# GENERATION OF SPEAKER MIXTURE VOICE USING SPECTRUM MORPHING

*Kohei Furuya, Tsuyoshi Moriyama, Shinji Ozawa*

Keio University
Department of Information and Computer Science,
3–14–1 Hiyoshi, Yokohama-shi, Kanagawa 223–8522, Japan
kpeg@ozawa.ics.keio.ac.jp

## ABSTRACT

We propose a method for synthesizing a "speaker mixture voice" that has both of two speakers' individualities. We define the "speaker mixture voice" as such that 50 percent of the subjects who listen to the voice would identify either speaker A or speaker B in ABX listening test that instructs them to identify the speaker. To synthesize the speaker mixture voice, we parameterize the spectrum envelope with respect to peaks and valleys, find the correspondence between two spectrum envelopes from independent speakers using DP matching, morph one to the other, and generate waveforms using TD-PSOLA[1]. Listening experiments showed 60 percent out of 56 synthesized voices were recognized as the speaker mixture voices. The primary application of the proposed method would be individualization of the characters in online games where multiple users play a single character as separate individuals at the same time.

## 1. INTRODUCTION

Online games lately provide a function that players can join them as one of the limited number of prepared characters. Many of such games use a voicing function with which users talk each other using the character's voice instead of their own voices. When multiple users want to share the same single character in the game, the host needs to make them identifiable.

Voice morphing is one of the technologies that make it possible. Morphing technique has originally been developed for gradually deforming the shape of an object in a picture into that of the other object. Analogously, voice morphing deforms one speaker's voice to the other speaker's. By mixing (morphing) the character's voice with each user's, the users in the game can talk with distinguishable voices with each other.

Such a synthesized voice is required to contain both the character's individuality and the user's. Though there are several papers about morphing[2][3], they only focused on smooth change from speaker A to speaker B.

We propose method for generating speaker mixture voice for online games by using spectrum morphing. Speaker mixture voice refers to the voice that makes the listeners equally perceive two of the original speakers used in morphing. Results of the listening experiment demonstrated that the proposed method can synthesize speech samples that contain two separate speakers individualities.

## 2. SYNTHESIS OF SPEAKER MIXTURE VOICE

Fig.1 is the proposed method for synthesizing the speaker mixture voice. Given two utterances from speaker A and speaker B, the prosody pattern of the speaker B is normalized by using TD-PSOLA. The system represents the spectra of both the voice from speaker A and the regularized voice from speaker B by STRAIGHT[4][5] and generate the spectrum of the speaker mixture voice by morphing one to the other. DP matching assures the formant structure of the target speech avoids corrupting in morphing.

### 2.1. Preparation

We recorded a couple of speech spoken by two individual speakers, e.g., a character A and user B, that call A and B in the following. We compute prosodic parameters from those speech. The prosodic parameters include fundamental frequency, power, and phoneme labels that contain durations. We use cepstrum for detecting frequency and Julius[6] for obtaining phoneme labels. Julius automatically gives durations of the phonemes when the set of phonemes are known.

**Table 1**. Analysis condition

| | |
|---|---|
| Sampling rate | 16kHz |
| Analysis window | Hanning window |
| Window length | 32msec(512point) |
| Frame period | 2msec(32point) |

This study was performed through Special Coordination Funds of the Ministry of Education, Culture, Sports, Science and Technology of the Japanese Government.

