UNSUPERVISED AND INCREMENTAL SPEAKER ADAPTATION UNDER ADVERSE ENVIRONMENTAL CONDITIONS

Keizaburo TAKAGI, Koichi SHINODA, Hiroaki HATTORI, and Takao WATANABE

Information Technology Research Laboratories, NEC Corporation
4-1-1 Miyazaki, Miyamae-ku, Kawasaki 216, JAPAN

ABSTRACT

A new speaker adaptation method is described. In practical applications of speaker adaptation, adaptation and testing environments change significantly and are unknown beforehand. In such cases, since the speaker adaptation adapts a reference pattern to the adaptation utterances in regard to differences in both environment and speaker at the same time, performance in speaker adaptation would be degraded. To cope with this problem, our proposed method first eliminates the environmental differences between each input utterance and a reference pattern by using a rapid environment adaptation algorithm based on spectrum equalization (REALISE) [2]. Then we apply an unsupervised and incremental speaker adaptation with autonomous control using tree structure pdfs (ACTS) [1] to the environmentally adapted reference pattern. By combining these two methods, the resulting system is expected to perform well under adverse environmental conditions and to show a stable improvement regardless of the amount of adaptation data. Evaluation experiments were carried out for utterances under three vehicle speed conditions. Recognition rates for a 100-Japanese-word recognition task after 100-word adaptation were improved from 92% (ACTS alone) to 95% (proposed method).

1. INTRODUCTION

Various speaker-independent (SI) speech recognition systems have widely been studied in recent years because they show good performance on average owing to their capability of including a wide variety of speaker individualities. However, their performance is still lower than that of well-trained speaker-dependent (SD) speech recognition systems. Speaker adaptation technique has been one breakthrough regarding this problem, and has been applied alongside SI speech recognition systems. By using a small number of adaptation utterances from a new speaker, speaker adaptation shows stable improvement for a long period of time.

We previously proposed a speaker adaptation method with autonomous control using tree structure probability density functions (ACTS) [1]. This method autonomously controlled adaptation complexity on the basis of the amount of adaptation data by using tree structure probability density functions (pdfs), and showed constant, stable improvement regardless of the amount of available adaptation data. Therefore, it is expected to perform well when applied in an unsupervised and incremental speaker adaptation framework that enables testing utterances themselves to be used for adaptation. In this study, unsupervised adaptation means that the speech recognition system itself determines the supervising signal by using the system's recognition results. A resulting speech recognition / adaptation system would not require any extra adaptation utterances beforehand.

In practical applications of such methods, however, testing and adaptation environments may change significantly for each utterance and are unknown beforehand (for example, in speech recognition / adaptation in mobile communications). In such cases, since speaker adaptation adapts a reference pattern to the adaptation utterances in regard to differences in both environment and speaker at the same time, performance would be degraded. The environmental sources which degrade performance in speaker adaptation and speech recognition are classified into two types: additive noise and multiplicative noise. Therefore, a new speaker adaptation method is required, which first eliminates the environmental differences between each utterance and a reference pattern in regard to both types of noises, and then adapts the environmentally adapted reference pattern to a new speaker. In preceding studies [3][4], the cepstral bias removal (CBR) technique was first applied to adaptation utterances as environment normalization, and then speaker adaptation was carried out on the environmentally normalized utterances. Although CBR eliminates differences in multiplicative noises, it cannot cope with additive noise.

In this paper, we propose a speaker adaptation method which first eliminates the environmental differences between each testing utterance and a reference pattern by using a rapid environment adaptation algorithm based on spectrum equalization (REALISE) [2]. REALISE assumes that clean speech X is contaminated by multiplicative noise A and additive noise B, as AX+B in the spectral domain. It then eliminates differences in the two types of noise between each single utterance and a reference pattern, thus eliminating environmental differences. Then we apply unsupervised and incremental speaker adaptation using ACTS to the environmentally adapted reference pattern. Since REALISE can treat both types of noise at the same time, a resulting system is expected to perform well under adverse environmental conditions. Moreover, since ACTS shows a constant, stable improvement regardless of the amount of available adaptation data, it is also expected to show a high performance

when applied to unsupervised and incremental frameworks.

