APPLYING PITCH TARGET MODEL TO CONVERT F0 CONTOUR FOR EXPRESSIVE MANDARIN SPEECH SYNTHESIS

[1]*Yongguo Kang,* [1]*Jianhua Tao,* [2]*Bo Xu*

[1]National Laboratory of Pattern Recognition, [2]High Technology Innovation Center
Institute of Automation, Chinese Academy of Sciences
{ygkang,jhtao}@nlpr.ia.ac.cn, xubo@hitic.ia.ac.cn

ABSTRACT

In the paper, pitch target model is employed to represent and convert F0 contour for synthesizing an emotional Mandarin speech from a neutral speech. Compared with conventional F0 transforming methods, the proposed method converts F0 patterns described by pitch target parameters rather than F0 contours themselves, and uses Gaussian Mixture Model(GMM) and Classification and Regression Trees (CART) methods to build mapping functions for well-chosen pitch target parameters. Other prosodic parameters such as duration and intensity are also converted. Listening tests prove that these converted speeches express corresponding emotional states.

1. INTRODUCTION

Expressive speech synthesis (ESS) has become a hot topic of speech research field because it can offer a more real circumstance in human-machine interaction. Differences among emotional speeches mainly manifest on a number of acoustic features such as fundamental frequency(F0), phonetic duration, voice quality, and pitch accent. Traditional methods[1] on expressive speech synthesis consist of formant synthesis, diphone concatenation and unit selection as well as speech synthesis. An approach to ESS is to obtain prosodic variations between neutral speech and emotional speech, and then make the synthesized emotional speech to acquire these prosodic variations.

A new method based on voice conversion framework[2] has been proposed for ESS. In this framework, a neutral speech is converted to an emotional speech using mapping functions of spectrum, F0 and other prosodic parameters. However in the literatures, spectral conversion is primarily considered while prosodic features such as F0 is simply converted using the standard linear mean-variance transformation. Prosodic features are very important to synthesize the emotional speech with a sustaining fact that the classification of emotional speech is mainly dependent on different prosodic parameters such as F0 mean, F0 level, duration and etc. Experiments listed in [2] also illustrate that prosodic features mainly contribute to emotional expression in speech. So the paper focuses on converting prosodic features especially F0 contour to implement emotional conversion.

The paper is supported by National Natural Science Foundation of China (No. 60575032).

In the paper, we would like to use a parametric F0 model to explore underlying relations between source and target F0 contours, and the pitch target model is selected for its capability of describing Mandarin F0 contour and its convenience for parametric alignment. The structure of this paper is organized as follows: the second section analyzes current F0 conversion methods and then describes pitch target model originally designed for Mandarin F0 contour; In the third section GMM and CART methods are employed to convert pitch target parameters; and then the fourth section presents the complete emotional conversion experiments based on voice conversion framework and the evaluation on this system, finally the last section gives the conclusion.

2. PARAMETRIC MODEL FOR F0 CONVERSION

2.1. Analysis of F0 conversion

There are various studies which propose many complex algorithms(listed in [3]) for modelling and converting F0 contours, and a common ground of these methods is that the F0 conversion is directly performed on F0 contour itself. However, there is serious uncertainty in prosody even when a speaker utters the same transcription in different time[4], in other words, the purpose of F0 conversion should not absolutely simulate target F0 contour. In addition, as mentioned by Yi Xu [5]: "observed F0 contours are not linguistic units per se. Rather, they are the surface realizations of linguistically functional units such as tone or pitch accent.". Thus we argue that the emphasis of F0 conversion should convert the underlying F0 pattern dependent on the speaker rather than F0 contour itself.

Some parametric F0 models such as Fujisaki model[6], STML[7], and pitch target model, can be used to explore the underlying relation between source and target speakers' F0 patterns. Among these parametric F0 models, pitch target model[5] originally designed for Mandarin speech is used in our F0 conversion framework for its capability of describing Mandarin F0 contour and its convenience for parametric alignment.

2.2. Pitch target model

According to the pitch target theory[5], the framework of the model consists of definitions for pitch targets and rules of their implementation. Pitch targets are defined as the smallest

operable units associated with linguistically functional pitch units, and these targets may be static(a register specification, [high] or [low]) or dynamic(a linear movement specification, [rise] or [fall]). The implementation rules are based on possible articulatory constraints on the production of surface F0 contours. The production of surface F0 contours are a process of continuous approximation of the target throughout the tone-carrying syllable, and when the syllable boundary is reached, the approximation in the next syllable begins to the new pitch target.

