Prosody Conversion From Neutral Speech to Emotional Speech

Jianhua Tao, *Member, IEEE*, Yongguo Kang, and Aijun Li

***Abstract*—Emotion is an important element in expressive speech synthesis. Unlike traditional discrete emotion simulations, this paper attempts to synthesize emotional speech by using "strong," "medium," and "weak" classifications. This paper tests different models, a linear modification model (LMM), a Gaussian mixture model (GMM), and a classification and regression tree (CART) model. The linear modification model makes direct modification of sentence F0 contours and syllabic durations from acoustic distributions of emotional speech, such as, F0 topline, F0 baseline, durations, and intensities. Further analysis shows that emotional speech is also related to stress and linguistic information. Unlike the linear modification method, the GMM and CART models try to map the subtle prosody distributions between neutral and emotional speech. While the GMM just uses the features, the CART model integrates linguistic features into the mapping. A pitch target model which is optimized to describe Mandarin F0 contours is also introduced. For all conversion methods, a deviation of perceived expressiveness (DPE) measure is created to evaluate the expressiveness of the output speech. The results show that the LMM gives the worst results among the three methods. The GMM method is more suitable for a small training set, while the CART method gives the better emotional speech output if trained with a large context-balanced corpus. The methods discussed in this paper indicate ways to generate emotional speech in speech synthesis. The objective and subjective evaluation processes are also analyzed. These results support the use of a neutral semantic content text in databases for emotional speech synthesis.**

***Index Terms*—Emotional speech, prosody analysis, speech synthesis.**

I. Introduction

RECENTLY, more and more efforts have been made in the research for expressive speech synthesis, among which emotion is a very important element [18], [19]. Some prosody features, such as pitch variables (F0 level, range, contour, and jitter), and speaking rate have already been analyzed [22], [23]. There are also some implementations in emotional speech synthesis. For instance, Mozziconacci [7] added emotion control parameters on the basis of tune methods, resulting in higher performance. Cahn [8], by means of a visual acoustic parameters editor, achieved the output of emotional speech via direct manual manipulation. Recently, some efforts have been made using a large corpus. A typical system was produced by Campbell [9], who created an expressive speech synthesis from a corpus gathered over five years and gave impressive synthesis results. Schroeder [10] and Eide [11] generated an expressive text-to-speech (TTS) engine which can be directed, via an extended speech synthesis markup language, to use a variety of expressive styles from about 10 h of "neutral" sentences. Furthermore, rules translating certain expressive elements to ToBI markup have been manually derived. Chuang [12] and Tao [4] used emotional keywords and emotion trigger words to generate an emotional TTS system. The final emotion state is determined based on the emotion outputs from text-content module.

The previous work was mostly focused on the simulation of discrete basic emotions. The rules and unit selection methods formed the basic previous research. Actually, such discrete emotional expression is just a simplification which we seldom encounter in normal life. On the other hand, emotional states can be thought of as zones along an emotional vector [2]. The vector might be a cline of emotions shading into one another, with extremes at either end. In the vector approach, expression would be a reflection of the intensity in a particular zone. With this idea, unlike the traditional methods, we allow the labeler to label the emotional training and testing corpus with four degrees, "strong," "normal," "weak," and "unlike" among emotions—"happiness," "sadness," "fear," and "anger." So-called "neutral" speech is used as the reference source for conversion, and was not labeled with degrees.

With this method, this paper tests different prosody conversion methods which aim at the transformation of the prosodic parameters, e.g., F0, duration, and intensity of the given utterance, to generate emotional speech. A linear modification model (LMM), a Gaussian mixture model (GMM) method and a classification and regression tree (CART) method were tried. The LMM makes direct modification of F0 contours (F0 top, F0 bottom, and F0 mean), syllabic durations, and intensities from the acoustic distribution analysis results. Twelve patterns (four emotions with three degrees, "strong," "medium," and "weak") were deduced from the training set of the corpus. In order to evaluate the expressiveness of the emotional speech output, this paper introduces a perception correlation calculation method. The evaluation results show that the LMM method does not produce very good output. Further analysis shows that the expression of emotion does not just influence these general prosody features, but also affects the sentence-stress and more subtle prosodic features. The GMM and CART models are then applied to solve the problem. The GMM method attempts to map the prosody features distribution from a "neutral" state to the various emotions, while the CART model links linguistic features to the prosody conversion.

Manuscript received November 20, 2005; revised March 27, 2006. This work was supported in part by the National Natural Science Foundation of China under Grant (60575032). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Thierry Dutoit.

J. Tao and Y. Kang are with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100080, China (e-mail: jhtao@nlpr.ia.ac.cn; ygkang@nlpr.ia.ac.cn).

A. Li is with the Institute of Linguistics, Chinese Academy of Social Sciences, Beijing 100732, China (e-mail: liaj@cass.org.cn).

Digital Object Identifier 10.1109/TASL.2006.876113

Unlike LMM, the GMM and CART models cannot directly use F0 contours, so a pitch target model has been introduced. The model is based on the assumption that "observed F0 contours are not linguistic units *per se*. Rather, they are the surface realizations of linguistically functional units such as tones or pitch accents." [5] In the model, variations in surface F0 contours result not only from the underlying pitch units but also from the articulatory constraints that determine how these units can be implemented. It is, therefore, extremely suitable for F0 pattern training and simulation in Mandarin speech.

The final analysis results show that the direct LMM method gives us the worst results among the three methods. The GMM method is more suitable for a small training set, while the CART method provides the best emotional speech output if it is trained with a large-coverage corpus. Though the conversion method has been widely used for the research on voice conversion, the methods discussed in this paper provide a new way to produce emotional speech synthesis; however, there is still lots of work to be done for real expressive speech synthesis.

This paper is composed of seven major parts. Section II introduces the corpus with emotion labeling. The acoustic features of the emotions were also analyzed. In Section III, this paper describes a linear modification model which uses prosody patterns from the acoustic mapping results directly. A perception correlation calculation method is also introduced for the evaluation of expressiveness in the synthesized emotions. Further analysis on emotion and stress reveals that emotions are closely related to subtle prosody distributions. Section IV describes the GMM and CART models which are used to convert the prosody features from "neutral" to emotional speech. The pitch target model is also introduced in this section. Some analysis on methods comparing context influences for emotional speech is provided for the model analysis. In Section V, this paper provides more discussion on comparison of the three methods and other acoustic factors which might influence emotional prosody conversion. Section VI provides a conclusion for this paper.