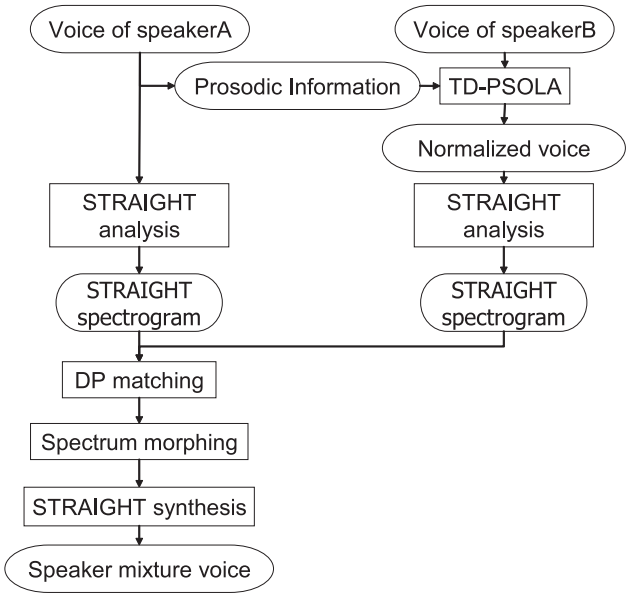



**Fig. 1**. Synthesis of speaker mixture voice

### 2.2. Prosody normalization

In order to mix individuality for each phoneme, its duration must be normalized. We do it by replacing B's prosody with A's, using TD-PSOLA(Fig.2).

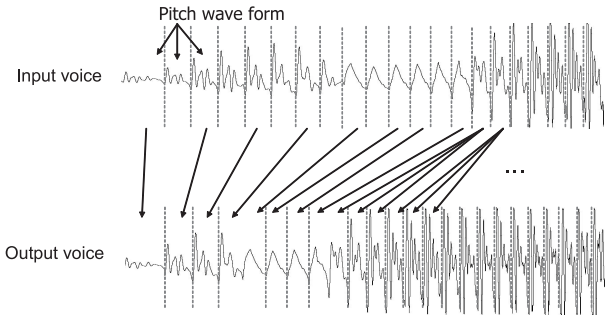
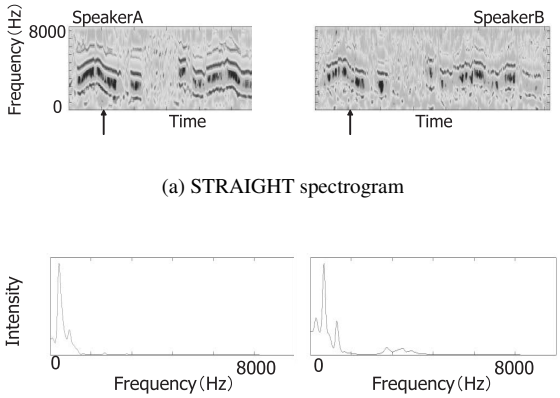



**Fig. 2**. TD-PSOLA

### 2.3. Spectral representation by STRAIGHT

To mix the individuality, it is necessary to express the speech by some parameters. The parameter needs to enable morphing without further degradation. "STRAIGHT spectrogram" is suitable as such parameter. It allows manipulation of speech parameters without further degradation. The STRAIGHT spectrogram that corresponds to the sound spectrogram is obtained by the STRAIGHT analysis(Fig.3).

Peaks and valleys exist in the spectrogram. The peak is the point where the spectrum intensity is especially high, and the valley is the point where the spectrum intensity is especially low. The positions of peaks and valleys existing in this




(a) STRAIGHT spectrogram

(b) Spectrum envelope of the position indicated above

**Fig. 3**. STRAIGHT spectrogram and spectrum envelope

spectrogram strongly influence the individuality[7] so that, we morph the individuality by modifying STRAIGHT spectrogram.

### 2.4. Correspondence between two spectrum envelopes

Correspondence between two spectrum envelopes from independent speakers is found by using DP matching that uses spectrum intensity as the distance measure.

The optimal matching between those spectrum envelopes gives correspondence between frequency components from $k_A$ of A's and $k_B$ of B's as:

$$k_B = \theta(k_A) \quad (0 <= k_A, k_B <= N-1) \tag{1}$$

where $\theta$ minimizes the difference of spectrum intensity as shown in Fig.4 and $N$ is the highest frequency in the signal.

Using the result of DP matching, peaks and valleys in the spectrogram of A's are associated with the counterparts of B's by the nonlinear stretching as shown in Fig.5.

