PAPER

A novel prosody adaptation method for Mandarin concatenation-based text-to-speech system*

Jian Yu[†] and Jianhua Tao[‡]

*National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences*

(Received 8 October 2007, Accepted for publication 5 June 2008)

Abstract: The paper presents a prosody adaptation method which is able to adapt the prosody model of text to speech (TTS) to a new style with a small training corpus. Unlike the conventional prosody mapping between two parallel prosody features, the paper tries to integrate the prosody conversion into the prosody generation model of TTS. In the paper, we use a template-based prosody model which consists of two major parts: the prosody template library and the template parameter prediction trees for TTS system. With this model, the prosody adaptation is realized by the following two steps: converting the prosody template library to the target speaker's prosody based on the mapping methods, re-training prosody prediction trees with the small target training set. In the model, some transformation algorithms, including linear regression, Gaussian Mixture Model (GMM) and Classification and Regression Tree (CART) are involved. Experimental results show that the prosody adaptation system can generate synthesized speech which is much similar with the target speaker.

Keywords: Text to speech, Prosody adaptation, Prosody conversion

PACS number: 43.72.Ja [doi:10.1250/ast.30.33]

1. INTRODUCTION

As the recent development of corpus-based TTS technology, clear and natural synthesized speech can be easily generated by computer. However, the synthesized result is still a little monotone, which limits the application of TTS system in our daily lives. Most of current TTS systems can only output speech with some specific pre-trained styles, while users may expect that the system is able to simulate the features of their own. This expectation leads to many research interests extended from TTS research, such as voice conversion (VC), which enables a user to transform a source speaker's speech into another target speaker's speech with distinct characteristics while preserving the original content (meaning). Typical methods include codebook mapping, and Gaussian Mixture Model (GMM) [1–5]. Most of voice conversions mainly focus on the spectrum transformation between different voices. Besides, some prosody conversions were also tried by using stochastic transformation method [6], codebook-based transformation method [7], tilt model-based trans-

formation method [8] and so on. Both spectrum conversion and prosody conversion are based on the direct acoustic features mapping. The parallel training corpus is necessary to guarantee the model training.

The objective of prosody adaptation is to build a complete prosody generation model for TTS system based on the small target training corpus. Unlike the voice conversion, we can get the context information from the text analysis part of TTS system as the additional feature for prosody adaptation. In that situation, if we have an enough target training corpus, it will be much easier to adapt the prosody model by simply re-training the model with the new corpus. The question is nobody wishes to use a large target corpus to do the adaptation. So, our work is just focusing on how to construct a well-performed target prosody model based on a small target corpus. For that purpose, the original, phonetically balanced large corpus could be helpful.

With that question, we separate the original training corpus for TTS system into two parts: the basic part and the extended part. The basic part is a small set, but covers all phonemes (or syllables in Mandarin) and most of frequently used context features. The extended part is used to generate more candidates for large corpus based unit selection TTS system. The target training corpus shares the

*Special issue on JCA2007, Paper number: SS-2-3

[†]e-mail: jyu@nlpr.ia.ac.cn

[‡]e-mail: jhtao@nlpr.ia.ac.cn

Fig. 1 The framework of prosody adaptation for TTS.

same transcription as that in the basic part of the original training corpus. With this idea, we can get a parallel training corpus between the source and the target. Such corpora are relatively small, but we have context information to facilitate the work. After we build the transform model between the source and the target, we transform all of syllabic F_0 contours from source corpus into the target styles. Then, we use the syllabic F_0 contours as the basic prosody template to create the prosody generation model for TTS system. We call it the template based prosody generation model [9,10]. Further adaptation will be made by re-training the prosody model in new target corpus with new adapted prosody templates. The prosody adaptation method presented in the paper is designed for concatenation-based TTS system because the concatenation method is still the most popular method in the commercial TTS system. Figure 1 shows an overview of the prosody adaptation framework.

The rest of the paper is organized as follows. Section two introduces the template-based prosody generation model, which is the foundation of prosody adaptation framework. Then, section three introduces the prosody adaptation framework in detail, which mainly includes the adaptation of prosody template library and the adaptation of prediction trees. Those two parts constitute a whole prosody model. Section four is experiment which mainly compares performances between the prosody adaptation and conventional prosody conversion. Section five introduces several applications about the prosody adaptation framework. Finally, there is a conclusion in section six.

