiligkeit synthetischer Sprache - Welche "Sprecherziehung brauchen maschinelle Sprecher? Vortrag Jahrestagung *Gesellschaft f. Angewandte Linguistik.*

Trouvain, J., Barry, W. J., Nielsen, C. & Andersen, O. (1998): Implications of energy declinations for speech synthesis. Proc. ESCA/COCOSDA *Workshop on Speech Synthesis,* Jenolan Caves, Australia, pp. 47-52.

Trouvain, J. & Grice, M. (1999): The effect of tempo on prosodic structure. *Proc. ICPhS* (2), San Franscisco, pp. 1067-1070.

Trouvain, J. & Barry, William J. (2000): The prosody of excitement in horse race commentaries. Proc. *ISCA-Workshop on "Speech and Emotion"*, Newcastle, Northern Ireland, pp. 86-91.

Trouvain, J., Koreman, J., Erriquez, A. & Braun, B. (2001): Artculation rate measures and their relations to phone classification of spontaneous and read German speech. *Proc. ISCA Workshop on Adaptation Methods for Speech Recognition*, August 2001, Sophia Antipolis, France.

Tsao,Y.-C. & Weismer, G. (1997). Interspeaker variation in habitual speaking rate: evidence for a neuromuscular component. *J Sp Lang Hear Res* **40**, pp. 858-866.

Turk, A.E. (1999). Structural influences on boundary-related lengthening in English. *Proc. ICPhS* San Francisco.

Turk, A.E. & Sawusch, J.R. (1997). The domain of accentual lengthening in American English. *J Phonetics***.**

Turk, A.E. & White, L. (1999). Structural influences on accentual lengthening in English. *J Phonetics* **27**, pp. 171-206.

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. *J Sp Hear Res* **39**, pp. 494-509.

Uhmann, S. (1989). On some forms and functions of speech rate changes in everyday conversation. *Working Paper "Kontextualisierung durch Rhythmus und Intonation"* **7**, University of Konstanz.

Vaane, E. (1982). Subjective Estimation of Speech Rate. *Phonetica* **39**, pp. 136-149.

Walker, J.F., Archibald, L.M.D., Cherniak, S.R. & Fish, V.G. (1992). Articulation rate in 3- and 5-year-old children. *J Speech and Hearing Research* **35**, pp. 4-13.

Wells, B. & Peppé, S. (1996). Endind up in Ulster: prosody and turn-taking in English dialects. In: Couper-Kuhlen, E. & Selting, M. (eds). *Prosody in Conversation. Interactional Studies*. Cambridge: Cambridge University Press, pp. 101-130.

Weijer, J. van de (1997). Language input to a prelingual infant. *Proc. GALA '97* Conf. on Language Acquisition. Edinburgh, Scotland, pp. 290—293.

Whiteside, S. (1996). Temporal-based acoustic-phonetic patterns in read speech: some evidence for speaker sex differences. *JIPA* **26** (1), pp. 23-40.

Whiteside, S.P. & Hodgson, C. (2000). Speech patterns of children and adults elicited via a picture-naming task: An acoustic study. *Speech Communication* **32** (4), pp. 267-285.

Wiese, R. (1983). *Psycholinguistische Aspekte der Sprachproduktion.* Hamburg: Buske.

Wightman, C., Shattuck-Hufnagel, S., Ostendorf, M. & Price, P.J. (1992). Segmental durations in the vicinity of prosodic phrase boundaries. *J Acoust. Soc. Am.* **91** (3), pp. 1707-1717.

Wood, S. (1973). Speech tempo. *Working Papers Phonetics Lund* **9**, pp. 99-147.

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

Appendix

Text Chapter 6 Production Experiment 1

Ein Sechzehnjähriger hat am Morgen einen 80-Meter-Sturz in den Silbersee bei Neunkirchen überlebt. Nach Auskunft der Polizei war der Junge bei einer Klettertour an einem Steilhang ausgerutscht. Im Fall schlug er mit dem Kopf auf und fiel bewusstlos in den See. Ein Camper holte den Jungen aus dem Wasser und alarmierte den Rettungsdienst. Der Sechzehnjährige ist außer Lebensgefahr.

Text Chapter 6 Production Experiment 2

Einst stritten sich Nordwind und Sonne, wer von ihnen beiden wohl der Stärkere wäre, als ein Wanderer, der in einen warmen Mantel gehüllt war, des Weges daherkam. Sie wurden einig, dass derjenige für den Stärkeren gelten sollte, der den Wanderer zwingen würde, seinen Mantel abzunehmen.

Der Nordwind blies mit aller Macht, aber je mehr er blies, desto fester hüllte sich der Wanderer in seinen Mantel ein. Endlich gab der Nordwind seinen Kampf auf. Nun erwärmte die Sonne die Luft mit ihren freundlichen Strahlen, und schon nach wenigen Augenblicken zog der Wanderer seinen Mantel aus. Da musste der Nordwind zugeben, dass die Sonne von ihnen beiden der Stärkere war.

Text Chapter 7 Perception Experiment 1

Die SPD hat bestätigt, in der Kölner Spendenaffäre doch sämtliche Namen der Empfänger gefälschter Spendenquittungen zu kennen. Die Partei teilte in Düsseldorf und Berlin mit, die Liste sei am 10. April eingetroffen. Bislang hatte die Partei behauptet, die Namen nicht vollständig zu kennen.

Text Chapter 6 Perception Experiment 2

Bezugsdauer und Höhe des Arbeitslosengeldes sollen nicht pauschal gekürzt werden. Einschnitte soll es aber für alle geben, die nicht bereit sind, für einen Arbeitsplatz umzuziehen oder nach einer bestimmten Frist einen Job bei einer Leiharbeitsfirma anzunehmen.