Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 57 (2014) 181–208

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Toward invariant functional representations of variable surface fundamental frequency contours: Synthesizing speech melody via model-based stochastic learning

Yi Xu [a], Santitham Prom-on [a,b,*]

[a] Department of Speech, Hearing and Phonetic Sciences, University College London, London WC1N 1PF, United Kingdom
[b] Department of Computer Engineering, Faculty of Engineering, King Mongkut's University of Technology Thonburi, Bangkok 10140, Thailand

Received 3 January 2013; received in revised form 7 September 2013; accepted 30 September 2013
Available online 15 October 2013

Abstract

Variability has been one of the major challenges for both theoretical understanding and computer synthesis of speech prosody. In this paper we show that economical representation of variability is the key to effective modeling of prosody. Specifically, we report the development of PENTAtrainer—A trainable yet deterministic prosody synthesizer based on an articulatory–functional view of speech. We show with testing results on Thai, Mandarin and English that it is possible to achieve high-accuracy predictive synthesis of fundamental frequency contours with very small sets of parameters obtained through stochastic learning from real speech data. The first key component of this system is syllable-synchronized sequential target approximation—implemented as the qTA model, which is designed to simulate, for each tonal unit, a wide range of contextual variability with a single invariant target. The second key component is the automatic learning of function-specific targets through stochastic global optimization, guided by a layered pseudo-hierarchical functional annotation scheme, which requires the manual labeling of only the temporal domains of the functional units. The results in terms of synthesis accuracy demonstrate that effective modeling of the contextual variability is the key also to effective modeling of function-related variability. Additionally, we show that, being both theory-based and trainable (hence data-driven), computational systems like PENTAtrainer can serve as an effective modeling tool in basic research, with which the level of falsifiability in theory testing can be raised, and also a closer link between basic and applied research in speech science can be developed.

Keywords: Prosody modeling; Target approximation; Parallel encoding; Analysis-by-synthesis; Simulated annealing

1. Introduction

Like the segmental aspects of speech (Perkell and Klatt, 1986), and perhaps to an even greater extent, speech prosody exhibits extensive variability and uncertainty, which makes its computational modeling extremely difficult. Among the various aspects of prosody, fundamental frequency (F_0) is by far the most challenging, and has attracted most of the research effort. Many theories and computational models of F_0 patterns have been proposed over the years (Anderson et al., 1984; Bailly and Holm, 2005; Black and Hunt, 1996; Fujisaki et al., 2005; Grabe et al., 2007; Hirst, 2005, 2011; Jilka et al., 1999; Kochanski and Shih, 2003; Mixdorff et al., 2003; Pierrehumbert, 1980, 1981; Prom-on et al., 2009; Taylor, 2000; van Santen and MÎbius, 2000; Xu and Wang, 2001; Xu, 2005), and a large number of empirical studies have been conducted (as reviewed by Wagner and Watson, 2010; Shattuck-Hufnagel and Turk, 1996; Xu, 2011). Despite the extensive effort,

* Corresponding author. Address: Department of Computer Engineering, Faculty of Engineering, King Mongkut's University of Technology Thonburi, 126 Prachauthit Road, Bangmod, Thungkhru, Bangkok 10140, Thailand. Tel.: +66 (0) 2470 9081; fax: +66 (0) 2872 5050.
E-mail addresses: yi.xu@ucl.ac.uk (Y. Xu), santitham@cpe.kmutt.ac.th (S. Prom-on).

0167-6393/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.specom.2013.09.013

however, most of the critical issues still remain unresolved and some are still under heated debate (Arvaniti and Ladd, 2009; Ladd, 2008; Wagner and Watson, 2010; Wightman, 2002; Xu, 2011). This lack of consensus has been an obstacle to linking basic prosody research to applied areas, resulting in slow advances in developing applications with capabilities for processing prosody.

One way to foster significant advances in prosody research is to develop computational models that can be used for theory testing. Such models would allow the translation of theories and empirical findings into algorithms that can predict fully continuous prosodic patterns, which can be directly compared to real speech data. Furthermore, and perhaps more importantly, such computational models would enable theories to predict phonetic details beyond the specific phenomena for which they were originally proposed. Testing such predictive powers would not only help demonstrate theories' generalizability, but also make them readily applicable to speech technology once the test results are positive. The present study is part of our continued effort in this direction, with a significant extension from our previous work (Prom-on et al., 2009), and with particular focus on the problem of variability. Before describing our current work, however, we will first discuss the main sources of prosodic variability and review how they have been addressed so far.

