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.