Speaker Adaptation Using ICA-Based Feature Transformation

Ho-Young Jung, Mansoo Park, Hoi-Rin Kim, and Minsoo Hahn

ABSTRACT—— Speaker adaptation techniques are generally used to reduce speaker differences in speech recognition. In this work, we focus on the features fitted to a linear regression-based speaker adaptation. These are obtained by feature transformation based on independent component analysis (ICA), and the feature transformation matrices are estimated from the training data and adaptation data. Since the adaptation data is not sufficient to reliably estimate the ICA-based feature transformation matrix, it is necessary to adjust the ICA-based feature transformation matrix estimated from a new speaker utterance. To cope with this problem, we propose a smoothing method through a linear interpolation between the speaker-independent (SI) feature transformation matrix and the speaker-dependent (SD) feature transformation matrix. From our experiments, we observed that the proposed method is more effective in the mismatched case. In the mismatched case, the adaptation performance is improved because the smoothed feature transformation matrix makes speaker adaptation using noisy speech more robust.

I. INTRODUCTION

An adaptation technique is useful in a speech recognition system because it reduces mismatches of environments or speakers. To solve the problem of different speaker variations, the maximum a posteriori (MAP) technique or maximum likelihood linear regression (MLLR) technique [1]-[3] has been widely used. These methods focus on modifying the model parameters, specifically Gaussian mean vectors in hidden Markov model (HMM) states. With a large amount of adaptation data, MAP can adapt model parameters to be converged to the corresponding SD model parameters. However, when there is only a small amount of adaptation data, this adaptation method does not show a good performance. On the other hand, MLLR is helpful when there is not sufficient adaptation data. The MLLR proposed by Leggetter and Woodland [4] applies a linear transformation on HMM model parameters that are usually Gaussian mean vectors. We use the MLLR technique to adapt a new condition, a new speaker configuration.

In this work, the proposed feature transformation technique, based on independent component analysis (ICA)[5], is used to provide the features fitted to speaker adaptation using a linear regression frame. It can make the estimation of the feature transformation matrix more robust even when there is a small amount of adaptation data.

II. ICA-BASED FEATURE TRANSFORMATION

1. The Infomax Algorithm for ICA

ICA is one of the solutions for blind source separation (BSS) [6] that recovers independent sources given only sensor observations that are linear mixtures of the independent sources. Blind means that both the sources and the mixing channel are unknown. ICA finds the unmixing system so that the output signals are statistically independent. In other words, ICA performs a linear transformation that makes the observed signals as statistically independent of each other as possible.

Figure 1 shows the infomax algorithm [7], which is a simple learning algorithm that recovers independent sources using information maximization. $g(\mathbf{u})$ denotes a non-linearity

Manuscript received June 19, 2002; revised Oct. 7, 2002.

Ho-Young Jung (phone: +82 42 860 1328, e-mail: hjung@etri.re.kr) is with Speech Technology Research Center, ETRI, Daejeon, Korea.

Mansoo Park (e-mail: mansoo@icu.ac.kr), Hoi-Rin Kim (e-mail: hrkim@icu.ac.kr), and Minsoo Hahn (e-mail: mshahn@icu.ac.kr) are with School of Engineering, Information and Communications University, Daejeon, Korea.

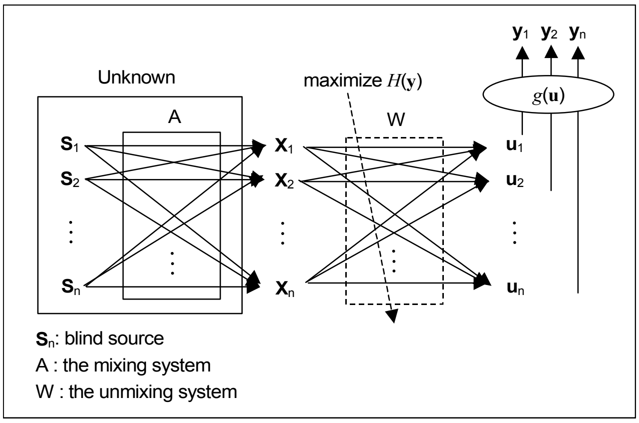

Fig. 1. Infomax learning rule process.

mapping function that makes the probability distribution of the output uniform and maximizes the joint entropy $H(\mathbf{y})$ of the output, that is, minimizes the mutual information of the output.

