Speech Communication 89 (2017) 92–102

Contents lists available at ScienceDirect

Speech Communication

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

# Quantitative intonation modeling of interrogative sentences for Mandarin speech synthesis

Ya Li[a,*], Jianhua Tao[a,b,c], Wei Lai[a,d], Xiaoying Xu[a,d]

[a] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[b] CAS Center for Excellence in Brain Science and Intelligence Technology,
[c] School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing, China
[d] Department of Chinese Language and Literature, Beijing Normal University, Beijing, China

ARTICLE INFO

*Article history:*
Received 4 September 2015
Revised 10 March 2017
Accepted 21 March 2017
Available online 22 March 2017

*Keywords:*
$F_0$ declination
Intonation
Interrogative sentences
Final lowering
Prosody

ABSTRACT

Previous intonational research on Mandarin has mainly focused on the prosody modeling of statements or the prosody analysis of interrogative sentences. To support related speech technologies, e.g., Text-to-Speech, the quantitative modeling of intonation of interrogative sentences with a large-scale corpus still deserves attention. This paper summarizes our work on the quantitative prosody modeling of interrogative sentence in Mandarin. A large-scale natural speech corpus was used in this study. By extracting the pitch contours and fitting the intonation curves, we found that $F_0$ declination and final lowering both existed in interrogative sentences, while they were claimed to be absent in Mandarin in some previous studies. In addition, the declination function could be modeled linearly, and the bearing unit of final lowering in Mandarin was found to be the last prosodic word in the utterance, regardless of its length, rather than a fixed duration range. It was argued in this study that the difference between this finding and the commonly believed rising intonation of the interrogative sentences resulted from the nonlinear relationship between prosody production and perception. The underlying mechanism for the existence of $F_0$ declination and final lowering in interrogative sentences is also discussed.



## 1. Introduction

By contrast to the long history of work on incorporating intonation models into Text-to-Speech systems to generate natural $F_0$ contours for synthesized English declarative and interrogative utterances (Mattingly, 1966; Qian et al., 2011; Santen and Moebius, 2000; Taylor, 2000), research on $F_0$ contour generation in Mandarin TTS systems has focused much more on syllable-level tone contrasts, including models of tonal coarticulation across syllables and interaction with phrase-level downtrends in declarative statements (Prom-on et al., 2009; Qian et al., 2011; Shih and Sproat, 1998; Yuan and Liberman, 2014).

Unlike the literature on English interrogative intonation patterns, research on Chinese interrogative intonation is complicated because of the conflict on pitch perception between tone and intonation (Ren et al., 2013). Moreover, much of the instrumental literature on $F_0$ contours in Mandarin questions has focused on declarative questions – i.e., questions where there is no morphosyntactic marking of the interrogative function. Previous instrumental studies model the differences between the intonational effects on declarative and interrogative utterances in terms of the following aspects. a) Trends of top/base lines; Gårding (1987) modeled Chinese intonation with grids, which qualitatively marks a time-varying pitch range that is different for interrogative intonation and declarative intonation. Shen adjusted Gårding's grid model and suggested a gentle fall on the top line and a slight rise on the base line for interrogative sentences (Gasrding, 1987; Shen, 1994). Lee (2005) found a rise on the top line and significant expansion of the pitch range are the crucial localized $F_0$ cues in yes-no questions. The expansion of pitch range was also found crucial in Yang (1995). b) Starting point: Shen (1990) proposed that all types of questions begin with a higher pitch register than statements. c) Boundary tone: Lin adopted autosegmental–metrical theory and emphasized the role of boundary tone in the distinction between interrogative and declarative intonation (Lin, 2004, 2006). d) Phrase curve and strength: Yuan suggested that an overall higher phrase curve and higher strength on final syllables accounts for interrogative intonation in Chinese (Yuan et al., 2002). For questions with question words or final particles, the interrogative intonation would have more variations (Wang, 2008).

* Corresponding author.
*E-mail addresses:* yli@nlpr.ia.ac.cn (Y. Li), jhtao@nlpr.ia.ac.cn (J. Tao), laiwei_0508@126.com (W. Lai), xuxiaoying2000@bnu.edu.cn (X. Xu).

http://dx.doi.org/10.1016/j.specom.2017.03.002
0167-6393/

Although considerable studies have been carried out on interrogative intonation in Mandarin, they cannot fully satisfy the needs of Text-to-Speech. For one thing, these studies focus on the general intonation trend of questions (Shen, 1994, 1990; Yuan et al., 2002), whereas more concrete details of pitch variation within the whole sentence are still eagerly needed to support Text-to-Speech. For instance, is pitch variation distributed evenly over the whole sentence, or is it mainly realized by specific parts of the sentence? Such issues are still under debate. For another thing, to generate natural prosody, conclusions from experimental corpus should be verified in large-scale natural speech. Since Chinese is a tonal language, previous studies on intonation tended to adopt designed stimuli to control the effect of tone (Shih, 2000; Wang et al., 2012; Xu, 2006; Xu and Wang, 1997; Yuan, 2004; Yuan et al., 2002); there are only a few studies in the literature about the interrogative intonation of Chinese in large-scale natural speech (Lee, 2005). To make up for these inadequacies, this paper particularly examined the global and utterance-final $F_0$ trends in a corpus of recorded utterances of question-answer pairs selected from online transcriptions of interviews in the Chinese University of Communication Media Language Corpus, to investigate the detailed prosody variations, and thereby contributes to the generation of natural prosody in Text-to-Speech.

