RAPID ENVIRONMENT ADAPTATION FOR ROBUST SPEECH RECOGNITION

Keizaburo TAKAGI, Hiroaki HATTORI and Takao WATANABE

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

ABSTRACT

This paper proposes a *rapid environment adaptation algorithm based on spectrum equalization* (REALISE). In practical speech recognition applications, differences between training and testing environments often seriously diminish recognition accuracy. These environmental differences can be classified into two types: difference in additive noise and difference in multiplicative noise in the spectral domain. The proposed method calculates time-alignment between a testing utterance and the closest reference pattern to it, and then calculates the noise differences between the two according to the time-alignment. Then, we adapt all reference patterns to the testing environment using the differences. Finally, the testing utterance is recognized using the adapted reference patterns. In a 250 Japanese word recognition task, in which the training and testing microphones were of two different types, REALISE improved recognition accuracy from 87% to 96%.

1. INTRODUCTION

Stochastic approaches, like Hidden Markov Models (HMMs), to automatic speaker-independent speech recognition have been widely used in recent years. In order to accommodate the naturally occurring wide distribution of individual speaker-characteristics, they require a huge volume of training utterances provided by a large number of speakers. However, it is well known that differences between training and testing environments seriously diminish recognition accuracy. If a large number of HMMs could be trained on all possible environments, this problem might be overcome, but the collection of great volumes of utterances for all possible environments is not feasible in practical terms.

Several approaches based on speaker adaptation techniques have been proposed for eliminating environmental differences (for one such example, see [1]) and reported that they are effective when the adaptation data in the testing environment is available beforehand. However, testing environments are not always known beforehand (as in, for example, speech recognition over telephone lines). Additionally, even when speeches having the same content are uttered by the same speaker, the acoustics may vary according to the physical and emotional conditions. Therefore, for practical applications, a new framework, which uses testing utterances themselves for adaptation, is effective to cope with such variation.

The framework needs a new adaptation algorithm which satisfies three specific requirements: (1) it is effective under unsupervised conditions, (2) it performs effectively even with a single adaptation utterance, and (3) its computational cost is low. CDCN [3] satisfies the first and the second requirements. However, its high computation costs, especially relating to its requirement for the iterative calculations, render it impractical in the framework[4].

In this paper we propose a *rapid environment adaptation algorithm based on spectrum equalization* (REALISE)[2]. This algorithm assumes that the environmental differences can be classified into two types of differences: differences in additive noise and in multiplicative noise in the spectral domain. Based on the assumption, this algorithm extracts the two types of noise differences according to the time-alignment between a testing utterance and the closest reference pattern to it. All reference patterns are then adapted to the testing environment using the differences and the input is recognized again using the adapted reference patterns. Because the parameters to be estimated are very few, and because the algorithm utilizes spectral averages for both the speech portion and the noise portion of utterances in the adaptation process, this approach is expected to perform well with a single testing utterance and to offer stability under unsupervised conditions. Moreover, because its computational cost is low, these three features allow testing utterances themselves to be used for adaptation.

This paper is organized as follows: in Section 2, we describe the new environment adaptation algorithm REALISE in detail. In Section 3, we report evaluation experiments. In Section 4, we briefly discuss these results.

2. ENVIRONMENT ADAPTATION

2.1. An Environmnetal Model

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 noises, 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 introduce models in which both an input speech and a reference pattern are distorted by their own additive noise B and multiplicative noise A. Assuming that A and B are constant within an utterance, we

0-7803-2431-5/95 $4.00

have

$$\begin{cases} \mathbf{V}(k) = \mathbf{A}_v \tilde{\mathbf{V}}(k) + \mathbf{B}_v \\ \mathbf{W}(k) = \mathbf{A}_w \tilde{\mathbf{W}}(k) + \mathbf{B}_w, \end{cases} \quad (1)$$

where k indicates the 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.

2.2. Rapid Environment Adaptation Algorithm based on Spectrum Equalization (REALISE)

The goal of REALISE is to estimate spectra which are newly distorted by the input environment. From Eq. (1), we formulate the distorted spectrum $\hat{\mathbf{W}}(k)$ as follows:

$$\begin{aligned} \hat{\mathbf{W}}(k) &= \mathbf{A}_v \tilde{\mathbf{W}}(k) + \mathbf{B}_v \\ &= \mathbf{A}_v \mathbf{A}_w^{-1} (\mathbf{W}(k) - \mathbf{B}_w) + \mathbf{B}_v. \end{aligned} \quad (2)$$

