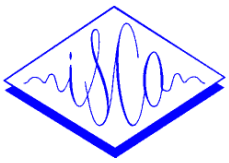

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

SPECTRAL MAPPING FOR VOICE CONVERSION USING SPEAKER SELECTION AND VECTOR FIELD SMOOTHING

Makoto HASHIMOTO and Norio HIGUCHI

e-mail: hasimoto@itl.atr.co.jp
ATR Interpreting Telecommunications Research Labs.
2-2 Hikaridai, Seika-cho, Soraku-gun, 619-02 Kyoto, Japan

ABSTRACT

This paper proposes a spectral mapping method for voice conversion using speaker selection and vector field smoothing. With this method, the spectral distance between transformed speech and the speech of a target speaker was reduced by 25% in mean value for eight target speakers (four males and four females) in comparison with the distance between the speech of a speaker who was selected from among multiple reference speakers and that of a target speaker using only one word as training data. Transformed speech samples of one male and one female were judged as closer to their target speakers than their selected speakers by 67% and 65%, respectively, in a hearing test.

1. INTRODUCTION

In the design of a high-quality synthetic speech and general-use speech synthesis system, it is important for the system to be able to output a variety of synthetic speech. Voice conversion is one of the required techniques to do this. On the other hand, a speech translation system used by multiple speakers had better be able to reproduce the vocal characteristics of a speaker in the synthesized output with a minimum of training data. Accordingly, we are studying a voice conversion technique aimed at achieving a speech synthesis system able to output a variety of synthetic speech. Previous studies of Japanese-to-Japanese voice conversion [1] and [2] are difficult to use in practice because of a large amount of training data requirements and complicated spectral mapping, respectively.

In contrast, a method that considers the transfer vector from the acoustical space of one speaker to that of another speaker as a mapping function can produce good results for speaker adaptation with a small amount of training data. This method is called Vector Field Smoothing (VFS) [3][4]. A study that uses a method similar to VFS for voice conversion exists [5], but the conversion between two speakers separated by a long spectral distance was described as difficult.

In this paper, we describe a spectral mapping method that employs Speaker Selection and VFS (SSVFS) for voice conversion [6]. We also confirm that spectral mapping is possible with a small amount of training data.

2. SPECTRAL MAPPING METHOD

In this paper, we define the following:

- Pre-Stored Speakers: multiple speakers whose spectrum codebooks and acoustical parameters are memorized in advance;
- Target Speaker: target speaker of transformation;
- Selected Speaker: speaker selected from pre-stored speakers whose spectrum is closest to the spectrum of the target speaker.

This proposed method consists of a training stage and a spectral mapping stage. In the training stage, two main steps are carried out: (1) Speaker Selection, in which a speaker is selected from pre-stored speakers, and (2) Transfer Vector Calculation, in which a transfer vector to map the acoustical feature space from the selected speaker to the target speaker is calculated. The spectral mapping stage involves, (3) Spectral Mapping, in which a mapping is carried out from the fuzzy vector quantized spectrum sequences of the selected speaker coded by his/her codebook to the acoustical feature space of the target speaker using the transfer vector calculated in the training stage. A block diagram of the proposed method is shown in Figure 1. In the following, we describe these steps.

2.1. Speaker Selection

In this step, after spectrum sequences of the same utterance made by the pre-stored speakers and the target speaker are time-aligned by DTW, the spectral distance between each pre-stored speaker and the target speaker is calculated. One speaker whose spectrum is closest to the spectrum of the target speaker is selected from the pre-stored speakers.