II. Corpus and Analysis

A. Corpus Preparation

To create the model, a corpus which contains 1500 sentences was produced by searching the appropriate data from 10 years' Reader's Digests via a Greedy Algorithm [3]. The following factors form the focus of our investigation:

1) identity of the current syllable;
2) identity of the current tone;
3) identity of the final in the previous syllable;
4) identity of the previous tone;
5) identity of the initial in the following syllable;
6) identity of the following tone;
7) number of preceding syllables in the word;
8) number of following syllables in the word;
9) number of preceding syllables in the phrase;
10) number of following syllables in the phrase;
11) number of preceding syllables in the utterance;
12) number of following syllables in the utterance.

Factor 1) has 417 values that correspond to the 417 syllable types in Chinese. Factors 2), 4), and 6) have each five values that correspond to the four full tones and the neutral tone (0). Factor 3) contains 20 values (initial types) and factor 5) contains 41 values (final types). Factors 7)–10) have three values each, 0, 1, and 2, where 0 means that the segment lies at the boundary, 1 means that it is one syllable away, and 2 means that it is 2 or more syllables away from the boundary. Factors 11) and 12) have two values each, 0 and 1, where 0 means that the segment lies at the boundary and 1 means that it is 1 syllable or more away from the boundary.

During the text selection phase, phrasing was coded solely on the basis of punctuation. After the text was selected and the database recorded, phrasing was recoded to correspond to pauses. Each utterance in our database contains at least two phrases. There were 3649 phrases and 14 453 syllables in total, so on average each utterance contained two phrases.

After the corpus was designed, each sentence was recorded in five emotions, "neutral," "happiness," "sadness," "fear," "anger," by a professional actress in a professional recording studio with a large membrane microphone. A laryngograph signal was also recorded in parallel to obtain accurate pitch information. Two persons assisted in the recording. One accompanied the speaker to provide hints on how to speak. One was outside the recording space for technical control. The speaker was asked to simulate the emotions based on her own experience. The accompanying person made the final judgment. Recording was not stopped until satisfactory results were obtained. After the recording, all the utterances were segmentally and prosodically annotated with break index and stress index information [14]. The F0 values were also processed and manually checked.

B. Labeling and Analysis

We presented the sentences in a randomized order to a group of 15 subjects, graduate students of engineering who were asked to participate in the experiment. Each sentence was played back to the subjects two times with a 3-s interval. The subjects were asked to annotate the perceived emotion with four levels, "strong" (degree 3), "medium" (degree 2), "weak" (degree 1), and "unlike" (degree 0).

Different listeners perceived different aspects of emotion realized by the speech and it was difficult to reach a consensus. One useful representation of the labeling results is the mean emotion degree over all subjects. The mean degrees were rounded off into integer values, and thus corresponded to the four degrees, i.e., "strong," "medium," "weak," and "unlike."

Out of the 1000 sentences in the corpus, 700 were used for analysis or training and the remaining 300 were used for testing. Table I shows the means and standard deviations of prosody parameters of the training sentences at different emotion levels.

In the table, Δ indicates the standard deviation, the other values indicate the means of F0 mean $(F0_{\mathrm{mean}})$, F0 topline $(F0_{\mathrm{top}})$, F0 baseline $(F0_{\mathrm{bottom}})$, syllabic duration (D_{syllable}), and intensity (E). The table partly confirms the previous research [18] that "happiness" and "anger" yield a high F0, while "sadness" generates lower F0 than "neutral," and "fear" is quite close to "sadness." The overlap of F0 mean and F0 topline in different emotions is less than that of F0 baseline. It seems that the F0 mean and topline provide better "resolving power" for perception than the F0 baseline.

TABLE I
DISTRIBUTION OF PROSODIC PARAMETERS IN DIFFERENT EMOTIONS

		Fear	Sadness	Anger	Happiness
$F0_{mean}$ (Hz)	Strong	172.2	160.7	195.1	193.4
	Medium	169.1	160.9	188.1	187.1
	Weak	160.5	161.9	179.5	178.5
$\Delta F0_{mean}$ (Hz)	Strong	44.6	43.0	51.6	54.1
	Medium	38.3	44.5	44.8	50.3
	Weak	37.9	48.2	43.3	45.2
$F0_{bottom}$ (Hz)	Strong	136.1	134.4	156.5	154.8
	Medium	136.5	136.3	149.6	148.6
	Weak	140.0	139.0	143.8	145.4
$\Delta F0_{bottom}$ (Hz)	Strong	32.3	37.7	48.7	49.3
	Medium	27.8	38.7	42.3	48.0
	Weak	31.5	42.0	41.8	43.6
$F0_{top}$ (Hz)	Strong	195.8	182.6	220.0	218.5
	Medium	194.2	181.1	213.0	213.5
	Weak	177.9	178.6	202.5	202.2
$\Delta F0_{top}$ (Hz)	Strong	70.4	62.9	63.5	67.4
	Medium	54.3	59.7	60.1	59.1
	Weak	52.6	59.2	55.7	58.4
$D_{syllable}$ (ms)	Strong	217.3	237.9	178.8	194.6
	Medium	206.8	225.5	179.9	188.6
	Weak	200.4	211.0	182.3	187.2
$\Delta D_{syllable}$ (ms)	Strong	65.2	80.2	57.2	61.3
	Medium	63.1	70.5	54.8	60.1
	Weak	60.5	60.8	53.0	58.5
E (DB)	Strong	84.5	81.2	88.4	85.5
	Medium	83.1	81.7	84.0	82.2
	Weak	81.5	84.1	80.6	81.0
ΔE (DB)	Strong	10.0	9.2	10.4	10.4
	Medium	8.5	8.5	9.9	8.4
	Weak	8.3	8.4	8.8	8.2

It is a complicated task to convert "neutral" speech into emotional speech because the emotional speech differs from the "neutral" speech in various aspects, including intonation, speaking rate and intensities, etc. From the very small standard deviation of mean syllabic duration and intensity, we find speaking rate and intensity to be well distributed in the different emotions. The method of linear ratio modification was used for these parameters. The subsequent discussion will be focused on the conversion of F0 contours among the emotions.