2. ICA-Based Feature Transformation

The assumptions of ICA conform to the framework of homomorphic analysis [5], [8]. The cepstrum vectors have the property of summation such that the functions of the glottal pulse, vocal tract, mouth radiation, and transmission line distortion can be added in the log-spectrum domain [5]. Since these filters result in a different dependency among cepstrum components for each speaker, the final cepstral feature may increase the complexity of the speaker adaptation problem. Thus, we would like to exclude a useless variation and to simplify the problem by disentangling this unknown dependency. The cepstral feature is orthogonal, and thus is comparatively effective for an independent projection. It is formulated by the following term, $\mathbf{C'} = \mathbf{W} \cdot \mathbf{C}$, where $\mathbf{C'}$ is a projected feature vector, $\mathbf{C}$ is an original feature vector, and $\mathbf{W}$ is an ICA-based feature transformation matrix.

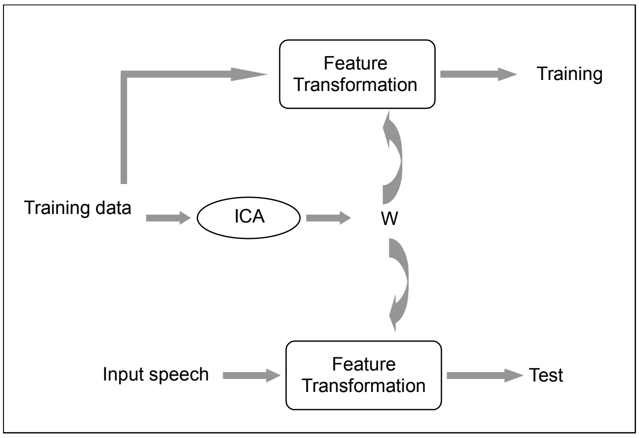

Fig. 2. ICA-based feature transformation process.

Figure 2 shows the ICA-based feature transformation process in an SI recognition system that can be achieved with the feature transformation matrix estimated from speaker independent training data.

III. THE PROPOSED FEATURE TRANSFORMATION FITTED TO SPEAKER ADAPTATION

Independent feature vectors can be found by applying ICA on the cepstrum domain. In general, the independent feature vectors enhance the pattern classification capability, and furthermore these characteristics will be more effective for a particular purpose, such as speaker or noise adaptation, by containing environmental cues without complex dependency among feature components. In the MLLR approach with a full regression matrix, this may solve an inaccurate estimation of off-diagonal terms relating the interdependencies among components that is due to a small amount of adaptation data. Particularly, an ICA-based feature is very effective with the MLLR approach with a diagonal regression matrix. Compared to a full regression matrix, the diagonal regression matrix considerably reduces the computational load, but yields a low performance in conventional features [4]. The proposed ICA-based independent feature can satisfy the assumption of diagonal regression and may be profitable to an online adaptation requiring a small amount of adaptation data.

The transformation matrix $\mathbf{W}_{\mathrm{SD}}$ is estimated from adaptation data of a new speaker (Fig. 3). However, it may not always be reliable due to information loss by the biased-specific condition corresponding to a small amount of adaptation data. To cope with this problem, we adjusted the W_{SD} transformation matrix for adaptation and recognition and proposed a smoothing method by a linear interpolation between $\mathbf{W}_{\mathrm{SI}}$ and $\mathbf{W}_{\mathrm{SD}}$,

$$\mathbf{W}_{\mathrm{smooth}} = (1-\alpha)\mathbf{W}_{\mathrm{SI}} + \alpha\mathbf{W}_{\mathrm{SD}}, \quad 0 \le \alpha \le 1. \qquad (1)$$

A loss of important information can be avoided because the interpolation process can make the estimation of the feature transformation matrix more robust even when there is a small amount of adaptation data. The smoothed feature transformation matrix also makes speaker adaptation by noisy speech more robust.

