Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 74 (2015) 76–92

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Optimization-based modeling of speech timing ☆

Andreas Windmann [a,*], Juraj Šimko [b], Petra Wagner [a]

[a] Faculty of Linguistics and Literary Studies, Bielefeld University, Germany
[b] Institute of Behavioural Sciences, University of Helsinki, Finland

Received 16 January 2015; received in revised form 14 September 2015; accepted 16 September 2015
Available online 25 September 2015

Abstract

We present a model of suprasegmental speech timing based on the assumption that speech patterns are shaped by global and local adjustments of trade-offs between conflicting demands of minimizing production effort and maximizing perceptual clarity. The model uses an optimization procedure to determine durations of suprasegmental constituents of simulated utterances by minimizing an independently motivated composite cost function. The cost is a function of the constituent durations and encompasses different components that represent independently derived measures of speaker-based production effort, listener-oriented perceptual clarity as well as time conceptualized as a resource shared between both parties, linked to transmission efficiency. The trade-offs between these influences can be globally and locally adjusted by weights assigned to individual cost components within the composite cost function. We show that this approach facilitates modeling a hierarchy of interacting prosodic features of utterances, such as different degrees of prominence or effects of speaking rate and overall requirements of clarity. We outline the theoretical foundations and the architecture of the model and present results of simulation experiments, demonstrating that the model correctly predicts a range of suprasegmental timing phenomena in stress-accent languages that have not been addressed by a unified model. Results underline the model's capacity to account for several empirical observations regarding durational variation in speech.

Keywords: Suprasegmental speech timing; Production–perception trade-offs; H&H theory; Optimization

1. Introduction

One of the fundamental challenges of speech science is to uncover the underlying mechanisms that give rise to the enormous variability in human speech. A promising platform for explaining at least some of this variation is provided by Hyper- and Hypoarticulation (H&H) theory (Lindblom, 1990). According to this theory, speech patterns arise as a consequence of a trade-off between speaker- and listener-oriented influences: the desire of the speaker to act economically and save effort (hypoarticulation) is set off against the necessity to communicate a message to a listener, for which some effort, however defined, is necessary (hyperarticulation). The variability of speech, according to H&H theory, comes about because different conditions reinforce or weaken either of the two hypothetical constraints on speech production, prompting the speaker to invest more in order to secure successful communication, or, conversely, allowing him or her to save effort if the situation permits it.

Importantly, the alleged trade-off between drives towards hyper- and hypoarticulation can be observed at different levels of speech description. For example, numerous studies have elicited forms of "clear", hyperarticulated

☆ This article is a revised and extended version of work previously published in Windmann et al. (2013, 2014b).

* Corresponding author at: Faculty of Linguistics and Literary Studies, Bielefeld University, Post Box 100131, 33501 Bielefeld, Germany. Tel.: +49 521 106 3537.
E-mail addresses: andreas.windmann@uni-bielefeld.de (A. Windmann), juraj.simko@helsinki.fi (J. Šimko), petra.wagner@uni-bielefeld.de (P. Wagner).

http://dx.doi.org/10.1016/j.specom.2015.09.007
0167-6393/

speech at a *global* level, i.e., throughout entire utterances or discourses, by giving explicit instructions to subjects or creating external conditions that prompt hyperarticulation (see Uchanski, 2005 for an overview). On the other hand, the famous example of *a stitch in time saves nine* vs. *the next number is nine* (Lieberman, 1963) illustrates that the same mechanism may apply also at a *local* level: on this view, the acoustic–phonetic differences between the two instances of *nine* would be interpreted as a consequence of a greater impetus towards hyperarticulation localized at the word level in the latter case, due to the lower predictability of the word based on the preceding context. This notion of local H&H variation is particularly interesting because it allows for re-interpreting well-described linguistic phenomena in terms of the H&H theory. An instructive example of this is provided by De Jong (1995)'s account of prosodic prominence as "localized hyperarticulation." It maintains that prominent syllables or words are highlighted because they are particularly critical for communicative success. Indeed, lexical stress tends to fall on root morphemes in many languages (Echols and Newport, 1992), and it provides a strong potential cue for word recognition and segmentation, particularly in adverse listening conditions (Bond, 1981; Cutler, 1991). Prominence relations between words, in turn, are used to mark words in an utterance which are semantically or pragmatically very important, often coinciding with discourse-new information. Changing the prominence relations between words, or reallocating the prosodic *focus* of an utterance typically results in major changes in its interpretation (Bolinger, 1958; Ladd, 2008; Schmitz, 2008). The greater *relative* acoustic salience of prominent syllables or words compared to their environment is thus interpreted as a local shift in favor of perceptual clarity, so as to draw listener's attention towards these important units.