It has also been pointed out that F0 jitter is an important parameter for emotional speech [13]. For F0 jitter, normally, a quadratic curve is fitted to the acoustic measurement with a moving window covering five successive F0 values. The curve is then subtracted from the acoustic measurements. F0 Jitter was calculated as the mean pitch period-to-period variation in the residual F0 values. Table II shows the results from "strong" emotions.

With the results, we can see that "happiness" has the highest F0 jitter while "sadness" contains the minimum F0 jitter distribution. During speech synthesis, F0 jitter is realized by a random variation in the length of the pitch periods with an amplitude in accordance to the parameters value. This random variation is controlled by a white noise signal filtered by a one pole lowpass filter. The F0 jitter is applied in all of the following conversion methods and will not be specifically mentioned again.

TABLE II
AVERAGE RESULTS OF F0 JITTER OF "STRONG" EMOTIONS

Emotions	fear	Sadness	anger	happiness
F0 jitter (HZ)	6.5	6.1	7.5	9.1

TABLE III
TRANSFORM SCALES OF PROSODIC PARAMETERS
FROM "NEUTRAL" TO "STRONG" EMOTIONS

Parameters	fear	sadness	anger	happiness
$F0_{top}$	-12.5%	-15.8%	+32.6%	+35.6%
$F0_{bottom}$	-0.3%	-8.1%	+17.7%	+24.5%
$F0_{mean}$	-12.4%	-13.9%	+23.3%	+37.2%
$D_{syllable}$	-3.1%	+4.0%	-9.4%	-2.4%
E	-5.1%	-6.8%	+2.0%	+3.3%

III. LINEAR MODIFICATION MODEL

A. Linear Modification

Among all prosody conversion methods, linear modification (LMM) seems to be the most intuitive. We select prosody modification patterns directly from the prosody features distribution among emotions

$$y_{n,i} = \alpha_{n,i} \cdot x.$$

Here, x indicates the input prosodic parameters: F0 topline, F0 baseline, F0 mean, syllabic duration and intensity. y denotes their outputs among different emotions. α is the transform scale of the parallel prosodic parameters between "neutral" and emotions as calculated from the training set of the corpus. n denotes the emotional state, i.e., "fear," "sadness," "anger," and "happiness," i indexes the emotion level, i.e., "strong," "medium," and "weak." Table III shows the transform scales for the simulation of "strong" emotions. "$+$" means "increasing by" and "-" means "decreasing by" with respect to the parameters of the "neutral" state. A group of transform scales form a transform pattern of emotion simulation. There are 12 patterns four emotions with three degrees, "strong," "medium," and "weak") in total.

B. Deviation of Perceived Expressiveness (DPE)

To evaluate the conversion method, 300 sentences from the test set of the corpus are used. All of the transform patterns are applied to convert the "neutral" speech into emotional speech via a synthesizer which used STRAIGHT [15] as the acoustic processing model. The results are compared to the corresponding natural recording.

Traditionally, an ABX test is commonly used for performance evaluation of voice conversion methods [16]. In an ABX test, the listener is required to judge whether the unknown speech sample X sounds closer to the reference sample A or B. For the evaluation of expressiveness in emotional speech, an ABX test is not easy to use because an emotion state cannot be easily defined, especially with the distinction among "strong," "medium," and "weak" degrees. The forced one-to-one match is also unsuitable.

TABLE IV
EXPRESSIVENESS DEVIATION BASED ON LMM METHOD

Degrees	Emotions	$\bar{d}_{n,i}$	$\bar{d}'_{n,i}$	$\Delta_{n,i}$	$\bar{\Delta}_i$
Strong	fear	1.89	2.81	-0.92	-0.76
	sadness	2.10	2.69	-0.59	
	anger	1.77	2.91	-1.14	
	happiness	2.51	2.89	-0.38	
Normal	fear	1.23	1.87	-0.64	-0.75
	sadness	1.41	1.99	-0.58	
	anger	1.02	2.11	-1.09	
	happiness	1.31	2.01	-0.70	
Weak	fear	0.25	1.01	-0.76	-0.60
	sadness	0.78	1.11	-0.33	
	anger	0.51	1.23	-0.72	
	happiness	0.43	1.02	-0.59	

To evaluate emotional simulation results, we, therefore, proposed a deviation of perceived expressiveness (DPE) method.

The DPE experiment involves the same group of 15 subjects who took part in the emotion labeling process. The subjects are asked to annotate 3600 synthesized utterances (300 test sentences with four emotions at three levels) by using the same method as described in Section II.

The error rate of a certain emotion simulation is measured by

$$\Delta_{n,i} = \sum_{i=0}^{I} (\bar{d}_{n,i} - \bar{d}'_{n,i}) \tag{1}$$

where, n denotes the emotion state, i.e., "fear," "sadness," "anger," and "happiness," i indexes the emotion level, i.e., "strong," "medium," and "weak." $\bar{d}_{n,i}$ represents the mean level labeled for the synthesized emotion, $\bar{d}'_{n,i}$ is the mean level labeled for the original speech. Table IV shows the results from DPE test.

Here, $\bar{\Delta}_i$ denotes the mean errors of emotion level i. From the tables, it is noted that "strong" emotions can hardly be synthesized with the LMM method, except for "happiness." Most of the synthesized emotions were perceived to be less expressive than the original natural ones. Some of them were even perceived as the "neutral," such as "weak fear," "weak happiness," though the prosody parameters have been modified substantially. This probably indicates that simple modification of prosodic parameters with a group of constant transform scales is not adequate to reflect the effect of emotion. Many detailed prosody features inside the utterance might have been lost.

C. Emotion and Stress

Previous research has found that there exists a strong relationship between emotions and stresses [32]. Stress refers to the most prominent element perceived in an utterance rather than literal "semantic focus" which is used to express speaker's attitudes.

In our corpus, most of the intonations of "neutral" utterances have a decreasing tendency from the start to the end. The sentence stresses normally appear at the beginning. To understand more about various locations of stresses among emotions, we carried out additional perceptual experiments on the corpus.

