An improved spectrum mapping applied to speaker adaptive Korean word recognition

Hiroshi Matsumoto [1], Yong-Ju Lee [2], Hoi-Rim Kim [2] and Ken'iti Kido [3]

[1] Dept. of Electrical and Electronic Eng., Shinshu University (Japan),
[2] Electrical and Telecommunications Research Institute (Korea), and
[3] Dept. of Information Eng., Chiba Institute of Technology (Japan)

ABSTRACT This paper improves the previously proposed spectral mapping method for supervised speaker adaptation in which a mapped spectrum is interpolated from speaker difference vectors at typical spectra based on a minimized distortion criterion. In estimating these difference vectors, it is important to find an appropriate number of typical points. The previous method empirically adjusts the number of typical points, while the present method optimizes the effective number by rank reduction of normal equation. This algorithm was applied to a supervised speaker adaptation for Korean word recognition using the templates from a prototype male speaker. The result showed that the rank reduction technique not only can automatically determine an optimal number of code vectors, but also slightly improves the recognition scores compared with those obtained by the previous method.

1. INTRODUCTION

Speaker adaptation or normalization is one of the important issues in a large-vocabulary speech recognition system for use by unrestricted speakers. The inter-speaker differences in static characteristics of spectra as well as in dynamic ones is a major factor which degrades speech recognition performance. In order to eliminate the effects of the inter-speaker differences, the vector quantization-based spectral mapping techniques have been successfully applied to speaker adaptation in speech recognition [1][2]. However, these methods not only distort the mapped spectra due to vector quantization, but also require relatively large amount of training data to obtain reliable spectral correspondence between the input and target code vectors.

In order to improve these defects, we have previously proposed a spectral mapping method based on interpolation of speaker difference vectors at typical spectral points, i.e., code vectors, under a minimized spectral distortion criterion. This mapping method includes the codebook mapping method by Shikano et al. as a special case [3]. Although this method is applicable to a small amount of training data, it is required to adjust empirically the optimal number of typical points depending on both amount and content of training data.

The present paper makes further improvement the previous method so that both number of typical points and the spectral distortion are simultaneously minimized for a given training data by the rank reduction of normal equation. Following the description of algorithm, the effect of rank reduction is examined through the supervised speaker adaptation in small-vocabulary Korean word recognition experiments.

2. SUPERVISED SPECTRAL MAPPING

2.1. Spectral Mapping Model

Considering that inter-speaker differences between speakers consist of phoneme-dependent ones in addition to phoneme-independent ones, the present method maps an input spectral vector onto that of the input speaker by interpolating the estimated speaker difference vectors, $\{ \Delta_1, \Delta_2, \ldots, \Delta_M \}$, between the reference and input speakers at the typical points, $\{ \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_M \}$, in the reference spectral space:

$$\hat{\mathbf{y}}_i = \mathbf{x}_i + \sum_{k=1}^{M} w_{ik} \Delta_k \,, \tag{1}$$

where the weighting coefficient w_{ik} is determined by the distance $d\,(\mathbf{x}_j\,,\,\mathbf{v}_k)$ between $\mathbf{x}_i$ and $\mathbf{v}_k$ as follows:

$$w_{ik} = \frac{d\,(\mathbf{x}_i\,,\,\mathbf{v}_k)^{-p}}{\sum_{k=1}^{M} d\,(\mathbf{x}_i\,,\,\mathbf{v}_k)^{-p}} \,. \tag{2}$$

The parameter p, which will be referred to as an interpolation parameter, adjusts the continuity of w_{ik} as a function of $d\,(\mathbf{x}_i\,,\,\mathbf{v}_k)$. This interpolation formula for p=1.0 was first proposed by Y.Niimi, et al. [4][5], and was generalized by introducing the smoothing parameter p in the unsupervised adaptation method [6].

2.2. Mapping Algorithm

Given the spectral sequences $\{\mathbf{x}_i\}$ and $\{\mathbf{y}_j\}$ for a source and target speakers as training samples, the mapped spectrum $\hat{\mathbf{y}}_i$ from the source to the target spectra is estimated by minimizing the following spectral distortion along the spectral correspondence, $\{c(l)\}=\{i(l),j(l)\}$:

$$J\,(c(1)\,,\,\ldots\,,\,c(N),\Delta_1\,,\,\ldots\,,\,\Delta_M) = \sum_{l=1}^{N} \| \hat{\mathbf{y}}_{i(l)} - \mathbf{y}_{j(l)} \|^2 \,. \tag{3}$$