2. TEMPLATE BASED PROSODY GENERATION MODEL

Before the introduction of the prosody adaptation method, we need to introduce the prosody model first. Because Mandarin is a syllable-based language, the prosody model is constructed based on prosody template. The model consists of two parts: the prosody template library and prosody template parameter prediction trees for

Fig. 2 Seven parameters used in the template extraction.

TTS system. Both parts need to be adapted in the prosody adaptation procedure.

2.1. Prosody Template

Prosody templates are extracted from real speech and represent syllabic F_0 contours of the corpus. For Mandarin speech, each syllable has a relatively constant F_0 pattern. However, the syllabic F_0 contour will be distorted in various context environments. Such kind of patterns can be represented by seven parameters, as shown in Fig. 2, which shows an example of syllabic F_0 contour. Among these, F_{0M} , F_{0B} and F_{0T} denote the F_0 register and the F_0 range, which reflect the overall trend of F_0 contours. While F_{0S} , F_{0E} , F_{0SD} and F_{0ED} can be considered as boundary features, which can be used to measure the naturalness of F_0 contours in local area.

With these seven parameters, we use each syllable F_0 pattern as a template, and create a template library which contains all syllable F_0 patterns from the training corpus. All seven parameters are used for the direction of template selection, and among these, five of them (F_{0T} and F_{0B} are not included) are used for the F_0 contour parameterization. In the mean time, we also keep the original context information for each template in the template library, which allows us to select and concatenate the templates to form an intonation during the speech synthesis procedure. In the following procedure, the template library is an important part which needs to be adapted.

2.2. Prosody Template Selection

The most important issue of prosody template selection and concatenation is how to reasonably define the cost function. The cost definition consists two parts: the target cost and the concatenation cost.

2.2.1. Target cost

With the template library, we train three Classification and Regression Trees (CART) to predict F_{0M} , F_{0B} and F_{0T} based on context information. To select the most appropriate prosody template for speech synthesis, the target cost of the selection is defined as the difference between the

Fig. 3 Examples showing the impact of adjacent syllables' F0 contours on the current one.

predicted values and the candidates' real values of these three parameters, noted as DF_{0M} , DF_{0B} and DF_{0T} .

$$target_cost = w_1 * DF_{0M} + w_2 * DF_{0B} + w_3 * DF_{0T} \quad (1)$$

Where, w_1 , w_2 and w_3 are weights for the parameters DF_{0M} , DF_{0B} and DF_{0T} , which are set based on experience.

The target cost is used to depict the overall trend of the F_0 contour. For example, F_0 declination in naturally read discourses can be realized by all of the three parameters descend as the sentence approaches the end. Minimizing the target cost can make the output F_0 contour similar to that in the training corpus. In summary, three template parameter prediction trees are constructed in that procedure.

2.2.2. Concatenation cost

A reasonable definition of the concatenation cost should be applied to the measurement of the naturalness of F_0 contours between adjacent syllables. Because the larynx can not change its state instantaneously in the speaking process, there are some strong relationships between the current syllable's template parameter value and the adjacent syllable's F_0 contour. Even in the situation where there are pitch jumps across the syllable boundary, it seems that the F_0 contour is virtually connected across the silence and voiceless initial, as Fig. 3 shows.

Based on these facts, our definition of the concatenation cost makes full use of these prosodic features, especially adjacent syllables' F_0 contours. Template parameters involved in the concatenation cost include F_{0S} , F_{0E} , F_{0SD} and F_{0ED} , which can be considered as boundary features of a syllable's F_0 contour. According to the above analysis, when we predict these four template parameters, the prosodic features listed in Table 1 should be involved, which will greatly improve the predicted results. In that table, the frequently used text information includes tone identity, initial/final identity and prosody structure information, etc. Classification and Regression Tree (CART) is also used as the predicting model. In that procedure, other four template parameter prediction trees are constructed. All seven prediction trees are the kernel part of the prosody model, which need to be adapted in the following procedure.

Table 1 Features used in predicting F_{0S} , F_{0E} , F_{0SD} and F_{0ED} .

features in predicting F_{0S} and F_{0SD}	features in predicting F_{0E} and F_{0ED}
frequently used text information (tone, initial/final identity, prosody structure, etc)	frequently used text information (tone, initial/final identity, prosody structure, etc)
previous syllable's F_{0E} and F_{0ED}	following syllable's F_{0S} and F_{0SD}
pause length before current syllable	pause length after current syllable
current syllable's initial length	following syllable's initial length

Fig. 4 The calculation procedure of the concatenation cost.