2. ADAPTATION ALGORITHMS

2.1. Rapid Environment Adaptation Algorithm based on Spectral Equalization (REALISE)

We assume there are two types of environmental noise sources which degrade speech recognition performance: additive noise and multiplicative noise in the spectral domain. Additive noise is caused by various user environments (e. g. machinery noise, speech from others, etc.), and multiplicative noise is caused by filtering processes (e. g. microphones, transmission channels, the vocal tracts of individual speakers, etc.). In this study, we assume that original speech signal $x(t)$ is first distorted by a linear filter $a(t)$ and then corrupted by an uncorrelated additive noise $b(t)$.

The Fourier power spectrum for observed speech $y(t)$ is given by

$$Y(\omega) = A(\omega)X(\omega) + B(\omega), \tag{1}$$

where $A(\omega)$, $X(\omega)$, and $B(\omega)$ represent the Fourier power spectra for $a(t)$, $x(t)$, and $b(t)$, respectively.

From this relation, we introduce models in which both an input speech and a reference pattern are distorted by their own additive noise $\mathbf{B}$ and multiplicative noise $\mathbf{A}$. Assuming that $\mathbf{A}$ and $\mathbf{B}$ are constant within an utterance, we have

$$\begin{cases} \mathbf{V}(k) = \mathbf{A}_v \tilde{\mathbf{V}}(k) + \mathbf{B}_v \\ \mathbf{W}(k) = \mathbf{A}_w \tilde{\mathbf{W}}(k) + \mathbf{B}_w, \end{cases} \tag{2}$$

where k indicates frame number, $\mathbf{V}(k)$, $\mathbf{W}(k)$, $\tilde{\mathbf{V}}(k)$, and $\tilde{\mathbf{W}}(k)$ are the observed spectra for the input and the reference pattern, and the undistorted spectra for the input and the reference pattern, respectively. Suffixes v and w indicate the input and the reference, respectively. Multiplicative noises $\mathbf{A}_v$ and $\mathbf{A}_w$ are diagonal matrices.

The goal of REALISE is to estimate spectra of the reference pattern which are newly distorted by the input environment. From Eq. (2), we formulate the distorted spectrum $\hat{\mathbf{W}}(k)$ as follows:

$$\begin{aligned} \hat{\mathbf{W}}(k) &= \mathbf{A}_v \tilde{\mathbf{W}}(k) + \mathbf{B}_v \\ &= \mathbf{A}_v \mathbf{A}_w^{-1} (\mathbf{W}(k) - \mathbf{B}_w) + \mathbf{B}_v. \end{aligned} \tag{3}$$

Finally, we obtain the following adaptation equation:

$$\hat{w}^i(k) \simeq \frac{s_v^i - n_v^i}{s_w^i - n_w^i} (w^i(k) - n_w^i) + n_v^i, \tag{4}$$

where $\hat{w}^i(k)$ and $w^i(k)$ are element-wise representation of $\hat{\mathbf{W}}(k)$ and $\mathbf{W}(k)$, respectively, s_v^i, s_w^i, n_v^i, and n_w^i are average spectra of speech portions for the input and reference pattern, and average spectra of noise portions for input and reference pattern, respectively, and superscript i indicates the ith element of the vectors. More detailed derivations of this equation are described in [2].

REALISE implementation into cepstrum-based features consists of the following three steps:

1. **Preliminary recognition** - determines the closest reference pattern to an input and obtains the time-alignment between the two.
2. **Environmental difference estimation** - calculates four cepstral averages corresponding to the four portions according to the time-alignment. Then the four cepstral averages are converted into the spectral averages, i. e. s_v^i, n_v^i, s_w^i, and n_w^i.
3. **Adaptation** - first converts mean cepstral vectors of all pdfs into the spectral domain. Then it adapts all reference patterns to the input environment by using Eq. (4). Finally, it converts the adapted spectra into the cepstral domain.

