TRAINING DATA SELECTION FOR VOICE CONVERSION USING SPEAKER SELECTION AND VECTOR FIELD SMOOTHING

Makoto HASHIMOTO and Norio HIGUCHI

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

ABSTRACT

We have previously proposed a spectral mapping method (SSVFS), for the purpose of voice conversion with a small amount of training data using speaker selection and vector field smoothing techniques. It has already been shown that SSVFS is effective for spectral mapping by both objective and subjective evaluations, and that it can operate with a very small amount of training data – as little as only one word [1].

This paper proposes a criterion for selecting effective training data for SSVFS. We defined coverage of parameter space with respect to training procedure of SSVFS as the criterion. This criterion is useful not only for the selection of effective training samples, which is important for the efficient learning of spectral characteristics, but also for the estimation of the degree to which learning is carried out.

To evaluate the validity of the proposed criterion, we measured the correlation between spectral resemblance and coverage. The result showed that the mean correlation coefficient for eight target speakers is -0.74 with the proposed criterion, and -0.59 without consideration of the training procedure. We conclude that the proposed criterion is useful in selecting effective training samples for SSVFS.

1. INTRODUCTION

In designing a high-quality speech synthesis system which can output personalized synthetic speech, it is important for the system to be able to generate a variety of different voices. Voice conversion is one of the ways to do this. In the case of a speech translation system used by multiple speakers, it has to be able to reproduce the vocal characteristics of a speaker for the synthesized output using a minimum of training data. Accordingly, we are studying a voice conversion technique aimed at achieving a speech synthesis system that can produce a variety of synthetic voices in order to reproduce the vocal characteristics of any given speaker.

There are previous studies on voice conversion for Japanese [2],[3]. The former needs a large amount of training data to perform a precise mapping. The latter uses interpolation of multiple speaker's spectra to produce new spectrum vectors close to a target speaker's using a small amount of training data, but needs a complicated process to get high performance. 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) [4],[5]. A study that uses a method similar to VFS for voice conversion exists [6], but the conversion between two speakers separated by a large spectral distance is described as difficult.

We have previously proposed a spectral mapping method (SSVFS), for the purpose of high quality voice conversion with a small amount of training data combining Speaker Selection and VFS techniques. It has already been shown that SSVFS is effective for spectral mapping in both objective and subjective evaluations, and can operate with a very small amount of training data - as little as only one word [1]. But a study focussing on selection of training data to get high performance has not previously been carried out.

In this paper, we describe a criterion for selecting training data in order to get higher performance in SSVFS. SSVFS is spectral mapping method based on codebook mapping. In the training process of SSVFS, mapping between a reference speaker's data and the target speaker's spectrum in the training data is carried out by DTW, and a transfer vector is calculated. The transfer vector for a codebook entry which has not been directly matched is calculated by interpolation. The proposed criterion is defined with consideration of a reliability to both matched and unmatched entries.

We also confirm that the proposed criterion is useful in the selection of effective training samples.

2. OUTLINE OF SSVFS SPECTRAL MAPPING METHOD

In this paper, we define the following:

- Pre-stored Data: VQ codebooks from spectral data and acoustical parameters of reference speakers;
- Target Speaker: speaker whose voice is the target for transformation;
- Selected Speaker: speaker selected from database whose spectrum is closest to that of the target speaker.

The 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 reference speaker is selected from pre-stored data based on minimum spectral distance, and (2) Transfer Vector Calculation, in which a transfer vector to map from the acoustical feature space of the selected speaker to that of the target speaker is calculated using VFS techniques.

The final stage involves (3) Spectral Mapping, in which the spectrum sequences of the output utterances of the selected speaker are coded by his/her codebook using fuzzy vector quantization. A mapping carried out from the fuzzy quantized vectors to the acoustical feature space of the target speaker using the transfer vector, and transformed speech is generated. A block diagram of SSVFS is shown in Fig.1.