Worth particular mentioning is that the predicted value by this method can be considered as the expected value by adjacent syllables, so the difference between these predicted values and real values can be used to measure the naturalness of F_0 contours between adjacent syllables. As Eq. (2) demonstrates, the value of the concatenation cost is defined as the weighted sum of differences between predicted values and real values of these four template parameters, noted by DF_{0S} , DF_{0E} , DF_{0SD} and DF_{0ED} respectively. Figure 4 schematically illustrated the calculation procedure of the concatenation cost. This definition is much reasonable because it makes full use of the internal dependence of F_0 contour.

$$concatenation_cost = w_4 * DF_{0S} + w_5 * DF_{0E} + w_6 * DF_{0SD} + w_7 * DF_{0ED} \quad (2)$$

Similar as Eq. (1), here, the parameters w_4 , w_5 , w_6 and w_7 are the weights for DF_{0S} , DF_{0E} , DF_{0SD} and DF_{0ED} . All the weights are set based on experience.

2.3. The Pitch Contour Generation

Each syllable in the synthesized sentence has many template candidates. For every single template candidate, a target cost is calculated, and for every two templates, which belong to two adjacent syllables respectively, a

concatenation cost is calculated. Based on these cost definitions, the Viterbi search algorithm is used to select the best template series with the minimum general cost.

Then, the last step is the generation of pitch contour based on the selected template series. In that process, the parameters F_{0S} , F_{0E} , F_{0SD} , F_{0ED} and F_{0M} are used. Suppose the pitch contour within a syllable can be precisely depicted by a function $f(x)$, an equation group can be listed based on the values of prosodic parameters. That is:

$$f(s) = F_{0S} \quad (3)$$

$$f(e) = F_{0E} \quad (4)$$

$$f'(s) = F_{0SD} \quad (5)$$

$$f'(e) = F_{0ED} \quad (6)$$

$$\int_s^e f(x)dx = (e - s) * F_{0M} \quad (7)$$

Where e and s are the starting time and ending time of the pitch contour within a syllable, respectively.

Two kinds of formulas are supposed to represent the function $f(x)$: third order polynomial function and exponential function, as follows:

$$1: f(x) = ax^3 + bx^2 + cx + d \quad (8)$$

$$2: f(x) = a * e^{bx} + cx + d \quad (9)$$

Among these two formulas, the second is based on the theory of PENTA model [11], and it is much easier to explain in phonetics. However, it can not precisely depict the pitch contour with more than one polar, just like the pitch contour plotted in Fig. 2. Therefore, the first one is adopted in our pitch model.

In addition, both two functions only have four coefficients, but the number of equations is five. Actually, this equation group has no solutions. To solve this problem, a nonlinear least squares minimization algorithm named Levenberg-Marquardt [12] is used to calculate similar solutions. At boundaries between finals and voiced initials, the F_0 values are smoothed by a moving average method, which is used to make sure the smoothness of the pitch contour.

An experiment is made to verify the effectivity of the proposed parameterization method. In that experiment, we calculate the prosody template parameters, and then re-synthesize the pitch contour directly based on the above method. The F_0 distance between the original and parameterized pitch contour is used to measure the effectivity of that parameterization method. The result are listed in Table 2, from that table, we can see that the parameterization error is quite small compared with the F_0 mean value, which show that the parameterization method is acceptable in the pitch contour generation model.

Under various constrains, this model outputs very natural F_0 contours. Because its good performance, that

Table 2 The error rate of the parameterization method.

Sentences involved	F_0 mean (Hz)	F_0 variance (Hz)	F_0 error (Hz)
1,000	250	61	10

model is adopted as the foundation of the prosody adaptation framework. As presented before, that model consists two parts: the template library and seven prediction trees. Both parts need to be adapted respectively in the prosody adaptation framework.

3. ADAPTATION OF THE PROSODY MODEL

As introduced in Sect. 1, the procedure of prosody adaptation can be separated into two steps:

Step 1: do the prosody conversion for all prosody templates in the basic parallel part between the source corpus and the target corpus, and then extend the conversion to all templates in the whole source corpus.

Step 2: re-train the template based prosody parameter prediction model in the new target corpus.

