Estimation of intonation variation with constrained tone transformations

Jinfu Ni a, Hisashi Kawai a,b, and Keikichi Hirose c

a) ATR Spoken Language Translation Research Laboratories, Japan
b) KDDI R&D Laboratories Inc., Japan
c) Department of Information and Communication Engineering, University of Tokyo, Japan

Abstract

This paper presents a method for quantitatively estimating intonation variation in Mandarin speech. Intonation variation is relative to identical lexical tone structures, and its estimation is performed on two sets of fundamental frequency (F0) contours: one for normal F0 contours and the other for variants. This is done by transforming target F0 contours in pairs from the norms to the variants in which the prosodic contribution to these F0 contours is analyzed as sequences of targets, all of which are confined to the basic elements of the underlying lexical tone structures. The tone transformations are constrained under an assumption of the structural formulation of F0 contours proposed previously. When the norms take the base values of the four lexical tones measured from isolated words in a neutral mood and voice, this method solves acoustic correlations of tone and intonation from the observed F0 contours. The method was implemented on a computer, and its capability of estimating intonation variation was shown through the analysis and synthesis of F0 contours.

1. Introduction

In spoken Chinese, the fundamental frequency (F0) contours of the voice can simultaneously manifest syllabic tones and sentence intonation. In Mandarin, four lexical tones exist traditionally called the first, second, third, and fourth tones (henceforth, Tones 1 to 4), each of which has its own features distinct from others as well as a neutral tone (Tone 0) without distinctiveness. For example, there are five distinct words with the same phonemic string: /ma1/ (mother), /ma2/ (numb), /ma3/ (horse), /ma4/ (to curse), and /maU/ (a question particle); each is distinguished from others by a unique type of intonation. Meanwhile, a sentence, for example, can be uttered as a question or a statement; and one part of the utterance can be emphasized. Moreover, the attitudes and emotions of the speaker can be added to the cognitive meaning of the words. Thus, Mandarin has similar intonation conveyed communicative functions found in such non-tonal languages as English. This poses an important problem: how to deal with the acoustic correlations of tone and intonation. Solutions to the problem are critical for synthesizing F0 contours in the next generation of text-to-speech systems applied to conversational systems, for example, where marked information may be available for generating expressive mood besides normal or neutral intonation.

From an analytical point of view, as remarked in [1], the key issue in the description of concrete intonation in Mandarin and in drawing general conclusions from the description is the appropriate way of separating the complex prosodic modifications of a string of one or more syllabic tones from the final prosodic shape of sentences. In recent years, several approaches have been proposed for estimating intonation variation: [2] analyzed the pitch ranges of phrases and [3] measured prosodic strength. Other interesting approaches are omitted for space limitations. However, existing methods seem limited in handling this issue. For instance, no reliable method is available yet for measuring the prosodic differences between two observed F0 contours, given identical underlying lexical tone structures. The term a lexical tone structure indicates the final form of a sequence of syllabic tones. The term sandhi rules. A well-known sandhi rule is that Tone 3 changes into Tone 2 if followed by another Tone 3. The approach presented here is based on an advanced theory of Tone [4] and Intonation [1] that estimates intonation variation from observed F0 contours. First, the four lexical tones themselves constitute a system for the function of distinguishing meaning. They can be sparsely specified by a few pitch targets on a five-point scale [4]. Second, intonation exists in the form of specific modifications of a string of one or more tones [1]. From such points of view, the underlying lexical tone structure of an observed F0 contour constructs a framework whereby both tone and intonation are acoustically articulated together. Moreover, the base values of lexical tones could be specific in an individual voice. Thus, intonation variation between two sets of F0 contours would be involved in observed target F0 values in pairs confined to the basic elements of the underlying lexical tone structures. In other words, intonation variation can be regarded as a series of local tone variation in F0 constrained by tone transformations. Section 2 introduces a structural formulation of F0 contours [5] that underlies a tone transformation suited for this purpose. Section 3 presents an algorithm for estimating intonation variation thereof. Experimental confirmation is described in Section 4, and Section 5 gives conclusions.