The next two subsections review the relevant previous literature on $F_0$ declination and final lowering, which are the two constructs that have been proposed as relevant in differentiating statements from questions in many languages. Previous studies found a downtrend of the overall pitch through the utterance for many languages, especially in statements (Pierrehumbert, 1979). There are many factors that can lead to a pitch downtrend. One is the declination effect, which refers to the tendency of $F_0$ to decline over the course of an utterance and which has been observed in many languages (Cohen et al., 1982; Ladd, 1984; Pierrehumbert, 1979; Shih, 2000; Umeda, 1980). Another is final lowering, referring to an additional lowering effect near the end of the sentence (Liberman and Pierrehumbert, 1984).

### 1.1. $F_0$ declination

$F_0$ declination has been found in many languages and for a long time has been treated as a universal property of speech intonation. Many intonation models adopt declination as a baseline upon which more local $F_0$ patterns reside.

The trend line of $F_0$ declination has been modeled in different ways by previous studies, for example, explicit (Lieberman et al., 1985) and implicit modeling of $F_0$ declination (Fujisaki, 1983). Explicit $F_0$ declination is described by fitting trend lines to the $F_0$ contour of the utterance or to certain salience points in the contour. Some studies characterized declination with a single line, e.g., a linear regression line fitted to all $F_0$ points (Lieberman et al., 1985). Others suggested that two separate reference lines are required to characterize $F_0$ declination, i.e., the topline that connects the peak values of $F_0$ and the baseline that connects the valley values, because the two lines show different patterns in coding linguistic structures (Cooper and Sorensen, 2012; Gårding, 1979; Maeda, 1976; O'Shaughnessy, 1976). By contrast, the implicit modeling of declination has been mostly obtained by mathematical modeling instead of empirical observation, for instance, in Fujisaki model (Fujisaki, 1983) accent command responses are superimposed onto the exponentially decaying phrase component, and the $F_0$ contour is the sum of the responses from the phrase components and the accent components.

The above characteristics of declination reveal another dichotomy: whether declination is straight or bent. Fitting straight lines to $F_0$ values facilitates the observation of the trend and the slope of declination and is widely used in intonational research, (e.g., Hart and Collier, 1979; Maeda, 1976). However, researchers have found that the declining rate is not always consistent within the whole utterance. Cooper and Sorensen (2012) modeled the declination by two successive straight lines, the steeper one is prioritized, indicating that the declining rate is faster at the beginning and becomes slower approaching the end of the utterance. In both Pierrehumbert's model (Pierrehumbert, 1980) and Fujisaki's model, the declination line decays exponentially as a function of time. Yuan and Liberman (2014) compared the topline and baseline of Chinese and English, and found that in English, both topline and baseline show an initial plateau, a middle declination and a final lowering, while in Chinese Mandarin, the topline is similar to English and the baseline is close to a straight line.

The declination trend was found to be closely related to utterance duration in plenty of previous research. Briefly speaking, a longer utterance starts at a higher pitch and contains a shallower slope (Cooper and Sorensen, 2012; Gårding, 1979; Hart and Collier, 1979; Maeda, 1976; Swerts et al., 1996). This relationship is considered as a sign of the existence of phrase-scale preplanning. Thorsen (1980) reported that statements have the steepest falling contours; syntactically unmarked questions have the gentlest ones, and in between these two endpoints are other types of questions and non-terminal statements. When more factors such as inter-speaker variation are taken into consideration, the relationship between declination trend and utterance duration turned out to be unclear in some previous studies (Heuven, 2004; Laniran and Clements, 2003; Liberman and Pierrehumbert, 1984; Prieto et al., 2006; Prieto et al., 1996).

The physical mechanism of declination has been investigated as well and several mechanisms/principles have been put forward, e.g., a drop in subglottal air pressure (Collier, 1975; Gelfer et al., 1983; Lieberman, 1967), a "tracheal pull" which lowers the sternum and the larynx (Maeda, 1976), or a "laziness principle" (Vaissière, 1983). However, it is still under debate whether declination is totally passive or whether it somehow involves active control of the respiratory muscles to signal linguistic meaning. If declination also involves some kind of phonological process, then this phonological nature lies in the fact that declination can be used for normalization in speech perception. For example, listeners would perceive two peaks as equally high in pitch when the second is acoustically lower (Pierrehumbert, 1979; Tseng, 2006). Another argument for its linguistic nature is that declination is dependent on sentence type; particularly, it is often suppressed in interrogative sentences (Thorsen, 1980).

As has been mentioned above, most previous research on intonation modeling was based on statements and we still do not know whether these modeling methods are applicable to interrogative sentences. Since the major prosodic difference between statements and interrogative sentences is the intonation, an accurate intonation model of interrogative sentences is still pressing need to the related speech technologies.

### 1.2. Final lowering

The second factor for $F_0$ drift down is final lowering, which indicates an additional lowering near the end of the utterance and has been found in many languages, e.g., Greek (Arvaniti and Godjevac, 2003), Danish (Thorsen, 1984), Dutch (Gussenhoven and Rietveld, 1988), Yoruba (Connell and Ladd, 1990; Laniran, 1993), Spanish (Prieto et al., 1996), German (Truckenbrodt, 2004), Japanese (Pierrehumbert and Beckman, 1988; Poser, 1984), Kipare (Herman, 1996) and Chinese (Lai et al., 2014a).

Potential physiological triggers of final lowering were first discussed by Liberman and Pierrehumbert (1984). They thought that a drop in the subglottal pressure and relaxation of the laryngeal muscles were necessary articulatory correlates to final lowering.