JCA2007 4-6 June
The Japan-China Joint Conference of Acoustics 2007

PROSODY ADAPTATION OF MANDARIN TEXT TO SPEECH SYSTEM

Jianhua Tao Jian Yu
National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences, Beijing, China, 100080
{jhtao, jyu}@nlpr.ia.ac.cn

ABSTRACT

The paper constructs a prosody adaptation model for Mandarin text to speech system by using the prosody template transformation method. Two classic transformation algorithms, including the Gaussian Mixture Model (GMM) and the Classification and Regression Tree (CART) are tested. Both methods use the linguistic context information as the part of input parameters. Experimental results show that the prosody adaptation model proposed in the paper can adapt the text to speech to a new speaker's prosody style with the high performance. Since the context information is simplified into some types for each syllabic tone in GMM, and the CART model uses much more detailed context features, the CART model get better performance than the GMM. It is proved by the listening tests in the paper.

INTRODUCTION

As the recent development of corpus-based TTS technology, clear and natural synthesized speech can be easily generated by the computer. But most of current TTS systems can only generate the speech with some certain styles, while users may expect the prosody of the system can be adaptable. So far, the most widely employed method for carrying out this adaptation is based on adjusting the pitch range of the source speaker to match the target while keeping the shape of the contour unchanged. The disadvantage of this method is many tiny variations in pitch contour are neglected during the conversion, which leads to dissimilar output pitch contours. In addition, some prosody conversion methods based on statistics and machine learning algorithm, such as stochastic transformation [4], codebook-based transformation [5], tilt model-based transformation [6] and so on, are tried. All these methods tried to directly map the pitch contours between two speakers without any context information. Results are quite limited in the simulation of new speakers.

Lots of previous work has proved that prosody features are strongly correlated to the context information which can be easily generated by the text analysis module of the TTS system. To use such features, a group of context information is integrated in our work. Both Gaussian Mixture Model (GMM) and Classification and Regression Tree (CART) method are tried in the paper. Since the context information is used, we would more like to use the name of "prosody adaptation" rather than "prosody conversion". Unlike the traditional prosody conversion which uses parallel training corpus, we use a small subset of the speech synthesis corpus for the training. Although outputs of both methods are quite similar, the input of the prosody conversion contains only the prosody

information, while both context information (from the text) and the prosody information are integrated in our prosody adaptation model. Fig. 1 shows the differences between prosody adaptation and prosody conversion.

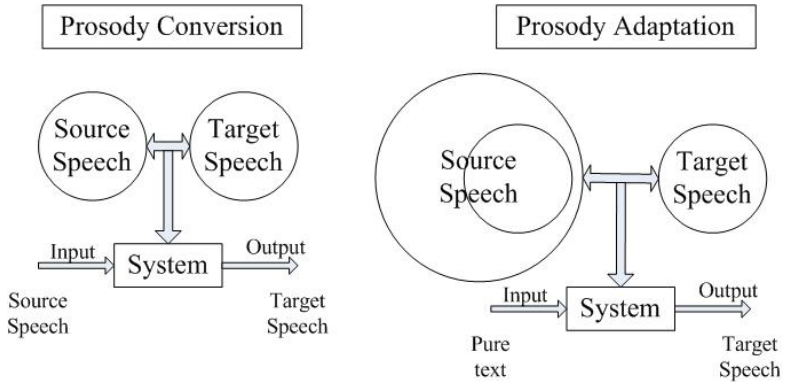

Fig. 1. The difference between prosody conversion and prosody adaptation

The essential idea of the prosody adaptation is to construct a new prosody template set by prosody parameter transformation. In the paper, we use the following steps:

At first, based on the parallel subset of source speech and target speech, the relationship between source and target prosodic parameters can be obtained. Then, by applying that relationship to the original prosody template set, a new set which has the target speaker's characteristics can be constructed. Based on this new template set, the speech with the target speaker's prosody style can be synthesized.

The rest of the paper is organized as follows. Section II briefly introduces pitch parameters used for prosodic templates generation. Section III describes the prosody adaptation method based on prosody template transformation in detail. Then, section IV is the experiment about that method. The conclusion is made in the final section V.

