Character Conversion Based on Eigenvoice Technique*

Teeraphon Pongkittiphan, Nobuaki Minematsu,
Daisuke Saito, Keikichi Hirose (University of Tokyo)

1 Introduction

Voice conversion [1] is a method to modify one speaker's speech so that it can be perceived as being spoken by another speaker without changing any linguistic contents. This technique is a potential tool for synthesizing speech with various kinds of speaker identity. There are many applications including those for non-speech media, such as speaker identity modification in Text-to-Speech (TTS) system [1], hand-motion to speech conversion [2], cross language voice conversion [3][4].

Several statistical techniques were proposed to estimate the conversion function. Among them, joint GMM approach, proposed by A. Kain et al. [1], is one of the most famous and efficient methods. This process requires a parallel set of utterances spoken by a source speaker and a target speaker. This set contains their utterances of the same sentences. However, each estimated model provides a specific transformation between only that designed speaker pair.

Recently, T. Toda et al. [5] proposed eigen-voice conversion based on GMM (EVC-GMM) that allows a flexible control of speaker characteristics. This eigenvoice technique is similar to eigenvoice based speech recognition [6]. To train an eigenvoice GMM (EV-GMM), plenty of parallel utterance pairs between one reference speaker and many target speakers were used as prior knowledge. Using this data, the joint GMM between one reference speaker and arbitrary target speakers can be estimated. In the eigenvoice speaker space, the identity of an arbitrary speaker can be represented as a unique weight vector, where a speaker characteristic is controlled by adjusting its weight vector.

One of the remaining challenges is an attempt to expand the diversity of character and personality within a single speaker. Let us consider a scenario in voice acting industry, only one single voice actor can professionally perform voices for two or more animated characters.

In this paper, we applied the EV-GMM approach not to speaker conversion, but to character conversion. Character conversion is the conversion from a source character to another character while keeping speaker identity. In other words, the conversion tries to generate various kinds of characters from a single speaker using training data of the various characters created by a single skilled voice actor/actress. This method is based on eigenvoice space built from 273 speakers and uses the voices of some different characters given from another single voice actor/actress. The experimental results demon-strate that our proposed character conversion method can work well comparing to only F_0 based conversion.

The remainder of this paper is organized as follows. Section 2 describes the basic EVC technique. In Section 3, a framework of character conversion is described. Section 4 describes the experimental evaluations. Finally, the paper is summarized in Section 5.

2 Eigenvoice conversion (EVC)

2.1 Eigenvoice GMM (EV-GMM)

The 2D-dimenstional feature vectors $X_t = [x_t^\top, \Delta x_t^\top]^\top, Y_t = [y_t^\top, \Delta y_t^\top]^\top$ are the static and dynamic features of source and target speakers, respectively, where $\top$ denotes the transposition of a vector. The EV-GMM on joint probability density is trained in advance with the source and target vectors as follows:

* Eigenvoice に基づくキャラクター変換、ポンキッティパン　ティーラポン、峯松信明、齋藤大輔、広瀬啓吉（東京大学）

$$P(X,Y\,|\lambda,w) = \sum_{m=1}^{M}\alpha_m N(X_t^\top, Y_t^\top; \mu_m^{(z)}(w), \Sigma_m^{(z)}) \quad (1)$$

$$\mu_m^{(z)}(w) = \begin{bmatrix} \mu_m^{(x)} \\ \mu_m^{(y)}(w) \end{bmatrix} \quad \Sigma_m^{(z)} = \begin{bmatrix} \Sigma_m^{(XX)} & \Sigma_m^{(XY)} \\ \Sigma_m^{(YX)} & \Sigma_m^{(YY)} \end{bmatrix}, \quad (2)$$

where M denotes the number of mixtures, $N(x;\mu_m,\sum_m)$ denotes the normal distribution for m[th]-mixture with mean vector μ_m and covariance matrix $\sum_m$, and α_m is the weight of m[th]-mixture.

To build the eigenvoice speaker space, first, we train a target independent joint GMM (TI-GMM) using all S[th] pre-store speakers' feature vectors. Then, target dependent GMMs (TD-GMM) for each speaker-pair is separately trained by updating the target means of TI-GMM.

We prepare a 2DM×S matrix, of which each column is a 2D×M dimensional supervector that is the concatenation of target means of each TD-GMM. To extract a small number of basis vectors, which can be used to represent any supervector, PCA is done. Finally, the target mean vector $\mu_m^{(y)}(w)$ can be represented as the linear combination of bias vector $b_m^{(0)}$ and $B_m = \left[b_m^{(1)}, \ldots, b_m^{(K)}\right]$, which is matrix of basis vectors, where K < S.

$$\mu_m^{(y)}(w) = B_m w + b_m^{(0)}. \quad (3)$$

The flexibility control of speaker individuality can be done by adjusting the K-dimensional weight vector w.