We call the first step “the adaptation of the template library.” In the paper, the prosody prediction consists of two cost functions which are realized by seven prediction trees. So, we call the second step “the adaptation of template parameter prediction trees.” Two steps are achieved as follows:

3.1. The Adaptation of the Template Library

In the adaptation of the template library, some algorithms which work in the prosody conversion field can be used. The difference is that, the small target corpus is only a subset of the large source corpus. The extended part of the source corpus could be used to enlarge the coverage of context information. Besides, if there is more than one source corpus, multiple transformation relationships can be constructed. Among these, the one with highest transformation precision is adopted. After getting the transformation relationship, we applied it to the corresponding large corpus and a new template library which has the same contextual coverage can be constructed. By this technique, the less coverage problem can be resolved.

Two techniques are used to get the transformation relationship. 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. Among these, the GMM just uses the prosodic features, while the CART model integrates

linguistic features into the mapping. The following part will introduce both two methods in detail.

(a) GMM-based Transformation

The GMM has proved to be very useful in the research of voice conversion. They work particularly well in spectrum smoothing, which assumes the probability distribution of the observed parameters to take the following form,

$$p(x) = \sum_{q=1}^Q \alpha_q N(x; \mu_q; \Sigma_q), \quad \sum_{q=1}^Q \alpha_q = 1, \quad \alpha_q \geq 0 \quad (10)$$

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,

$$N(x) = \frac{1}{(2\pi)^{D/2} |\Sigma_q|^{1/2}} \exp \left[-\frac{1}{2} (x - \mu_q)^T \Sigma_q^{-1} (x - \mu_q) \right] \quad (11)$$

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

$$F(x) = \sum_{q=1}^Q p_q(x) [\mu_q^Y + \Sigma_q^{YX} (\Sigma_q^{XX})^{-1} (x - \mu_q^X)] \quad (12)$$

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

$$p_q(x) = \frac{\alpha_q N(x; \mu_q^X; \Sigma_q^X)}{\sum_{p=1}^Q \alpha_p N(x; \mu_p^X; \Sigma_p^X)} \quad (13)$$

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

$N(x; \mu_q^X; \Sigma_q^X)$ denotes a normal distribution with mean vector μ_q^X and covariance matrix Σ_q^X .

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 [13]. It can lead to a more judicious allocation of mixture components and avoids certain numerical problems. For each prosody template parameter, 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 $Z_k = [X_k \ Y_k]^T$, $k = 1, \dots, N$ is used to estimate the GMM parameters.

(b) CART-based Transformation

Classification and Regression Trees (CART) have been successfully used in prosodic parameter prediction, such as duration, prosody phrase boundaries, etc. They efficiently integrate the context information which can be obtained from the TTS engine into prediction. In our research,

Fig. 5 The framework of CART-based prosody transformation.

the framework of the CART-based prosody conversion is shown in Fig. 5.

In this model, the input parameters of the CART contain:

- 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 output parameters are the differences between target and source prosody template parameters.

Wagon toolkit [14], with full CART function, is used in our work. Similar to the GMM-based method in training procedure, source and target F_0 contours from parallel corpus are aligned according to syllable boundaries, and then prosodic parameters are extracted from each syllable's F_0 contour. Finally, the mapping functions of parameters are estimated using the CART regression. For conversion, the prosodic parameters estimated from source F_0 contours are transformed by the mapping functions obtained in the training procedure and then the converted prosodic parameters are used to generate new F_0 contours associated with the target characteristics.

In the adaptation framework, both methods are adopted. At the last, the transformation with highest precision is adopted to apply in the corresponding large source corpus. By that way, a new template library is constructed, which has not only target speaker's prosodic personality but also reasonable coverage of different context information.

3.2. The Adaptation of Template Parameter Prediction Trees

In Sect. 2, we have built seven template parameter prediction trees in the target cost and concatenation cost definition. If the training corpus is not large enough, the

Template parameter: F0 start

Fig. 6 An incomplete tree and a well-performed tree.

constructed trees will be incomplete, and the prediction results will be unnatural and imprecise. Figure 6 shows a comparison between an incomplete prediction tree and a well-constructed prediction tree. The task of that section is not only to make the adapted prediction tree have target speaker's personality, but also make the incomplete tree be precise enough.