3. MEASURE OF COVERAGE IN VECTOR SPACE

It is important to select effective training samples to get optimal performance in SSVFS. Moreover, when considering incremental learning, it is necessary to estimate the progress of learning.

In this section, the proposed measure to select effective training samples for SSVFS is described.

In SSVFS, the calculation of a transfer vector is carried out by Vector Field Smoothing techniques. In Fig.2, the principle of the training procedure in VFS is shown. VFS assumes that the correspondence between feature vectors from different speakers can be viewed as a smooth vector field. Based on this assumption, first, a transfer vector is calculated for the codeword corresponding in the training data with the following equation.

$$V_m = \frac{1}{N_m} \sum_{x \in M} x - C_m^s \qquad (1)$$

where N_m is the number of vectors of the target speaker corresponding to the mth codeword of selected speaker C_m^s, and M is a set of vectors corresponding to C_m^s.

Next, for untrained codebook entries which have not been matched in the training sample, a transfer vector is calculated by interpolating with the transfer vectors obtained in the first step based on the k-nearest

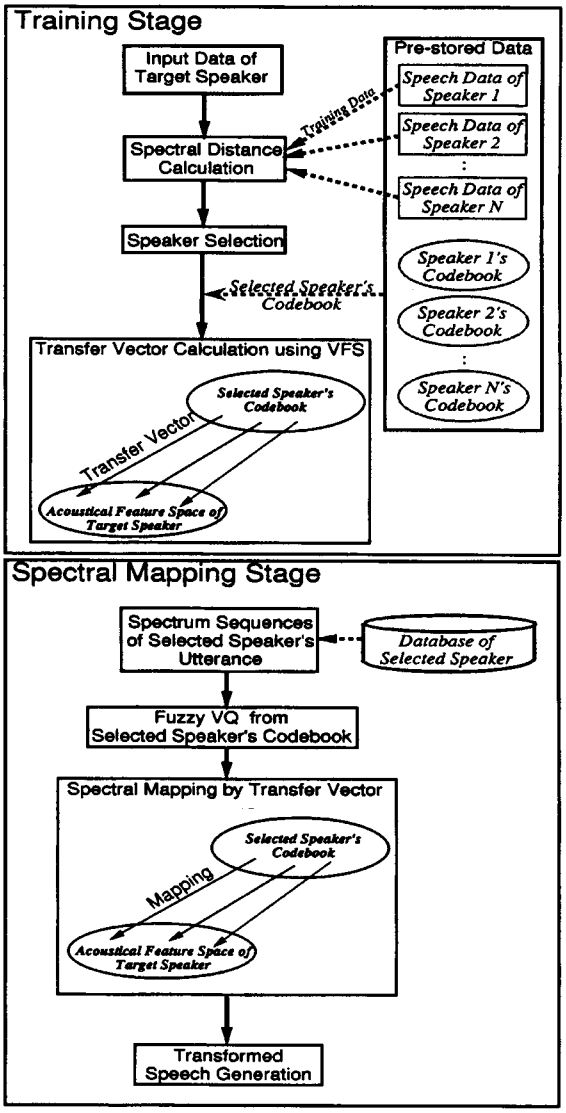

Figure 1: Block diagram of SSVFS

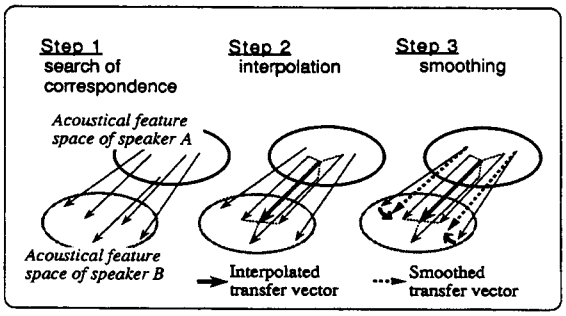

Figure 2: Principle of training procedure in VFS

neighbor principle, and smoothing for all transfer vectors is carried out to decrease estimation errors of the transfer vector.

To generate a measure for the selection of effective training samples for the VFS algorithm, the following two assumptions are used here.