2.2 Adaptation of arbitrary target speakers

The adaptation of w to any given target speaker Y is done by applying the maximum likelihood eigen-decomposition to the weight vector estimation in EVC as follow:

$$\hat{w} = \underset{w}{argmax} \int P(X,Y\,|\lambda,w)dX \quad (4)$$

We use EM-algorithm to find the weight vector for target speaker Y, which can be written as

$$\hat{w} = \left\{\sum_{m=1}^{M}\bar{\gamma}_m B_m^\top \Sigma_m^{(YY)^{-1}} B_m\right\}^{-1} \sum_{m=1}^{M} B_m^\top \Sigma_m^{(YY)^{-1}} \bar{Y}_m, \quad (5)$$

$$\bar{\gamma}_m = \sum_{t=1}^{T}\gamma_{m,t} \quad (6)$$

$$\bar{Y}_m = \sum_{t=1}^{T}\gamma_{m,t}(Y_t - b_m^{(0)}) \quad (7)$$

$$\gamma_{m,t} = P(m|Y_t,\lambda,w). \quad (8)$$

3 Character conversion

3.1 Delta-weight vector calculation

To realize a conversion from "*character-A*" to "*character-B*" on a given source speaker, a parallel set of *character-A* and *B* utterances performed by a single voice actor were used to estimate both weight vectors of *character-A* and *B*. Instead of choosing two characters from two different speakers, we collected those from the same speaker to avoid an error that may be caused by an articulation variation between different speakers. Note that, in this paper, *character-A* is often the original voice of that voice actor.

Then, the *"delta-weight vector A → B"* ($\Delta w_{A\to B}$), which indicates a change from *character-A* to *character-B* on the same voice actor, is determined by following function,

$$\Delta w_{A\to B} = w_{(B)} - w_{(A)}. \quad (10)$$

The $\Delta w_{A\to B}$ is later added to another speaker's weight vector.

$$w_{(A\to B\ on\ src)} = w_{(src)} + \alpha\Delta w_{A\to B}, \quad (11)$$

where $w_{(src)}$ is a weight vector of source speaker, $w_{(A\to B\ on\ src)}$ is a estimated weight vector of the expected *character-B-like* voice of that source speaker.

With this modified weight vector, the expected "*character-B-like*" voice of that source speaker can be realized. Moreover, a coefficient α is included to control the degree of delta-weight vector (Δw). The more α is, the more change to target character B will be applied on the source speaker.

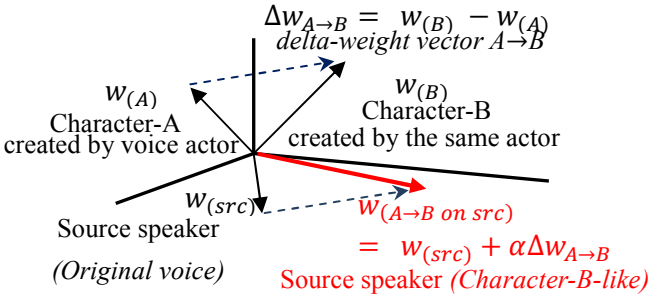

Fig. 1 Illustration of character conversion.

3.2 F0 and spectral conversion

The spectral conversion can be done directly by our proposed method, mentioned in 3.1. In the case of F_0 conversion, we use a linear transformation to project the F_0 of the source speaker to that of the same speaker of different character with the following equation,

$$F_0' = \mu_y + \frac{\sigma_y}{\sigma_x}(x - \mu_x) \qquad (12)$$

where F_0' is a converted F_0, x is an input source speech F_0, μ_x is the F_0 mean of source speaker, μ_y is the F_0 mean of target character, σ_x, σ_y is the standard deviation of F_0 of source speaker and that of target character, respectively.

4 Experimental evaluation

4.1 Speaker space construction

By using eigenvoice GMM, a speaker space is built as a weight space using a reference speaker and many target speakers. This construction was done using conditions as shown in table 1.

4.2 Characters and speakers selection

We selected source characters from a well-performed voice actor and actress, who individually gives 3 distinct character voices, shown in table 2, and selected another source speaker on whose voice character conversion will be applied. According to the source characters of these two voice artists, we decided to conduct a within-gender character conversion to avoid the unexpected errors that may be caused by articulation variation found in speakers of different genders. We selected 2 male and 2 female as source speaker for a closed-gender experiment as shown in table 3.

Table 1 EV-GMM training condition

Statistical model	*Joint GMM (128Mixtures)*
Feature vectors	*24thMCEP, ⊿24thMCEP*
Training data	*273 speakers in JNAS database*
Reference speaker	*MTS in ATR database (MTS and 273 speakers read the same sentence sets)*

Table 2: F_0 mean and standard deviation (s.d.) of source characters of voice artists

speaker	character	mean	s.d.
Voice actor	A: original	159.31	39.95
	B: elderly man	117.88	27.20
	C: cheerful boy	200.90	48.79
Voice actress	A: original	238.89	58.75
	B: elderly woman	296.01	70.08
	C: young girl	230.69	44.86

Table 3: F_0 mean and standard deviation (s.d.) of source speaker

source speaker	mean	s.d.
male #1	144.66	48.01
male #2	141.97	37.82
female #1	237.16	46.20
female #2	231.80	54.62

Table 4: Three different comparative conversions

method	F_0 conversion	spectral conversion	coefficient α
F0	○	×	×
X1	○	○	α =1
X2	○	○	α =2

4.3 Design of evaluation

We decided the evaluation strategy by conducting 3 kinds of comparative conversions, named '*F0', 'X1'* and *'X2'* as shown in table 4.