2. Background assumption: an F0 model

In Mandarin, an F0 contour in an individual voice for a sentence can simultaneously manifest both tone and intonation. Let F0(t) represent an F0 contour as a function of time t, and parameters f0b and f0t indicate the bottom and top frequencies of the vocal range, respectively. Assume a function Λ(t) to indicate a sequence of virtual tone graphs hypothetically independent of the individual voice and sentence intonation in a so-called λ-time space. The term Λ indicates square frequency ratios of forced vibrating systems applied to the physical interpretation of the physiological-based body-cover hypothesis [6] for a more quantitative look at the mechanics of F0 control. Additionally, assume a latent scale ζ(t) to characterize sentence intonation, a hypothetical construct invoked to explain observed co-variation in tonal behavior. ζ(t) is specified by damping ratios of the forced vibrating systems. Thus, the F0 contour on the logarithmic scale of fundamental frequency is expressed as a scale transformation of the sequence of tone graphs from λ-
time space to frequency-time space, corresponding to the syllabic tones fitting themselves with sentence intonation in the vocal range. In a mathematical term [5],

$$\frac{\ln f_0(t) - \ln f_{0b}}{\ln f_{1b} - \ln f_{0b}} = \frac{A(\lambda(t), \zeta(t)) - A(\lambda_0, \zeta_0)}{A(\lambda_0, \zeta_0) - A(\lambda_0, \zeta(t))} \quad t \geq 0,$$

(1)

where

$$A(\lambda, \zeta) = \frac{1}{(1 + (1 - 2\lambda^2)\lambda^2 + 4\zeta^2(1 - 2\zeta^2))^{1/2}} \quad \lambda \geq 1.$$  

(2)

Here, \(A(\lambda, \zeta)\) indicates amplifying coefficients of forced vibrating systems (also known as resonance curves); and \(\zeta \in (0, 0.707)\) indicates damping ratios of the systems. Parameters \(\lambda_0\) and \(\lambda_b\) can be commonly fixed at 1 and 2, respectively. In other words, the top \(f_b\) is realized at the minimum \(\lambda = 1\) and the bottom \(f_{0b}\) at \(\lambda = 2\): the larger the \(f_0\) in \([f_{0b}, f_b]\), the smaller the \(\lambda\).

Equations (1)-(2) jointly indicate a structural formulation of the hidden process of articulating syllabic tones and sentence intonation into a final sentence melody. In Mandarin, the syllabic tones can be determined from text with a few tone sandhi rules; each syllable has a tone, and sentence intonation may vary from one utterance to another. Equation (1) states that \(F_0\) contour \(F_0(t)\) is a transformation of a sequence of virtual tone graphs \(A(t)\) on a latent scale \(\zeta(t)\). Equation (2) expresses how to perform the transformations when given the latent scale. Technically, \(\zeta(t)\) anchors these tone graphs in the vocal range and simultaneously adjusts their height and ranges in \(F_0\).

A testable consequence of the structural formulation would be a function of input \(A(t)\) to output \(F_0(t)\), if different values set to \(\zeta(t)\) could lead to the same input/output function. This shall make sense of parameter estimation as a first step to optimal approximations of an observed \(F_0\) contour by simply fixing \(\zeta(t)\) at default values. For this purpose, \(A(t)\) is eventually expressed as the concatenation of \(n\) parametric mountain-shaped patterns lined up in a series on the time axis:

$$A(t) = A_{r_i}(t) + \sum_{i=1}^{n-1} \min(\lambda_{f_i}A_{r_{i+1}}(t) + A_{f_i}(t)).$$

(3)

Here, \(A_{r_i}(t)\) and \(A_{f_i}(t)\) respectively indicate the rising and falling components of the \(i\)th mountain-shaped pattern, namely:

$$A_{r_i}(t) = \begin{cases} \lambda_{r_i} + \Delta \lambda_{r_i}(1 - D_{r_i}(t, p_i, t)), & t \leq p_i, \\ 0, & \text{otherwise}, \end{cases}$$

(4)

$$A_{f_i}(t) = \begin{cases} \lambda_{f_i} + \Delta \lambda_{f_i}(1 - D_{f_i}(t, p_i, t)), & t \geq p_i, \\ 0, & \text{otherwise}, \end{cases}$$

(5)

where

$$D_{r_i}(t) = (1 + \frac{48}{\Delta \lambda_{r_i}})e^{-\frac{48}{\Delta \lambda_{r_i}}}, \quad x \in \{r, f\}.$$  

(6)

The model parameters through Eqs. (3)-(6) indicate

\(n\): number of mountain-shaped patterns,

\(\{p_i, \lambda_{r_i}\}: \text{ith peak coordinate in the time-space},

\(\Delta \lambda_{r_i}, \Delta \lambda_{r_i}: \text{response time for the ith rising component,}

\Delta \lambda_{f_i}, \Delta \lambda_{f_i}: \text{amplitude of the ith rising component,}

\Delta t_f \text{ith time for the ith falling component,}

\Delta \lambda_f: \text{amplitude of the ith falling component,} \quad i = 1, \ldots, n.$$