2.2. Speaker Adaptation with Autonomous Control Using Tree Structure (ACTS)

In this study, speaker adaptation for continuous densities, mixture Gaussian HMMs is attempted by a simple spectral mapping technique as follows:

$$\hat{\mu}_i = \mu_i + \delta_i, \qquad i = 1, 2, ..., K, \tag{5}$$

where i is a index to each pdf, μ_i is an original mean vector of each pdf for the SI model, $\hat{\mu}_i$ is an adapted mean vector of each pdf for the SD model, and δ_i is an adaptation vector associated with each pdf, which is calculated by Viterbi-alignment between HMMs and adaptation utterances. Since the total number of pdfs K is very large, the performance for speaker adaptation would be degraded when the amount of adaptation data is small. One possible solution to overcome this problem is reducing the number of free parameters by tying the adaptation vectors. However, a small number of free parameters leads to less improvement as the amount of adaptation data increases. Therefore, the number of free parameters should be controlled according to the amount of adaptation data.

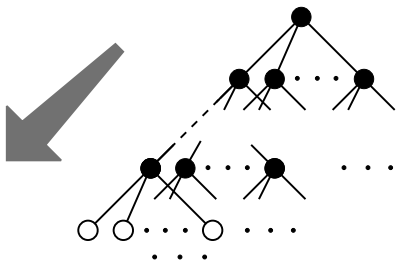

Figure 1: Tree Structure of pdfs

ACTS is way of creating such a mechanism by using static tree structure pdfs (see Fig. 1). The tree structure was constructed in a top-down manner by k-means algorithm, in which the distance measure is defined by Kullback-divergence between pdfs in the SI model [5]. In this tree structure (s indicates depth), each pdf i belongs to one leaf node (S,l), and a tied adaptation vector $\Delta(s,l)$ is defined for each node (s,l). $\Delta(s,l)$ is calculated by the following equation:

$$\begin{cases} \Delta(s,l) = \frac{1}{N(s,l)} \sum_{i=\Psi(s,l)} n_i \delta_i \\ N(s,l) = \sum_{i=\Psi(s,l)} n_i, \end{cases} \quad (6)$$

where $\Psi(s,l)$ is a function to acquire an entire index set of pdfs which are descendants of node (s,l), and n_i is the number of frames in the adaptation data which are aligned to the pdf i. An algorithm to pick up one tied adaptation vector $\Delta(s,l)$ for each pdf i is as follows:

1 Let D be a threshold value for the number of data frames. If n_i is larger than D, δ_i is an adaptation vector for pdf i. Otherwise, go to step 2.

2 Go up to the parent node. If the depth is equal to 1, $\Delta(1,1)$ is an adaptation vector for pdf i. Otherwise, if $N(s,l)$ is larger than D, $\Delta(s,l)$ is an adaptation vector for pdf i. Otherwise, go to step 2.

As described above, since ACTS autonomously controls the number of free parameters according to the amount of adaptation data, it showed a stable improvement regardless of the amount of data.

2.3. Unsupervised and Incremental Speaker Adaptation using REALISE and ACTS

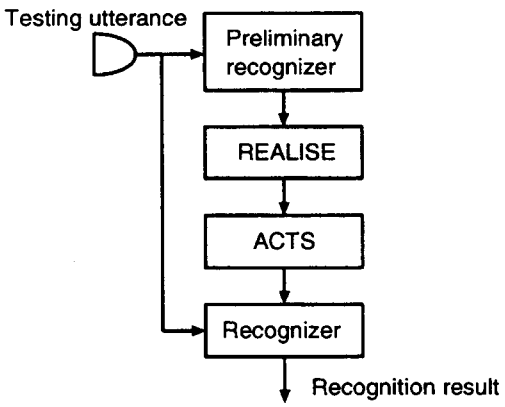

Figure 2: Unsupervised and Incremental Speaker Adaptation

In practical applications of speaker adaptation, adaptation utterances are not always available beforehand. Therefore, we mainly study an unsupervised and incremental speaker adaptation which enables testing utterances to be used for both adaptation and recognition, and which facilitates implementation using limited cpu power and storage.

