American Journal of Signal Processing 2012, 2(5): 134-138
DOI: 10.5923/j.ajsp.20120205.06

GMM-Based Emotional Voice Conversion Using Spectrum and Prosody Features

Ryo Aihara*, Ryoichi Takashima, Tetsuya Takiguchi, Yasuo Ariki

Department of Computer Science and Systems Engineering, Kobe University, Japan

Abstract We propose Gaussian Mixture Model (GMM)-based emotional voice conversion using spectrum and prosody features. In recent years, speech recognition and synthesis techniques have been developed, and an emotional voice conversion technique is required for synthesizing more expressive voices. The common emotional conversion was based on transformation of neutral prosody to emotional prosody by using huge speech corpus. In this paper, we convert a neutral voice to an emotional voice using GMMs. GMM-based spectrum conversion is widely used to modify non linguistic information such as voice characteristics while keeping linguistic information unchanged. Because the conventional method converts either prosody or voice quality (spectrum), some emotions are not converted well. In our method, both prosody and voice quality are used for converting a neutral voice to an emotional voice, and it is able to obtain more expressive voices in comparison with conventional methods, such as prosody or spectrum conversion.

Keywords Voice Conversion, GMM, Emotion, Spectrum, Prosody

1. Introduction

In recent years, speech synthesis techniques have been well developed; e.g., text reading system, speech-oriented guidance system, and synthesizing singing voices[1-3]. However, in these systems only the linguistic information is synthesized, and they cannot handle human emotion.

The conventional method in emotional speech synthesis was replacing prosody using huge speech corpus. This method requires enormous time and effort to convert emotional prosody. Mori et al[4] proposed an F0 synthesis method for using subspace constraint in prosody. In this method, principal components analysis is adopted to reduce the dimensionality of prosodic components, which also allows us to generate new speeches that are similar to training samples. Wo et al[5] proposed a hierarchical prosody conversation. The pitch contour of the source speech is decomposed into a hierarchical prosodic structure consisting of sentence, prosodic word, and subsyllable levels. Veaux et al[6] proposed an F0 conversion system based on a Gaussian mixture model (GMM). A GMM is used to map the prosodic features between neutral and expressive speech, and the converted F0 contour is generated under dynamic features constraints. However, these methods do not include conversion of voice quality (spectrum), and, hence, some emotions were not converted well.

A GMM is widely used in spectrum conversion to modify non linguistic information such as voice characteristics while keeping linguistic information unchanged[7-9]. Toda et al adopted this method to articulatory speech synthesis[10] and speaking-aid system for laryngectomees[11].

In this paper, we propose an emotional voice conversion method that includes both voice quality and prosody. Voice quality is synthesized by spectrum conversion using the GMM and maximum likelihood method[7-9]. Prosody is also converted using GMM-based F0 conversion. Our result demonstrates that emotions are synthesized sufficiently by converting both F0 and spectrum.

The rest of this paper is organized as follows: In Sec. 2, GMM-based voice conversion is introduced; our proposed method is developed in Sec. 3; the experimental results are described in Sec. 4; and the final section is devoted to our conclusions.

2. GMM-Based Voice Conversion

2.1. Spectrum Conversion

2.1.1. Mapping Function[7]

Let x_t and y_t be the source and target feature vectors at the t-th frame, respectively. We treat the dynamic features $\Delta x_t, \Delta y_t$ as additional states; i.e., the augmented states $X_t = [x_t^T, \Delta x_t^T]^T$ and $Y_t = [y_t^T, \Delta y_t^T]^T$ are introduced.