1.1. Two types of prosodic variability

Like in the case of segmental aspect of speech (Ladefoged, 1967; Peterson and Barney, 1952), the nature of prosodic variability is best highlighted by controlled comparisons. Fig. 1 displays two very different types of F_0 variability with previously reported empirical data (Liu and Xu, 2005; Xu, 1997). The first type is contextual variability, defined as the varying F_0 manifestation of a tonal category as a function of its adjacent tones. As shown in Fig. 1A, contextual variability is mostly assimilatory: when the same tone in the second syllable of each graph is preceded by four different tones in the first syllable, its F_0 contour varies extensively, espe-

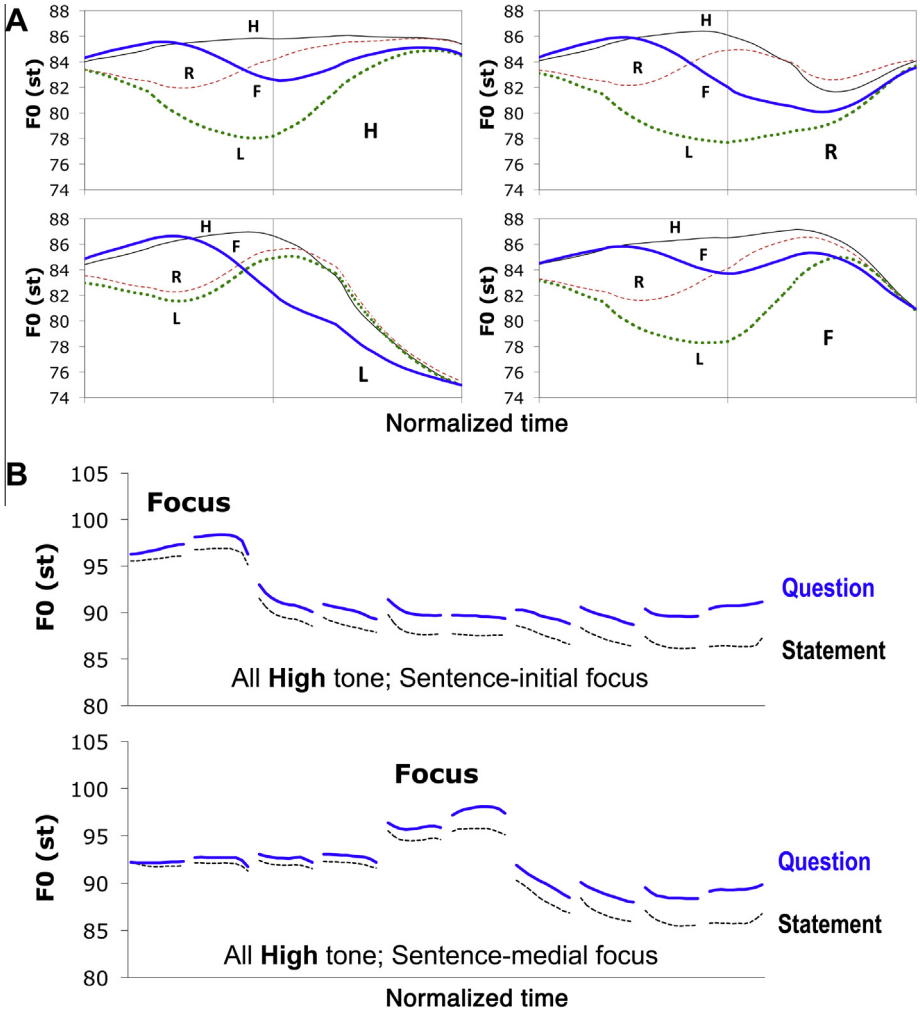

Fig. 1. (A) Mean F_0 contours of Mandarin tones in disyllabic sequences (mama) spoken by eight male speakers (data from Xu, 1997). In each plot the tone of the second syllable is held constant while that of the first syllable alternates across four tones. (B) Mean F_0 contours of Mandarin sentence (Zhangwei danxin Xiaoying kaiche fayun [Zhangwei is concerned that Xiaoying may get dizzy when driving]), spoken by eight speakers (four females and four males) as statement or question and with focus on the first or third disyllabic word (data from Liu and Xu, 2005).