Modeling with optimization algorithms is a well-suited method for exploring the assumptions of H&H theory. This approach typically entails an explicit quantification of the degree to which any given speech act (e.g., an utterance) succeeds with respect to each of the hypothesized influences. For each influence considered in the model, this degree is usually expressed in terms of a partial (component) cost function. For example, the requirement of economy can be quantified as a physiological cost associated with producing the given rendition of the speech act, the demand of communicative success as a cost associated with potential failure on the part of listener to comprehend the rendition as intended, etc. An optimization algorithm can then be used to find the form of the utterance that incurs the lowest *overall* cost that combines all the partial cost components in a parallel fashion. This optimal solution represents the best compromise among the multiple requirements imposed on communication by the model. Importantly, the relative influence of individual requirements can be explicitly quantified by globally or locally adjusting the premium placed on separate cost components within the overall cost function. For example, by increasing or decreasing the weight assigned to the partial cost representing the risk of communicative failure, the modeler can invoke conditions favoring hyper- or hypoarticulation, respectively. Optimization models that are based on similar principles have provided convincing accounts of various acoustic–phonetic phenomena, including formant transitions in CV sequences (Flemming, 2001), intrasyllabic timing effects at the segmental level (Flemming, 2001; Katz, 2010) and aspects of intonation patterns (Kochanski and Shih, 2003). Related optimization approaches have been successful in explaining universal tendencies in vowel systems (Liljencrants and Lindblom, 1972; Schwartz et al., 1997), and have also inspired phonological theory (Boersma, 1998; Flemming, 2001; Katz, 2010; Kirchner, 2013).

Recently, an optimization approach explicitly inspired by H&H theory has been applied in an embodied model of articulation (Šimko and Cummins, 2010), demonstrating capability to generate plausible gestural scores and to replicate intricate gestural timing phenomena linked to stable CV-phasing (Šimko and Cummins, 2011), phrasal boundary effects (Beňuš and Šimko, 2014) and emergence of phonological quantity contrast (Šimko et al., 2014b). In particular the last two works share many crucial aspects with the present model and are to a large extent compatible with the approach advocated here.

The optimization-based model of suprasegmental speech timing presented in this paper implements and tests the hypothesis discussed above, namely that durational patterns at the suprasegmental level emerge from the resolution of conflicting tendencies to achieve maximal communicative success with minimal effort. The model is broadly based on the principles outlined in Šimko and Cummins (2010), but operates on acoustic durations of prosodic constituents rather then individual articulatory gestures. A key feature of the model is the distinction between global and local variation of the relative influences of partial cost components as outlined above. We introduce three such cost functions, which for a given durational patterns quantify, in a rather abstract way, the cost associated with speaker's effort, the perceptual cost on part of the listener, and the cost of a resource shared by both parties–overall time spent with generating and comprehending the sequence. Then we will motivate local and global adjustments of trade-off parameters weighting these partial influences within the joint overall cost function in terms of localized and global H&H variation. Finally, we will show that the durational patterns that are optimal with respect to this overall cost function qualitatively reproduce a range of suprasegmental timing effects observed in stress-based languages, grounding them in a cognitively plausible model architecture.

Currently, the prevalent class of computational models of suprasegmental speech timing utilizes coupled oscillators in order to account for durational patterns in speech (O'Dell and Nieminen, 1999; Barbosa, 2007; Saltzman et al., 2008; Tilsen, 2009), and previous versions of the present model have incorporated similar assumptions