Prosody Conversion for Emotional Mandarin Speech Synthesis Using the Tone Nucleus Model

Miaomiao WEN[1], *Miaomiao WANG*[1], *Keikichi HIROSE*[2], *Nobuaki MINEMATSU*[2]

[1]Department of Electrical Engineering and Information Systems, the University of Tokyo, Japan
[2]Department of Information and Communication Engineering, the University of Tokyo, Japan
{wenm,wangm,hirose,mine}@gavo.t.u-tokyo.ac.jp

Abstract

In this paper, tone nucleus model is employed to represent and convert F_0 contour for synthesizing an emotional Mandarin speech from a neutral speech. Compared with previous prosody transforming methods, the proposed method 1) only converts the tone nucleus part of each syllable rather than the whole F_0 contour to avoid the data sparseness problems; 2) builds mapping functions for well-chosen tone nucleus model parameters to better capture Mandarin tonal information. Using only a modest amount of training data, the perceptual accuracy achieved by our method was shown to be comparable to that obtained by a professional speaker.

Index Terms: emotion conversion, expressive speech synthesis, prosody modeling, Mandarin

1. Introduction

With the intelligibility of synthetic speech approaching that of human speech, the need for increased naturalness and expressiveness becomes more palpable. However, there has been a lack of emotional affect in the synthetic speech of the state-of-art TTS systems. This is largely due to the fact that the prosodic modules in these systems are unable to predict prosody from text accurately for emotional speech. Differences among emotional speeches are mainly represented by a number of acoustic features such as fundamental frequency (F_0), phonetic duration and voice quality. Previous methods of expressive speech synthesis consist of formant synthesis, diphone concatenation, unit selection or HMM-based methods [1] [2]. The quality of the data-driven methods all heavily relies on the size of the emotional speech corpus, which takes great effort to build. Another expressive speech synthesis approach is to obtain prosodic variations between neutral speech and emotional speech, and then make the synthesized emotional speech acquire these prosodic variations. As prosody prediction model for neutral speech has been extensively studied and implemented as robust prosodic modules in current state-of-the-art TTS systems, it would be beneficial to build the prosody prediction model for emotional speech upon these existing systems, such as prosody conversion systems. [3] used GMM and CART-based F_0 conversion methods for mapping neutral prosody to emotional Mandarin prosody. [4] adopted a difference approach to predict the prosody of emotional speech, where the prosody variation parameters (F_0, duration and intensity) are predicted for each phoneme. [5] proposed a segment selection emotion conversion approach that utilized a concatenative framework to directly search for F_0 segments in the training corpus. [6] investigates voice conversion and modification techniques to reduce database collection and processing efforts while maintaining acceptable quality and naturalness.

Mandarin is the standard Chinese. It is a typical tonal language and each syllable with the same phoneme sequence has up to four tone types, each indicating different meaning. The process of a Mandarin speech synthesis system could be generally similar to that of an English system, but may also be implemented in different ways to account for its special features. Tones constitute a special set of additional features which should be reproduced by a speech synthesis system. F_0 contours of utterances should include these local tonal features in addition to the sentential intonation corresponding to higher-level structures. Tone nucleus model [7] suggests that a syllable F_0 contour may consist of three segments: onset course, tone nucleus and offset course. Among the three segments, only the tone nucleus is obligatory. The other two are optional and non-deliberately produced articulatory transition F_0 loci. Tone nucleus model has been proved useful in both Mandarin tone recognition [7] and Mandarin speech synthesis [8]. These findings led us to propose a prosody conversion method for emotional Mandarin speech based on the tone nucleus model. The basic assumption is that we only model the prosodic features of the tone nucleus part of each syllable, instead of directly convert the whole syllable F_0 contour, which will contain two much redundancy and cause the data sparseness problem. In this paper, a data-driven, tone nucleus model-based prosody conversion method is utilized to predict the prosodic variations between neutral and emotional Mandarin speech.

The remainder of this paper is organized as follows. The second section describes tone nucleus model and adapts the original model to our prosody conversion purpose. In the third section CART models are employed to convert tone nucleus model parameters. In Section 4, experiments and results are described and discussed. Finally, the last section gives the conclusion.