[*] Corresponding author:
aihara@me.cs.scitec.kobe-u.ac.jp (Ryo Aihara)

Published online at http://journal.sapub.org/ajsp

Defining $Z_t = [X_t^T, Y_t^T]^T$, its joint probability density is set as

$$P(Z_t \mid \lambda^{(Z)}) = \sum_{m=1}^{M} \alpha_m N(Z_t; \mu_m^{(Z)}, \Sigma_m^{(Z)}), \quad (1)$$

where $\lambda^{(Z)}$ is a parameter set of the weights α_m, source mean vector $\mu_m^{(X)}$, target mean vector $\mu_m^{(Y)}$, and covariance matrices $\Sigma_m^{(Z)}$, and they are given by

$$\Sigma_m^{(Z)} = \begin{bmatrix} \Sigma_m^{(XX)} & \Sigma_m^{(XY)} \\ \Sigma_m^{(YX)} & \Sigma_m^{(YY)} \end{bmatrix}, \quad \mu_m^{(Z)} = \begin{bmatrix} \mu_m^{(X)} \\ \mu_m^{(Y)} \end{bmatrix}.$$

These parameters are estimated by using the EM-algorithm.

2.1.2. Maximum Likelihood Estimation[8]

Defining the time sequences of the source and target feature vectors as

$$X = [X_0, X_1, \dots, X_t, \dots, X_T]^T$$
$$Y = [Y_0, Y_1, \dots, Y_t, \dots, Y_T]^T,$$

the likelihood function is given by

$$P(Y \mid X, \lambda^{(Z)}) = \sum_{m} P(m \mid X, \lambda^{(Z)}) P(Y \mid X, m, \lambda^{(Z)}), \quad (2)$$

where $m = \{m_1, m_2, \dots, m_t, \dots, m_T\}$ is a mixture component sequence. The *m*-th conditional probability distribution is given by

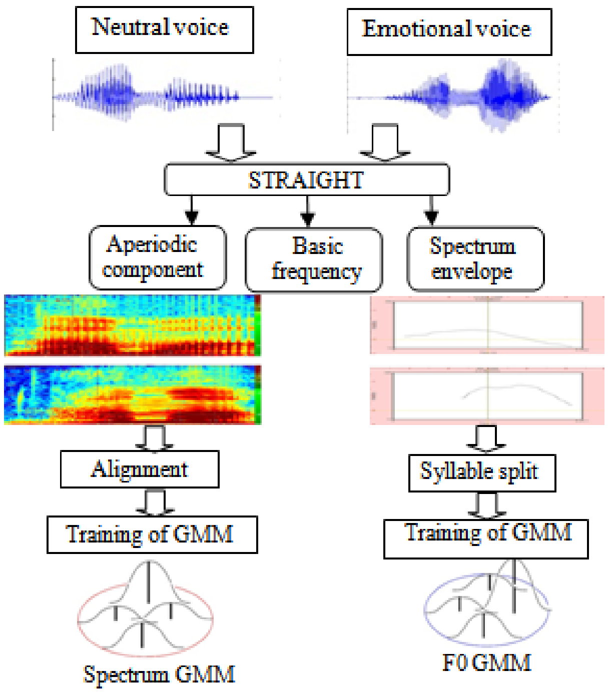

Figure 1. Training process

$$P(Y_t \mid X_t, m, \lambda^{(Z)}) = N(Y_t; E_{m,t}^{(Y)}, D_m^{(Y)}) \quad (3)$$

$$E_{m,t}^{(Y)} = \mu_m^{(Y)} + \Sigma_m^{(YX)} \left(\Sigma_m^{(XX)}\right)^{-1} \left(X_t - \mu_m^{(Y)}\right) \quad (4)$$

$$D_m^{(Y)} = \Sigma_m^{(YY)} - \Sigma_m^{(YX)} \left(\Sigma_m^{(XX)}\right)^{-1} \Sigma_m^{(XY)} \quad (5)$$

Instead of estimating Y directly, we estimate the converted *K*-dimensional static target feature y, which satisfies $Y = Wy$ where W is a 2*KT*-dimensional square matrix[8]. Hence we seek

$$\hat{y} = \arg\max P(Y \mid X, \lambda^{(Z)}) \ \mathit{subject\ \ to}\ Y = Wy. \quad (6)$$