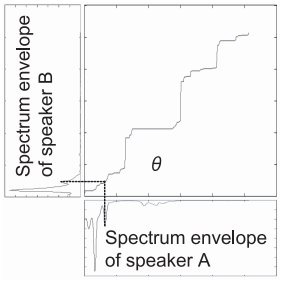



**Fig. 4**. DP matching

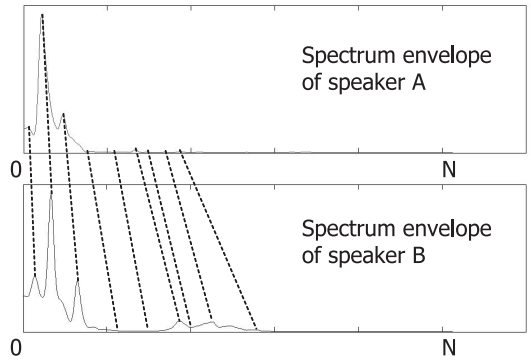



**Fig. 5**. Correspondence between two spectrum envelopes

### 2.5. Spectrum morphing

The Spectrum morphing is done by morphing spectrum position and spectrum intensity. Morphed spectrum intensity $E_M(k)$ is obtained by spectrum intensity $E_A(k_A)$ of A's and $E_B(k_B)$ of B's as:

$$E_M(k) = (1-r)E_A(k_A) + rE_B(k_B) \quad (0 <= k <= N-1) \tag{2}$$

where $r$ is morphing rate. Morphed spectrum position $k$ is obtained by spectrum position $k_A$ of A's and $k_B$ of B's as:

$$k = (1-r)k_A + rk_B \quad (0 <= k_A <= N-1) \tag{3}$$

The purpose of our method is to synthesize the voice that makes the listeners to equally perceive two of the original speakers used in morphing, so $r$ was defined as $r = 0.5$.

Obtained $k$ is not an integer value, so it is interpolated from the value before and behind that with a linear interpolation. A synthetic voice is obtained by the STRAIGHT synthesis by using obtained $E_M(k)$.

By such processing, formant can be prevented being increased and decreased by doing morphing. Therefore, morphing can be done without breaking the formant structure of a synthetic voice(Fig.6).

## 3. EXPERIMENTAL RESULTS

We evaluated the speaker mixture voice synthesized in our method by listening experiment using ABX method.

### 3.1. Conditions

The synthesis used the voices pronounced by speaker A and B as "Arayuru Genzitsuwo Subete Zibunno Houe Nezimagetanoda".

The subjects were first presented the prosodic arranged voice of A's and B's, and a synthetic voice was presented afterwards. Then, they answered whether A's or B's pronounced the voice.

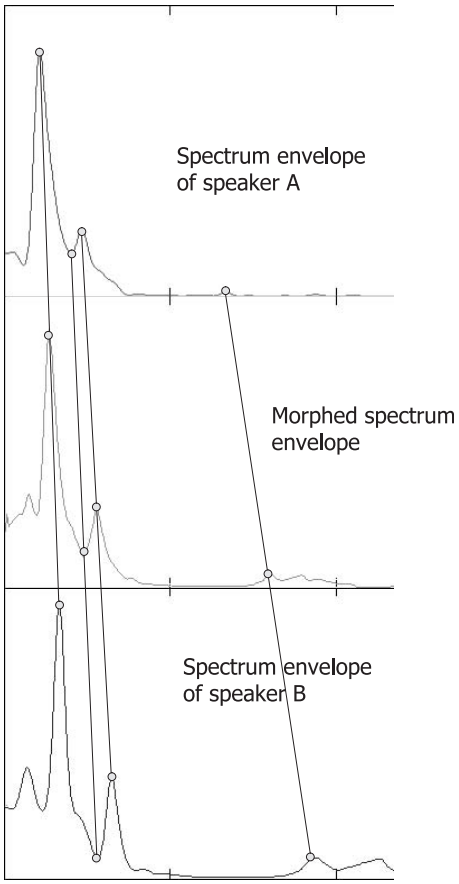



**Fig. 6**. Spectrum morphing

The order of making the voice of A's and B's heard was decided at random.

Subjects were 19 people, and 56 synthesized voices were prepared. The subjects were native Japanese speakers, and synthesized voices were presented by using headphone.

### 3.2. Results and Evaluation

Fig.7 shows the selection rate of speaker A as a speaker of the synthetic voice.

It was seen to be distributed to selection rate 50 percent and about 60 percent a lot, and the voice that the selection rate became 40 percent to 60 percent was the 60 percent. So 60 percent out of 56 synthesized voices were recognized as the speaker mixture voice.