PROSODIC TEMPLATES

What kinds of prosodic parameter should be selected to be transformed is very critical, which determines the final performance of prosody adaptation model. The most simple and direct method is to transform the pitch contour itself. Although its realization is quite easy, its application is very inflexible and the result is not so good. It seems that the continuous pitch contour must be converted to a pattern represented by a certain number of parameters. In the paper, we apply a model shown in Fig. 2. In this method, the pitch contour of a Chinese syllable is parameterized into five parameters. Among these, $F0_M$ denote the pitch register, which reflects the overall trend of pitch contours, while $F0_S, F0_E, F0_{SD}$ and $F0_{ED}$ can be considered as boundary features, which can be used to measure the naturalness of pitch contours in local area.

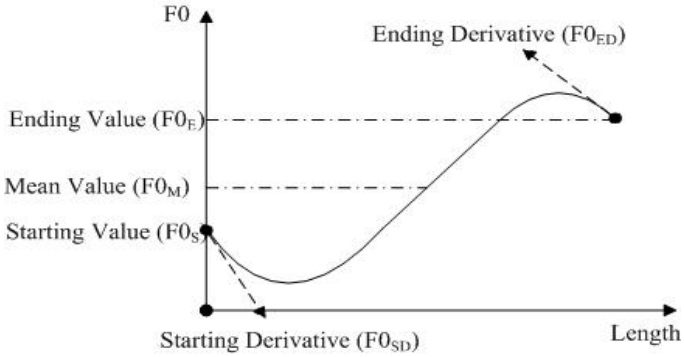

Fig. 2. Five parameters used in the template extraction

PROSODY TRANSFORM

Gaussian Mixture Models (GMMs) and Classification And Regression Tree (CART) are two popular methods in machine learning research, which can be used to find the relationship between source and target prosodic features. Unlike the simplest linear modification method, the GMM and CART models try to map the subtle prosody distributions. The following part will introduce both two methods in detail.

GMM based method

The GMM assumes the probability distribution of the observed parameters to take the following form,

$$p(x)=\sum_{q=1}^{Q}\alpha_q N(x;\mu_q;\Sigma_q),\sum_{q=1}^{Q}\alpha_q=1,\alpha_q\geq 0 \tag{4}$$

where Q is the number of Gaussian components, α_q is the normalized positive scalar weight, and $N(x;\mu_q;\Sigma_q)$ denotes a D-dimensional normal distribution with mean vector μ_q and covariance matrix Σ_q, and can be described as,

$$N(x)=\frac{1}{(2\pi)^{D/2}\left|\Sigma_q\right|^{1/2}}\exp[-\frac{1}{2}(x-\mu_q)^T\Sigma_q^{-1}(x-\mu_q)] \tag{5}$$

The parameters of the conversion function are determined by the joint density of source and target features. It has been confirmed that the joint density performs better than the source density. It can lead to a more judicious allocation of mixture components and avoids certain numerical problems. For each prosodic parameter $F0_M$, $F0_S, F0_E, F0_{SD}$ and $F0_{ED}$, source and target parameters are assumed to be Gaussian distributed, and then the combination of source (marked as x) and target (marked as y) vectors $Z_k=\begin{bmatrix}X_k & Y_k\end{bmatrix}^T, k=1,...,N$ is used to estimate the GMM parameters.

To integrate the context information, we built several GMMs in different context information which contains tone identity (including current, previous and following tones, with 5 categories), initial identity (including current and following syllables' initial types, with 8 categories), final identity (including current and previous syllables' final types, with 4 categories), position in sentence, part of speech (including current, previous and following words, with 30 categories). The model is based on the syllable level. To simply the work, we only use only 20 GMMs for each syllabic tone.

The parameters (α, μ, Σ) are estimated with the expectation-maximization (EM) algorithm, and the conversion function can be found by using regression

$$F(x) = \sum_{q=1}^{Q} p_q(x)[\mu_q^Y + \Sigma_q^{YX}(\Sigma_q^{XX})^{-1}(x - \mu_q^X)] \tag{6}$$

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

$$p_q(x) = \frac{\alpha_q N(x; \mu_q^X; \Sigma_q^X)}{\sum_{p=1}^{Q} \alpha_p N(x; \mu_p^X; \Sigma_p^X)} \tag{7}$$

here $\Sigma_q = \begin{bmatrix} \Sigma_q^{XX} & \Sigma_q^{YX} \\ \Sigma_q^{XY} & \Sigma_q^{YY} \end{bmatrix}; \mu_q = \begin{bmatrix} \mu_q^X \\ \mu_q^Y \end{bmatrix}$. $N(x; \mu_q^X; \Sigma_q^X)$ denotes a normal distribution with mean vector μ_q and covariance matrix Σ_q.