Assumption 1 The coverage in vector space for a trained code vector is optimal and the coverage for an untrained code vector can be obtained from the distance between the untrained vector and the k-nearest neighbor trained vectors. In brief, the coverage in vector space grows as the distribution of trained codebook entries becomes broad.

Assumption 2 From Equation 1, the reliability of the trained entry increases as the number of corresponding vectors of the target speaker's speech increases.

From the above assumptions, measure C, the criterion for selecting effective training samples, can be obtained from the following function.

$$C = \frac{1}{N_{cb}}\{\sum_i C_{trained}(i) + \sum_j C_{untrained}(j)\} \quad (2)$$

where N_{cb} is the number of clusters of the codebook, $C_{trained}$ is the coverage in vector space for codebook entries matched in the training procedure, and $C_{untrained}$ is the coverage for entries which have not been matched. $C_{trained}$ and $C_{untrained}$ are defined as the following functions, respectively.

$$C_{trained}(i) = 1 - exp(-\alpha N_i) \quad (3)$$

$$C_{untrained}(j) = \max_k[exp(-\beta D_{j,k}) \times C_{trained}(k)] \quad (4)$$

where N_i is the number of target speaker's vectors that correspond to the ith code vector. $D_{j,k}$ is the Euclidean distance between the jth code vector and the k-nearest neighbor vectors, and α and β are weight parameters which control C.

Measure C means that the coverage in vector space for unmatched entries decreases as the distance between the untrained code vector and the nearest neighbor trained code vector increases. C is considered to reflect the distribution and reliability of the trained entries. When $\alpha = \infty$ and $\beta = \infty$, this means that the measure is only the ratio of the number of corresponding entries to N_{cb}.

4. EXPERIMENTAL CONDITIONS

An evaluation of the proposed measure was carried out using the ATR speech database [7],[8]. The coverage in vector space and the resemblance (to the target speaker) for seven different training words were calculated, and the evaluation was carried out based on the correlation between the coverage in vector space and the degree of resemblance. Training words are shown in Table 1. The resemblance was indicated by mean cepstrum distance over fifty test words between speech spoken by a target speaker and speech transformed from the selected speaker to that of the target speaker. The reference speakers were eight announcers (four males and four females) and the target speakers were eight other announcers (four males and four females). The closest reference speaker was decided using the word */uchiawase/*.

Table 1: Words used for training

/ai/, /ou/, /megane/, /ukeau/, /wagamama/, /kewashii/, /uchiawase/

Table 2: Experimental conditions

Speech Material	ATR speech database
Analysis Condition	Sampling frequency: 12kHz Window: Blackman (21.3ms) Frame shift: 5ms Feature: 30-dimensional FFT cepstrum
Mapping Condition	Target speaker: 4 males and 4 females Reference speaker: 4 males and 4 females Data used for codebook generation: 503 phonetically balanced sentences The number of clusters: 512 VFS k-nearest neighbors: 4

The spectrum codebooks of the pre-stored speakers were made from 503 phonetically balanced sentences in advance using ESPS toolkit [9]. The number of clusters was 512, and k for k-nearest neighbors was 4. The feature vector was a 30-dimensional cepstrum coefficients vector. The distance D of the cepstrum was calculated by the following equation.

$$D = \frac{1}{N_{fr}}\sum_{ij}(c^r_{ij} - c^t_{ij})^2 \quad (5)$$

where c^r_{ij} is the DTWed ith frame jth cepstrum of the reference speaker, c^t_{ij} is ith frame jth cepstrum of the target speaker, and N_{fr} is the number of frames. The experimental conditions are listed in Table 2.

5. EXPERIMENTAL RESULTS AND DISCUSSIONS

To confirm the validity of the proposed measure, the correlation coefficient between the coverage in vector space and the mean cepstrum distance was calculated, and compared with the correlation coefficient for the case where only the number of matched entries was used as a measure.