Introducing the following approximation;

$$P(Y \mid X, \lambda^{(Z)}) \cong P(m \mid X, \lambda^{(Z)}) P(Y \mid X, m, m, \lambda^{(Z)}) \quad (7)$$

we obtain the suboptimum mixture component sequence $\hat{m}$ and the converted static feature vector $\hat{y}$ as follows:

$$m = \underset{m}{\arg\max}\, P(m \mid X, \lambda^{(Z)}) \quad (8)$$

$$\hat{y} = \left(W^T D_{\hat{m}}^{(Y)-1} W\right)^{-1} W^T D_{\hat{m}}^{(Y)-1} E_{\hat{m}}^{(Y)}, \quad (9)$$

where

$$E_{\hat{m}}^{(Y)} = [E_{\hat{m}1,1}^{(Y)\,T}, E_{\hat{m}2,2}^{(Y)\ \ T}, \dots, E_{\hat{m}t.t}^{(Y)T}, \dots, E_{\hat{m}T,T}^{(Y)\ \ T}] \quad (10)$$

and

$$D_{\hat{m}}^{(Y)-1} = diag[D_{\hat{m}1}^{(Y)-1}, D_{\hat{m}2}^{(Y)-1}, \dots, D_{\hat{m}t}^{(Y)-1}, \dots, D_{\hat{m}T}^{(Y)-1}]. \quad (11)$$

2.2. Prosody Conversion[6]

Prosody conversion is performed applying the conversion method described in Sec. 2.1 to F0. $s = [s_0^{(i)}, s_1^{(i)}, \dots, s_{L-1}^{(i)}]$ denotes an F0 contour of length *L*, which is extracted over the i-th syllable. This contour is represented by its first *N* Discrete Cosine Transform (DCT) coefficients normalized by $1/\sqrt{L}$. Normalized DCT coefficients are written as $[c_0^{(i)}, c_1^{(i)}, \dots, c_{N-1}^{(i)}]$. The inverse DCT is defined as

$$s_l^{(i)} = c_0^{(i)} + \sqrt{2} \sum_{n=1}^{N-1} c_n^{(i)} \cos\left[\frac{\pi}{L} n \left(l + \frac{1}{2}\right)\right] \quad (12)$$

The target static feature vector of the *i*-th syllable is represented as $x_i = [c_0^{(i)}, c_1^{(i)}, \dots, c_{N-1}^{(i)}]$. The dynamic feature is calculated from the static feature, and X_i represents the static and dynamic features of the *i*-th syllable. The source vector and the target vector are augmented as $Z_i = \left[X_i^T, Y_i^T\right]^T$. Hence, we obtain $\hat{y}$ from Eq. (9).

3. GMM-Based Emotional Conversion

In this paper, both spectrum envelope and basic frequency which are extracted from a neutral voice are converted to those of emotional voices, where target emotions are “Anger”, “Sadness” and “Joy”, and the GMM is constructed for each emotion. Our system has two phases: the training phase and the conversion phase.

The outline of the training phase is shown in Fig. 1. The neutral voice word is the same as that of the emotional voice. These are spoken by the same speaker. The spectrum envelope, basic frequency, and aperiodic component are extracted from these two voices using the STRAIGHT analysis method[12-15]. The aperiodic component is not used in our method.

The outline of the conversion method is shown in Fig. 2. The extracted basic frequency is divided into syllables and converted using the F0 GMM trained in Fig. 1. The spectrum envelope is converted using the spectrum GMM. The emotional voice is synthesized from the converted F0, spectrum envelope and source speaker's aperiodic envelope using the STRAIGHT synthesis method.

3.1. Spectrum Conversion

The spectrum envelope, which is extracted using the STRAIGHT analysis method, is converted using the GMM described in Sec. 2.1. The duration of the source and target spectrum must be modified by the DP-matching algorithm. To reduce the dimension of the envelope, static features are represented by its first 12 DCT coefficients in Eq. (12).

Dynamic features are defined as follows:

$$\Delta x = \frac{1}{2}\left(x_{t+1} - x_{t-1}\right) \qquad (13)$$

These features are modelled using Eq. (1).