'*F0'* is a representative of the conventional F_0-based voice conversion where only F_0 of a source speaker is projected to that of target speaker with linear transformation.

X1' and *'X2'* are the character conversion done by our proposed method. They both apply F_0 conversion, and spectral conversion with coefficient α =1 and α =2, respectively.

Namely, all samples done by 3 conversion methods have the same F_0 patterns, while there are only differences in spectral features.

4.4 Resynthesized speech preparation

We estimated a weight vector using 24 utterances of each corresponding speaker, described in 4.2. Using this weight vector, the sample speech of that speaker was resynthesized with the eigen- voice method using the reference speaker of MTS. We have to admit that the naturalness of the resynthesized samples is considerably lower than the original and natural one, but we have to accept the fact these resynthesized voices show the best performance given by the eigenvoice method.

Since we have 3 different characters, there are 6 possible character conversion patterns; A→B, A→C, B→C and vice versa. For each conversion pattern, 3 different conversion methods (*F0, X1 and X2*) were applied to synthesize the samples for subjective evaluation. To conclude, there are 18 (6x3) sample pairs of a source character and target character of the same speaker, and 72 sample pairs as a total for 4 source speakers.

4.5 Listening test

To evaluate the conversion performance, 72 pairs of a source character and target character were presented to 6 Japanese and 7 Thai subjects. In the first half of 36 sample pairs of the 2 male source speakers, each subject had to listen to the 3 characters of resynthesized speech of the voice actor and pay very careful attention to how each character changes to another character in 6 combinations. Finally, for each of the 72 sample pairs, a listener selected 1 from 6 conversion combinations that best matches with the character conversion of that sample pair. Similarly, for the second half, each subject now compared another 3 characters of resynthesized samples of the voice actress and judged the similarity of each of the remaining 36 sample pairs.

4.6 Experiment results

The subjective evaluation results are shown in table 5 as correctness of subjects' judgment. Since the number of candidates used for each selection is 6, the chance level is 16.7%. Considering this, we can say that the subjects' performance is very high and the proposed character conversion works well. As shown in table 2, each character has its own F_0 range and only F_0 conversion can change the character of a speaker. Table 2 shows, however, that spectral conversion in addition to F_0 conversion improves the conversion performance both in the cases of male and female speakers.

In the case of conversion among the voice actress' characters, the F_0-based approach has much lower performance because F_0 means of original(A) and 2nd character (C) are considerably equal to each other , which is thought to make it difficult for the subjects to distinguish between these two characters. That means that our proposed method is still effective with a very small change in F_0. This demonstrates that only converting spectral feature is also possible to increase the diversity of speaker character and personality.

When considering the cross-culture perspective, it is quite obvious that the native speakers have better perception and more precise discrimination among distinct character voices.

Table 5: Results of subjective evaluation

	Conversion method	Subject	
		Japanese	Thai
Overall	*F0*	69.4%	55.4%
	X1	78.5%	68.5%
	X2	84.0%	63.1%
Male (voice actor)	*F0*	77.8%	65.5%
	X1	75.0%	78.6%
	X2	83.3%	71.4%
Female (voice actress)	*F0*	61.1%	45.2%
	X1	81.9%	58.3%
	X2	84.7%	54.8%

5 Conclusions

We propose a new method of character conversion by applying EVC-GMM technique. Using training data of the various characters created by a single skilled voice actor/actress, the conversion can generate various kinds of characters from a single speaker, while keeping the speaker identity. From the results of the listening test, our proposed method has better performance in character conversion than the general F_0-based conversion, and still effective even using a very small change in F_0. This indicates that using only spectral conversion is possible to expand the variability of speaker character and personality.

References

[1] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," Proc. ICASSP, vol. 1, pp. 285-288, 1998.

[2] A. Kunikoshi, Y. Qiao, N. Minematsu, and K. Hirose, "Speech generation from hand gestures based on space mapping," Proc. INTERSPEECH, pp. 308-311, 2009.

[3] D. Sundermann, H Hoge, A. Bonafonte, H. Ney and J. Hirschbreg, "Text-Independent Cross-Language Voice Conversion," Proc. INTER-SPEECH, 2006.

[4] M. Charlier, Y. Ohtani, T. Toda, A. Moinet and T. Dutoit, "Cross-Language Voice Conversion Based on Eigenvoices,"

[5] T. Toda, Y. Ohtani, K. Shikano, "Eigenvoice Conversion Based on Gaussian Mixture Model," Proc. INTERSPEECH, 2006.

[6] R. Kuhn, J-C. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in Eigenvoice space," IEEE Trans. Speech and Audio Processing, vol.8, 2000