Speech Communication 88 (2017) 17–38

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Statistical modelling of phonetic and phonologised perturbation effects in tonal and non-tonal languages

Si Chen[a,*], Caicai Zhang[a,b], Adam G. McCollum[c], Ratree Wayland[d]

[a] Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong SAR, China
[b] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
[c] Department of Linguistics, University of California, San Diego, CA, US
[d] Department of Linguistics, University of Florida, FL, US

ARTICLE INFO

Article history:
Received 2 June 2016
Revised 8 December 2016
Accepted 6 January 2017
Available online 11 January 2017

Keywords:
Phonologisation
Perturbation
Functional data analysis
Growth curve analysis
Underlying pitch targets

ABSTRACT

This study statistically models perturbation effects of consonants on f0 values of the following vowel in order to quantify the differences between phonetic perturbation effects (i.e., phonetic variation) and phonologised perturbation effects (i.e., tone distinctions). We investigated perturbation effects in a non-tonal language, Japanese and a tonal language, Chongming Chinese. Traditional methods of modelling cannot distinguish phonetic and phonologised effects on surface f0 contours, as variation caused by both effects reached statistical significance. We therefore statistically modelled and tested the differences in underlying pitch targets, which successfully distinguished between phonetic and phonologised effects, and is robust to data variability. The methods used in this study can be further applied to examine perturbation effects cross-linguistically and shed light on the development of tones and stages of phonologisation more broadly.

1. Introduction

1.1. Gradient and categorical phenomena in speech

The distinction between phonetics and phonology and their relationship has been a subject of an on-going debate (Chomsky and Halle, 1968; Ohala, 1990; Keating, 1996; Steriade, 2000; Flemming, 2001; Keyser and Stevens, 2001; Arvaniti, 2007; Cohn, 2007; Kingston, 2007; Hyman, 2013). According to a modular view, phonology and phonetics are two distinct components of the grammar of sounds in a language; the former deals with discrete and categorical entities (phonological representations), and the latter deals with continuous and gradient phenomena (phonetic implementation) (Chomsky and Halle, 1968). Hyman (2013: 4) summarizes the characteristics that distinguish phonology and phonetics as the following: categorical vs. gradient, discrete vs. continuous, qualitative vs. quantitative, symbolic vs. physical, digital vs. analog, and syntactic vs. semantic. Under this view, the nature of the representation and the operatives within phonology and phonetics are fundamentally distinct. According to Jakobson and Halle's work (1962), a phonological feature refers to a phonetic (articulatory or auditory) property that serves to distinguish a lexical contrast. Physical differences between sounds such as a release burst is not considered a feature because no language has a phonemic distinction between released and unreleased stops (Steriade, 2000). In other words, standard phonological representations only include a subset of physical properties of sounds containing in a word or a phrase, corresponding roughly to its broad transcription (Flemming, 2001). Non-contrastive, fine-grained phonetic details such as segmental duration, timing, precision, coordination, etc., are assumed to be a consequence of universal principles (Chomsky and Halle, 1968), or are supplied by language-specific phonetic component of grammar (Keating, 1990,1994). In language evolution, phonetic variation is interpreted as phonological processes and phonetically motivated sound change leads to recurrent synchronic sound patterns (Blevins, 2004).

A number of criticisms have been raised against compartmentalizing phonetics and phonology to two separate components of the grammar. For example, Steriade (2000) argues that the distinction is neither productive nor enforceable. It is unproductive because phonological patterns cannot be understood without references to their physical manifestations, and it cannot be "coherently enforced" because multiple physical features are simultaneously required to implement most lexical contrasts; to single one out as the phonological feature necessarily involves some degree of

* Corresponding author.
E-mail addresses: sarah.chen@polyu.edu.hk, qinxi3@gmail.com (S. Chen), caicai.zhang@polyu.edu.hk (C. Zhang), agmccoll@ucsd.edu (A.G. McCollum), ratree@ufl.edu (R. Wayland).

http://dx.doi.org/10.1016/j.specom.2017.01.006
0167-6393/

arbitrariness (Steriade, 2000: 314). Similarly, Flemming (2001:10) points out that since phonological features are phonetically specified, it would appear that "sounds are represented twice in the grammar, once at the coarse level of detail in the phonology, and then again at the finer grain in the phonetics" . In addition, obvious parallels have been noted on so called 'phonological' (e.g. assimilation) and 'phonetic' (e.g. coarticulation) phenomena.

These criticisms are evident in more unified, constraint-based models of phonetics and phonology found in some contemporary phonology (e.g., Steriade, 2000; Flemming, 2001; McCarthy and Prince, 1993; Prince and Smolensky, 1993). In Flemming (2001), both categorical (phonological) and gradient (phonetics) phenomena are derived within the same component of grammar, and are subject to the same set of phonetics (speech production) constraints (e.g., minimize articulatory effort), resulting in their observed similarities. However, more integrated models often derive categorical and gradient phenomena differently. For example, in Optimality Theory (OT, McCarthy and Prince 1993; Prince and Smolensky, 1993), which represents an only partially unified model, categorical, non-phonetic contrast-maintenance faithfulness constraints such as "Don't deviate from inputs, IDENT" interact with phonetic constraints to yield an optimal output. In unified models, like Flemming (2001), there is little to differentiate categorical from gradient, and both are derived entirely from phonetic constraints (see Cohn, 2007 for a more modular account of categorical and gradient phenomena). In both Flemming's model and OT, outputs that optimally satisfy conflicting constraints are selected. However, unlike OT, constraint conflicts are resolved by a weighting system rather than a strict dominance system in Flemming's (2001) model (see also weighted Optimality theoretic grammars, e.g. Pater, 2009).