This minimization with respect to $\{ c(l) \}$ and $\{ \Delta_k \}$ is solved iteratively fixing one of these variable sets in turn. With the fixed $\{ \Delta_k \}$, the best match correspondence $\{ c(l) \}$ is obtained by means of a dynamic time warping (DTW). With the fixed $\{ c(l) \}$, the speaker difference vectors are given by the following normal equation:

$$\sum_{r=1}^{M} W_{kr} \cdot \Delta_r = \mathbf{E}_k \,, \qquad (\, k = 1\,,\,\ldots\,,\,M\,) \tag{4}$$

where

$$W_{kr} = \sum_{l=1}^{N} w_{ik} \cdot w_{ir} \,, \tag{5}$$

$$\mathbf{E}_k = \sum_{l=1}^{N} w_{i(l)k} \cdot (\, \mathbf{x}_{i(l)} - \mathbf{y}_{j(l)} \,) \,. \tag{6}$$

In solving the normal equation (4), it is important to choose an optimal number of typical points for given training samples, since too many typical points might result

in an ill conditioned inverse matrix of **W**. Instead of adjusting the set of typical points, a rank reduction technique [7] can be applied to get an effective number of typical points as follows. The normal equations (4) are rewritten in matrix-vector form as

$$\mathbf{W}\Delta^{(j)} = \mathbf{E}^{(j)}, \qquad (j = 1, \ldots, p) \tag{7}$$

with

$$\mathbf{W} = [\, W_{kr} \,]_{M\times M}, \tag{8}$$

$$\Delta^{(j)} = [d_{1j}, d_{2j}, \ldots, d_{Mj}]^T, \tag{9}$$

$$\mathbf{E}^{(j)} = [e_{1j}, e_{2j}, \ldots, e_{Mj}]^T, \tag{10}$$

where d_{kj} and e_{kj} are the jth elements of Δ_k and $\mathbf{E}_k$, respectively. The $M\times M$ real symmetric matrix **W** is decomposed as

$$\mathbf{W} = \mathbf{S}\cdot\Lambda\cdot\mathbf{S}^T \tag{11}$$

with

$$\mathbf{S} = [\mathbf{u}_1, \mathbf{u}_2, \cdots \mathbf{u}_M] \tag{12}$$

$$\Lambda = diag[\lambda_1, \lambda_2, \cdots \lambda_M] \tag{13}$$

where $\mathbf{u}_r$ is the eigenvector of **W** associated with the rth largest eigenvalue λ_r. Then, by neglecting nonsignificant eigenvalues $\{\lambda_{r+1}, \lambda_{r+2}, \cdots \lambda_M\}$, the minimum-norm linear least square estimate of $\Delta^{(j)}$ is obtained by

$$\Delta^{(j)} = \mathbf{S}^T\cdot\Lambda_r^{-1}\cdot\mathbf{S}\cdot\mathbf{E}^{(j)} \tag{14}$$

where

$$\Lambda_r^{-1} = diag[\, 1/\lambda_1, \cdots 1/\lambda_r, 0, \cdots 0] \tag{15}$$

and the rank r of **W** is determined by a threshold θ such that

$$\lambda_r \geq \theta\lambda_1 > \lambda_{r+1}. \tag{16}$$

The rank reduction threshold θ will be experimentally determined.

2.3. Relationship to Codebook Mapping

As a special case, if the spectra $\mathbf{x}_i$ from reference speaker had been quantized using the set of typical points, $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_M\}$, as a codebook, the weighting coefficient w_{ik} turns to δ_{kr}, which is the Kronecker delta. Thus, the mapped typical points $\hat{\mathbf{v}}_k$ is given by

$$\hat{\mathbf{v}}_k = \sum_{l=1}^{N} \delta_{jk}\cdot\mathbf{y}_{j(l)} \;/\; \sum_{l=1}^{N} \delta_{jk}\,. \tag{17}$$

Therefore, if y_j was quantized with the codebook for the target speaker, $\hat{v}_k$ is equivalent to the mapped code vector by Shikano [6].

3. EVALUATION

3.1. Data Base and Recognition Procedure

The speech data consist of four repetitions of 40 words uttered by 10 male speakers and 10 female speakers. The speech signals were digitized at a sampling frequency of 16 kHz with the frequency band limited to 7 kHz. The linear predictive autocorrelation method with 24 poles was applied to these speech data with a constant frame of 20 ms, a frame shift of 10 ms, first-order backward differences for preemphasis, and a Hamming window. The 1st to 16th LPC mel-cepstral coefficients were used as the components of the feature vector. The 1- to 256-vector codebooks from the prototype speaker were obtained by the standard LBG algorithm using the cepstral distortion measure [8]. The set of typical points, $\{ v_k \}$, consisted of the nearest codeword in the 128-vector codebook to each codeword in a small-size codebook of M vectors.