In the paper, we define the tree trained from original source corpus as the source tree, and define the tree with the target corpus as the target tree. To makeup the small target tree, the source tree and the target tree should be aligned first. Normally, the source tree has much more nodes than the target tree. In the above example, there are only three nodes in the small target tree but six nodes in the large source tree. Besides, different nodes' input features can not be exactly matched directly. To resolve those problems, the following strategy is adopted:

First, a group of input vectors which can almost cover all context information are created. In the inputs of prediction trees, there are both discrete features and continual features. For every discrete input feature with N possible values, the sample space is divided into N parts. For instance, for the input feature "tone," the sample space is divided into 5 parts, each of which corresponds to one specific tone in Mandarin. For the continual input feature, such as the previous syllable's ending F_0 value, first we divided them into N parts, and then treated them as discrete input features too. The value of N is set based on experience: for F_0 values, N is set as 10 and for F_0 derivative values, N is set as 5. That setting can not only make a good coverage of the sample space, but also make sure the number of input vectors is not too huge. By that way, a group of input vectors which cover all the sample space is created, which is used for the alignment of source tree and target tree.

Then, all the input features are put into both the large source tree and the small target tree, and two kinds of output value are obtained. These two kinds of output are the aligned result, as Fig. 7 shows. Based on that, the transformation relationship between the source tree and the target tree can be obtained.

Two kinds of techniques are used, one is simple linear regression and another is GMM-based transformation. Just like the template transformation, the transformation rela-

Fig. 7 The alignment of target tree and source tree.

tionship which has the highest precision is adopted to apply the corresponding large source tree. That procedure can be considered as a re-training procedure for the target corpus. The source tree's structure is kept while the predictive values of leaf nodes have been adapted. By that way, the new generated target tree not only has the target speaker's personality, but also the same structure of the large source tree, whose result is precise enough. Based on the new trees and the new template library, the TTS system can synthesize speech whose prosody style sounds like the target speaker.

4. EXPERIMENT

To evaluate the prosody adaptation method presented in the paper, we prepared four large source corpora and two small target corpora for the testing procedure. Each of the source corpora contains 6,000 phonetically balanced Mandarin sentences, and among the two target corpora, one contains 250 sentences (200 for training and 50 for testing) and another contains 550 sentences (500 for training and 50 for testing). The transcript of the target corpus is that of the source corpus's basic part. All six speakers are female.

4.1. The Experiment of Template Library Adaptation

To evaluate the performance of the template library adaptation, the following criterion is designed:

$$c = \sum_{n=0}^N \sum_{i=0}^7 \frac{|P_{ni}^{\text{target}} - P_{ni}^{\text{transformed}}|}{P_{ni}^{\text{target}}} \times \frac{1}{7 \times N} \quad (14)$$

Where, N is the number of prosody templates. P_{ni} represents the value of template parameter, as introduced in Sect. 2. $P_{ni}^{\text{transformed}}$ represents the value transformed from source template parameter and P_{ni}^{target} represents the original value of target template parameter. The P_{ni}^{target} could be F_0 values or F_0 derivative values. Therefore, P_{ni}^{target} is put as a denominator for evaluate the method as a whole. There are two methods involved in the template library adaptation. Table 3 lists the adaptation precision comparison of those two methods in different speaker combinations. For the

Table 3 The adaptation precision comparison between GMM and CART.

		Source 1	Source 2	Source 3	Source 4
Target 1	GMM	15.1%	18.2%	12.6%	19.7%
	CART	14.8%	15.6%	14.5%	20.1%
Target 2	GMM	11.1%	13.3%	12.1%	15.6%
	CART	9.9%	10.7%	10.4%	12.5%

CART method, the node stops splitting when the sample number is lower than 25. For the GMM method, the number of Gaussian model is set as 32.

From this table, we can see that when the size of target corpus grows, the performance gets better: the performance of target 2 is much better than that of the target 1. In the transformation for target 1 which contains 200 sentences, we can't see that one method performs better than another. However, in the transformation for target 2 which contains 550 sentences, the CART-based method is better than the GMM-based method. As we know, GMM and CART have their own advantages and disadvantages. The GMM method is applied purely on prosody features while the CART method take the context information into accounted. When the target corpus can not cover enough context information, the GMM method performs better. The CART method can obtain better results with a relatively large training corpus.