Regardless of theoretical accounts on how speech is underlyingly represented and derived, it is commonly acknowledged that surface, physical manifestation of speech exhibits a great deal of variation, simultaneously signalling its targets, functional, contrastive units while satisfying articulatory constraints. Thus, the central question in speech perception is (and has been for nearly six decades), how listeners separate 'substance' from 'form' in the physical signals. Broadly, two theoretical approaches have been proposed to provide an answer. The first approach, represented by the Motor Theory, suggests that speaker's intended, invariant neuromotor commands associated with underlying articulatory targets is retrieved from the acoustic signals by a specialized, speech-specific neural network (Liberman and Mattingly, 1985; Liberman et al., 1967). In contrast, the general auditory approach argues that objects of speech perception are auditory or acoustic events present in the speech signals. Relying on the same auditory and cognitive mechanisms evolved to perceive other sounds in the environment, the humans' auditory processing system is sensitive to statistical regularities in the distributions of acoustic properties as they co-vary with phonemic distinctions in different contexts (e.g., Diehl, Lotto and Holt, 2004). Evidence of either intrinsic or extrinsic normalization processes during speech perception lends support to this latter account of speech perception mechanism (e.g., Johnson, 2008; Zhang and Chen, 2016).

The overall goal of this current study is to better understand the differences between categorical and gradient effects on the physical realization of speech. Specifically, the study attempts to show that these two effects may be separated using statistical modelling (see Shih, 2005 for an analysis of Mandarin Tone 2 sandhi), thus allowing underlying functional targets of speech to be directly extracted from its physical, acoustic signals and compared statistically. If successful, such approach would not only provide new insights into the human's speech perception mechanism, but also significantly improve computerized speech recognition systems.

The study has 3 specific aims. The first aim is to explore categorical and gradient perturbation effects in a tone (Chongming Chinese) and a pitch-accent (Japanese) language. The second goal is to determine statistical modelling procedures that can most effectively differentiate gradient and categorical perturbation effects in both languages. The third goal is to compare results of different methods of statistical modelling and to extend the models for future investigations of categorical and gradient pitch (f0) phenomena in the world's languages.

The remainder of the paper is organized as follows. Sections 1.2–1.5 introduce phonetic and phonologised perturbation effects, statistical modelling of surface f0 contours and underlying pitch targets, and present hypothesized situations where modelling procedure may differentiate between the two. Section 2 evaluates all proposed statistical methods with Japanese data. Section 3 further applies them to Chongming Chinese data. Sections 4 and 5 provide discussion and conclusion of the results, as well as their implications for future studies of phonetic and phonological perturbation.

1.2. Phonologisation of perturbation effects

It is generally acknowledged that surface representations of underlying segmental units such as vowels and consonants show a great deal of variation (e.g., Lindblom, 1963; Öhman, 1966; Steven and House, 1963). A similar conclusion has been reached for the physical manifestation of suprasegmental phenomena in speech including tone, pitch accent and intonation, whose main acoustic correlate is f0 level or f0 contour (e.g., Xu and Wang, 2001). A number of articulatory constraints contributing to observed f0 variations have been documented including vowel intrinsic f0 (e.g., Lehiste and Peterson, 1961; Shi and Zhang, 1987; Whalen and Levitt, 1995) and initial consonant f0 perturbation (Hombert, 1978; Howie, 1974; Lehiste, 1975; Lehisted and Peterson, 1961)

Perturbation effects of preceding consonants on f0 are widely noted in many languages, both tonal and non-tonal (Abramson and Lisker, 1985; Gandour, 1974; Hyman, 1973a, 1973b). Some consonants tend to raise f0 and others lower it. For example, fricatives exert a greater f0 raising effect than stops in Mandarin Chinese (Shih, 2001). Moreover, initial voiced consonants exhibit an f0 lowering effect on the following vowels whereas voiceless consonants exert the opposite effect. These effects have been attested in a number of languages including Yoruba (Hombert, 1978), Siamese (Gandour, 1974), Yucatec Maya (Frazier, 2009) and Phuthi (Donnelly, 2009). The effects of voiceless unaspirated vs. voiceless aspirated consonants on pitch are also reported, though not as consistent as the effects of onset voicing (see Chen, 2011 for a summary).

These perturbation effects have been claimed to play a role in the phonologisation of f0 and ultimately the development of lexical tone contrasts. For instance, f0 perturbation caused by consonants formed the basis for the widely adopted theory of tonogenesis (Haudricourt, 1954; Chen, 2000; Hombert et al., 1979; Matisoff, 1973; Rose, 2002; Svantesson and House, 2006). In his analysis of the origin of lexical tones in Vietnamese, Haudricourt (1954) proposes that f0 perturbation effects of initial and final consonants on the following and the preceding vowels play a direct role in the development of the six lexical tones in Vietnamese: proto initials determine pitch height or register (high vs. low) whereas proto final consonants determine pitch contour (level, falling and rising). Thurgood (2002, 2007) replaces Haudricourt's consonant-based account with a laryngeal-based account of tonogenesis, arguing for an intermediary stage of voice quality distinctions (e.g., breathy, clear and creaky) developed after initial proto voiced and voiceless, and proto final voiced sonorants, stops and voiceless fricatives, which are responsible for pitch height and pitch contour distinctions in Vietnamese. In other words, Thurgood argues