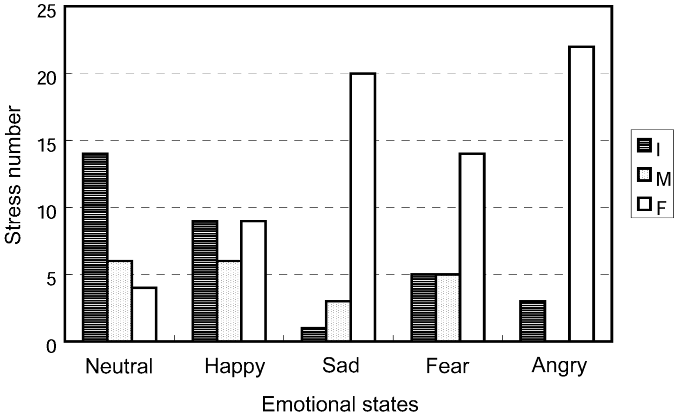

Fig. 1. Perceived stress pattern.

Three subjects were asked to annotate sentence stress on the most prominent syllable (or prosodic word) after they listened to the utterances played in random order. The results were rated by checking the stressed words. Three points was given to a syllable (or word) if it is perceived by all of the three listeners to bear the sentence stress; two points was given when only two listeners had that consensus. Zero points means that none of the listeners perceived the same stress. For example, in one of the "fear" utterances, the word "yin2 hang2" (bank) was perceived as having sentence stress by listeners 1 and 3, but listener 2 assigned the most prominent stress to another word, so the score of this "fear" utterance was evaluated as two points. Based on the perceptual results, the sentence stress is assigned to a prosodic word which gets two or three points.

Fig. 1 shows the perceived stress patterns among five "strong" emotions. "I" means the sentence stress is located in the first word of the utterance, while "F" means the final word, and "M" means any middle words. We found that the stress is shifting in the sentence among different emotions. In our corpus, the shifting amplitude is pertinent to the emotional states, from big to small: "anger" > "sadness" > "fear" > "happiness."

Though stresses might be shifted among emotions, the stress distribution varies with different content, different situation and different person. Sometimes, the stress in "happiness" keeps the stress pattern as in the "neutral" states. Also, when a sentence stress is sentence-medial, the corresponding stress in other emotions might be shifted to the sentence-final words. This transform is so complicated that the LMM is unable to model it.

IV. PROSODY CONVERSION METHOD

To solve the problem mentioned previously, some other mapping methods (GMM and CART) were considered. The underlying meaning of the mapping is to establish a relation between two sets of multidimensional vectors, which correspond to the source speech and the converted speech respectively. To use a more complex conversion model, it is difficult to deal with the raw F0 contour. A suitable parametric model for describing contours is needed.

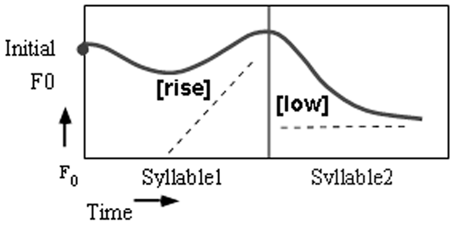

Fig. 2. Pitch target model [5].

A. F0 Model

Mandarin is a typical tonal language, in which a syllable with different tone types can represent different morphemes. There are four tone types referred to be "high," "rising," "low," and "falling" [5]. They are mainly manifested by the F0 contours. There have been numerous studies on the tones and intonation of Mandarin. Several quantitative representations have been proposed to describe continuous F0 contours, such as the Fujisaki model [6], The Soft Template Mark-Up Language (STEM-ML) model [17] and the pitch target model [5].

The Fujisaki model is a command-response model for F0 contours. It uses two types of commands: the impulse-shaped phrase commands giving rise to the phrase-level component for global intonation, and the pedestal-shaped accent commands giving rise to accent components for local undulation due to word accent. The STEM-ML proposed by Bell Labs is a tagging system, in which F0 contours are described by mark-up tags, including both stress tags for local tone shapes and step and slope tags for global phrase curves. Both models have the ability of representing F0 contours in Mandarin. Their common problem is that it is difficult to establish the relation among the model commands (or tags) of different utterances.

In the pitch target model, variations in surface F0 contours result not only from the underlying pitch units (syllables for Mandarin), but also from the articulatory constraints. Pitch targets are defined as the smallest operable units associated with linguistically functional pitch units, and these targets may be static (e.g., a register specification, [high] or [low]) or dynamic (e.g., a movement specification, [rise] or [fall]). Among these models, the features of the pitch target model are quite suitable for prosody conversion.

Fig. 2 gives a schematic illustration of hypothetical pitch targets (dashed lines) and their surface realization (solid curved line). The three vertical lines represent the boundaries of the two consecutive pitch target-carrying units. The level dashed line on the right of the figure represents a static pitch target [low]. The oblique dashed line on the left represents a dynamic pitch target [rise]. In both cases, the targets are asymptotically approximated.

The implementation rules are based on possible articulatory constraints on the production of surface F0 contours. The production of surface F0 contours is a process of continuous approximations of the targets throughout tone-carrying syllables. When the syllable boundary is reached, it starts the new approximation for the next syllable with the new pitch target.

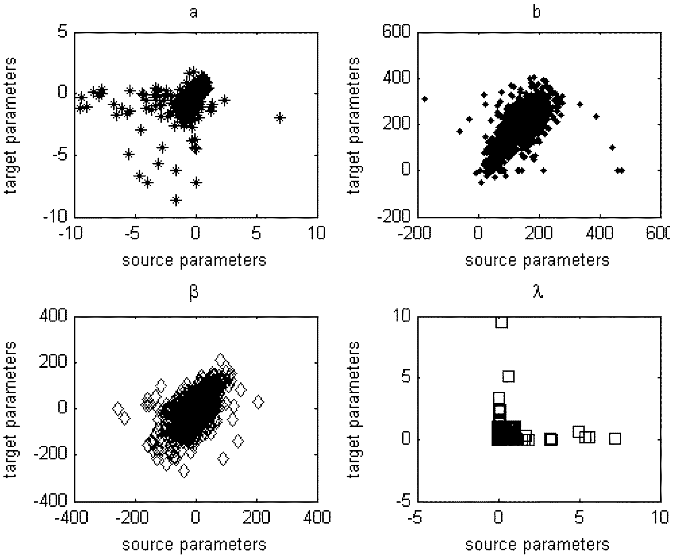

Fig. 3. Scatter plots of four pitch target parameters in "neutral" to "strong happiness" conversion.

