Emotional Voice Conversion for Mandarin using Tone Nucleus Model – Small Corpus and High Efficiency

Miaomiao Wang[1], *Miaomiao Wen*[2], *Keikichi Hirose*[3], *Nobuaki Minematsu*[3]

[1] Toshiba (China) R&D Center
[2] Language Technologies Institute, Carnegie Mellon University, USA
[3] Graduate School of Information Science and Technology, University of Tokyo, Japan

wangmiaomiao@rdc.toshiba.com.cn, mwen@cs.cmu.edu, {hirose, mine}@gavo.t.u-tokyo.ac.jp

Abstract

The GMM-based spectral conversion techniques were applied to emotion conversion but it was found that spectral transformation alone is not sufficient for conveying the required target emotion. In this paper, we adopt the tone nucleus model to carry the most important information of tones and represent F_0 contour for Mandarin speech. And then tone nucleus part is converted to emotional speech from neutral ones. The tone nuclei variations are modeled by the classification and regression tree (CART) and dynamic programming. Compared with previous prosody transforming methods, the proposed method 1) uses only the tone nucleus part of each syllable rather than the whole F_0 contour to avoid the data sparseness problems in emotion conversion; 2) builds mapping functions for well-chosen tone nucleus model parameters to better capture Mandarin tonal and emotional 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: Emotional voice conversion, Mandarin, Tone nucleus

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 Text-To-Speech (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. 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. In [3] Gaussian Mixture Model (GMM) and CART-based F_0 conversion methods are used for mapping neutral prosody to emotional Mandarin prosody. In [4] a difference approach is adopted to predict the prosody of emotional speech, where the prosody variation parameters are predicted for each phoneme. The GMM-based spectral conversion techniques were applied to emotion conversion but it was found that spectral transformation alone is not sufficient for conveying the required target emotion. All the researches depend on correct understanding and modeling of prosodic features including F_0 contours, duration and intensity.

F_0 modeling is important for emotional voice conversion for any languages, but it is critical to Mandarin. Mandarin, the standard Chinese, is a well-known tonal language which means the word meaning depends crucially on shape and register distinctions among four highly stylized syllable F_0 contour types. But on the same time, Mandarin also allows some range of F_0 variations to express emotion, mood and attention. Thus F_0 modeling will be more complex than non-tonal languages, such as English and Japanese. In the case of Mandarin, F_0 variations show larger undulations than those in the non-tonal languages. The lexical tones show consistent tonal F_0 patterns when uttered in isolation, but show complex variations in continuous speech [5, 6]. The invariance problem is the difficulty of giving a physical phonetic definition of a given linguistic category that is constant and always free of context [7].

Tone nucleus model suggest that a syllable F_0 contours can be divided into three segments: onset course, tone nucleus and offset course. The tone nucleus of a syllable is assumed to be the target F_0 of the associated lexical tone. The other two are optional and non-deliberately produced articulatory transition F_0 loci. Tone nucleus usually conforms more likely to the standard tone pattern than the articulatory transitions. Tone nucleus model has improved the tone recognition rate in [8] to show that tone nucleus keeps the important discriminant information between tonal F_0 patterns and underlying tone type and successfully improved the tone recognition rate. Those findings lead us to the idea that we can apply the tone nucleus model to improve the naturalness and expressiveness for speech synthesis system, which has been successfully applied for Mandarin.

So in this paper, firstly we will introduce Mandarin tones and tone nucleus model. Then instead of directly convert the whole syllable F_0 contour, which will contain two much redundancy and cause the data sparseness problems; a data-driven, tone nucleus model-based prosody conversion method is utilized to predict the prosodic variations between neutral and emotional Mandarin speech. A GMM of the joint probability density of source and target features is employed for performing spectral conversion between emotions. Finally we will discuss experiments and results.

2. F_0 conversion using tone nucleus model

2.1. Tone nucleus

There are four basic lexical tones (referred to as T1, 2, 3, 4) and a neutral tone for each Mandarin syllable. The four basic lexical tones are characterized by their perceptually distinctive

pitch patterns which are conventionally called by linguists as: high-level (T1), high-rising (T2), low-dipping (T3) and high-falling (T4) tones [9]. The neutral tone, according to [9] does not have any specific pitch pattern, and is highly dependent on the preceding tone and usually perceived to be temporally short and zero F_0 range.

For a syllable F_0 contour, as pointed out in [8], lexical tone is not evenly distributed in a syllable because of F_0 variations in a syllable F_0 contour in various phonetic contexts. 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 nothing to the tonality. From these considerations, we can classify a syllable F_0 contour into underlying target and articulatory transitions:

1) Underlying target represents the target F_0 and serves as the main acoustic cue for tone perception.

2) Articulatory transitions are the F_0 variations occurring as the transitions to or from the target F_0.