3.2. Prosody Conversion

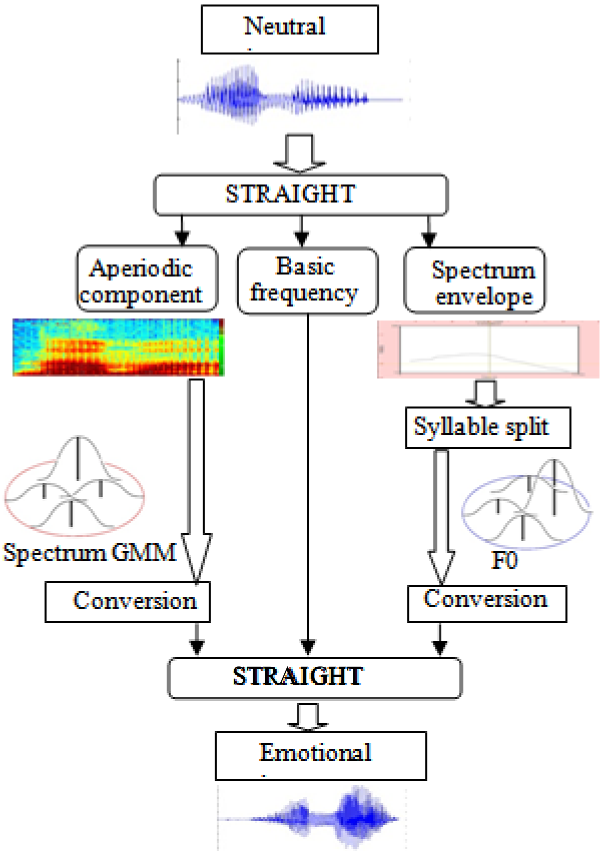

Figure 2. Conversion process

The basic frequency, which is extracted using the STRAIGHT, is also converted using the GMM in Sec. 2.2. Fig. 3 shows how to extract the prosody feature from a Japanese word "AMAGAERUWA". The basic frequency cannot be converted on each frame because the basic frequency is the 1-dimensional vector. Therefore, the word is divided into syllables to obtain the prosody feature. In this paper, the contour of a syllable is represented by its first 5 DCT coefficients. When the contour length is defined as *L*, the coefficients are normalized by $1/\sqrt{L}$. The dynamic features are calculated using Eq. (13).

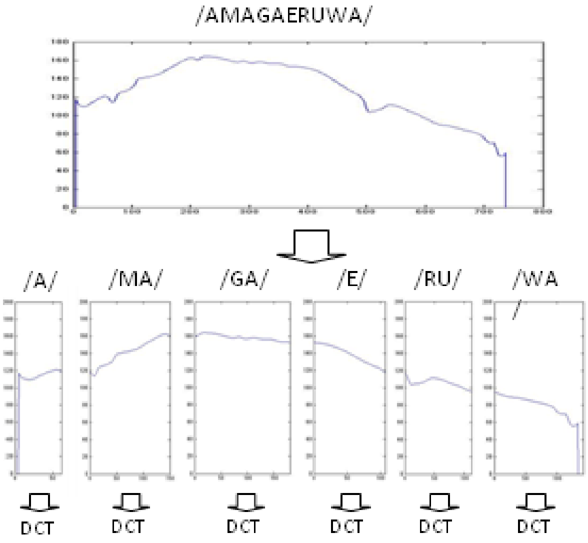

Figure 3. Syllable split of basic frequency

4. Experiments

In our experiment, neutral words are converted to emotional words. We performed five types of experiments as shown in Table 1. In experiments (a) and (b), the spectrum envelope or the F0 are converted. To show the effectiveness of conversion, the neutral spectrum envelope, or F0, is replaced with emotional ones in experiments (d) and (e).

Table 1. Five types of experiments

	Spectrum	F0
(a)	conversion	source
(b)	source	conversion
(c)	conversion	conversion
(d)	target	source
(e)	source	target

4.1. Experimental Conditions

The "Keio University Japanese Emotional Speech Database" was used in our experiment. A male Japanese speaker with acting experience recorded 47 emotions for each 20 words. We used three emotions: "Neutral", "Anger", "Joy" and "Sadness" from the database.

The speech data was directly recorded into the hard disk drive through a microphone connected to the computer in a sound-proof room. Waveforms were digitized by 16-kHz sampling and 16-bit quantization.

In our experiment, we converted "Neutral" to "Anger", "Joy", and "Sadness". Training and converting were conducted separately for each emotion. All 20 recorded words were used as training data, and we converted the same 20 words. The number of mixtures of GMM is set at 64 in spectrum and F0 conversion.