Let the syllable boundary be $[0, D]$. The pitch target model uses the following equations [20]:

$$T(t) = at + b \tag{2}$$

$$y(t) = \beta \exp(-\lambda t) + at + b$$
$$0 \leqslant t \leqslant D, \lambda \geqslant 0 \tag{3}$$

where $T(t)$ is the underlying pitch target, and $y(t)$ is the surface F0 contour. The parameters a and b are the slope and intercept of the underlying pitch target, respectively. These two parameters describe an intended intonational goal of the speaker, which can be very different from the surface F0 contour. The coefficient β is a parameter measuring the distance between the F0 contour and the underlying pitch target at $t = 0$. λ describes how fast the underlying pitch target is approached. The greater the value of λ is, the faster the speed. A pitch target model of one syllable can be represented by a set of parameters (a, b, β, λ).

As described in [20], (a, b, β, λ) can be estimated by nonlinear regression process with expected-value parameters at initial and middle points of each syllable's F0 contour. The Levenberg–Marquardt algorithm [20] is used for estimation as a nonlinear regression process.

Fig. 3 shows the scatter plots of the mapping from "neutral" to "strong happiness" for the four model parameters. The correlation coefficient of (a, b, β, λ) between "neutral" to "strong happiness" is (0.3592, 0.5282, 0.3564, 0.0676). It can be observed that a, b, and β exhibit more correlation between source speech and target speech than λ.

B. GMM-Based Prosody Conversion

GMMs have proved to be very useful in the research of voice conversion [21]. They work particularly well in spectrum smoothing, which assumes the probability distribution of the observed parameters to take the following form:

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

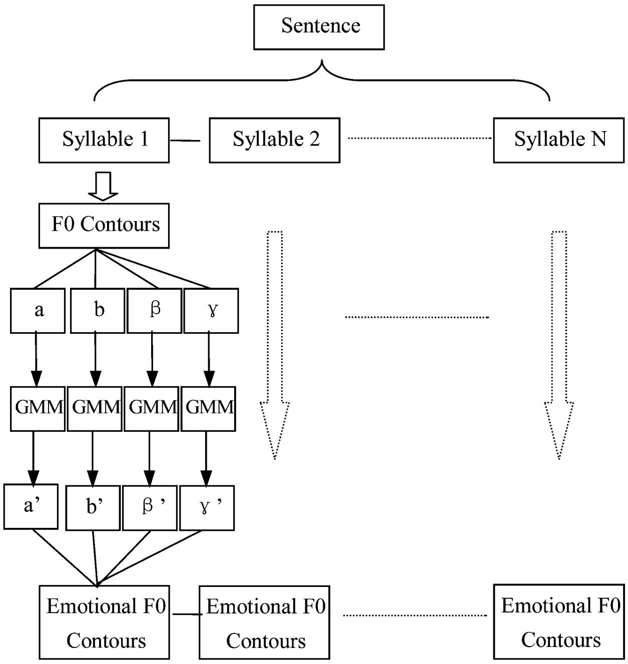

Fig. 4. Framework of GMM-based emotional prosody conversion.

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

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

The parameters (α, μ, Σ) are estimated with the expectation-maximization (EM) algorithm [24].

The conversion function can be found using regression

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

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

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

Here

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

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

The parameters of the conversion function are determined by the joint density of source and target features [5]. In Kain's work [16], it was shown 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 [16]. For each pitch target parameter a, b, β, and λ, 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 $Z_k = [X_k \quad Y_k]^T, k = 1, \ldots, N$ is used to estimate the GMM parameters.

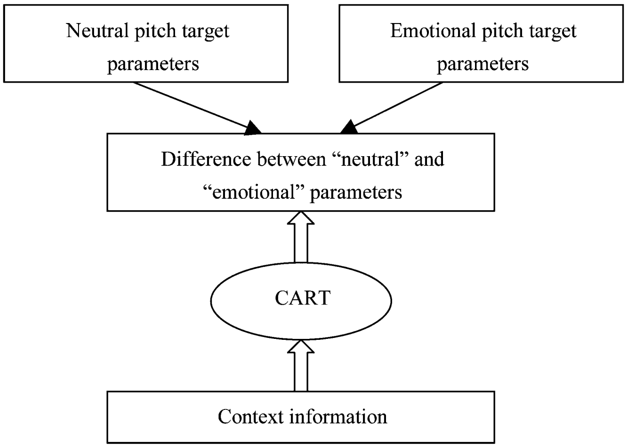
Neutral pitch target parameters
Emotional pitch target parameters
Difference between "neutral" and "emotional" parameters
CART
Context information

Fig. 5. Framework of CART-based emotional prosody conversion.

Fig. 4 shows the framework of the GMM based prosody conversion.

A GMM was trained for the conversion from "neutral" to each of the four emotions at three levels. This method facilitates continuous and smooth conversion without discontinuities with incremental learning, but it is a pure numerical algorithm; in other words, the GMM mapping method fails to incorporate any linguistic information.

C. *CART-Based Model*

CARTs have been successfully used in prosody prediction, such as duration, prosody phrase boundaries, etc. They efficiently integrate the contextual information into prediction [29]. In our research, the framework of the CART based prosody conversion is shown in Fig. 5.

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

- tone identity (including current, previous and following tones, with five categories);
- initial identity (including current and following syllables' initial types, with 8 categories);
- final identity (including current and previous syllables' final types, with four categories);
- position in sentence;
- part of speech (including current, previous and following words, with 30 categories).

The output parameters are the differences of pitch target parameters a, b, β, and λ between "neutral" and emotional parameters.