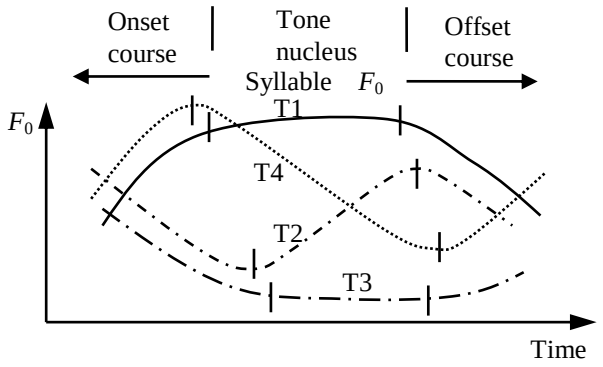


Figure 1: *Tonal F_0 contours with possible articulatory transitions and their tone nucleus*

Figure 1 illustrates some typically observed tonal F_0 variations in continuous speech and their tone nuclei notations. The three segments are called onset course, tone nucleus, and offset course, respectively, which are defined in [8]. The tone nucleus part will conform more likely to the standard tone pattern. Tone-onset and tone-offset indicate the F_0 values, which takes either low (L) or high (H) value, at the tone onset and offset, respectively. These F_0 values serve as distinctive features characterizing the four basic lexical tones.

2.2. Automatic tone nucleus extraction

To apply the tone nucleus model for speech synthesis and emotion conversion, it is necessary to automatically estimate tone nucleus parameters 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 in [8]. When segmentation becomes available, "which segment is tone nucleus" is decided according to the rules as discussed in [10]. Figure 2 shows an example of extracted tone nucleus parts of syllables of an example sentence in different emotions. As compared in Figure 2, we see that the tone nucleus parts extracted by our method remain more stable than the original syllable F_0.

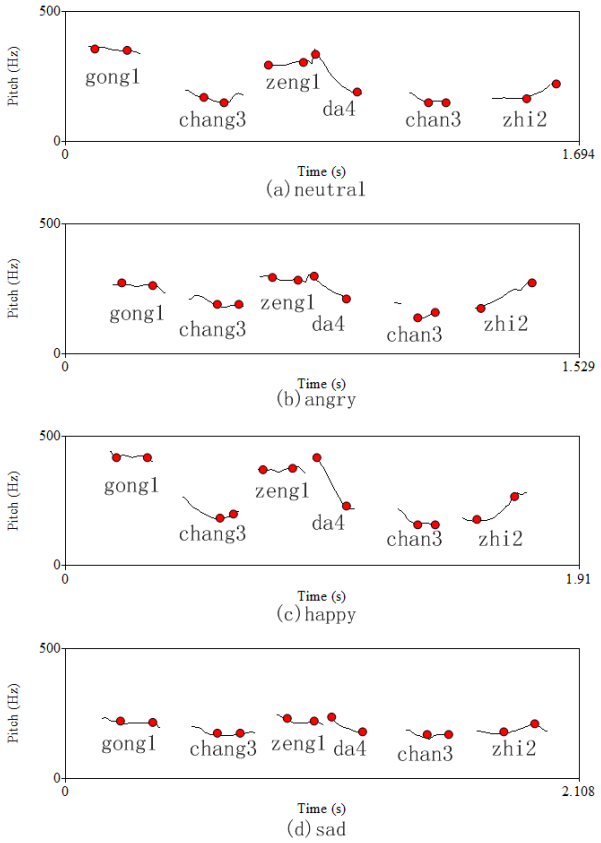


Figure 2: An *example sentence in different emotions: the segment between the dots indicates the tone nucleus part of the syllable.*

2.3. F_0 conversion

F_0 conversion is to convert a neutral F_0 contour into an emotional F_0 contour 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 neutral and emotional 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, each with five categories

Initial identity, including current and following syllables' initial types, each with five categories

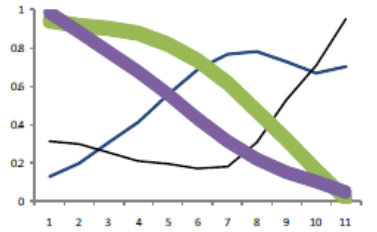
Final identity, including current and previous syllables' final types, each with two categories

Position of the current word in the current word foot/prosodic word/sentence

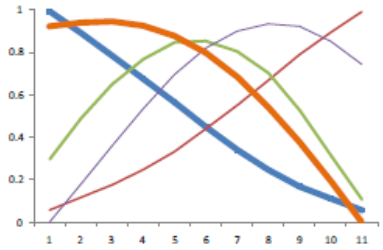
Part of speech (POS), including the current, previous and following words, each with 30 categories

Figure 3(a) shows the templates for T4 nuclei in angry utterances. The width of the line represents the percentage of this cluster out of all the angry T4 syllables. For comparison, we averagely divide each syllable into three segments, and

then normalize the center segment. The clustered templates for these T4 center segments in angry utterances are shown in Figure 3(b). The vertical and horizontal axes are normalized frequency and time respectively. 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 rising shape for T2) and are easier to predict. It should be noticed that 12% of the extracted angry 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 [6], e.g. T4 turned into neutral tone because it is unstressed; and tone nucleus extraction error.