For simple citations, hereafter let the following expressions be called the respective solutions of \(f_0\), \(\lambda\), and \(\zeta\) subjected to Eqs. (1)-(2), given any two of them and vocal range \([f_{0b}, f_b]\):

$$f_0 = T_{\lambda\zeta}(\lambda, \zeta), \text{computing } f_0 \text{ from both } \lambda \text{ and } \zeta\text{ values.}$$  

(7)

$$\lambda = T_\lambda(f_0, \zeta), \text{computing } \lambda \text{ from both } f_0 \text{ and } \zeta \text{ values.}$$  

(8)

$$\zeta = T_\zeta(\lambda, f_0), \text{computing } \zeta \text{ from both } \lambda \text{ and } f_0 \text{ values.}$$  

(9)

Equation (7) can be directly calculated by using Eqs. (1)-(2), while Eqs. (8) or (9) can be solved by an iteration process.

3. Outline of the method

3.1. Tone transformation technique

Transformations are the essence of a generative theory. A tone transformation shall specify procedures that, when applied to the basic elements of a tone structure, will produce all the more complex forms from its base values. The term tone transformation here means a computational method for such a purpose.

Figure 1 presents schematic representations of the four lexical tones where lines indicate tone patterns. The basic elements of tone structures are displayed by white and black circles and marked by \(\text{Pj}(\text{or Vj})\), \(j \in \{1, 2, 3, 4\}\), and \(i \in \{1, 2\}\): each lexical tone has a primary target (black circles).

Let \(f_{0i}\) indicate the norm values of the \(i\)th target (a peak or a valley) of a lexical tone, \(i \in \{1, 2, 3\}\); \(f_{0i}\) may take its base values, as in (12), which can be measured from neutral-mood, isolated words. Also, \(f_{0i}\) may take observed values from any \(F_0\) contour used as norms. By using Eq. (8), each \(f_{0i}\) can be converted to a \(\lambda\) value with respect to \(\zeta_0\) (hereafter, \(\zeta_0 = 0.156\)). Inversely, the \(\lambda\) value can be transformed to numerous \(F_0\) values with \(\zeta\) varying from 0 to 0.707, denoted by \(f_{0i}\) (hereafter, a variable with \(\zeta\) shows its values relative to transformed values). By using Eq. (7), a tone transformation is expressed as:

$$f_0 = T_{\lambda\zeta}(\lambda, \zeta), \quad \zeta \in \{1, 2, 3\}.$$  

(10)

To balance \(\zeta\) on both sides of default value \(\zeta_0\), normalized damping ratios \(\zeta_n\) are defined as a closed set \([-1, 1]\):

$$\zeta_n = \begin{cases} \frac{(\zeta - \zeta_0)}{(0.7 - \zeta_0)}, & \zeta \geq \zeta_0, \\ \frac{(\zeta - \zeta_0)}{\zeta_0 - \zeta_0}, & \zeta < \zeta_0. \end{cases}$$  

(11)

Figure 2 demonstrates tone transformation applied to the base values of the four lexical tones listed in (12) uttered by a female narrator (WL) having vocal range \([100, 500]\). Panel (a) shows the citation forms of four lexical tones (a form of virtual tone graphs) with an assumed temporal structure. The citation forms are superimposed together over time and repeated six times. Panel (b) shows citations \(\zeta_n(t) = 0\), i.e., \(\zeta(t) = \zeta_0\). When \(\zeta_n(t)\) varies linearly from 0 to -1 over a two-second interval shown in (c), these citation forms in (a) will change into the transformation forms shown in (d). While \(\zeta_n(t)\) appears as curves shown in (e), corresponding transformation forms are displayed in (f).

This demonstration shows that it is possible to transform the base values of the four lexical tones to the possible variants in the vocal range. Suppose that these citation forms were drawn on an elastic transparent sheet, as described in (4); stretching the sheet vertically or letting it shrink would alter their ranges and height in the vocal range. The relationships between the citation forms of the four lexical tones and their transformation forms can be compared with a Chinese kanji character and its handwritten form, where its internal structure maintains unchanged in the sense that a human can still recognize the character. It is also important for any lexical tone to distinguish itself from others regardless of sentence intonation.

Figure 1: Schematic of lexical tone patterns represented by peaks and valleys (circles); black circles are primary targets.
3.2. Estimation of intonation variation

An algorithm includes three steps for estimating intonation variation from norms \( F_{0u} (t) \) to variants \( F_{0v} (t) \), given that both the norms and variants have identical lexical tone structures.