Wagon toolkit,[1] with full CART function was used in our work. Similar to the GMM method in training procedure, source and target pitch contours from parallel corpus are aligned according to labeled syllable boundaries, and then pitch target parameters are extracted from each syllable's pitch contour, fi-

[1] [Online] Available: http://festvox.org/docs/speech_tools-1.2.0/x3475.htm

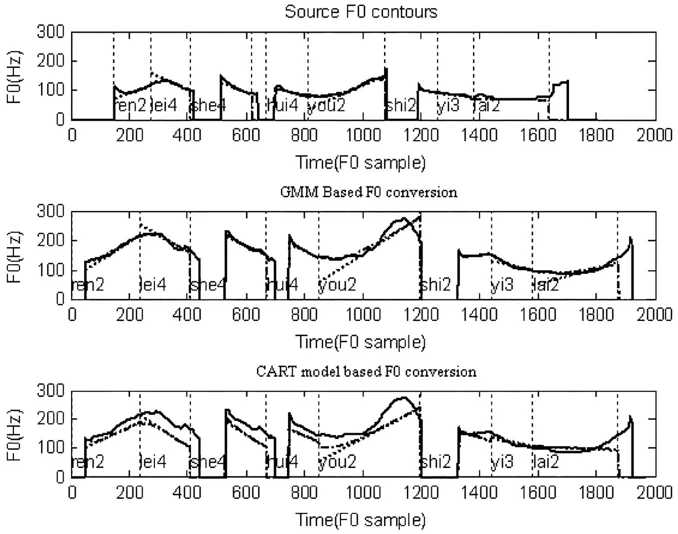

Fig. 6. Example of F0 conversion using the pitch target model in "neutral" to "strong happiness" conversion.

nally mapping functions of parameters a, b, β, and λ are estimated using the CART regression. Again, there were totally 12 CART models trained with different "neutral" and emotion mappings. For conversion, the pitch target parameters estimated from source pitch contours are transformed by the mapping functions obtained in the training procedure, and then the converted pitch target parameters generate new pitch contours associated with the target characteristics.

D. Accuracy Analysis

An example of prosody conversion is given in Fig. 6, in which the "neutral" pitch contours are converted into the "strong happiness" with the GMM method and the CART method.

To get more statistical information of conversion results, four emotional conversions (marked as "neutral—strong happiness," "neutral—strong anger," "neutral—strong sadness," and "neutral—strong fear") were conducted in the experiment. To compare the GMM and CART mapping methods, root mean square errors are shown in Table V.

TABLE V
MAPPING ERRORS OF GMM AND CART METHOD

Methods		neutral - strong happiness	neutral - strong anger	neutral-strong sadness	neutral-strong fear
GMM	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
CART	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

From the table, the performance of the GMM method has better results than that of the CART method. However, the conventional GMM-based conversion tends to generate overly smoothed prosody.

E. Combined With Spectral Conversion

Though it is commonly believed that the prosody features are very important for emotional speech classification, the modification of prosody features alone might not be sufficient to generate an expected emotion. An emotional speech utterance differs from a neutral one not only in prosodic features but also in spectral features. Parameters describing laryngeal processes on voice quality were also taken into account [25], [26]. It has been pointed out that strong feelings often literally distort the physical vocal tract [27]. 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." Spectral conversion is necessary to implement an emotional conversion especially from "neutral" to "negative" emotions. There are a lot of mapping methods available, such as codebook mapping [29], [30], linear multivariate regression (LMR) [36], neural networks [31], GMMs [33], [34], and hidden Markov models (HMMs) [35]. Among these mapping methods, codebook mapping and GMM methods are two representative and popular mapping algorithms.

In our paper, we integrate GMM and codebook mapping [37]. This method encodes the basic spectral envelope using GMM and converts spectral details using an offset codebook mapping method. By this means, the problems of smoothing and discontinuity can be counteracted. We finally use STRAIGHT to synthesize the speech.

F. Expressiveness Analysis by Combining Prosody Conversion and Spectral Conversion

With the integration of prosody conversion and spectral conversion, more results of DPE tests are shown in Tables VI and VII.

Table V–VII seem to confirm that the GMM mapping method is better than the CART method, though the DPE tests did not show much difference. The results are not consistent with the previous analysis that the prosody patterns of emotions are closely related to content and linguistic features. A possible reason might be that the training data in the experiment was not enough to cover most linguistic information when using the CART method. The GMM method is applied purely on prosody features, The CART method may obtain better result with a larger training corpus, which still needs to be confirmed in our future research.

V. DISCUSSION

Though plenty of analysis has been performed on the acoustic distributions among emotions, these emotions have not actually been clearly defined via perception. Even for the same emotion there are still various expression methods. One speaker may increase F0 jitter for "happiness," rather than increasing the overall pitch level. The locations of sentence stress in "anger" utterances can also vary according to differences in content and linguistic emphasis. In most case, it is located in the word which the speaker wants to emphasize. These various methods of emotional expression increase our difficulties in emotion simulations, since the acoustic features can be widely distributed.

TABLE VI
EXPRESSIVENESS DEVIATION BASED ON GMM-BASED PROSODY CONVERSION

Degrees	Emotions	$\bar{d}_{n,i}$	$\bar{d}'_{n,i}$	$\Delta_{n,i}$	$\bar{\Delta}_i$
Strong	fear	1.71	2.81	-1.10	-0.56
	sadness	2.53	2.69	-0.16	
	anger	2.34	2.91	-0.57	
	happiness	2.50	2.89	-0.39	
Normal	fear	1.51	1.87	-0.36	-0.29
	sadness	1.40	1.99	-0.59	
	anger	1.92	2.11	-0.19	
	happiness	1.99	2.01	-0.02	
Weak	fear	0.61	1.01	-0.40	-0.23
	sadness	0.82	1.11	-0.29	
	anger	1.02	1.23	-0.21	
	happiness	1.00	1.02	-0.02	

TABLE VII
EXPRESSIVENESS DEVIATION FROM CART-BASED PROSODY CONVERSION

Degrees	Emotions	$\bar{d}_{n,i}$	$\bar{d}'_{n,i}$	$\Delta_{n,i}$	$\bar{\Delta}_i$
Strong	fear	2.01	2.81	-0.80	-0.60
	sadness	2.23	2.69	-0.46	
	anger	2.56	2.91	-0.35	
	happiness	2.11	2.89	-0.78	
Normal	fear	1.32	1.87	-0.55	-0.43
	sadness	1.41	1.99	-0.58	
	anger	2.02	2.11	-0.09	
	happiness	1.53	2.01	-0.48	
Weak	fear	0.58	1.01	-0.43	-0.27
	sadness	0.78	1.11	-0.33	
	anger	1.12	1.23	-0.11	
	happiness	0.83	1.02	-0.19	

Context information from linguistic features is very important for emotion expression. Sometimes it is not necessary to change any prosodic parameters to express the emotion, if some functional emotional keyword was inserted into the utterance [4]. Such as, "I'm really angry about what you did" sufficiently shows the "anger" emotion by use of the functional word "angry" alone. Besides, emotion may mislead the listener into perceiving a different sentence act if the context is not provided. Sometimes, the listeners may perceive some "happy" voices with a raising end intonation as an echo question, as shown in Fig. 7. Here, the high boundary tone (with a final stressed pattern) is used to express a "happy" emotion. Since the high boundary tone is one of the major features of an interrogative sentence, it is not strange for the listener to recognize an emotional statement as a question, without context.