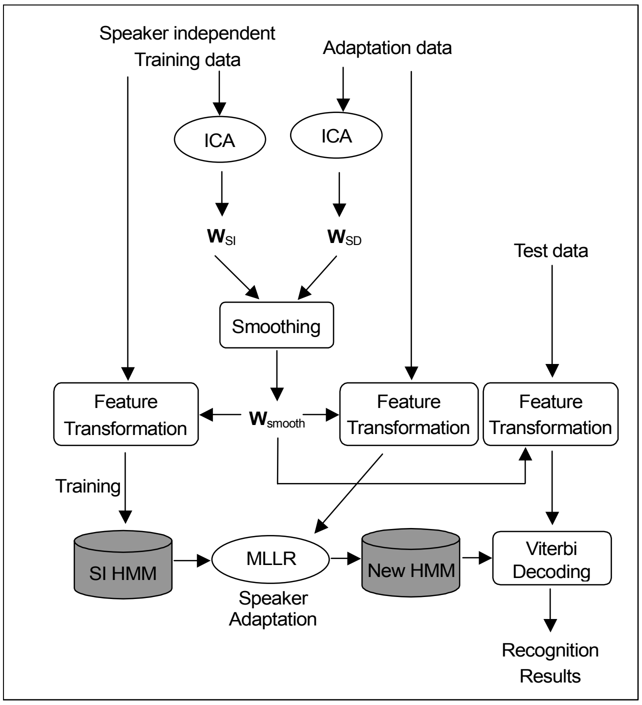

Fig. 3. Speaker adaptation process using the smoothed feature transformation matrix.

IV. EXPERIMENTS

1. Database

PBW452 DB consists of 452 Korean isolated words spoken two times by 70 speakers, 38 males and 32 females. The speech signal was sampled at 16 kHz and quantized with 16 bits and was down-sampled to 8 kHz to cope with the telecommunication channel. The speech data obtained from 63 speakers were used for training, and the remainder was reserved for speaker adaptation and performance evaluation. The first utterances of each speaker among the 7 reserved speakers, including 4 males and 3 females, were used as adaptation data, and the second utterances were for evaluation. To estimate the SI feature transformation matrix $\mathbf{W}_{SI}$, we used 6300 words from the training data. They consisted of 100 words of each speaker. The SD feature transformation matrix $\mathbf{W}_{SD}$ was obtained from all the adaptation data of each speaker.

The speech signal was parameterized as the 39th order feature vectors every frame with an 8 ms rate. In the baseline system, the feature vector consisted of 12 mel frequency cepstral coefficients (MFCC), their first-order temporal regression coefficients, their second-order temporal regression coefficients, and the log-energy and its first- and second-order temporal regression coefficients. In order to remove the recording condition dependencies, we performed cepstral mean normalization (CMN). Then, the ICA-based projected feature was the 12th order feature vectors obtained from the MFCC-CMN, their first-order temporal regression coefficients, their second-order temporal regression coefficients, and the log-energy and its first- and second-order temporal regression coefficients. In this work, model parameters are based on a three-states left-to-right continuous density HMM, and tri-phone unit with 1 mixture.

2. Speaker Adaptation in the Matched Case: Clean Speech

In this experiment, we compared the performance of speaker adaptation according to different feature transformation matrices. From the initial experiments, the SI recognition performance with no feature transformation or no speaker adaptation was 98.7%. The feature transformation technique with $\mathbf{W}_{SI}$ was applied to the training process, and the smoothed feature transformation matrix was applied to the speaker adaptation and test processes.

Table 1 shows that the best result was obtained by the SI or smoothed feature transformation matrix based on ICA. Actually, the smoothing process is insignificant since there are no odd independent components in the matched case. In the case of PCA-based feature transformation, it is worse than the baseline system, and the smoothing method is not necessary because the eigenvector matrix to find the principal axes in the matched case is estimated from sufficient training data.

Strictly speaking, it is hard to say that the effectiveness of PCA or ICA is absolutely good since the performance of speaker adaptation in the baseline system is sufficiently high only when clean speech data are used in the matched case.

Table 1. Average error reduction rate (ERR) with the smoothed feature transformation matrix in the matched case.

Feature Transformation	Accuracy	Average ERR
MLLR only	99.46%	-
PCA-based $\mathbf{W}_{SI}$	99.43%	–5.6%
ICA-based $\mathbf{W}_{SI}$	99.59%	24.1%
ICA-based $\mathbf{W}_{smooth}$ ($\alpha = 0.25$)	99.59%	24.1%