(a)



(b)

Figure 3: *Syllable F_0 templates for (a) angry T4 by tone nucleus. (b) angry T4 by syllable length F_0 contours.*

3. Spectrum Conversion based on GMM

Spectrum conversion can be thought of as just another form of voice conversion. The neutral speech could be regard as the source speaker, while the target emotion speech could be regard as the target speaker. In practice, voice conversion techniques have focused on the transformation of the vocal tract spectra, as it has been pointed out that strong feelings often literally distort the physical vocal tract. For example, "anger" often involves a physical tension which can be felt throughout the body and certainly has an effect on the tenseness of the speech organs, which in turn creates a distinct acoustic effect. Similarly, "happiness" might involve a less total physical change, often just a smile which is "talked through" [3]. In [11] and [12], GMM-based spectral conversion techniques were applied to emotion conversion but it was found that spectral transformation alone is not sufficient for conveying the required target emotion. In [3], an integrated method that based on GMM and codebook mapping is used.

The GMM based voice conversion is first introduced by [14]. A GMM of the joint probability density of source and target features is employed for performing spectral conversion between speakers. Stylianou's method is to converts spectral parameters frame by frame based on the minimum mean square error. Although this method is reasonably effective, the deterioration of speech quality is caused by some problems: 1) appropriate spectral movements are not always caused by the frame-based conversion process, and 2) the converted spectra are excessively smoothed by statistical modeling. To address these problems, we use the Toda'01 method [13], which is based on the maximum likelihood estimation of a spectral parameter trajectory. Not only static but also dynamic feature statistics are used for realizing the appropriate converted spectrum sequence.

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 experiment uses 40 order cepstrum feature, number of Gaussian mixture is 64. The STRAIGHT analysis and synthesis method [15] were employed for spectral extraction and speech synthesis, respectively.

In the training procedure, neutral and other three emotional F_0 contours from the parallel corpus are firstly aligned according to syllable boundaries. Then tone nucleus model parameters are extracted from each syllable's F_0 contour and mapping functions of the parameters are obtained. As for duration conversion, we use relative prediction which predicts a scaling factor to be applied to the neutral phone duration. The same feature set is used to train a relative regression tree. After that, the converted tone nucleus parameters are used to generate the emotional F_0 contours.

As for comparison, in the listening test, the one converted using original syllable F_0 will have tone errors as some of the syllable will sound as other tones. Thus it is greatly change the sentence meanings. Two native speakers checked the converted sentences using original syllable F_0 contours and found that the syllable tone error rate as shown in Table 1, while our method using tone nucleus model doesn't have such kind of errors.

Table 1. *Tone error rate for emotion conversion using original syllable F_0 contour*

	Angry	Happy	Sad
Tone error rate	9.71%	4.37%	5.34%

Figure 4, 5, 6 respectively shows the perceptual results of the synthesized emotional speech utterances with the neutral utterances. The four groups of utterances are listed below.

(1) **Natural speech (NS+NP)** the original recorded utterance

(2)**Converted emotional spectrum with linearly converted prosody (CS+LCP)** Spectrum is converted by GMM-based method. F_0 linearly converted from neutral F_0 contour using the following equation, where $p_t(x)$and $p_t(y)$ are input and converted F_0 values, respectively. $\mu(\cdot)$ and $\sigma(\cdot)$ are the mean and the standard deviation of F_0, respectively.

$$p_t(y) = \frac{p_t(x) - \mu(x)}{\sigma(x)} \times \sigma(y) + \mu(y) \qquad (1)$$

(3)**Natural spectrum with converted emotional prosody (NS+CP)** Converted emotional prosody using Tone Nucleus model is given to original recorded speech.

(4) **Converted spectrum with converted prosody (CS+CP)** Spectrum is converted to that of target emotion from neutral speech and converted emotional prosody using Tone Nucleus model is given to it.

As the result of subjective experiment clearly shows, prosodic features mainly dominate emotional expression. The prosody converted using Tone Nucleus model performs better than the ones using linearly converted from neutral F_0 contour. The happy and sad emotions indicate that spectrum conversion will lower the perception rate. This may due to a lot of u/v errors and unnaturalness caused by spectrum conversion.