4.2. The Adaptation of Prediction Trees

The adapted target in that task is the corresponding value in the leaf node. So the criterion used in the adaptation of template library can be still useful. As introduced in Sect. 3.2, there are two methods involved: GMM-based method (32 mixtures are used) and linear regression-based method. Linear regression is a form of regression analysis in which observational data are modeled by a straight line. Table 4 lists the transformation result. From this table, we can see that the GMM-based method performs better than the linear regression-based method in most situations. Still, the target 2 with 500

Table 4 The adaptation precision comparison between GMM and LR.

		Source 1	Source 2	Source 3	Source 4
Target 1	GMM	12.5%	16.3%	11.1%	16.7%
	LR	17.8%	18.6%	17.5%	16.1%
Target 2	GMM	11.4%	10.3%	10%	12.1%
	LR	17.1%	15.2%	13.6%	15.4%

sentences performs much better than the target 1 with 250 sentences.

4.3. Overall Evaluations

To evaluate the prosody adaptation framework as a whole, the prosody adaptation method presented in this paper is compared with conventional prosody conversion methods, in which the relation between parallel source speech and target speech is directly obtained based on GMM or CART. An ABX test, which is commonly used for performance evaluation of voice conversion methods, is involved. In an ABX test, the listener is required to judge whether the unknown transformed speech sample X sounds closer to the reference sample A or B. If the transformed speech sounds like the target speaker, the score is 1, and if that speech sounds like source speaker, the score is 0. Both the GMM-based conversion and CART-based conversion are involved. Besides, another objective criterion is involved, which is the correlation between the original target pitch contour and the one which is generated by adaptation or conversion method.

For removing the influence of spectral characteristics, the output prosody is integrated with a specific speaker's spectrum and duration structure generated by a HTS system [15,16]. Worth particularly mentioned is that, HTS is a well known prior adaptation method with a fairly small corpus. However, our method is specifically designed for concatenation-based TTS because concatenation-based method is still very popular in the commercial TTS system.

The HTS system's speaker is a female speaker which is beyond both source speakers and target speakers. By that technique, the evaluation will focus on the difference of pitch contours. If there is more than one source corpus, multiple transformation relationships could be obtained. Among these, the one which performs best based on the criterion represented by Eq. (14) is adopted to be applied in the corresponding source speaker's prosody model. By that technique, more precise result could be obtained. 15 persons, all of them under-graduate university students, are asked to provide a rating for the 50 test sentences. Table 5 shows the comparison result. From this table, we

Table 5 The ABX and correlation comparison between the prosody adaptation system and the conventional prosody conversion system.

		ABX	Correlation
Target 1	Adaptation	41/50	0.74
	GMM-based conversion	31/50	0.66
	CART-based conversion	32/50	0.61
Target 2	Adaptation	43/50	0.78
	GMM-based conversion	35/50	0.65
	CART-based conversion	37/50	0.70

can see that by the inclusion of the prosody adaptation framework, the generated speech sounds much more like the target speaker.

5. APPLICATION OF PROSODY ADAPTATION

Because of the better performance compared with conventional prosody conversion and the integration of TTS system, the prosody adaptation framework has many applications in the daily life. In this section, two general and potential applications are introduced. Among these, one is about reading mobile short message in specific voice, and another is used in the online talking.

By now, the speech synthesis technology has been widely used in some mobiles, so the mobile can read short message and the caller's name. However, under current technology, the mobile can only utter as one or several specific voices, while users may expect the mobile can read it using their own voice, or the message sender's voices. The speaker's personality is conveyed by both spectral and prosody features. The prosody adaptation framework presented in the paper can be used to simulate the target speaker's prosody features. Based on that framework, the user only needs to speak several hundred specific sentences and the system will construct a new template library and new prediction trees. By the integration with the TTS engine and voice conversion technique, the mobile can utter as the user's voice. With the integration of the voice conversion technology, the output voice sounds like the target speaker not only in the spectral aspect, but also in the speaking style.

Another application is in the online talking, such as MSN, GTalk and so on. In this application, the TTS engine and the prosody adaptation framework can utter the talking content using kinds of voices. Besides, some technology such as voice transformation can make the talking process much funnier.

6. CONCLUSIONS

The paper constructs a prosody adaptation system, which is an effective integration of conventional prosody conversion and TTS system. That system can adapt the prosody model of TTS to a new style with a small training corpus. The adapted prosody model not only has the target speaker's personality, but also the reasonable coverage of different context information. Besides, with the help of TTS engine, some text information such as pronunciation and prosody structure can be obtained. Experiments prove that the prosody adaptation system performs much better than the conventional prosody conversion system, so it has a much wider applications.

