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.