In this study, we assume the environment changes for each utterance. Hence the environmental difference between the utterance and reference pattern should be eliminated for each single utterance. Figure 2 shows the overall unsupervised and incremental speaker adaptation / recognition procedure. For each testing utterance, REALISE is first applied and an environmentally adapted reference pattern is obtained. From Eq. (4), a cepstral representation of each pdf i can be calculated by

$$\mu_i = C[\hat{\mathbf{W}}(i)], \quad (7)$$

where $\hat{\mathbf{W}}(i)$ is an environmentally adapted spectral representation for pdf i, $C[.]$ is a function which converts a spectrum into a cepstrum. Thus we obtain an entire pdf set in which the environmental differences are eliminated. Then we calculate adaptation vectors for each pdf for ACTS. Let us assume the user speaks the wth utterance, then the adaptation vector $\delta_i^{(w)}$ for the pdf i in an incremental manner is

$$\begin{cases} n_i^{(w)} = n_i' + n_i^{(w-1)} \\ \delta_i^{(w)} = \frac{n_i^{(w-1)} \delta_i^{(w-1)} + n_i'(x_i - \mu_i)}{n_i^{(w)}}, \end{cases} \quad (8)$$

where x_i is the mean cepstrum which is aligned to the pdf i in the wth testing utterance, n_i' is the number of frames for x_i, and $n_i^{(w)}$ is the total number of frames up to the wth utterance, which are aligned to the pdf i. By using Eq. (8), we can calculate Eq. (6) and carry out the entire adaptation process in ACTS. We store the adaptation vectors $\delta_i^{(w)}$ and the total number of frames $n_i^{(w)}$ for the next $w+1$th utterance. Finally, the wth utterance is recognized again using a speaker adapted reference pattern using ACTS.

As described above, since the environmental differences between each testing / adaptation utterance and reference pattern are eliminated by REALISE, adaptation vectors are not affected by the input environment, and the resulting system is expected to perform well under adverse environmental conditions.

3. EVALUATION EXPERIMENTS

Evaluation was carried out using a demi-syllable HMM [6] based 100 Japanese-word recognition task. The speaker-independent HMM was trained using 250 Japanese phonetically balanced words spoken by 43 speakers, which were recorded in a quiet room through a vocal microphone. We used a 3-state left-to-right HMM with two Gaussian mixture densities in each state for representing each demi-syllable.

The evaluation utterances were recorded at three vehicle speed conditions (idling, 50 km/h, and 100 km/h), and six speakers uttered 100 Japanese words in each condition. The utterances were recorded through uni-directional microphone

set on a sun visor. The distance between the microphone and the speaker's mouth was about 30 cm. The utterances were sampled at 16 kHz, spectral subtraction (SS) [7] was applied in the FFT domain, and then ten mel-scaled cepstrum coefficients were calculated every 16 ms. Finally, we used 21-dimensional feature vectors which consist of ten mel-cepstrum coefficients, ten first-order time derivatives of the mel-cepstrum coefficients and one-dimensional delta-power. In both adaptation process, only 10 mel-cepstrum coefficients were adapted, and the other coefficients were not adapted.

We measured, in advance, mean signal-to-noise ratios (SNRs) for each condition. SNRs for idling, 50 km/h, and 100 km/h were 13.6 dB, 6.2 dB, and 1.6 dB, respectively. This means that the whole utterance was very noisy and that SNRs differ drastically depending on the conditions.