Since $\mathbf{A}_v$ and $\mathbf{A}_w$ are diagonal matrices, we can rewrite Eq. (2) in elementwise representation as

$$\hat{w}^i(k) = \frac{a_v^{ii}}{a_w^{ii}} (w^i(k) - b_w^i) + b_v^i, \quad (3)$$

where superscript i indicates the ith element of the vector and the matrix. By taking a time-alignment, using Dynamic Programming (DP) matching, between a testing utterance and its closest reference pattern, we attempted to approximate the four noises, a_v^{ii}, a_w^{ii}, b_w^i, and b_v^i. Two additive noises, b_v^i and b_w^i, can be calculated directly from the noise portion of the input and the reference pattern. Taking the average of the noise portion for the input, which is decided from the time-alignment, b_v^i is approximated as

$$b_v^i \simeq \frac{1}{K_\nu} \sum_{k \in \nu} v^i(k) \doteq n_v^i, \quad (4)$$

where ν indicates a frame set which is aligned to the noise portion of the reference pattern, K_ν denotes the number of frames among ν, and $v^i(k)$ is an elementwise representation of the vector $\mathbf{V}(k)$. Similarly, we obtain additive noise for the reference pattern: $b_w^i \simeq n_w^i$.

On the other hand, a_v^{ii} and a_w^{ii} cannot be calculated directly, but can be related to the averages of speech portions. Taking the average of the speech portion for the input, the relation between a_v^{ii} and the average is given as

$$a_v^{ii} \bar{v}^i + n_v^i \simeq \frac{1}{K_\sigma} \sum_{k \in \sigma} v^i(k) \doteq s_v^i, \quad (5)$$

where σ indicates a frame set, which is aligned to the speech portion for the reference pattern, K_σ denotes the number of frames among σ, and $\bar{v}^i$ is the average speech portion for the undistorted input. We obtain a similar expression for the reference: $a_w^{ii} \bar{w}^i + n_w^i \simeq s_w^i$, where $\bar{w}^i$ is the average speech portion for the undistorted reference.

Finally, by substituting the four noises, we transform Eq. (3) as follows:

$$\begin{aligned} \hat{w}^i(k) &= \frac{a_v^{ii}}{a_w^{ii}} (w^i(k) - b_w^i) + b_v^i \\ &= \frac{s_v^i - n_v^i}{s_w^i - n_w^i} \frac{\bar{w}^i}{\bar{v}^i} (w^i(k) - n_w^i) + n_v^i \\ &\simeq \frac{s_v^i - n_v^i}{s_w^i - n_w^i} (w^i(k) - n_w^i) + n_v^i. \end{aligned} \quad (6)$$

Two averages of undistorted speech, $\bar{w}^i$ and $\bar{v}^i$, retain information regarding both utterance contents and speaker individualities. The derivations of this equation imply that they have no significant difference. Because we use the closest reference pattern to the testing utterance, they are expected to be similar in regard to their contents. There is no significant difference in the speaker individualities, because speaker-independent speech recognition covers variations in the speaker individualities.

Implementation of REALISE 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 (unsupervised condition). When the correct supervising signal is given, i. e. supervised adaptation, only the time-alignment between the input and reference pattern which is assigned by the supervising signal is obtained in this part.
2. **Environmental difference estimation** - calculates the spectral averages, i. e. s_v^i, n_v^i, s_w^i, and n_w^i, according to the time-alignment.
3. **Adaptation** - adapts all reference patterns to the input environment by using Eq. (6). In Eq. (6), there are three subtraction parts, and we apply a flooring rule similar to spectral subtraction[6] to avoid the subtraction results becoming zero or a negative value.

Although we showed DP matching based implementation of REALISE here, it can also be implemented in continuous densities HMMs. In this case, some modifications are required. In the preliminary recognition step, we select the best one Gaussian probability density function (pdf) from mixture densities at each state for obtaining the time-alignment. In the environmental difference estimation step, the two spectral averages for reference pattern, i. e. s_w^i and n_w^i, are calculated by averaging the mean vectors of the pdfs which are decided in the preliminary recognition step. Fig. 1 shows the portions on which the four spectral averages, $\mathbf{S}_v = [s_v^i]$, $\mathbf{N}_v = [n_v^i]$, $\mathbf{S}_w = [s_w^i]$, and $\mathbf{N}_w = [n_w^i]$, are calculated.