However, there are still several disadvantages in the prosody adaptation framework. First of all, in the prosody

adaptation framework, we only simulate the target speaker's characteristic in the F_0 contour, but assume that the underlying prosody structure of source and target speaker is identical. Actually, different speaker tends to use different underlying rhythm organization, but there are still some useful rules. As presented in [17], given prosody chunks which are the basic units in the rhythm organization of Chinese utterances, multiple break solutions are acceptable. Future work will focus on how to adapt the rhythm organization of different speakers.

Another question in the presented prosody adaptation framework is the number of source target speakers. In the experiment, there are only four female source corpora involved. The performance may not be very good if the target corpus is male. In the future work, more corpora including some male corpora will be included. As we can imagine, the more the source corpora, the better the adaptation performance would be. However, constructing a source corpus which can be used as TTS system is a very time-consuming work, so future work will find a tradeoff between the source corpus size and the adaptation performance.

ACKNOWLEDGEMENTS

The work was supported by the National Natural Science Foundation of China (No. 60575032, No. 70611120555), 863 Programs (No. 2006AA01Z138, No. 2006AA01Z194) and Nokia funding.

REFERENCES

- [1] Y. Stylianou, O. Cappe and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech Audio Process.*, **6**, 131–142 (1998).
- [2] M. Abe, S. Nakamura, K. Shikano and H. Kuwabara, "Voice conversion through vector quantization," *Proc. ICASSP 88*, Vol. 1, pp. 655–658 (1988).
- [3] Z. X. Wang, R. H. Wang, Z. W. Shuang and Z. H. Ling, "A novel voice conversion system based on codebook mapping with phoneme-tied weighting," *Proc. ICSLP 2004*, Jeju Island, Korea, pp. 1197–1200 (2004).
- [4] Y. Kang, Z. Shuang, J. Tao, W. Zhang and B. Xu, "A hybrid GMM and codebook mapping method for spectral conversion," *Proc. 1st Int. Conf. Affective Computing & Intelligent Interaction*, Beijing, China, No. 3784, pp. 303–310 (2005).
- [5] Y. Chen, M. Chu, E. Chang, J. Liu and R. Liu, "Voice conversion with smoothed GMM and MAP adaptation," *Proc. Eurospeech 2003*, Geneva, pp. 2413–2416 (2003).
- [6] T. Ceyssens, W. Verhelst and P. Wambacq, "On the construction of a pitch conversion system," *Proc. EUSIPCO 2002*, Toulouse, France, Vol. I, pp. 423–426 (2002).
- [7] D. Chappell and J. Hansen, "Speaker-specific pitch contour modeling and modification," *Proc. ICASSP 98*, USA, Vol. 2, pp. 885–888 (1998).
- [8] P. Taylor, "Analysis and synthesis of intonation using the tilt model," *J. Acoust. Soc. Am.*, **107**, 1697–1714 (2000).
- [9] J. Yu, W. Zhang and J. Tao, "A new pitch generation model based on internal dependence of pitch contour for Mandarin TTS system," *Proc. ICASSP 2006*, Toulouse, France (2006).

- [10] J. Yu, J. Tao and X. Wang, "Pitch prediction for Mandarin TTS with mutual prosodic constraint," *Proc. ISCSLP 2006*, Singapore (2006).
- [11] Y. Xu and Q. E. Wang, "Pitch targets and their realization: Evidence from Mandarin Chinese," *Speech Commun.*, **33**, 319–337 (2001).
- [12] <http://www.ics.forth.gr/~lourakis/levmar/>
- [13] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," *Proc. ICASSP 98*, pp. 285–288 (1998).
- [14] <http://www.cstr.ed.ac.uk/projects/festival/>
- [15] J. Yu, M. Zhang, J. Tao and X. Wang, "A novel HMM-based TTS system using both continuous HMMs and discrete HMMs," *Proc. ICASSP 2007*, Honolulu, Hawaii, USA, Vol. 4, pp. IV-709–IV-712 (2007).
- [16] H. Zen and T. Toda, "An overview of Nitech HMM-based speech system for Blizzard Challenge 2005," *Proc. Inter-Speech 2005*, Lisbon, Portugal, pp. 93–96 (2005).
- [17] M. Chu, H. Dong and J. Tao, "A perceptual study on variability in break allocation with Chinese sentences," *Speech Prosody 2006*, Germany (2006).