In this paper, we tried to integrate some of these linguistic features to predict the emotional prosody, but the results were still far from our expectations. Emotional functional words were analyzed in our previous work for emotion control [4], but were not considered in this paper because we did not collect a large-enough training set for the CART model. Otherwise, we can consider the functional emotional keywords as a kind of special part of speech.

We have combined the prosody conversion and spectral conversion for emotion simulation. The work does enhance the expressiveness of emotions, especially for "happiness." When Ladd described the relations between paralanguage and intonation [27], he described the relations between paralinguistic and linguistic features and pointed out that lexical tone was also affected by paralinguistic expression. In fact, intonational and tonal ambiguities are caused by the stress patterns in expressing certain strong emotions or attitudes. The influences on stress patterns, tones, and intonations from paralinguistic features should also be considered in future work, otherwise, the utterance will not be able express emotions or attitudes as naturally as possible.

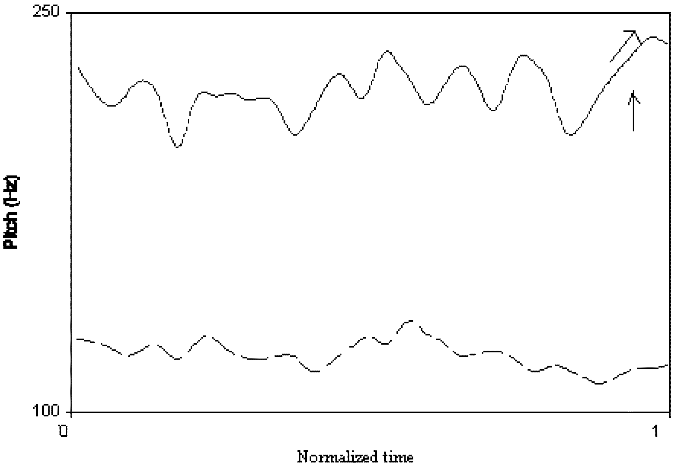

Fig. 7. F0 contours of a "happy" statement (upper curve) and its neutral counterpart (lower dashed curve). The utterance is "xi1 an1 bo1 yin1 qi1 san1 qi1 fei1 ji1." The happy statement is perceived as an echo question because both the contour and the register of the final syllable "ji1," bearing a high-level tone HH, are raised.

In this paper, we also proposed a DPE method to evaluate the conversion results. Unlike traditional ABX test which is normally used for voice conversion evaluation, but is hardly to be used for a uncertain judgment, DPE method uses different degrees for the perception, such as "strong," "medium," and "weak." The degrees give more flexible comparing among the emotions. In addition, DPE method adopted the advantage of MOS test, with the mean scores of the degrees from many listeners.

VI. CONCLUSION

This paper has described a perception experiment that was designed to make soft classification of emotional speech with different degrees of "strong," "medium," and "weak" expression. The classification results help us to achieve more subtle acoustic patterns when synthesizing emotional speech with various types of expressiveness.

When generating expressive speech synthesis, we are easily tempted to fall into the practice of using the acoustic patterns driven by the speech with emotion state with a linear modification approach. However, without a more detailed distribution of these acoustic patterns, it is hard for us to synthesize more expressive or less expressive speech. To solve this problem, this paper proposed using a GMM method and compared it with a similarly functioning CART method. Unlike the linear modification method, both the GMM and CART models efficiently map the subtle prosody distributions between neutral and emotional speech. While GMM just uses the acoustic features, the CART model allows us to integrate linguistic features into the mapping. A pitch target model which was designed to describe Mandarin F0 contours was also introduced. For all conversion methods, a DPE method was employed to evaluate the expressiveness of the resulting output speech. The results show that the linear modification model provides the worst results among the three methods. The GMM method is much more suitable for a small training set, while the CART method gives us the better emotional speech output if trained in a large context balanced corpus.

Acknowledgment

The authors would like to thank Prof. T. Lee, Dr. N. Campbell, and Dr. Y. Xu for their advice, and they would also like to thank especially the students of NLPR and CASS for their cooperation with the experiments.

References

[1] N. Campbell, "Perception of affect in speech—Toward an automatic processing of paralinguistic information in spoken conversation," in *Proc. ICSLP*, Jeju, Korea, Oct. 2004, pp. 881–884.

[2] A. Ortony, G. L. Clore, and A. Collins, *The Cognitive Structure of Emotions*. Cambridge, U.K.: Cambridge Univ. Press, 1988.

[3] J. P. H. van Santen, R. W. Sproat, J. P. Olive, and J. Hirschberg, Eds., *Progress in Speech Synthesis*. New York: Springer, 1997.

[4] J. Tao, "Emotion control of Chinese speech synthesis in natural environment," in *Proc. Eurospeech*, 2003, pp. 2349–2352.

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

[6] H. Fujisaki and K. Hirose, "Analysis of voice fundamental frequency contours for declarative sentence of Japanese," *J. Acoust. Soc. Jpn. (E)*, vol. 5, no. 4, pp. 233–242, 1984.

[7] S. J. L. Mozziconacci and D. J. Hermes, "Expression of emotion and attitude through temporal speech variations," in *Proc. ICSLP*, Beijing, China, 2000, pp. 373–378.

[8] J. E. Cahn, "The generation of affect in synthesized speech," *J. Amer. Voice I/O Soc.*, vol. 8, pp. 1–19, Jul. 1990.

[9] Synthesis units for conversational speech—Using phrasal segments, N. Campbell. [Online]. Available: http://feast.atr.jp/nick/refs.html

[10] M. Schröder and S. Breuer, "XML representation languages as a way of interconnecting TTS modules," in *Proc. ICSLP*, Jeju, Korea, 2004, pp. 1889–1892.

[11] E. Eide, A. Aaron, R. Bakis, W. Hamza, M. Picheny, and J. Pitrelli, "A corpus-based approach to $< ahem$ expressive speech synthesis," in *Proc. IEEE Speech Synthesis Workshop*, Santa Monica, CA, 2002, pp. 79–84.