The performance of the spectral adaptation is evaluated by recognition rates in DTW-based word recognition tests. In the speaker independent and the speaker adaptive word recognition, the patterns were provided by the first or the second utterances from a male speaker whose reference templates give the highest average recognition scores in cross speaker word recognition for 10 male speakers. In the subsequent experiments, the first and third utterances of each word were used as the training samples for adaptation or as the speaker dependent references. The second and the fourth utterances were used as test samples for the first and the third utterances as training or reference samples, respectively. Training sets of different duration were prepared by dividing the 40 word data into 5-, 10-, 20- and 30-word subsets, which provide 8, 4, 2 and 2 training sets, respectively.

3.2. Effect of Rank Reduction

In this section, the test speakers are the five female and five male speakers whose samples gave the lowest recognition scores in speaker independent word recognition. First, the optimum value for the interpolation parameter, p, was examined by supervised speaker adaptation experiments under the condition of M = 128 using the 30-word training sets for each speaker. As shown in figure 1, although the effect of p on the recognition scores is not clear, the scores tend to be highest around p = 2.0 for the male speakers. Therefore, in the subsequent experiments, the value of p will be set to 2.0 for computational simplicity.

The effect of the number of typical points was examined using the 5-, 10-, 20- and 30-word subsets of the training samples. Figures 2 shows the results without rank reduction, i.e., the previous method.

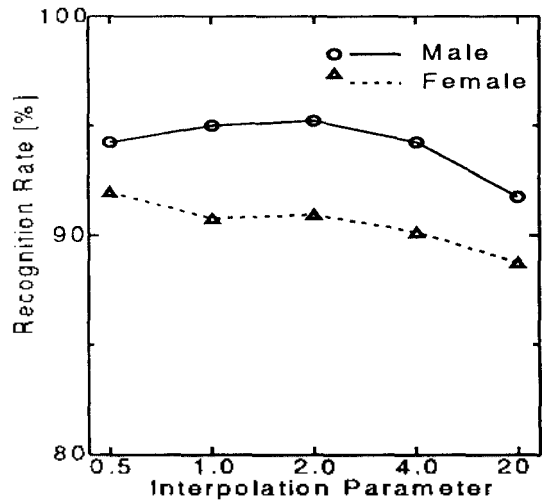

Fig.1 The average recognition rates for the 5 male and 5 female speakers as a function of the interpolation parameter p with M=128 and no rank reduction using 30 words of training data.

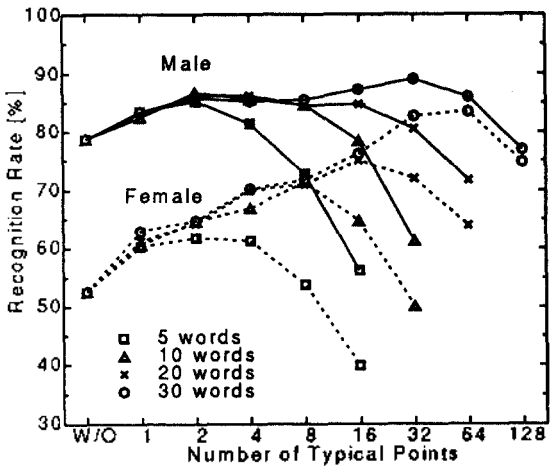

Fig.2 The average recognition rates for the 5 male and 5 female speakers as a function of the number of typical points with p=2.0 and no rank reduction for the two sets of training words.

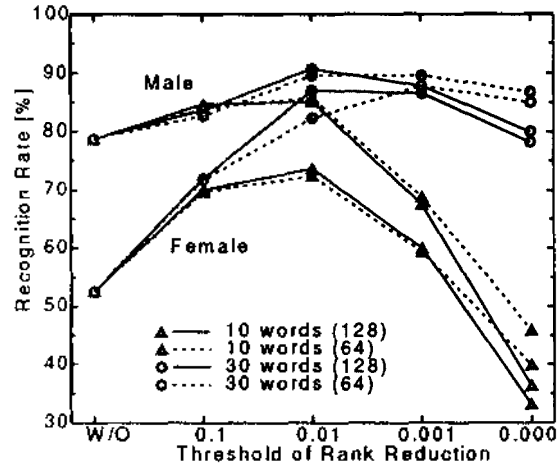

Fig.3 The average recognition rates for the 5 male and 5 female speakers as a function of the rank reduction threshold with M=64 or 128 for the two sets of training words.