Assuming the syllable boundary is [0,D], the following equations[8] are defined as

$$T(t) = at + b \tag{1}$$

$$y(t) = \beta \exp(-\lambda t) + at + b \tag{2}$$

$$0 \le t \le D, \lambda \ge 0$$

where $T()$ is the underlying pitch target, and $y()$ is the surface F0 contour. Parameters a and b are the slope and intercept of the underlying pitch target, respectively. These two parameters describe an intended intonational goal by the speaker, which can be very different from the surface F0 contour being observed. Coefficient β is a parameter measuring the distance between F0 contour and the underlying pitch target when $t = 0$. Parameter λ is a positive number representing the rate of decay of the exponential part. In other words, it describes how fast the underlying pitch target is approached. The greater the value of λ is, the faster the speed. Thus a pitch target model of one syllable can be represented by a set of parameters: (a, b, β, λ).

2.3. Estimating pitch target parameters

To apply the pitch target model for F0 conversion, it is necessary to automatically estimate pitch target parameters (a, b, β, λ) from real F0 contour. In current research, similar with the method in [8], pitch target parameters can be estimated by nonlinear regression routine with expected-value parameters at initial and middle points of each syllable's F0 contour. The Levenberg-Marquardt algorithm is used for estimation as the nonlinear regression routine. Because the estimating procedure is syllable by syllable, syllable boundaries need to be determined in advance.

There are some possible extensions to the above estimating method. A first-order linear function is given to describe the portion of F0 contour between the ending point of pitch target and syllable ending boundary. Because the purpose of F0 conversion is to map the main trend of F0 contour and too many parameters may induce more error in mapping procedure, the portion before the ending point of pitch target is represented by pitch target model and the rest of F0 contour remains. If there are more than one pitch contour segments in a syllable, the longest pitch contour is picked up to estimate pitch target model for representing the syllable.

3. F0 CONVERSION

F0 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 F0 contour, pitch target parameters estimated from F0 contour are employed to build mapping rules.

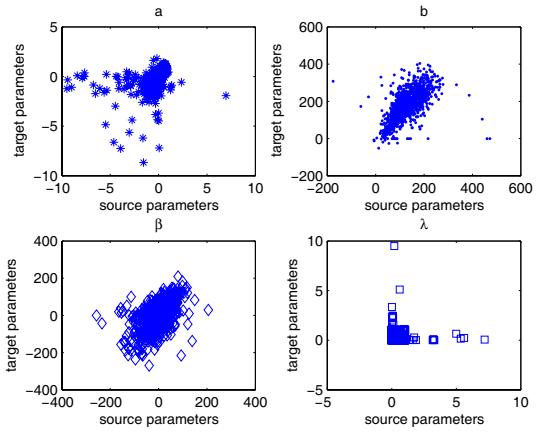

Fig. 1. The scatter plots of four pitch target parameters in neutral to joy conversion.

Pitch target parameters are estimated from aligned parallel F0 contours according to the estimating method above, and scatter plots of four pitch target parameters in neutral to joy conversion are given in Fig.1. It can be observed that parameter λ have worse correlation than parameters a, b and β between source and target pitch target models, so the parameter λ is not considered in the following conversion.

3.1. GMM

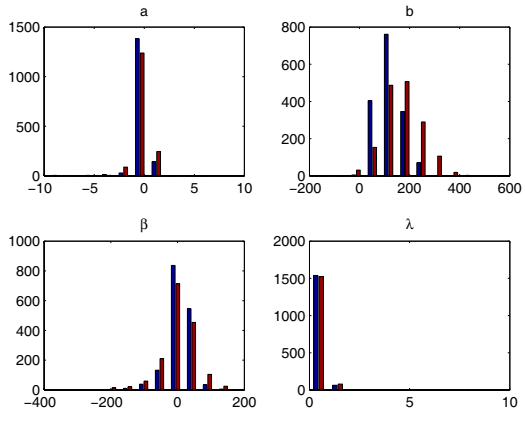

Fig. 2. Histogram plots of four pitch target parameters in neutral to joy conversion, which distinguish source parameters from target parameters using different colors.

Fig.2 shows histogram plots of four pitch target parameters, and it discloses that the parameters of a, b, and β approximately obey normal or Gaussian distributions. Inspired by these observations, Gaussian Mixture Model(GMM) is employed to convert pitch target parameters as well as used in spectral conversion. For each pitch target parameter a,

b and β, source and target parameters are treated as obeying joint Gaussian distributions, and then the combination of source(marked as x) and target(marked as y) vectors $z = [x^T y^T]^T$ is used to estimate GMM parameters. Thus the mapping function can be obtained by the regression[9].