CART based method

CART model is traditionally used for prosody prediction, such as F0s, durations, prosodic phrase boundaries, etc. They have been proved to be able to efficiently integrate the contextual information in the prosody processing. Similar as the F0 prediction of normal speech synthesis, we use the difference between the source and target prosody as the target for the training. The framework is shown in Fig. 3.

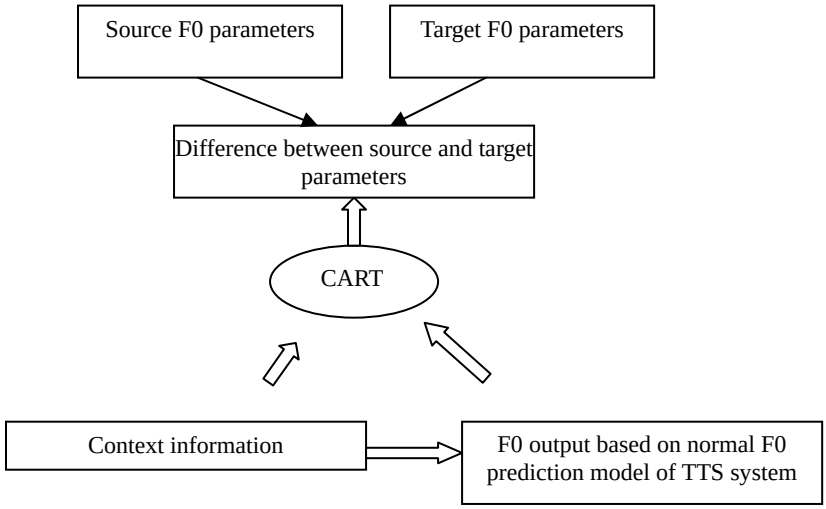

Fig.3. CART based prosody adaptation

In the model, the input parameters of the CART contain two parts, the context information which is mentioned in the above part, and the F0 output (contains five parameters, $F0_M$, $F0_S, F0_E, F0_{SD}$ and $F0_{ED}$) of the normal text to speech prosody prediction model which is described in [7]. The idea is to use the output of the normal prosody model as the reference for the comparison between the source and the target. The output parameters are the differences of F0 parameters, $F0_M$, $F0_S, F0_E, F0_{SD}$ and $F0_{ED}$ between the target and the source. Wagon toolkit, with full CART function, is used in our work. Similar to the GMM method in training procedure, source and target pitch contours from parallel corpus are aligned according to syllable boundaries, and then prosodic parameters are extracted from each syllable's pitch contour. Finally, parameters of the model are estimated by using the CART regression algorithm.

EXPERIMENTS

Database used for the training

During the experiment, the source speech corpus which is also used for the speech synthesis consists of 6000 phonetically balanced Mandarin sentences, uttered by a female professional speaker. The target speech corpus consists of 500 sentences which covers all syllables in Mandarin, uttered by a non-professional female speaker. All corpus is sampled at 16000 Hz, and the segmentation is done automatically first and then checked manually.

Evaluations

Three subjective experiments, including experiment on prosodic naturalness, speaker individuality and ABX test, are designed to evaluate the performance of the proposed methods, GMM method and CART method. There are 50 testing sentences and 6 listener in the evaluation. The evaluation is performed in the following ways:

Evaluation of prosody naturalness: This evaluation is to evaluate the naturalness of the output prosody via the MOS method. Due to the unclear output prosody arisen from overly smooth problem of GMM, listeners are specially required to compare the naturalness of the prosody at this point. The score of this experiment is ranged from “one” to “five”. “five” means excellent quality while “one” means very bad quality.

Evaluation of speaker individuality: This experiment is to evaluate speaker individuality of the output prosody and is measured using scores ranged from “one” to “five”. “five” means that the output prosody is very closed to target speaker's speaking style while “one” means that the output prosody has the very bad similarity.