4.2. Results

We performed a subjective emotional classification test. All the listeners were Japanese, and the number of listeners was 10. The listener classified a converted voice into one emotion from "Neutral", "Anger", "Joy" or "Sadness".

In Table 2, some results of subjective emotional classification for recorded words are given. The classification rate of 100% was obtained for all emotions, hence the corpus is sufficient for recognizing emotion.

Table 2. Results of Classification for Recorded (Original) Voices[%]

Tar. / Percept.	Anger	Sadness	Joy	Neutral
Anger	**100**	0	0	0
Sadness	0	**100**	0	0
Joy	0	0	**100**	0
Neutral	0	0	0	**100**

Table 3. Results of Classification for Converted Voices[%]

(a) Spectrum Conversion

Tar. / Percept.	Anger	Sadness	Joy	Neutral
Anger	**45**	0	5	50
Sadness	10	**5**	0	85
Joy	5	5	**5**	85

(b) F0 Conversion

Tar. / Percept.	Anger	Sadness	Joy	Neutral
Anger	**5**	15	20	60
Sadness	5	**80**	5	10
Joy	25	25	**20**	30

(c) Spectrum and F0 Conversion

Tar. / Percept.	Anger	Sadness	Joy	Neutral
Anger	**65**	0	10	25
Sadness	5	**80**	5	10
Joy	10	20	**45**	25

(d) Spectrum Exchange Using Target Voices

Tar. / Percept.	Anger	Sadness	Joy	Neutral
Anger	**65**	0	0	35
Sadness	0	**10**	5	85
Joy	0	0	**5**	95

(e) F0 Exchange Using Target Voices

Tar. / Percept.	Anger	Sadness	Joy	Neutral
Anger	**5**	15	5	75
Sadness	0	**95**	0	5
Joy	10	35	**25**	45

The classification results for the converted voices are shown in Table 3. Results of spectrum conversion only are shown in Table 3-(a). Almost half the listeners classified "Anger" correctly. However, the other emotions tended to be classified as "Neutral". Hence, the use of just spectral conversion is imperfect for emotional conversion.

Table 3-(b) shows the results of the conversion of basic frequency only. The classification rate of 80% was obtained for "Sadness". Therefore, "Sadness" can be expressed by the basic frequency conversion only. However, "Anger" tends to be classified as "Neutral". The result of "Joy" is not well classified. Hence, the conversion of basic frequency only is also imperfect for emotional conversion.

Our proposed method is shown in Table 3-(c). The classification rate of "Sadness" did not increase in comparison with Table 3-(b). Hence, spectrum conversion did not work on conversion of "Sadness". The classification rates of "Anger" and "Joy" greatly increased in comparison with Table 3-(a) and 3-(b). Hence, both spectrum and F0 conversion is much effective in conversion with "Anger" and "Sadness".

4.3. Discussion

We performed three types of conversions to three different emotions. In spectrum conversion, the half of listeners classified "Anger" correctly. The other two emotions tended to be classified as "Neutral". Table 3-(d) shows results by replacing neutral spectrum only with emotional spectrum using the target voice. The classification rates of "Sadness" and "Joy" in Table 3-(d) were almost the same as shown in Table 3-(a). The classification rate of 65% was obtained for "Anger", which is close to "Anger" in Table 3-(a). Hence, spectrum conversion has a significant influence on synthesizing "Anger". Moreover, the experiment results show that, in emotional conversion, only the conversion of the spectrum envelope is imperfect.

In F0 conversion, "Sadness" obtained a high classification rate; however, the other two emotions could not achieve a high rate. Table 3-(e) shows the results obtained when replacing neutral F0 only with emotional F0 using the target voice. "Sadness" in Table 3-(e) obtained a classification rate of almost 100%. Therefore, F0 conversion has a significant influence on synthesizing "Sadness". Also, the results in Table 3-(b) are similar to those in Table 3-(e). Hence, converting F0 in emotional conversion has sufficient accuracy.