[12] Z.-J. Chuang and C.-H. Wu, "Emotion recognition from textual input using an emotional semantic network," in *Proc. Int. Conf. Spoken Language Processing, ICSLP 2002*, Denver, CO, 2002, pp. 2033–2036.

[13] E. Rank and H. Pirker, "Generating emotional speech with a concatenative synthesizer," in *Proc. ICSLP*, 1998, pp. 671–674.

[14] A. Li, "Chinese prosody and prosodic labeling of spontaneous speech," in *Proc. Speech Prosody*, 2002, pp. 39–46.

[15] H. Kawahra and R. Akahane-Yamada, "Perceptual effects of spectral envelope and F0 manipulations using STRAIGHT method," *J. Acoust. Soc. Amer.*, pt. 2, vol. 103, no. 5, p. 2776, 1998. 1aSC27.

[16] A. B. Kain, "High-resolution voice transformation," Ph.D. dissertation, Oregon Health and Sci. Univ., Portland, Oct. 2001.

[17] G. P. Kochanski and C. Shih, "STEM-ML: Language independent prosody description," in *Proc. ICSLP*, Beijing, China, 2000, pp. 239–242.

[18] I. Murray and J. L. Arnott, "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion," *J. Acoust. Soc. Amer.*, pp. 1097–1108, 1993.

[19] R. M. Stibbard, "Vocal expression of emotions in non-laboratory speech: An investigation of the reading/leeds emotion in speech project annotation data," Ph.D. dissertation, Univ. Reading, Reading, U.K., 2001.

[20] X. Sun, "The determination, analysis, and synthesis of fundamental frequency," Ph.D. dissertation, Northwestern Univ., Evanston, IL, 2002.

[21] E. Moulines and Y. Sagisaka, "Voice conversion: State of the art and perspectives," *Speech Commun.*, vol. 16, no. 2, pp. 125–126, Feb. 1995.

[22] S. McGilloway, R. Cowie, E. Doulas-Cowie, S. Gielen, M. Westerdijk, and S. Stroeve, "Approaching automatic recognition of emotion from voice: A rough benchmark," in *Proc. ISCA workshop Speech Emotion*, 2000, pp. 207–212.

[23] N. Amir, "Classifying emotions in speech: A comparison of methods," in *Proc. Eurospeech*. Holon, Isreal, 2001, pp. 127–130.

[24] G. McLachlan and T. Krishnan, "The EM algorithm and extensions," in *Wiley Series in Probability and Statistics*, New York: Wiley, 1997.

[25] C. Gobl and A. N'1Chasaide, "The role of voice quality in communicating emotion, mood and attitude," *Speech Commun.*, vol. 40, pp. 189–212, 2003.

[26] R. Tato, R. Santos, R. Kompe, and J. M. Pardo, "Emotional space improves emotion recognition," in *Proc. ICSLP*, Denver, CO, Sep. 2002, pp. 2029–2032.

[27] V. A. Petrushin, "Emotion recognition in speech signal: Experimental study, development and application," in *Proc. ICSLP*, Beijing, China, 2000, pp. 222–225.

[28] B. Hayes, *Metrical Stress Theory: Principles and Case Studies*. Chicago, IL: Univ. Chicago Press, 1995.

[29] Z.-W. Shuang, Z.-X. Wang, Z.-H. Ling, and R.-H. Wang, "A novel voice conversion system based on codebook mapping with phoneme-tied weighting," in *Proc. ICSLP*, Jeju, Korea, Oct. 2004, pp. 1197–1200.

[30] L. M. Arslan and D. Talkin, "Voice conversion by codebook mapping of line spectral frequencies and excitation spectrum," in *Proc. Eurospeech*, Rhodes, Greece, 1997, pp. 1347–1350.

[31] T. Watanabe *et al.*, "Transformation of spectral envelope for voice conversion based on radial basis function networks," in *Proc. ICSLP*, Denver, CO, 2002, pp. 285–288.

[32] A. Li and H. Wang, "Friendly speech analysis and perception in standard chinese," in *Proc. ICSLP*, Jeju, Korea, 2004, pp. 897–900.

[33] T. Toda, A. W. Black, and K. Tokuda, "Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter," in *Proc. ICASSP*, 2005, pp. 9–12.

[34] Y. Chen *et al.*, "Voice conversion with smoothed GMM and map adaptation," in *Proc. Eurospeech*, Geneva, Switzerland, 2003, pp. 2413–2416.

[35] Y. Stylianou *et al.*, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 2, pp. 131–142, 1998.

[36] H Mizuno, H Mizuno, and M Abe, "Voice conversion based on piecewise linear conversions rules of formant frequency and spectrum tilt," *Speech Commun. 16*, pp. 153–164.

[37] Y. Kang, Z. Shuang, J. Tao, W. Zhang, and B. Xu, "A hybrid gmm and codebook mapping method for spectral conversion," in *Proc. 1st Int. Conf. Affective Comput. Intell. Interaction*, 2005, pp. 303–310.

Jianhua Tao (M'98) received the M.S. degree from Nanjing University, Nanjing, China, in 1996 and the Ph.D. degree from Tsinghua University, Beijing, China, in 2001.

He is currently an Associate Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing. His current research interests include speech synthesis, speaker simulation, affective computing, and multimedia integration. He has published more than 60 papers in major journals and proceedings.

Dr. Tao received several awards from the important conferences, such as Eurospeech, NCMMSC, etc. He was elected as the chair or program committee member for several major conferences. Currently, he is the Secretary of Speech Information Processing Committee of Chinese Information Processing Society and Secretary of the special interest group of Chinese Spoken Language Processing in ISCA.

Yongguo Kang received the B.Sc. degree in electrical engineering from Xi'an Jiaotong University, Shanghai, China, in 2000. He is currently pursuing the Ph.D. degree in the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China.

His main research interests include voice conversion, speech signal processing, and expressive speech synthesis.

Aijun Li received the B.Sc. degrees in both computer science and engineering and electronic engineering from Tianjin University, Tianjin, China, in 1991.

She is the Director of the Phonetic Laboratory in Institute of Linguistics, Chinese Academy of Social Sciences, Beijing, China. She is also the Secretary of the Phonetics Association of China. Her early research work was on phonetics-oriented and Klatt-liked speech synthesis for Standard Chinese. Recently, she has focused on Chinese speech prosody, speech corpus, and acoustic and phonetic aspects of Standard Chinese for L2 learners and regional accent learners.