Preferences:

Format: oral presentation
Wish to be considered for publication in an edited volume: yes

Optimization Modeling of Speech Timing

Andreas Windmann, Juraj Šimko, Britta Wrede, Petra Wagner

Bielefeld University, Germany
{andreas.windmann, juraj.simko, petra.wagner}@uni-bielefeld.de, bwrede@techfak.uni-bielefeld.de

Keywords: speech timing, computational modeling, optimization

1. INTRODUCTION

We present a novel paradigm for the computational modeling of speech timing. Our approach is based on two core assumptions: (1) the organization of speech is governed by the resolution of trade-offs between production and perception demands, and (2) isochrony at the syllable and stress group level, while not measurable in absolute terms in the speech signal, is present in a language-specific manner as an underlying tendency. These assumptions are implemented as an optimization procedure, minimizing a composite cost function whose components relate to durations of various prosodic constituents.

The theoretical idea that much of the variation in speech can be explained as emerging from trade-offs between minimizing effort and maximizing perceptual clarity has been advocated most prominently by [2]. A recent computational implementation of this idea shows that various temporal coordination phenomena at the level of articulatory gestures can be accounted for on these grounds [3].

2. MODELING AND RESULTS

In our model we have adapted the computational platform [3] to larger prosodic units. The model operates on sequences of consonant and vowel segments representing prosodic phrases. The temporal organization of these sequences emerges as a result of an optimization procedure resolving trade-offs between production and perception constraints. The production-related component cost simply increases linearly with the durations of segments, providing a crude measure of production effort. Simultaneously, a perception-related cost function provides an impetus to lengthen segments in a non-linear fashion, conjectured to approximate the inverse of the probability of recognizing a segment of a given duration [3]. Timing of higher-level prosodic units is evaluated as a cost related to standard deviation of syllable and inter-stress interval durations. These components of the cost function are combined as a weighted sum. This facilitates to model phenomena such as stress and final lengthening by locally varying premiums imposed on individual components. Stressed syllables, for example, are modeled by temporarily increasing the perception cost weight.

In a preliminary experiment, data from a German and an Italian speaker from the BonnTempo Corpus [1] have been simulated. Given appropriate weights for the higher-level prosodic cost functions motivated by traditional rhythmic characterizations of the individual languages, the model successfully reproduces language-specific timing phenomena, namely (1) regression coefficients for inter-stress interval duration as a function of the number of syllables (2) foot-level shortening and (3) frequency distributions of syllable and inter-stress interval durations. Moreover, we find substantial correlations between syllable and inter-stress interval durations in the data and generated by the model.

Importantly, despite the language-dependent settings of the higher-level prosodic cost functions, the surface temporal variability the data emerges from interactions with other characteristics, such as syllabic structure of a given language.

3. DISCUSSION

Our results show that the proposed approach holds promising prospects for the modeling of speech timing. In particular, it provides a strong platform for *explaining* timing phenomena, applying general principles that have been shown to hold in various other speech domains. The proposed approach thus represents a step towards the development of a unified account of the relationship between production-perception trade-offs and the variability encountered in natural speech.

4. REFERENCES

[1] Dellwo, V., Aschenberner, B., Wagner, P., Dankovicova, J., Steiner, I. 2004. BonnTempo corpus and tools. *Proceedings of Interspeech 2004*, 777-780.

[2] Lindblom, B. 1990. Explaining phonetic variation: a sketch of the H&H theory. In Hardcastle, W.J. & Marchal, A. (eds.), *Speech Production and Speech Modeling*. Dordrecht, Kluwer, 403-439.

[3] Simko, J., Cummins, F. 2010. Sequencing and optimization within an embodied task dynamic model. *Cognitive Science* 35(3), 527-562.