The results obtained using our proposed method, which is the combination of F0 and spectrum conversions, are shown in Table 3-(c). The classification rates of "Anger" and "Joy" increased over those in Table 3-(a) and (b). "Anger" obtained a classification rate of 65%, and it is the same as in Table 3-(d). Our method seems to recover the classification rates of emotion obtained by converting F0. "Joy" obtained 45% in Table 3-(c), and it is a higher rate than Table 3-(d) and (e). These rates show the effectiveness of our proposed method.

5. Conclusions

We proposed emotional conversion of spectrum and prosody. Experimental results show that both spectrum and prosody conversion is effective in synthesizing "Anger" and "Joy". "Sadness" could be synthesized using prosody conversion only, and spectrum conversion had no effect.

ACKNOWLEDGEMENTS

The authors are grateful for permission to use Keio University Japanese Emotional Database[3].

REFERENCES

[1] T. Cincarek, H. Kawanami, R. Nishimura, A. Lee, H. Saruwatari and K. Shikano "Development, long-term operation and portability of a real-environment speech-oriented guidance system", IEICE TRANSACTIONS on In-formation and Systems (2008).

[2] G. Bailly, N. Campbell and B. Mobius, "ISCA special session: Hot topics in speech synthesis," Eurospeech, pp. 37-40, (2003).

[3] J. Li, H. Yang, W. Zhang and L. Cai, "A Lyrics to Singing Voice Synthesis System with Variable Timbre," Communications in Computer and Information Science, 2011, Volume 225, Part 1, 186-193, (2011).

[4] S. Mori, T. Moriyama, and S. Ozawa, "Emotional speech synthesis using subspace constraints in prosody," In Proceedings of the IEEE Conference on Multimedia and Expo, pp. 1093–1096, (2006).

[5] C. H. Wu, C. C. Hsia, C. H. Lee, "Hierarchical Prosody Conversion Using Regression-Based Clustering for Emotional Synthesis," IEEE Trans. Audio, Speech and Lang Proc., (2010).

[6] C. Veaux, X. Robet, "Intonation conversion from neutral to expressive speech," INTERSPEECH, pp. 2765-2768, (2011).

[7] Y. Stylianou, O. Cappe and E. Moilines, "Statistical methods for voice quality transformation," Eurospeech, pp. 447-450, (1995).

[8] T. Toda, A.W. Black, K. Tokuda, "Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory," IEEE Trans. ASLP, Vol. 15, No.8, pp. 2222-2235, (2007).

[9] T.Toda, Y.Ohtani, K.Shikano, "One-to many and Many-to-one Voice Conversion Based on Eigenvoices", Acoustics, Speech and Signal Processing, (ICASSP 2007), Vol.IV, pp1249-1252, (2007).

[10] T. Toda, A.W. Black, K.Tokuda, "Mapping From Articulatory Movements to Vocal Tract Spectrum with Gaussian Mixture Model forArticulatory Speech Synthesis," Proc. 5th ISCA Speech Synthesis Workshop, pp. 31–36, Pittsburgh, USA, June 2004, (2004).

[11] K.Nakamura, T. Toda, Y. Nakajima, H.Saruwatari, K.Shikano, "Evaluation of Speaking-Aid System with Voice Conversion for Laryngectomees Toward Its Use in Practical Environments" INTERSPEECH, pp.2209-2212, (2008)

[12] H. Kawahara, "STRAIGHT, exploitation of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds," Acoustical Science and Technology, pp. 349-353, (2006).

[13] H. Kawahara and H. Matsui "Auditory morphing based on an elastic perceptual distance metric in an interference-free time-frequency representation," In Proc. 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2003), Vol. I, pp. 256-259, Hong Kong, (2003).

[14] H. Matsui and H. Kawahara. "Investigation of emotionally morphed speech perception and its structure using a high quality speech manipulation system," In Eurospeech'03, pp. 2113-2116, Geneva, (2003).

[15] H. Kawahara, A. de Cheveigné, H. Banno, T. Takahashi and T. Irino, "Nearly Defect-free F0 Trajectory Extraction for Expressive Speech Modifications based on STRAIGHT," In Interspeech'05, pp. 537-540, Lisboa, (2005).