$$F(x) = \sum_{q=1}^{Q} p_q(x)[\mu_q^Y + \Sigma_q^{YX}(\Sigma_q^{XX})^{-1}(x - \mu_q^X)] \quad (3)$$

where $p_q(x)$ is the conditional probability of a GMM class q given x:

$$p_q(x) = \frac{\alpha_q N(x; \mu_q^X, \Sigma_q^X)}{\sum_{p=1}^{Q} \alpha_p N(x; \mu_p^X, \Sigma_p^X)} \quad (4)$$

$$\Sigma_q = \begin{bmatrix} \Sigma_q^{XX} & \Sigma_q^{YX} \\ \Sigma_q^{XY} & \Sigma_q^{YY} \end{bmatrix}; \mu_q = \begin{bmatrix} \mu_q^X \\ \mu_q^Y \end{bmatrix}$$

$N(x; \mu_q, \Sigma_q)$ denotes a normal distribution with mean vector μ_q and covariance matrix Σ_q, α_q is normalized positive scalar weight, and Q is the number of GMM components.

3.2. CART

On the assumption that emotions are relevant to linguistic and contextual information, it is reasonable to use Classification and Regression Trees (CART) for building mapping rules between source and target pitch target parameters. Eight factors are considered in a CART's feature set, which are:

- Tone Identity (Including current, previous and next tones, with 5 categories).
- Initial Identity (Including current and next syllables' initial types, with 8 categories).
- Final Identity (Including current and previous syllables' final types, with 4 categories).
- Position in sentence.

For a emotional conversion, CART are used to predict differences between source and target pitch target parameters a, b and β.

In the training procedure, source and target pitch contours from parallel corpus is firstly aligned according to syllable boundaries, and then pitch target parameters are extracted from each syllable's pitch contour, finally mapping functions of parameters a, b, and β are obtained ; In the converting procedure, source pitch target parameters estimated from source pitch contours are transformed by these mapping functions obtained in training procedure and then converted pitch target parameters to generate new pitch contours associated with target characteristic. A converting example is given in Fig.3, in which source pitch contours(the top) in neutral state are converted into new pitch contours(the bottom) to simulate corresponding target pitch contours(the middle) in joy state.

4. EXPERIMENTS AND EVALUATIONS

In our emotional corpus, there are 300 sentences with average length of 7 syllables using the same text for different emotional performances. These sentences are produced by a professional actor in five basic emotional states: neutral, joy, sadness, fear, anger, and then each sentence is automatically segmented at the syllable level by a forced alignment procedure. 260 sentences including about 1600 syllables in each emotional category are used to train transforming functions and the rest are employed to test these conversions.

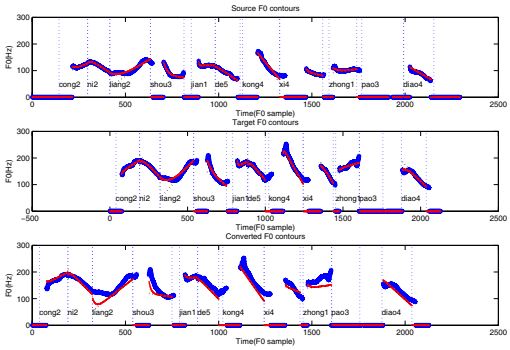

Fig. 3. An example of F0 conversion using the pitch target model in neutral to joy conversion.

Four emotional conversions (marked as neu-joy, neu-anger, neu-sadness and neu-fear) are conducted in the experiment. To compare GMM and CART (using Wagon toolkit) mapping methods, root mean square errors are shown in Table.1 and Table.2. From these tables, it can be observed that the GMM mapping method is better than the CART method. A possible reason is that training data in the experiment are not enough to cover most linguistic information when using the CART method. Another fact can be concluded from these tables is that both mapping methods need to be improved to implement the more accurate mapping.

	neu-joy	neu-sadness	neu-anger	neu-fear
a	1.96	2.59	1.70	1.92
b	105.60	71.03	79.64	46.09
β	43.03	29.93	42.93	27.04

Table 1. Mapping errors of GMM method

	neu-joy	neu-sadness	neu-anger	neu-fear
a	2.71	3.25	2.54	2.84
b	101.02	98.23	106.71	82.45
β	53.74	47.86	56.74	45.47

Table 2. Mapping errors of CART method

It is a complicated task to convert a neutral speech into an emotional speech because the emotional speech differs from the neutral speech on not only segmental cues such as spectral envelope, but also suprasegmental cues such as F0(intonation), speaking rate, loudness, etc. In the paper,

speaking rate(duration) and loudness(intensity) are also simply transformed, and converting ratios are picked up through experiential selection as shown in Table.3. Based on our previous work[9], spectral envelopes are represented by linear spectral pair (LSP) features and are transformed by a hybrid mapping method.