This approach is expected to offer stability under unsupervised conditions, because the spectral averages would be similar to those for a correct one, even when the preliminary recognition fails. In addition, this approach is expected to perform effectively with a single testing utterance, because it only requires estimating the four spectral averages. Moreover, the computational cost for this algorithm is low, because no iterative procedure is needed, unlike CDCN. These features enable a recognition system to use a testing utterance itself for adaptation.

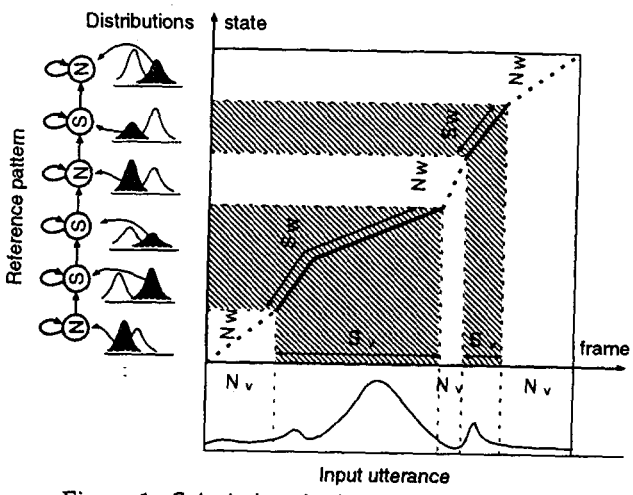

Figure 1: Calculating the four spectral averages

3. EVALUATION EXPERIMENTS

3.1. Experimental conditions

REALISE was evaluated in a demi-syllable HMM [8] based 250 Japanese-word recognition task. The HMM was trained using 250 Japanese phonetically balanced words spoken by 85 speakers, which were recorded in a quiet room through a vocal microphone. We used a 4-state left-to-right HMM with two Gaussian mixture densities at one state for representing each demi-syllable.

The evaluation utterances were recorded in an office room, and consist of different 250 Japanese words spoken by seven different speakers from the training database. The utterances were recorded through both a vocal microphone (Mic. A), which is the same microphone as that used in the training, and a desk-top boundary microphone (Mic. B). The utterances were sampled at 16kHz, and ten mel-scaled cepstrum coefficients (MCCs) were calculated every 10ms. 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 the adaptation process, only 10 mel-cepstrum coefficients were adapted, and the other coefficients were not adapted.

3.2. Comparison between supervised and unsupervised adaptation

We compared effectiveness of REALISE in supervised and unsupervised adaptation. In a real implementation of the supervised adaptation, a supervising signal can be obtained through a confirmation for recognition results. Therefore, we used a previous utterance for adaptation.

This scheme only assumes that the testing environment does not change during the two successive utterances: a time-alignment was calculated between a previous testing utterance and unadapted (initial) HMMs, and the unadapted HMMs were adapted to the utterance, then the next utterance was recognized using these adapted HMMs.

The results are shown in Table 1. Results for Mic. A and Mic. B utterances without REALISE are also shown for comparison. In addition, in order to investigate the effect of the additive noises, we evaluated all the results with and without spectral subtraction (SS)[7].

Table 1: Comparison between supervised and unsupervised

Mic.	Supervised/ Unsupervised	Accuracy	
		no SS	SS
A	–	96.9%	97.4%
B	–	56.9%	87.8%
B	Supervised	93.6%	96.1%
B	Unsupervised	90.5%	95.6%

Table 1 showed that the difference in environments seriously degrades the recognition accuracy (96.9% → 56.9%) without REALISE. SS gave a considerable improvement. However, the accuracy with SS was still lower than that for Mic. A utterances. This is because SS can cancel the difference in additive noises, but cannot cope with the multiplicative noises.

The overall recognition accuracies with REALISE were significantly improved from the baseline results. Use of SS together with REALISE improved the recognition performance, because the accuracy for the time-alignment was improved by SS.

Comparing unsupervised and supervised cases, there was no significant difference in the recognition accuracies. This is because the spectral averages are roughly the same for the incorrect reference pattern and for the correct one, even when the preliminary recognition fails. Additionally, the additive noises can be estimated correctly as long as only the time-alignment for the noise portions is correct.

These results shows that REALISE is effective with a single utterance under unsupervised condition.