We compared another environment adaptation technique [8] to REALISE, one which was based on cepstral mean equalization (CME). CME was implemented in such a way that mean cepstral differences for speech portions and noise portions are equalized independently by using the time-alignment in the preliminary recognition step.

To simulate a condition where the environments for adaptation and testing are different, we used 100 words at 100 km/h for adaptation, and tested 200 words when idling and at 50 km/h. Baseline recognition accuracy were 79.1% (SI), 85.8% (CME alone), and 90.3% (REALISE alone). We tested three topologies for tree structure: binarys tree with eleven stages (2p11), 3p7, and 4p6, and also tested three threshold values for the number of data frames D: $D = 10, 30, 50$. Figure 3 shows the speech recognition rates versus the number of adaptation utterances using the optimal topology and D for each method.

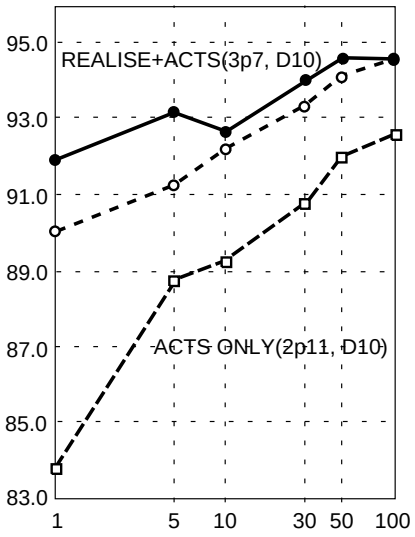

Figure 3: Result for Speech Recognition/Adaptation Experiment

In Fig. 3, overall recognition accuracy showed stable improvement as the number of adaptation utterances increased. This is because ACTS effectively controls the adaptation complexity. By applying environmental adaptation (i. e. CME or REALISE) together with ACTS, performance was considerably better than ACTS alone. This proves that the method in which environmental adaptation was applied ahead of speaker adaptation is more effective than that without such processing. In particular, applying REALISE together with ACTS performed better than applying CME. This is because CME only accommodates multiplicative noise and has no explicit modeling for additive noise, and the effect of additive noise cannot be ignored in very noisy vehicle environments.

4. CONCLUSION

In this paper we proposed an unsupervised and incremental speaker adaptation method. The evaluations proved that the method is effective even under very noisy vehicle environments.

ACKNOWLEDGMENTS

The authors wish to thank the members of the Human Language Research Laboratory for their continuous support.

5. REFERENCES

1 K. Shinoda and T. Watanabe: "Speaker Adaptation with Autonomous Control Using Tree Structure", EUROSPEECH95, pp. 1143-1146, 1995.

2 K. Takagi, H. Hattori, and T. Watanabe: "Rapid Environment Adaptation for Robust Speech Recognition", ICASSP95, pp. 149-152, 1995.

3 Y. Zhao: "Iterative Self-learning Speaker and Channel Adaptation under Various Initial Conditions", ICASSP95, pp. 712-715, 1995.

4 M. Feng: "Speaker Adaptation Based on Spectral Normalization and Dynamic HMM Parameter Adaptation", ICASSP95, pp. 704-707, 1995.

5 T. Watanabe, K. Shinoda, K. Takagi, and E. Yamada: "High Speed Speech Recognition Using Tree-Structured Probability Density Function", ICASSP95, pp. 556-559, 1995.

6 T. Watanabe, R. Isotani, and S. Tsukada: "Speaker Independent Speech Recognition Based on Hidden Markov Model Using Demi-Syllable Units", Trans. on IEICE(D-II), J75-D-II, 8, pp. 1281-1289, 1992 (*in Japanese*).

7 S. F. Boll: "Suppression of Acoustic Noise in Speech Using Spectral Subtraction", IEEE Trans. on ASSP, Vol. ASSP-27, No. 2, April 1979.

8 S. Lerner and B. Mazor: "Telephone Channel Normalization for Automatic Speech Recognition", ICASSP92, pp. 261-264, 1992.