	neu-joy	neu-sadness	neu-anger	neu-fear
duration	0.95	1.25	0.8	0.7
intensity	1.1	0.7	1.5	1.0

Table 3. Transforming ratios of duration and intensity

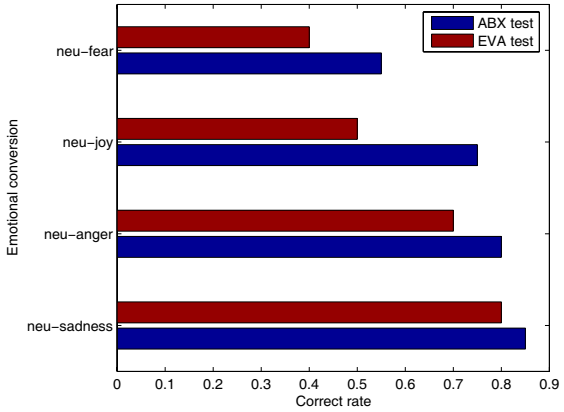

Fig. 4. Results of expressive converting evaluations.

There are 8 listeners to give a subjective evaluation on these test sentences. Two methods are conducted to evaluate the proposed emotional conversion:

- *ABX test:* ABX test in evaluating voice conversion is used in the evaluation. all listeners are required to judge whether a converted speech X sounded closer to a source neutral speech A or a target emotional speech B. This test confirms whether the conversion system is successful.
- *EVA test:* Only converted speeches are listened to and then the associated emotional state is given by these listeners. This test confirms whether the emotional conversion is successful.

Results of the evaluation are shown in Fig.4, in which X axis is the emotional state and Y axis is the mean correct rate(it is the ratio of judging X as B in ABX test, and considering the converted speech as the corresponding emotional speech in EVA test) of all listeners. ABX test has proved that the converted emotional speech possesses the corresponding emotional state compared with the source speech. Because in ABX tests conners can compare the converted speech with the source neutral speech, results of ABX tests are better than those of EVA tests. There are differences among emotional conversions in these perception tests, in which the neu-sadness and neu-fear conversions are respectively best and worst. It can be noted in the experiment that the corpus including 300 sentences is not enough to accurately implement emotional conversion, and a larger corpus may improve the expressive speech synthesis system.

5. CONCLUSIONS

This paper proposes to employ pitch target model to implement F0 conversion for expressive Mandarin speech synthesis. Advantages of the proposed method are that parametric F0 models such as pitch target model can provide an underlying linguistical or physiological description for surface F0 contour and it often can furnish several compact parameters to represent a long pitch contour. GMM and CART mapping methods are employed to generate transforming functions of pitch target parameters, and the GMM method is slightly better than the CART method in our current experiment. The current corpus is not enough to train perfect mapping function and a larger speech corpus is required in future work. The proposed F0 conversion method can be used to implement voice conversion too.

6. ACKNOWLEDGEMENTS

The author would like to thank Professor Yi Xu for his many constructive suggestions for this work.

7. REFERENCES

[1] Marc Schröer, "Emotional speech synthesis: A review," in *Proc. Eurospeech 2001*, 2001, pp. 561–564.

[2] H. Kawanami, Y. Iwami, and T. Toda, "Gmm-based voice conversion applied to emotional speech synthesis," in *Proc. Eurospeech 2003*, Geneva, Switzerland, 2003, pp. 2401–2404.

[3] Zeynep Inanoglu, "Transforming pitch in a voice conversion framework," M.S. thesis, University of Cambridge, July 2003.

[4] Min Chu, "The uncertainty in prosody of natural speech and its application in speech synthesis," *Joural of Chinese Information Processing*, vol. 18, no. 4, pp. 66–71, 2004.

[5] Yi Xu and Q. Emily Wang, "Pitch targets and their realization: Evidence from mandarin chinese," *Speech Communication*, vol. 33, pp. 319–337, 2001.

[6] H.Fujisaki, Changfu Wang, Sumio Ohno, and Wentao Gu, "Analysis and synthesis of fundamental frequency contours of standard chinese using the commandcresponse model," *Speech Communication*, vol. 47, pp. 59–70, 2005.

[7] Greg Korchanski and Chilin Shih, "Prosody modeling with soft templates," *Speech Communication*, vol. 39, pp. 311–352, 2003.

[8] Xuejing Sun, *The Determination, Analysis, and Synthesis of Fundamental Frequency*, Ph.D. thesis, NORTHWESTERN UNIVERSITY, 2002.

[9] Yongguo Kang, Zhiwei Shuang, Jianhua Tao, Wei Zhang, and Bo Xu, "A hybrid gmm and codebook mapping method for spectral conversion," in *Proc. The First International Conference on Affective Computing and Intelligent Interaction*, 2005.