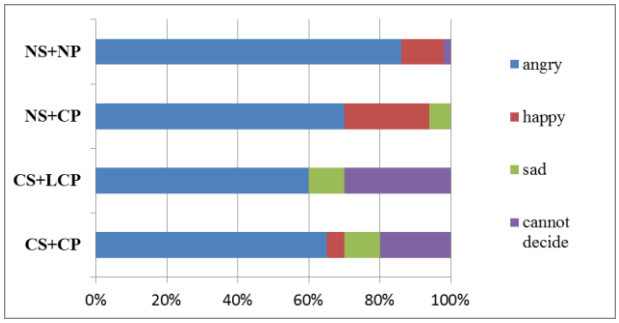


Figure 4: *Subjective evaluation of angry speech.*

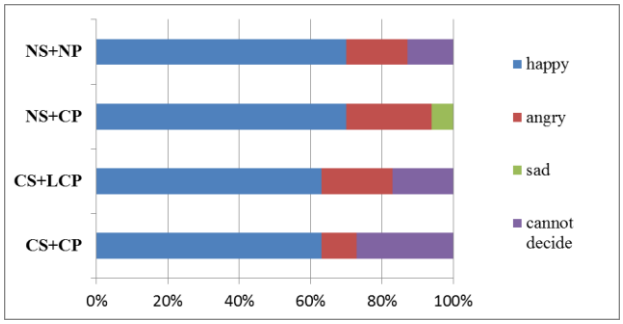


Figure 5: *Subjective evaluation of happy speech.*

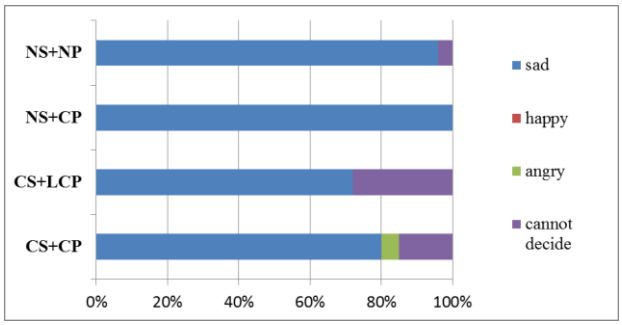


Figure 6: *Subjective evaluation of sad speech.*

5. Conclusions

This paper explains how 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. GMM-based spectral conversion techniques were also adapted to spectrum conversion. The subjective listening test shows that synthesized speech using predicted prosody parameters is able to present specific emotion.

6. Acknowledgements

The authors of this paper would like to thank Prof. Jianhua Tao in Chinese Academy of Sciences for offering us the emotional speech corpus and his kindly advice; and Prof. Jinsong Zhang in BLCU for his useful suggestions.

7. References

[1] Schröder, M., “Emotional Speech Synthesis: A Review”, In Proc. Eurospeech, pp. 561-564, 2001

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

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

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

[5] S.-H. Chen and Y.-R. Wang, "Tone recognition of continuous Mandarin speech based on Neural Networks", IEEE Trans. on SAP, Vol. 3, No. 2, 1995, pp.146-150.

[6] Xu, Y., Contextual tonal variations in Mandarin. J. Phonetics 25, 61–83, 1997.

[7] B. Lindblom, "Explaining phonetic variation: a sketch of the H&H theory", W.Hardcastle and A. Marchal (ed.), Speech Production and Speech Modelling. Kluwer Academic Publishers, 1990, pp.403-439.

[8] Zhang, J. and Hirose, K., "Tone nucleus modeling for Chinese lexical tone recognition," Speech Communication, Vol. 42, Nos. 3-4, pp. 447-466, 2004.

[9] Chao, Y.-R., 1968. A Grammar of Spoken Chinese. University of California Press, Berkeley.

[10] M. Wen, M. Wang, K. Hirose, N. Minematsu, "Prosody conversion for emotional Mandarin speech synthesis using the tone nucleus model," Proc. INTERSPEECH, pp.2797-2800, 2011

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

[12] H. Kawanami, Y. Iwami, T. Toda, H. Saruwatari, and K. Shikamo. 1999. GMM-based Voice Conversion Applied to Emotional Speech Synthesis, IEEE Trans. Speech and Audio Proc., 7(6):697-708.

[13] T. Toda, H. Saruwatari, and K. Shikano. 2001. Voice Conversion Algorithm based on Gaussian Mixture Model with Dynamic Frequency Warping of STRAIGHT spec-trum. In Proc. ICASSP, pp. 841–844, Salt Lake City, USA.

[14] Y. Stylianou. 1998. Continuous probabilistic transform for voice conversion, IEEE TSAP, no. 6, pp. 131-142.

[15] H. Kawahara, I. M. Katsuse, and A. D. Cheveigne, “Restructuring speech representations using a pitch adaptive time frequency smoothing and an instantaneous frequency-based F0 extraction: possible role of a repetitive structure in sounds”, Speech Communication, vol. 27, no. 3–4, pp. 187–207, 1999.