ABX test: ABX test ask listeners to make an opinion. Listeners are required to judge whether the output prosody X sounded closer to source A or target B. The test confirms whether the prosody adaptation is successful.

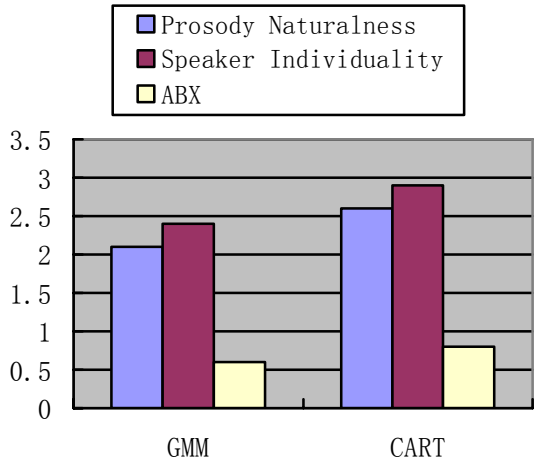

Fig. 4. Evaluation scores of two algorithms

The mean scores of all listeners are shown in Fig.4. The result of ABX is the proportion of the output prosody X which is closer to target B. It is observed that all two methods successfully convert the source prosodies to target ones. The results of experiments on speaker individuality and the prosody naturalness have similar distribution with ABX test. Because of the difficulty to model the relationship between two speakers, the overly smooth problem arises in converted features based on GMM method due to the simplification of context features. The CART method uses more

detailed context information than the GMM method and keeps more subtle information of the converted prosody. Experiment results prove that the performance of CART method is better than that of GMM method.

SUMMARY

In the paper, we introduce a prosody adaptation model for Mandarin text to speech system based on the prosody template transformation method. Unlike the prosody conversion, the prosody adaptation uses the context information as part of the input parameters. The training set is a subset of the whole speech synthesis corpus. By using the prosody model of the normal text to speech system as the baseline prosody output, prosody adaptation model can easily adapt the system to different speaker's prosody styles.

However, there are still some disadvantages of the methods proposed in the paper, such as there is no processing of duration and prosodic structures between two speakers, the result is still very encouraging for the prosody simulation of a specific person. The future work will try to solve the problem in these disadvantages.

ACKNOWLEDGMENT

The work is partly supported by National Natural Science Foundation of China (No. 60575032) and the 863 Program (No. 2006AA01Z138). The authors thank especially the students of NLPR for their cooperation with the experiments.

REFERENCES

[1] Zhiweng Shuang, E. A Novel Voice Conversion System based on Codebook Mapping with Phoneme-tied Weighting. in ICSLP2004. 2004. Jeju Island, Korea.

[2] Yongguo Kang, Z.S., Jianhua Tao, Wei Zhang, Bo Xu. A hybrid GMM and codebook mapping method for spectral conversion. in The First International Conference on Affective Computing & Intelligent Interaction. 2005. Beijing, China.

[3] Yining Chen, M.C. Voice Conversion with Smoothed GMM and MAP Adaptation. in Eurospeech 2003. 2003. Geneva.

[4] Ceyssens. On the Construction of a Pitch Conversion System. in EUSIPCO2002. 2002. Toulouse, France.

[5] Chappell, D., Hansen, J. Speaker-specific pitch contour modeling and modification. in ICASSP1998. 1998. USA.

[6] Taylor, P., Analysis and synthesis of intonation using the tilt model. Journal of Acoustic Society of America, 2000. 107: p. 1697-1714.

[7] Yu, J., W. Zhang, and J. Tao. A New Pitch Generation Model Based on Internal Dependence of Pitch Contour for Mandarin TTS System. in ICASSP. 2006. Toulouse, France.

[8] Yu, J., J. Tao, and X. Wang. Pitch Prediction for Mandarin TTS with Mutual Prosodic Constraint. in ISCSLP. 2006. Singapore.

[9] Yi Xu, Q.E.W., Pitch Targets and Their Realization: Evidence from Mandarin Chinese. Speech Communication, 2001. 33.

[10] Kang, Y., et al. A hybrid gmm and codebook mapping method for spectral conversion. in The First International Conference on Affective Computing and Intelligent Interaction. 2005. Beijing, China.