10.21437/Eurospeech.1995-117

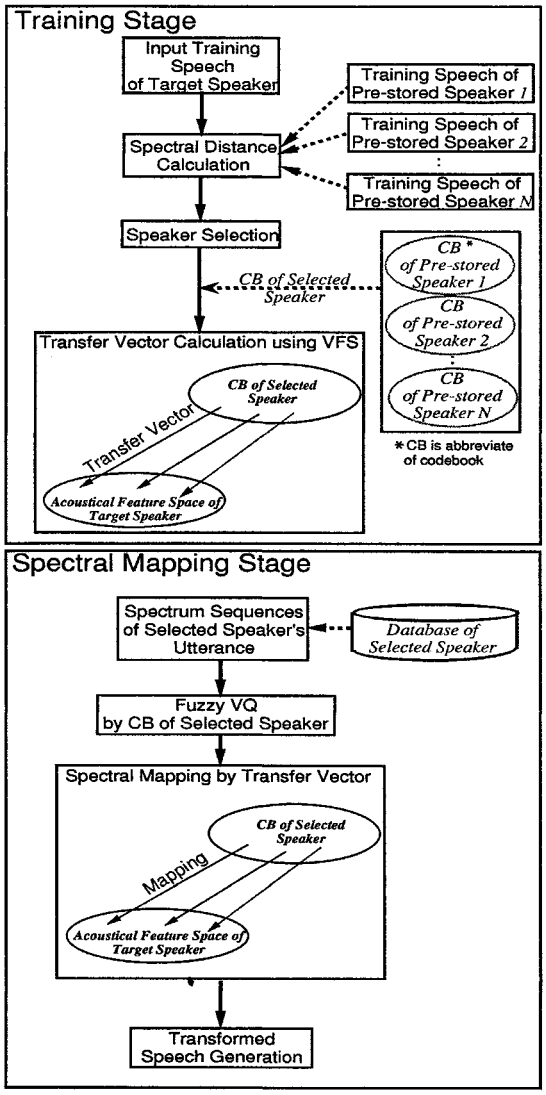

Figure 1: Block Diagram of SSVFS

2.2. Transfer Vector Calculation

In this step, transfer vectors V from the spectrum codebook of selected speaker C^s to the acoustical feature space of the target speaker are calculated by VFS. Then, a mapped codebook C^t from C^s to the space of the target speaker is obtained. VFS is a method of mapping the acoustical feature space from one speaker to another under the assumption that the correspondence between acoustical feature vectors of different speakers can be viewed as a smooth vector field. The algorithm is as follows.

[Interpolation Step]

(1) Initialize C^t using C^s.
(2) Quantize training spectrum sequences of the selected speaker using C^s.
(3) Time-align the quantized vector and the spectrum sequences of the target speaker by DTW.
(4) Calculate a transfer vector V_m between the mth codeword of selected speaker C^s_m and the average of the spectrum sequences of the target speaker corresponding to C^s_m.

$$V_m = \frac{1}{N_m} \sum_{x \in M} x - C^s_m$$

where N_m is the number of vectors of the target speaker corresponding to C^s_m, and M is a set of vectors corresponding to C^s_m.
(5) For untrained codevector C^s_n, calculate transfer vector V_n by interpolating the transfer vectors obtained in Step (4) based on the k-nearest neighbor rule. Then, update the mapped codebook using the obtained transfer vectors.

$$V_n = \sum_{k \in K} \mu_{n,k} V_k$$

$$C^t = C^s + V$$

where $\mu_{n,k}$, which is the weight of transfer vector V_k, is calculated with C^s_n and C^s_k which is made to correspond in Step (4). The following fuzzy membership function is used as $\mu_{n,k}$.

$$\mu_{n,k} = 1/\{\sum_{j \in K} (d_{n,k}/d_{n,j})^{1/(f-1)}\}$$

where $d_{n,k}$ is the distance between C^s_n and C^s_k, f is a control parameter called "fuzziness", and K is a set of codeword numbers whose transfer vector is obtained in Step (4).
(6) Iterate from Step (3) to Step (5) till the DTW distance between the transformed vector from the quantized vector in Step (2) to the acoustical feature space of the mapped codebook using the transfer vector and the spectrum sequences of the target speaker is converged.

[Smoothing Step]

In the above steps, estimation errors of transfer vectors remain when there is a small amount of training data. To decrease the amount of errors, we use the smoothing technique for transfer vectors.
(1) Calculate $\mu_{l,k}$ between C^s_l and the k-nearest neighbor vector C^s_k.
(2) Calculate the smoothed transfer vector V_l by smoothing.

$$V_l = \sum_k \mu_{l,k} \alpha V_k / \sum_k \mu_{l,k} \alpha$$

where α is a weight parameter. If k=l, $\mu_{l,k}$ holds 1.
(3) Update mapped codebook using V_l.

$$C^t_l = C^s_l + V_l$$

2.3. Spectral Mapping

In this step, speech decoding is carried out by the Fuzzy VQ of spectrum sequences of selected speaker

Table 1: Analysis Conditions

Sampling frequency	:	12kHz
Window	:	Blackman window
Window length	:	21.3ms
Frame shift	:	5ms

X^s using his/her codebook. We calculate $\mu_{p,q}$ between X_p^s which is the pth frame vector of X^s and C_q^s, then transformed spectrum X_p^t is obtained by C_q^t and $\mu_{p,q}$.

