Optimization based model of speech rhythm and timing

Andreas Windmann, Juraj Šimko, Britta Wrede, Petra Wagner

Bielefeld University, Germany
andreas.windmann@uni-bielefeld.de

In the last decades, Hypo- & Hyper-articulation (H&H) theory (Lindblom, 1990) has provided a promising explanatory platform for many speech phenomena, explaining the variation in speech as resulting from the resolution of conflicting demands related to efficiency of production and perceptual clarity. As for computational modeling, optimization techniques represent a well-suited tool for simulating H&H mechanisms, with quantified production and perception demands implemented as counteracting components of a composite cost function. Comparable techniques have been used within the framework of Stochastic Optimality Theory and related approaches (Boersma, 1998; Katz, 2010). Recently, optimization has been applied to the replication of gestural coordination phenomena within an embodied model of articulation by Simko and Cummins (2011).

In this paper we present an optimization-based model attempting to unify interlinked segmental and suprasegmental aspects of speech in a hierarchical model of timing and rhythm. At the core of our model is a composite cost function, implemented as a weighted sum of component cost functions that relate to the durations of various prosodic constituents in sequences of consonant and vowel segments. The variables of the cost function are temporal boundaries of segments in a simulated phrase. The optimization algorithm determines the boundaries minimizing the cost function, resulting in the optimal temporal description of the given phrase with respect to the model parameters. Weights assigned to component functions facilitate modeling a hierarchy of prosodic features of utterances ranging from speaking rate variations through phenomena related to stress and phrase final lengthening to language specific rhythmic properties of speech.

Two basic component cost functions operate at the segmental level and reflect the production-perception trade-offs. The function D_s increases linearly with the duration of segment s, and can be interpreted as a crude measure of articulatory effort. This is countered by the cost P_s of parsing the given segment s by a listener. The function P_s decreases with segmental duration in a non-linear fashion, declining rapidly for the first few milliseconds but asymptotically converging to zero. We thus assume that the shorter the segment, the greater demand it poses on the listener, the difficulty being especially high for very short durations (Grimm, 1966). Consonants and vowels are distinguished by different slope constants for the parsing cost function, reflecting their different statuses in speech parsing (Diehl et al., 1987). In addition, a cost function T increasing (logarithmically) with the duration of a whole sequence provides a control mechanism for speech tempo.

In addition to this, our model features cost functions aimed at eliciting rhythmic properties of speech. We assume that although perfect isochrony at any prosodic level is not present in the speech signal, there exist (language-specific) tendencies for individual levels of the prosodic hierarchy to dominate the temporal organization of speech. The model currently features two functions that impose costs on the "non-rhythmicity", or non-evenness of the duration of suprasegmental units: the functions S and F that are difference functions of syllable and inter-stress interval durations in the entire sequence, respectively. While the general architecture of our model is inherently language independent, adjusting the weights for the higher-level prosodic cost functions allows for specifying dominant levels of temporal organization in a given language.

The overall cost function C is thus defined as

$$C = \alpha_D \sum_s \delta_s D_s + \alpha_P \sum_s \pi_s P_s + \alpha_T T + \alpha_S S + \alpha_F F,$$

where the sums range over all segments s in an utterance, α's are weights assigned to component cost functions used to elicit variations in speaking rate (α_D and α_T), H&H scale (α_P) and in overall rhythmic properties of speech (α_S and α_F). Moreover, the premium placed on segment duration and parsing cost can be locally

adjusted for each segment, using the weighting factors δ_s and π_s. This mechanism allows for a principled treatment of prosodic conditions such as stress or final lengthening. Stressed syllables are modeled by increasing the weighting factor π_s for their constituent segments. As stressed syllables are particularly critical for decoding linguistic structure (Cutler et al., 1997), we assume that speakers put a premium on perceptual clarity and "care less" about minimizing effort and duration when producing them. Similarly, phrase final lengthening can be elicited by lowering the duration weighting factor δ_s, putting less premium, locally, on the durational aspect.

In a preliminary experiment, we have simulated data from a German and an Italian speaker from the Bonn-Tempo Corpus (Dellwo et al., 2004). Results show that the optimization procedure converges and produces meaningful results: the model successfully reproduces the reported difference in regression coefficients for inter-stress interval duration as a function of the number of syllables, with the intercept of the function being at roughly 100 ms for Italian and 200 ms for German (Eriksson, 1991). This is achieved by choosing a parameter setting with $\alpha_S < \alpha_F$ for German and $\alpha_S > \alpha_F$ for Italian, corresponding to the alleged syllable-timed and stress-timed characteristics of the two languages. Moreover, we find high correlations between real and simulated syllable and inter-stress interval durations. Crucially, the built-in interactions between syllabic structure and the different levels of rhythmic organization ensure that the model does not generate isochrony at any prosodic level, but reproduces realistic durational variability.

Our preliminary results suggest that the model is a promising candidate for accounting for the temporal organization of speech. The key strength of the model is its explanatory power, as its individual components as well as the overall structure are grounded in biologically and developmentally plausible principles. This, in our opinion, distinguishes the presented approach from similar hierarchical paradigms, e.g. the oscillator-based and task-dynamical accounts of timing in speech (O'Dell and Nieminen, 1999; Saltzman et al., 2008). We believe that the underlying assumptions of the model will open intriguing possibilities towards interpretation with regard to language evolution and acquisition in the course of further work.

References

Boersma, P. (1998). *Functional phonology: Formalizing the interactions between articulatory and perceptual drives*. PhD thesis, University of Amsterdam.

Cutler, A., Dahan, D., and van Donselaar, W. (1997). Prosody in the comprehension of spoken language: A literature review. *Language and Speech* 40(2), 141–201.

Dellwo, V., Aschenberner, B., Wagner, P., Dankovičová, J., and Steiner, I. (2004). BonnTempo-corpus and BonnTempo-tools: A database for the study of speech rhythm and rate. In *Proc. Interspeech* 2004, Jeju Island, Korea, 777–780.

Diehl, R., Kluender, K., Foss, D., Parker, E., and Gernsbacher, M. (1987). Vowels as islands of reliability. *Journal of Memory and Language* 26(5), 564–573.

Eriksson, A. (1991). *Aspects of Swedish speech rhythm*. PhD thesis, University of Gothenburg.

Grimm, W.A. (1966). Perception of segments of English-spoken consonant-vowel syllables. *Journal of the Acoustical Society of America* 40(6), 1454–1461.

Katz, J. (2010). *Compression effects, perceptual asymmetries, and the grammar of timing*. PhD thesis, Massachusetts Institute of Technology.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In Hardcastle, W. J. and Marchal, A., editors, *Speech Production and Speech Modelling*, Dordrecht, Kluwer Academic Publishers, 403–439.

O'Dell, M. and Nieminen, T. (1999). Coupled oscillator model of speech rhythm. In *Proc. 14th ICPhS*, San Francisco, 1075–1078.

Saltzman, E. L., Nam, H., Krivokapic, J., and Goldstein, L. (2008). A task-dynamic toolkit for modeling the effects of prosodic structure on articulation. In *Proc. Speech Prosody* 2008, Campinas, Brasil, 7–16.

Simko, J. and Cummins, F. (2011). Sequencing and optimization within an Embodied Task Dynamic model. *Cognitive Science* 35(3), 527–562.