**Input:**
- norms \( F_{0u} (t) \) and a variant \( F_{0v} (t) \)
- underlying lexical tone structure, and
- values of the vocal range \([f_{0u}, f_{0v}]\).

**Step 1:** Determine sparser specifications of the two \( F_{0} \) contours:
- make sparser specifications of the \( F_{0} \) contours confined to basic elements of the lexical tone structure; and
- an algorithm in [7] is basically suited for this purpose;
- if Tone 3 takes a half form, then only targets P31 and V31 exist; if a syllabic tone is Tone 0, it may be ignored.

**Step 2:** Determine target pairs \((f_{0u,j}, f_{0v,j})\), \( j = 1, \ldots, N \), where
- \( N \) indicates the number of target peaks and valleys;
- \( f_{0u,j} \) indicates the \( j \)th target \( F_{0u} \) value on \( F_{0u} (t) \); and
- \( f_{0v,j} \) indicates the \( j \)th target \( F_{0v} \) value on \( F_{0v} (t) \).

**Step 3:** Compute \( \zeta_{j} = T_{j} (T_{j} (f_{0u,j}, \zeta_{j}), f_{0v,j}), j = 1, \ldots, N \); and convert each \( \zeta_{j} \) to \( \zeta_{0u,j} \).

Output: \((t_{j}, \zeta_{0u,j})\), \( j = 1, \ldots, N \); \( t_{j} \) is the timing of the \( j \)th target.

Simply stated, the prosodic contribution to two sets of \( F_{0} \) contours is the analyzed sequences of targets for sparser representations of the \( F_{0} \) contours. Intonation variation between them is then measured as local tone variation from target \( F_{0} \) values \( f_{0u,j} \) to \( f_{0v,j} \), \( j = 1, \ldots, N \). Note that, if the norms take the base values of the four lexical tones as (12), no \( F_{0u} (t) \) is needed. Then \((t_{j}, \zeta_{0u,j})\), \( j = 1, \ldots, N \), indicate sentence intonation conveyed by \( F_{0u} (t) \) but excludes the lexical tone effects.

4. Experimental confirmation

The validity of our novel method can be tested by its ability to analyze observed samples in the real world. For this purpose, we select several speech samples from a Chinese speech corpus to evaluate the proposed method in two experiments. Experiment 1 demonstrates the reliability of the method. Experiment 2 is intended to solve the acoustic correlations of tone and intonation and test the regularity of separated intonation patterns.

4.1. Experiment 1: statement-question conversion

This is a demonstration of conversion between statements and questions using the tone transformation technique. The speech samples here are two pairs of utterances from a female native:

(1) /guo4lu4ke4zhao4xiang4/ (A passerby is taking photos.)
(2) /guo4lu4ke4zhao4xiang4/ (Is a passerby taking photos?)
(3) /hong2bi2tou2mei2quan2/ (The Red Nose has no power.)
(4) /hong2bi2tou2mei2quan2/ (Has the Red Nose no power?)

Each pair, (1) vs. (2) and (3) vs. (4), is characterized by identical lexical tone and syntactic structures. The observed \( F_{0u} \) contours are shown in Fig. 4 by “+” sequences. Let \( F_{0u} (t) \) and \( F_{0v} (t) \) denote a pair of observed \( F_{0} \) contours in statements and questions, respectively; vocal range \([f_{0u}, f_{0v}] = [100 \text{ Hz}, 500 \text{ Hz}]\).

The experimental procedure is roughly described as follows:

1. Calculate prosodic modifications, for instance, from \( F_{0u} (t) \) (the norms) to \( F_{0v} (t) \) (the variants), denoted by \( \zeta_{0u} (t) \).
2. Superimpose the prosodic modifications \( \zeta_{0u} (t) \) on \( F_{0v} (t) \), thus converting the statements to questions with the following synthetic \( F_{0} \) contours \( F_{0a} (t) \) by using Eqs. (7)-(8).
3. Perform listening tests of both samples \( F_{0a} (t) \) and \( F_{0b} (t) \).

By using the algorithm described in Sec. 3.2, intonation variation from the norms to the variants is estimated, focusing particularly on their sparser specifications confined to the basic elements of the underlying lexical tone structures, as sketched in Fig. 1. Here, either \( F_{0u} (t) \) or \( F_{0v} (t) \) is used as the norms of the underlying lexical tone structures. Figure 3 plots the results for application of variable norms to the estimation of intonation variation in \( \zeta_{0u} (t) \) between these statement and question samples. Triangles indicate intonation variation with statements as norms, and circles with questions as norms; the temporal structures are adopted from the corresponding norms.