2. Method

2.1. Tone nucleus model

In Mandarin, there are up to four lexical tones attached to each syllable. They are referred to as T1, T2, T3 and T4, which are characterized by high-level, high-rising, low dipping, and high-falling F_0 contours, respectively. For a syllable F_0 contour, only its later portion, approximately corresponding to the final vocalic part, is regarded to bear tonal information, whereas the early portion is regarded as physiological transition period from the previous tone. It was also found that there are often cases where voicing period in the ending portion of a syllable also forms a transition period of vocal vibration and contributes little to the tonality [7]. From these considerations, a tone nu-

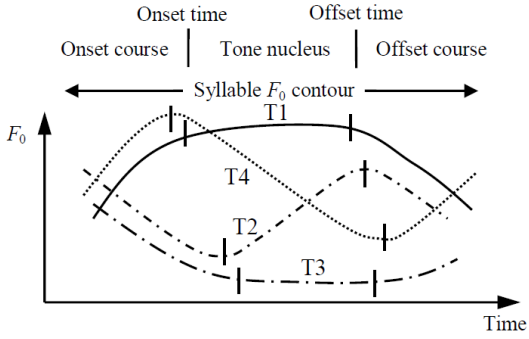

Figure 1: *Tone nuclei for the four lexical tones.*

cleus model, which divides a syllable F_0 contour into three segments according to their roles in the tone generation process, was proposed [7]. Figure 1 illustrates some typically observed tonal F_0 variations in continuous speech and their tone nuclei notations. The three segments, onset course, tone nucleus, and offset course, are defined as follows:

- Onset course is an F_0 transition from the preceding syllable to the onset target of the tone nucleus. This segment covers the initial consonant and the transition period of the final vocalic part.
- Tone nucleus is a portion where the F_0 contour keeps the basic pattern of the tone unless it is affected by high-level prosodic factors such as neutralization, contextual effect, focus, phrasing, and etc. This segment covers the nucleus of the final vocalic part.
- Offset course is an F_0 transition from the offset target of the tone nucleus to the following syllable. This segment holds the ending course of the final vocalic part.

2.2. Define tone nucleus model for prosody conversion

One specialty of Mandarin emotional speech synthesis is how the tone F_0 contours are preserved while expressing different emotions. For example, when a sentence is uttered sadly, the pitch contour of syllables can become very flat. The tone nucleus F_0 range of the T4 syllable ("da4" in Figure 2(d)) and the T2 syllable ("zhi2" in Figure 2(d)) is less than their neutral counterparts (Figure 2(a)). T1 syllable ("gong1" in Figure 2(d)) has a much lower average F_0. Almost opposite situation happens when the emotion state is happy (Figure 2(c)). But the tone patterns are maintained so that native Mandarin speaker could identify the tones of each syllable in the emotional utterance quite easily. Thus, different parameters, such as F_0 range and average F_0, should be predicted to characterize the conversion of four different tones.

We also observe that the shape or pattern of the tone nucleus part of the syllable (the F_0 segment between the red dots in Figure 2) remains rather stable while the emotional status changes. But the features (parameters) of the tone nucleus vary with the emotional status. For example, all T4 syllables have a similar falling tone nucleus F_0 contour, but it will have bigger average F_0 and F_0 range in the angry or happy utterance, and a smaller average F_0 and F_0 range in a sad utterance. Moreover, it is argued in [9] that pitch contour statistics such as mean, maximum, minimum, and range are more emotionally prominent than features describing the pitch shape. So instead of predicting the exact F_0 shape parameters like in [3], we use a few F_0 contour

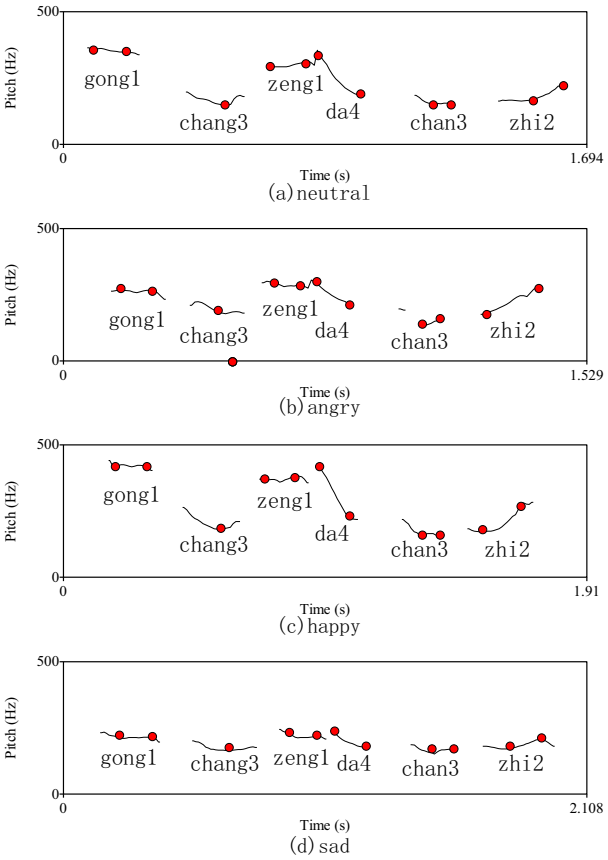

Figure 2: *An example sentence read with different emotions. The segment between the dots indicate the tone nucleus part of the syllable. This sentence means "The factory increases production".*