3. EXPERIMENTAL CONDITIONS

A transformation experiment was carried out using the proposed method. The ATR speech database [7][8] was used. One word "*uchiawase*" was used for training, and fifty words were used for testing. The pre-stored speakers were four males and four females. The target speakers were eight other persons (four males $M1, M2, M3, M4$ and four females $F1, F2, F3, F4$). The spectrum codebooks of the pre-stored speakers were made from 503 phonetically balanced sentences in advance using the LBG algorithm [9][10]. The codebook size was 512. Fuzziness in the mapping process was 1.5. α, the weight parameter for smoothing, was set at 1.0. k for k-nearest neighbors was 4. The feature set was a 30-dimensional cepstral coefficient vector. The distance D of the cepstrum was calculated by the following equation.

$$D = \frac{1}{fr} \sum_{ij} (cep_{ij}^s - cep_{ij}^t)^2$$

where cep_{ij}^s is the DTWed ith frame jth cepstrum of the selected speaker, cep_{ij}^t is ith frame jth cepstrum of the target speaker, and fr is the number of frames. The analysis conditions are listed in Table 1.

4. EVALUATION

4.1. Analysis of the Effect of Fuzziness

To confirm the basic performance of the proposed method and fuzziness in the VFS process, we calculated the cepstral distance between the transformed speech and the speech of the target speaker for $M1$ and $F1$ when the fuzziness was 1.5, 2, 3, 4 and 5, and between the speech of the selected speaker and that of the target speaker. Figure 2 shows the mean cepstral distance for fifty test words.

From Figure 2, the following results can be observed: 1) The distance between the transformed speech and the speech of the target speaker is reduced about 41% for $M1$ and 26% for $F1$ compared with the values of the selected speaker and target. 2) The distance decreases a little as the value of fuzziness increases. Based on this result, the following

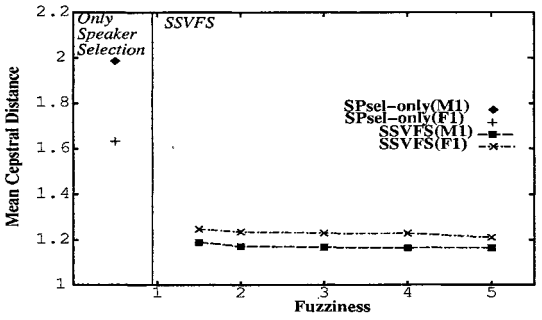

Figure 2: Evaluation of Fuzziness

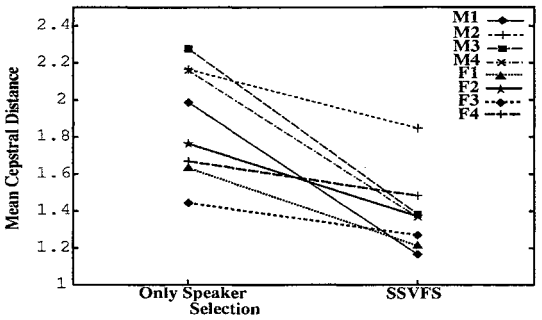

Figure 3: Evaluation of SSVFS

experiment was carried out at a fuzziness level of 5; this seemed to show the best performance.

4.2. Objective Evaluation in Cepstral Distance

Next, to confirm the performance of the proposed method, we calculated the cepstral distance between the transformed speech and the speech of the target speaker for all target speakers. Figure 3 shows the results.

In Figure 3 the following results can be observed: 3) The mean cepstral distance between the transformed speech and the speech of the target speaker is reduced for all target speakers in comparison with the distance between the selected speaker's speech and the target's. The reduction rate of the distance was about 25% in mean value (41% in max value). These results show that the SSVFS is effective for spectral mapping.