Moreover, if the speaker mixture voice was allowed from 30 percent to 70 percent then 80 percent out of 56 synthesized voices were recognized as the speaker mixture voice.

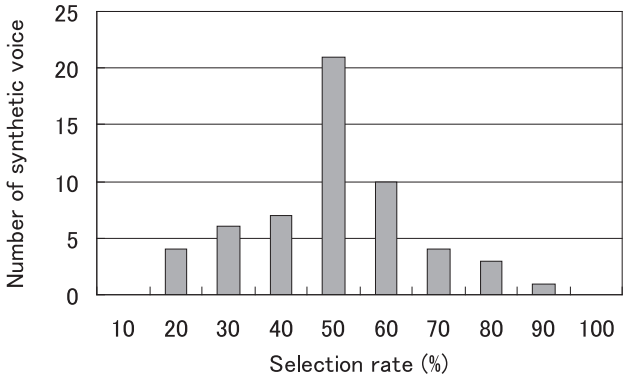



**Fig. 7**. Selection rate of speaker A

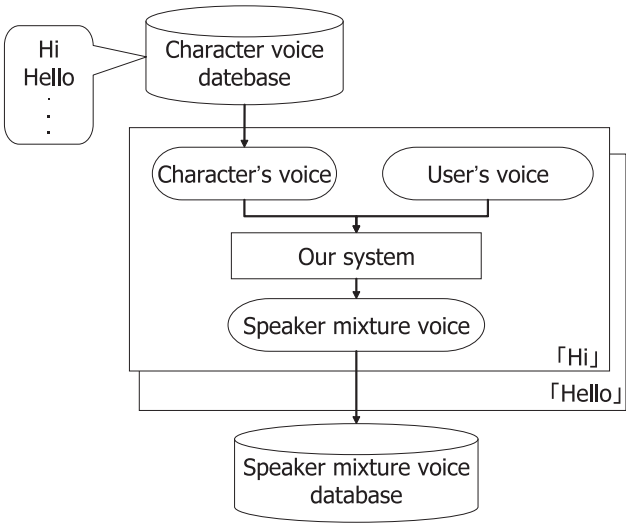



**Fig. 8**. Synthesis of the "speaker mixture voice" database

## 4. EXAMPLE APPLICATION

We are assuming this system can be used in the field of the entertainment.

For instance, when using it to the online game, the system needs user's voice at the beginning of the game, and system makes speaker mixture voice. Then, on the playing game, the user can use the database of the speaker mixture voice that the system made. Fig.8 illustrates the procedure.

## 5. CONCLUSIONS

We proposed the method for synthesizing a speaker mixture voice that has both of two speakers' individualities by morphing the STRAIGHT spectrogram using the DP matching. By using the DP matching, STRAIGHT spectrogram obtained by the STRAIGHT analysis can morph the position of the formant and the spectrum intensity without breaking its formant structure.

Listening experiments showed 60 percent out of 56 synthesized voices were recognized as the speaker mixture voice.

## 6. REFERENCES

[1] Christian Hamon, Eric Moulines, and Francis Charpentier, "A diphone synthesis system based on time-domein prosodic modifications of speech," in *Proc. ICASSP*, 1989, vol. 89, pp. 238–241.

[2] M. Abe, "Speech morphing by gradually changing spectrum parameter and fundamental frequency," in *Proc. ICSLP '96*, Philadelphia, PA, 1996, vol. 4, pp. 2235–2238.

[3] C. Orphanidou, I. M. Moroz, and S. J. Roberts, "Wavelet-based voice morphing," in *WSEAS Transactions on Systems*, December 2004, vol. 3, pp. 3297–3302.

[4] H. Kawahara, "Speech representation and transformation using adaptive interpolation of weighted spectrum: Vocoder revisited.," in *ICASSP*, April 1997, vol. 2, pp. 1303–1306.

[5] H. Kawahara, Jo Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system straight," in *MAVEBA*, 2001.

[6] A. Lee, T. Kawahara, and K. Shikano, "Julius – an open source real-time large vocabulary recognition engine," in *Proc. EUROSPEECH*, 2001.

[7] T. Kitamura and M. Akagi, "Speaker individualities in speech spectral envelopes," in *Proc. ICSLP*, 1994, vol. 3, pp. 1183–1186.