templates to represent the tone nucleus shape. These considerations lead us to an idea of generating F_0 contours only for tone nuclei, and to concatenate them to produce the whole sentence F_0 contour. Based on our observation, the tone nucleus are defined for each tone type as shown in Table 1. We use several tone nucleus templates to represent different T2 and T4 nucleus contour shapes. For T0, T1 and T3, tone nuclei are defined as a flat F_0, which is represented by a single parameter, i.e. average F_0 value.

Tone type	F_0 contour feature of tone nucleus	Parameters (to be predicted)
T1	Flat F_0 with high level	Average F_0
T2	Rising F_0	Average F_0, F_0 range, Template identity
T3	Flat F_0 with low level	Average F_0
T4	Falling F_0	Average F_0, F_0 range, Template identity
T0	No specific feature	Average F_0

Table 1: Definition of tone nucleus and output parameters of the predictor for each tone type. The tone nucleus onset/offset time is predicted for T1, T2, T3 and T4.

2.3. Tone nucleus parameters extraction

To apply the tone nucleus model for prosody conversion, it is necessary to automatically estimate tone nucleus parameters

(Table 1) from F_0 contour. For each syllable F_0, we use a robust tone nucleus segmentation and location method based on statistical means. The method has two steps: the first step is F_0 contour segmentation based on the iterative segmental K-means segmentation procedure, with which a T-Test based decision of segment amalgamation is combined [7]. When a segmentation becomes available, which segment is tone nucleus is decided according to the following rules in the second step: (1) For T1, the segment with the biggest average F_0. (2) For T2, the segment with the largest average ΔF_0. (3) For T3, the segment with the lowest average F_0. (4) For T4, the segment with the lowest average ΔF_0. Considering that the syllable's maximum and minimum F_0 points carry important information in expressing emotions, if the chosen segment fails to cover the maximum or the minimum F_0 point, it will be expanded to include these two critical points. Since T0 shows no inherent F_0 contour, a stable definition of tone nucleus for T0 is difficult. We assume the entire voiced segment of the syllable as the tone nucleus for T0.

After extraction, average F_0 is estimated for each T0, T1, T2, T3 and T4 tone nucleus. F_0 range is estimated for each T2 and T4 tone nucleus. To obtain the tone nucleus template identity for T2 and T4, we normalize the extracted tone nucleus F_0 contours in time and frequency. Let the nucleus part of the syllable represented by $O = (o_1, o_2, \ldots, o_{11})$, the vector o_i is a two component vector $(logF0_j, \Delta logF0_j)$. Then for T2 and T4, all Os are clustered into a few (less than 10) groups using XMeans clustering method [10]. For each group, an F_0 template is calculated as the average of the samples in the group. Figure 3(a) shows the templates for T4 nuclei in angry utterances. For comparison, we averagely divide each syllable into three segments, then normalize the center segment. The clustered templates for these T4 center segments in angry utterances are shown in Figure 3(b). It is clear that F_0 templates of the center segment are scattered and thus hard to predict. The extracted tone nucleus templates could better capture the tone F_0 shape (e.g. a falling shape for T4) and are easier to predict. It should be noticed that 12% of the extracted T4 nucleus have a rising shape, this may due to several possible reasons. First, when expressing angry, the speaker sometimes adopts a rhetorical mood. Then the ending part of the utterance will have a rising F_0. Also, these rising T4 nuclei might be caused by tone co-articulation [11] and tone nucleus extraction error.

3. F_0 Conversion

F_0 conversion is to convert a pitch contour (neutral) into a new pitch contour (emotional) using a mapping function. The mapping function is automatically learned from the parallel speech corpus. In this paper, instead of directly mapping surface F_0 contour, tone nucleus model parameters estimated from the F_0 contours are employed to build the mapping rules.