4.3. Subjective Evaluation in a Hearing Test

To investigate whether transformed speech is similar to its target in listening, an ABX hearing test was carried out for $M1$ and $F1$. In this experiment, A and B were cepstral analysis-synthesis speech of a target or selected speaker, and X was transformed speech or cepstral analysis-synthesis speech of a selected speaker. Three transformed words were selected for the perceptual test: one word had a reduction rate below the mean for the speaker, another was close to the mean, and the third was above the

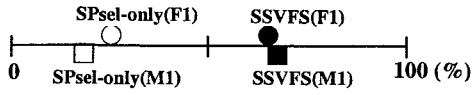

Figure 4: Results of ABX hearing test

mean. F0, duration and power factors were made to coincide with the target speaker's by DTW. The six listeners (three males and three females) serving as subjects had to judge the similarity between X and A,B. Each word was presented to the subjects four times (ABX: two times, BAX: two times) at random, provided that A,B and X were set as the same word. The evaluation of this experiment was done by judgment rate JR.

$$JR = \frac{P_j}{P_{all}} \times 100\ [\%]$$

where P_j is the number of times X was judged similar to the target speaker and P_{all} is the number of judgments. The results are shown in Figure 4.

We observed the following two results: 1) The JR of the transformed speaker being similar to the target speaker was about 66%, which is substantially higher than speaker selection without VFS. 2) The JR of the selected speaker being similar to the target speaker was about 22%. In other words, the JR of the target speaker being similar to the target speaker will be about 78%. Therefore, 66% can be considered a good value; this shows that the proposed method is also effective as a hearing test.

5. CONCLUSIONS

A spectral mapping method based on Speaker Selection and VFS (SSVFS) has been proposed for voice conversion with a small amount of training data. This method was applied to the transformation of a spectrum using only one word as training data. It was evaluated with fifty test words through the cepstral distance for eight target speakers (four males and four females) and an ABX hearing test for two target speakers (one male and one female). The following results were obtained:

- The cepstral distance between transformed speech and the speech of a target speaker was reduced by about 25% in mean value (41% in max value) in comparison with the distance between the selected speaker's speech and the target's.
- The judgment rate JR of the transformed speaker being similar to the target speaker was about 66%, which was a good value considering that the JR of the selected speaker being similar to the target speaker was about 22%.

These results showed that the SSVFS is effective for spectral mapping.

Future work needs to provide an improvement in precision by studying how to select the pre-stored speakers and training utterances. A study of mapping techniques for other acoustical features and for other languages is also needed.

[ACKNOWLEDGMENTS]
The authors wish to thank Y. Yamazaki for his support. We would also like to acknowledge K. Ohkura and the ITL members for their useful advice.

REFERENCES

[1] M. Abe, S. Nakamura, K. Shikano and H. Kuwabara: "Voice conversion through vector quantization", Proc. ICASSP'88, pp. 565-568 (1988)

[2] N. Iwahashi and Y. Sagisaka : "Voice Adaptation Using Multi-Functional Transformation with Weighting by Radial Basis Function Networks", ICSLP94, pp. 1599-1602 (1994)

[3] H. Hattori and S. Sagayama: "Speaker Adaptation based on Vector Field Smoothing", Tech. report of IEICE, SP92-15, pp. 15-22 (1992)

[4] K. Ohkura, M. Sugiyama and S. Sagayama: "Speaker Adaptation based on Transfer Vector Field Smoothing with Continuous Mixture Density HMMs", Proc. ICSLP'92, We.fPM.1.1, pp. 369-372 (1992)

[5] H. Matsumoto, Y. Maruyama and H. Inoue: "Voice quality conversion based on supervised/unsupervised spectral mapping", Journal of Acoustical Society of Japan, **50**, No.7, pp. 549-555 (1994) (in Japanese)

[6] M. Hashimoto and N. Higuchi: "Spectral Mapping for Voice Conversion using Speaker Selection and Vector Field Smoothing", Tech. report of IEICE, SP95-1, pp. 1-8 (1995) (in Japanese)

[7] K. Takeda, Y. Sagisaka, S. Katagiri, M. Abe and H. Kuwabara: "Speech Database User's Manual", Tech. report of ATR, TR-I-0028 (1988) (in Japanese)

[8] M. Abe, Y. Sagisaka, T. Umeda and H. Kuwabara: "Speech Database User's Manual", Tech. report of ATR, TR-I-0166 (1990) (in Japanese)

[9] Y. Linde, A. Buzo and R.M. Gray: "An Algorithm for Vector Quantizer Design", IEEE Trans. Commun., **COM-28**, 1, pp. 84-95 (1980)

[10] Entropic Research Lab., INC.: "ESPS programs", ESPS version 5.0 manual (1993)