At a glance of the experimental results, the questions have a higher level and a rising end compared to the statements. The differences in the end between Figs. 3(a) and 3(b) may demonstrate the existence of such tone type dependent tunes in Mandarin, too, as (Tune 4) described in the Thai language [9].

Figure 4 shows the synthetic \( F_{0} \) contours (triangle and circle sequences) by superimposing the approximations of intonation variation (the thin lines through the triangles or circles shown in Fig. 3) on corresponding norms (“+” sequences) using Eq. (13). By a STRAIGHT-based analysis-by-synthesis technique [8], synthetic contours \( F_{0a} (t) \) are used to replace natural values \( F_{0u} (t) \) and \( F_{0v} (t) \) to replace \( F_{0u} (t) \) for re-synthesizing speech samples, such that estimated intonation variation is reproduced on other speech samples (here, the norms).

An informal listening test was conducted on both the re-synthesized and original samples to evaluate their prosodic similarity. Basically, perceptual results show that the intonational characteristics of the re-synthesized samples are indistinguishable from the original except that questions with synthetic contours \( F_{0u} (t) \) have a relatively slow speech rate. Consequently, they sound less natural than questions with observed \( F_{0} \) contours \( F_{0a} (t) \).

4.2. Experiment 2: solving acoustic correlations

Solving acoustic correlations of tone and intonation is demonstrated by a case analysis of four conventional greetings:

(1) /ni3hao3./ (Hello.)
(2) /zen3me0yang4a0?/ (How are you doing?)
(3) /ni3 mang2ma0?/ (Are you busy?)
(4) /ni3 shen1ni3 hao3ma0?/ (How are you feeling?)

The four speech samples were produced by narrator WL.

The experimental procedure is roughly described as follows, using /ni3hao3./ as an example. Note that the base values of the four lexical tones by the speaker were expressed in (12). Hereafter, T0 to T4 indicate Tones 0 to 4, respectively.
(1) Determine sparser specifications \((t_{ij}, f_{ij})\), for example, 
\[(0.07, 243), (0.19, 367), (0.36, 197), (0.45, 152), (0.64, 204)\] 
for /ni3hao3/; \(t_{ij}\) is the \(i\)th target timing and \(f_{ij}\) the \(F_0\)'s.

(2) Prepare base values for T2T3 where T2 is a sandhi tone: V12 (250), P22(355), P31(235), V31(200), and P32(280) in (12).

(3) Compute \(\hat{\gamma}_i\) using the method described in Sec. 3.2 where 
- the norms take the base values of the four lexical tones;
- if a syllabic tone is Tone 0, its base values take the base values of the last target of the preceding tone;
- variants take \(F_0\) values from the sparser specifications.

Figures 5(a)-(d) show the observed \(F_0\) contours ("+" sequences) where the short vertical lines across the \(F_0\) contours indicate sparser specifications. Figures 5(e)-(h) show the separated intonation patterns by removing the base values of the underlying lexical tone structures from the observed values; underlying lexical tone structures, such as T2T3, are superimposed on this figure. The intonation patterns involved in these examples are probably identical because they have an identical pattern: "rising [level] falling level."

The regular characteristics among the four samples are the final "falling level." "Falling" indicates the nucleus of sentence intonation, while "level" indicates the speaker is producing the complete forms of the last lexical tones in each greeting, probably showing politeness. This intention is partially observable from Tone 0 in /ma0/ in Fig. 5(c); Tone 0 can transfer part of the preceding tone's patterns. Tone 0 continues the tendency of \(F_0\) movements of the preceding Tone 2 to a target (corresponding to P22 in Fig. 1) and then approximately maintains it until near the end, rather than continuing the rise as frequently observed in yes-no questions, or fall as usually observed in normal statements. The separated intonation patterns show considerable similarity to the high-low pitch accents used in descriptions of Japanese intonation and the so-called "hat pattern" in the intonation of some European languages as well as English.

5. Conclusions

This paper presented a proposed method for quantitative estimation of intonation variation based on a tone transformation technique, undertaken in the spirit of analysis-by-synthesis. Intonation variation was measured as a series of local tone variation in \(F_0\) from selected norms to variants. When the norms took observed target \(F_0\) values of lexical tone structures, this method could measure prosodic differences between utterances having identical lexical tone structures. This capability was tested by conversion between statements and questions undertaken by perceptual tests. When the norms took the base values of the four lexical tones measured from typical isolated words, this method could separate actual intonation from the observed \(F_0\) contours by excluding lexical tone effects. Experimental results showed the regularity of the separated intonation patterns.

6. Acknowledgements

This research was supported in part by the National Institute of Information and Communications Technology.

7. References