The average recognition rates for the male and the female speakers are improved as the number of the typical points increases, but tends to saturate or decrease beyond a small number of points for the shorter training samples. Thus, the number of typical points should be determined depending on the amount of training samples. Therefore, the effect of rank reduction was examined with the fixed number of typical points, which are set to 64 or 128. Figure 3 shows the average scores as a function of the threshold θ with the 10- and 30-word training sets. The rank reduction from the 128 typical points attains the highest recognition score at θ=0.01 independently of speaker's sex and the training data sets. In addition, these recognition scores are slightly higher than those obtained using the 64 typical points and also the best scores with full rank in figures 2.

3.3. Comparative Experiments

The recognition experiments were carried out for the 9 male speakers and the 10 female speakers under the following conditions:
(1) without adaptation,
(2) with supervised adaptation, and
(3) with speaker dependent references.
In the supervised speaker adaptation, the parameters M and θ were set to 128 and 0.01, respectively.

Figure 4 summarizes the average recognition rates as a function of the amount of training samples. In the supervised adaptation, the average recognition rate for the male speakers increased from 80.2% for no adaptation to 90.3% when all of the 40-word data is used. For the female speakers, the supervised speaker adaptation significantly improved the average

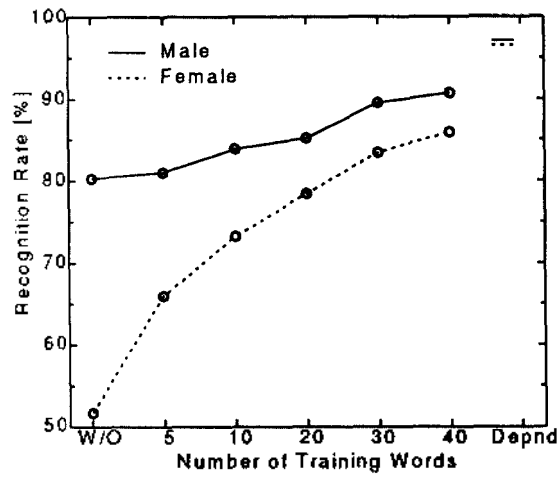

Fig.4 Average correct scores for the male and female speaker in speaker independent, speaker dependent, speaker adaptive recognition experiments with various amounts of training data.

recognition scores from 52.0% without adaptation to 86.0% for all the training data. However, the best scores in the supervised adaptation are still lower by 6 % for the male speakers and by 10 % for the female speakers than those in the speaker independent case. The inspection of recognition errors reveals that the most of errors are between very similar words, for instance, /eum?/ and /eum/, /il/ and /ye/, /nyuk/ and /yuk/, and /nyuk/ and /ryuk/. These errors seem to be mostly caused by inter-speaker differences in dynamic spectral characteristics.

4. CONCLUSION

This paper has presented an improved spectral mapping method by speaker difference interpolation with rank reduction for supervised speaker adaptation. The results of DTW-based word recognition experiments show that a minimized distortion criterion with the rank reduction not only improves the recognition scores, but also automatically determines the effective number of typical points depending on the training data.

In future work, it is important to improve the adaptation accuracy as well as to develop a method to eliminate speaker differences in dynamic characteristics due to coarticulation.

REFERENCES

1 K. Shikano, "Speaker adaptation through vector quantization," Proc. ICASSP-86, 2643-2646 (1986).

2 S. Nakamura and K. Shikano, "Spectrogram normalization using fuzzy vector quantization," J. Acoust. Soc. Jpn. (J) **45**, 107-114 (1989) (in Japanese).

3 H. Matsumoto and H. Inoue, "A minimum distortion spectral mapping applied to voice quality conversion," Proc. of ICSLP, 161-164 (1990).

4 Y. Niimi and Y. Kobayashi, "Speaker-adaptation of a code book of vector quantization," Proc. of European Conference on Speech Technology, Sep. 1987.

5 Y.Shiraki and M.Honda, "Piecewise-linear adaptive vector quantization and its application to speaker adaptation," Acoust. Soc. Jpn. Meeting in Spring, No.1-1-4, Mar. 1977 (in Japanese).

6 Y.Yamashita and H. Matsumoto, "Speaker adaptation for word recognition based on error vectors in VQ," IEICE Tech. Rep. SP87-118, 35-42 (1988) (in Japanese).

7 K. Konstantinos and K. Yao, "Statistical analysis of effective singular values in matrix rank determination," IEEE trans. on Acoustics, Speech, and Signal Processing, Vol.36, No.5, 757-763 (1988).

8 Y.Linde, A.Buzo and R.M.Gray, "An algorithm for vector quantizer design," IEEE trans. on Commun., **COM-28, 1**, 84-95 (1980).