The differences between the source and target tone nucleus parameters are modeled by classification and regression trees(CART). The input parameters of the CART contain the following:

- Tone identity (including current, previous and following tones, with five categories)
- Initial identity (including current and following syllables' initial types, with five categories [8])
- Final identity (including current and previous syllables' final types, with two categories [8])

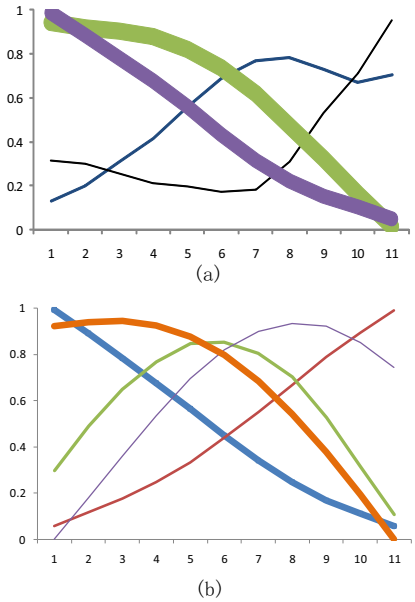

Figure 3: *(a) Syllable tone nucleus F_0 templates for angry T4. (b) Syllable center part F_0 templates for angry T4. The vertical and horizontal axes are normalized frequency and time respectively. The width of the line represents the percentage of this cluster out of all the angry T4 syllables.*

- Position of the current word in the current word foot/prosodic word/sentence
- Part of speech (including the current, previous and following words, with 30 categories)

4. Experiments and Results

Our emotional corpus contains 300 sentences with no obvious textual bias towards any of the expressive styles. A professional actor read each sentence in four basic emotional states: neutral, anger, joy and sadness. And then each sentence was automatically segmented at the syllable level by a forced alignment procedure. 270 sentences, including about 1700 syllables, are used to train transforming functions and the rest are employed to test our conversion method.

Our experiments contains the training, the converting and the evaluation procedures. In the training procedure, source and target pitch contours from the parallel corpus are firstly aligned according to syllable boundaries. Then tone nucleus model parameters (Table 1) are extracted from each syllable's pitch contour and mapping functions of the parameters are obtained. As for duration conversion, we use relative prediction which predict a scaling factor to be applied to the neutral phone duration. The same feature set is used to train a relative regression tree. In the converting procedure, source pitch target parameters estimated from source pitch contours are transformed by these mapping functions. Duration of phones are also converted. Then, the converted tone nucleus parameters are used to generate the target emotional contours. An example conversion result is shown in Figure 4, in which source pitch contour (Figure 4(a)) in neutral state is converted into the angry contour (Figure 4(b)), the happy contour (Figure 4(c)) and the sad contour (Figure 4(d)). Finally, we use the TD-PSOLA synthesis method for modifying pitch and timing of the original neutral utterance.

In the evaluation procedure, five native Mandarin speakers

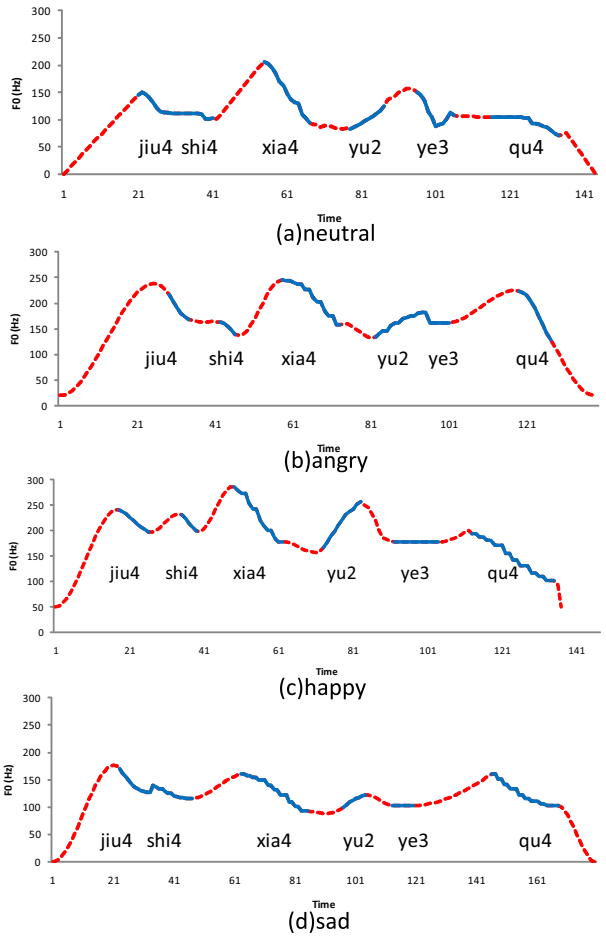

Figure 4: *An example of prosody conversion from (a)the original neutral utterance to (b)angry, (c)happy, (d)sad utterances. The tone nucleus part of the syllable appears in solid line. The dotted line is the result of interpolation. This sentence means "I will go there even if it rains".*