3. Speaker Adaptation in the Mismatched Case: Clean Speech vs. Noisy Speech (SNR: 15, 10, and 5 dB)

This experiment process is similar to the one above. Just additive white Gaussian noise (AWGN) was added to the test and adaptation data. Here, the signal to noise ratio (SNR) of the test and adaptation data was about 15 dB, 10 dB, and 5 dB. The results were similar to those in Table 1. Tables 2 to 4 show the adaptation results according to each noise level. The proposed feature transformation technique is more effective in the mismatched case with a 15 dB and 10 dB SNR. The optimum interpolation factor, α, was 0.73 in 15 dB and 0.67 in 10 dB. However, Table 4 shows that PCA-based feature transformation is a little better than ICA-based smoothing feature transformation in 5 dB. In the case of PCA-based feature transformation, the smoothing feature transformation method is much worse than the SI feature transformation method since there are different principal axes in the mismatched case, that is, the principal axes are different because of noise components. On the other hand, in the case of ICA-based feature transformation, the smoothing feature transformation method is significant since there is an odd independent component which is noise in the mismatched case.

Table 2. Average ERR with the smoothed feature transformation matrix in the mismatched case. (SNR: 15 dB)

Feature Transformation	Accuracy	Average ERR
MLLR only	96.05%	-
PCA-based $\mathbf{W}_{SI}$	97.34%	32.7%
ICA-based $\mathbf{W}_{smooth}$ (α= 0.73)	97.44%	35.2%

Table 3. Average ERR with the smoothed feature transformation matrix in the mismatched case. (SNR: 10 dB)

Feature Transformation	Accuracy	Average ERR
MLLR only	90.65%	-
PCA-based $\mathbf{W}_{SI}$	92.38%	18.5%
ICA-based $\mathbf{W}_{smooth}$ (α= 0.67)	92.64%	21.3%

Table 4. Average ERR with the smoothed feature transformation matrix in the mismatched case. (SNR: 5 dB)

Feature Transformation	Accuracy	Average ERR
MLLR only	75.76%	-
PCA-based $\mathbf{W}_{SI}$	78.60%	11.7%
ICA-based $\mathbf{W}_{smooth}$ (α= 0.55)	78.45%	11.1%

V. CONCLUSIONS

In this work, we proposed an ICA-based feature transformation method to obtain more effective features for speaker adaptation. This is also helpful to suit the feature parameter to speaker-independent speech recognition [9] and the HMM aligner with a diagonal covariance matrix [10]. An ICA-based feature transformation can overall improve the speaker adaptation performance even if the PCA-based feature transformation is a little better in the higher mismatched case with an SNR of 5 dB. According to the experiment results, we can say that the proposed technique, ICA-based smoothing feature transformation, is overall effective in finding the adaptation-fitted features in both matched and mismatched cases. A linear interpolation between the SI and the SD transformation matrices to avoid the loss of non-speaker-specific information in the mismatched case is needed.

REFERENCES

[1] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing*, Prentice Hall PTR, 2001.

[2] Steve Young, *The HTK BOOK* (*for HTK Version* 3.0), 2000.

[3] Sam-Joo Doh, *Enhancements to Transformation-Based Speaker Adaptation: Principal Component and Inter-Class Maximum Likelihood Linear Regression*, Ph.D. Thesis, Carnegie Mellon University, 2000.

[4] C.J. Leggetter and P.C. Woodland, “Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models,” *Computer Speech and Language*, vol. 9, 1995, pp. 171-185.

[5] Gil-Jin Jang, Seong-Jin Yun, and Yung-Hwan Oh, “Feature Vector Transformation Using Independent Component Analysis and Its application to speaker identification,” *Proc. of EUROSPEECH*, 1999, pp. 767-770.

[6] J.-F. Cardoso, “Blind Signal Separation: Statistical principles,” *Proc. of IEEE*, vol. 86, Oct. 1998, pp. 2009-2025.

[7] Te-Won Lee, *Independent Component Analysis: Theory and Applications*, Boston, MA, Kluwer, 1998.

[8] L. Potamitis, Fakotakis, and G. Kokkinakis, “Independent Component Analysis Applied to Feature Extraction for Robust Automatic Speech Recognition,” *Electronics Lett.*, vol. 36, no. 23, 2000, pp. 1977-1978.

[9] Youngjik Lee and Kyu-Woong Hwang, “Selecting Good Speech Features for Recognition,” *ETRI J.*, vol. 18, no. 1, Apr. 1996, pp. 29-40.

[10] Sanghun Kim, Youngjik Lee, and Keikichi Hirose, “Unit Generation Based on Phrase Break Strength and Pruning for Corpus-Based Text-to-Speech,” *ETRI J.*, vol. 23, no. 4, Dec. 2001, pp. 168-176.