From a preliminary experiment, with $\alpha = 0.01$ and $\beta = 50$, many cases of high correlation were obtained. The correlation coefficients are shown in Table 3, and

Table 3: Correlation coefficients between coverage in vector space and mean cepstrum distance between target and transformed speech for 50 words

Reference Speaker	Target Speaker	Correlation Coefficient	
		Proposed measure (α=0.01, β=50)	Only the number of corresponding codes
MMY	MAU	-0.96	-0.98
MHT	MXM	-0.60	-0.27
MSH	MTT	-0.84	-0.67
MTK	MTM	-0.77	-0.75
FKS	FAF	-0.96	-0.56
FKN	FYN	-0.62	-0.48
FTK	FSU	-0.76	-0.89
FYM	FFS	-0.43	-0.14

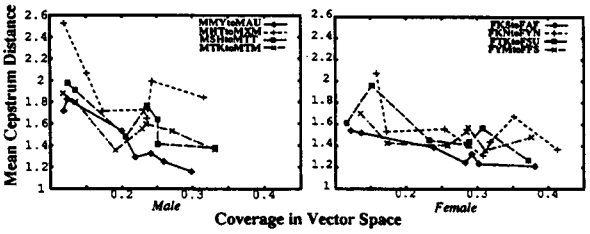

Figure 3: Proposed coverage in vector space *vs.* Mean cepstrum distance between target and transformed speech for 50 words ($\alpha = 0.01, \beta = 50$)

the relation between the mean cepstrum distance and the coverage in vector space is shown in Fig.3. The following results can be observed: 1) From Table 3, the proposed measure which takes account of the coverage in vector space for entries which are not matched and the reliability for corresponding entries shows higher correlation between the measurement and the cepstrum distance than the correlation for a measure based only on the number of corresponding entries. 2) From Table 3 and Fig.3, the correlation between the proposed measure of the coverage and the cepstrum distance is high in most cases.

6. CONCLUSIONS

A criterion for selecting effective training samples for SSVFS spectral mapping method combining Speaker Selection and VFS was tested. Coverage in the vector space was defined as taking account of the training procedure used in SSVFS.

To evaluate the definition of the coverage in vector space, a correlation coefficient for eight target speakers (four males and four females) between a measurement of the coverage and mean cepstrum distance over fifty test words was calculated and compared with a measurement based only on the number of codebook entries which corresponded in the training procedure. The following results were obtained:

- The proposed measure which takes account of the coverage in vector space for entries which are not matched by a corresponding vector of the target speaker's speech in the training, and the reliability for corresponding entries has higher correlation than the correlation for a measure defined based only on the number of corresponding entries.
- The correlation for the proposed measure is high in most cases.

These results show that the proposed measure is effective for the selection of training samples for SSVFS and also for the estimation of the degree to which learning is carried out. It may, however, be necessary to improve the precision of the measure because some cases still show a low correlation.

Future work involves building a full voice conversion system using SSVFS. Moreover, a study of mapping techniques for other acoustical features and other languages is needed.

[ACKNOWLEDGMENTS]
The authors wish to thank Y. Yamazaki for his support. We would also like to acknowledge Dr. Nick Campbell and other ITL colleagues for their useful advice.

References

[1] M. Hashimoto and N. Higuchi: "Spectral Mapping for Voice Conversion Using Speaker Selection and Vector Field Smoothing", *Proc. EUROSPEECH'95, pp.431-434, 1995.*

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

[3] N. Iwahashi and Y. Sagisaka : "Speech spectrum conversion based on speaker interpolation and multi-functional representation with weighting by radial basis function networks", *SPEECH COMMUNICATION, Vol.16, pp.139-151, 1995.*

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

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

[6] H. Matsumoto, Y. Maruyama and H. Inoue: "Voice quality conversion based on supervised/unsupervised spectral mapping", *J. Acoustical Society of Japan, Vol.50, No.7, pp.549-555, 1994.*(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] Entropic Research Lab., Inc.: "ESPS programs", *ESPS version 5.0 manual, 1993.*