(three male, two female) participate in the listening test. 30 synthesized sentences, each with three emotion states are presented to the listeners randomly. The listener is asked to judge whether the utterance conveys angry, happy, sad or "cannot decide". The confusion matrix is shown in Table 3. The conversion from neutral speech to sad and angry conversions are respectively best and worst. The conversion outputs convey happiness and sadness that is comparable to or even better than the original speech (Table 2), while the recognition rate for anger (67%) was lower than that of the naturally spoken anger (86%). The results suggest that sadness and happiness are recognized from the prosody. The angry emotion style is not recognized from prosody alone, energy also serves as an important cue, which is not modeled in our experiment. This is consistent with [5] and [6].

5. Conclusions

This paper proposes to employ tone nucleus model to implement F_0 conversion for expressive Mandarin speech synthesis. Advantages of the proposed method are that parametric F_0 models such as tone nucleus model can provide an underlying linguistically or physiological description for surface F_0 contour and it can furnish several compact parameters to represent a long pitch contour. CART mapping method is employed to generate transforming functions of tone nucleus model parameters. The subjective listening test shows that synthesized speech using predicted prosody parameters is able to present specific emotion. In our future work, we would like to further consider the concatenation cost between two candidate adjacent tone nuclei in our framework.

		Identified			
		angry	happy	sad	cannot decide
Intended	angry	0.86	0.12	0	0.02
	happy	0.19	0.77	0	0.14
	sad	0	0	0.96	0.04

Table 2: Confusion matrix for the emotion classification task of original emotional utterances.

		Identified			
		angry	happy	sad	cannot decide
Intended	angry	0.67	0.25	0.04	0.04
	happy	0	0.75	0	0.25
	sad	0	0	1	0

Table 3: Results of expressive conversion evaluation.

6. Acknowledgements

The authors of this paper would like to thank Prof. Jianhua Tao in Chinese Academy of Sciences for offering us the speech corpus and his kind advice.

7. References

[1] M. Schroder, "Emotional speech synthesis: A review", Proc. Eurospeech 2001, pp.561-564, 2001.

[2] J. Yamagishi, K. Onishi, T. Masuko and T. Kobayashi, "Modeling of various speaking styles and emotions for HMM-based speech synthesis", Proc. Eurospeech 2003, pp.2461-2464, 2003.

[3] J. Tao, Y. Kang, and A. Li, "Prosody conversion from neutral speech to emotional speech", IEEE Trans. Audio, Speech and Language Processing, vol.14: 1145-1153, 2006.

[4] H. Tang, X. Zhou, M. Odisio, M. Hasegawa-Johnson and T. Huang, "Two-Stage prosody prediction for emotional text-to-speech synthesis", Proc. Interspeech 2008, pp.2138-2141, 2008.

[5] Z. Inanoglu and S. Young, "Data-driven emotion conversion in spoken English", Speech Communication, 51, pp.268-283, 2009.

[6] O. Turk and M. Schroder, "Evaluation of expressive speech synthesis with voice conversion and copy resynthesis techniques", IEEE Trans. Audio, Speech and Language Processing, 18, pp.965-973, 2010.

[7] J. Zhang and K. Hirose, "Tone nucleus modeling for Chinese lexical tone recognition", Speech Communication, 42, pp.447-466, 2004.

[8] Q. Sun, K. Hirose, W. Gu and N. Minematsu, "Generation of fundamental frequency contours for Mandarin speech synthesis based on tone nucleus model", Proc. Eurospeech 2005, pp.3265-3268, 2005.

[9] C. Busso, S. Lee, and S. Narayanan. "Analysis of emotionally salient aspects of fundamental frequency for emotion detection", IEEE Transactions on Audio, Speech and Language Processing, 17(4):582 – 596, May 2009.

[10] R.O. Duda, P.E. Hart and D.G. Stork, Pattern Classification (2nd ed.), John Wiley and Sons, 2001.

[11] Y. Xu, "Contextual Tonal Variation in Mandarin Chinese", Ph.D. dissertation, The University of Connecticut, 1993.