3.3. Adaptation using a testing utterance

In this experiment, we present evaluation results for the performance of REALISE using a single testing utterance itself. In this case, only unsupervised adaptation is feasible in practical terms. We evaluated two alternative REALISE implementations: (INIT) - adaptation from unadapted (initial) HMMs for each testing utterance, and (PREC) - adaptation from adapted HMMs which are the adaptation results for a preceding testing utterance. INIT assumes the most severe condition wherein there is no a priori knowledge about an input environment, other than a single testing utterance itself, and that the environment changes for each testing utterance. On the other hand, PREC assumes a more relieved condition, considering that the environment does not change so rapidly through one session. SS was also applied. The other experimental conditions were the same as those in the previous evaluation. Results are shown in Table 2. By applying REALISE under INIT condition, performance was considerably improved from 87.8% to 92.9%. Under the PREC condition, further improvement for REALISE was observed (92.9% → 96.3%). This is because

Table 2: Adaptation using a testing utterance

Mic.	Cond.	Accuracy
B	INIT	92.9%
B	PREC	96.3%

testing utterances, which we evaluated here, had little environmental changes through one session, and also more correct supervising signal and more accurate time-alignment were obtained by using the adapted HMMs in the preliminary recognizer, when the testing environments does not change so rapidly through one session.

3.4. Use of REALISE together with speaker adaptation

In this section, in order to develop a high-performance and robust speech recognition system, we evaluated the use of REALISE together with the speaker adaptation method[9].

Once the system is adapted to a specific speaker by using speaker adaptation, it maintains a high and stable performance for the speaker. On the other hand, since the environment may change for each testing utterance, its change should be treated using testing utterances. Therefore, the use of REALISE together with the speaker adaptation method is expected to become environment-independent and to show high performance for the speaker.

In this experiment, speaker adaptation[5] was carried out using 100 utterances for each speaker, recorded through Mic. A and Mic. B. Then, testing utterances were recognized with or without REALISE. Testing utterances were the same as those used in the previous evaluation, recorded through Mic. B. REALISE was carried out using a testing utterance, under PREC condition. SS was applied to all the utterances. The results are shown in Fig. 2.

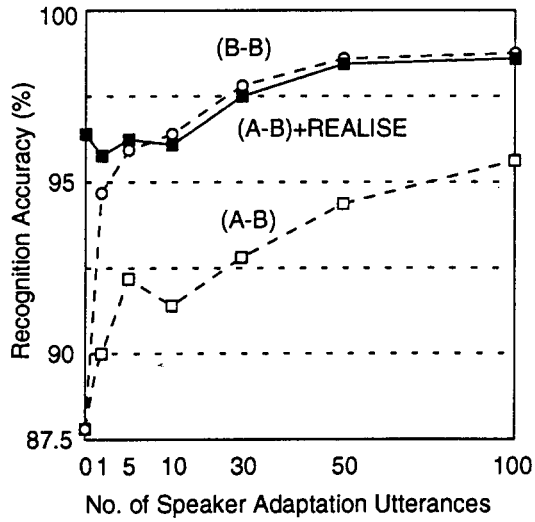

Figure 2: Use of REALISE together with speaker adaptation: (B-B) speaker adaptation with Mic. B and testing with Mic. B; (A-B) speaker adaptation with Mic. A and testing with Mic. B; ((A-B)+REALISE) REALISE for condition (A-B)

In Fig. 2, although the result for the same environment (B-B) showed a high recognition accuracy without REALISE, the result for the different environment (A-B) was degraded without REALISE. By applying REALISE to the degraded condition ((A-B)+REALISE), recognition accuracy was greatly improved and was comparable with the result for the same environment. These results mean that, once the system is speaker adapted in any single environment, the performance of the system is kept high in other environments as well as in the same environment.

4. DISCUSSION

Sections 3.2 and 3.3 showed that REALISE performs effectively even under unsupervised condition and using only a single testing utterance. Moreover, as described in Section 2.2, since there is no iterative procedure, the computational cost for REALISE is low. Hence, these three features enable adaptation using a single testing utterance itself.

Section 3.4 showed that, by using REALISE together with speaker adaptation, once the user prepares the utterances for speaker adaptation in any single environment, he does not have to speak any more utterances for speaker adaptation, regardless of a change in the testing environments.

Although we evaluated the change in microphone here, which was considered to be a typical example of the environmental change, REALISE will be evaluated in various testing environments in further work.

5. CONCLUSION

In this paper we proposed a *rapid environment adaptation algorithm based on spectrum equalization* (REALISE). The evaluations proved that the algorithm is effective even under unsupervised condition using a single testing utterance.

6. REFERENCES

[1] K. Takagi, K. Shinoda, and T. Watanabe: "Input Environment Adaptation for Speech Recognition", Proc. of ASJ Spring Meeting, 1-4-22, pp. 545-546, 1993(*in Japanese*).

[2] K. Takagi, H. Hattori, and T. Watanabe: "Speech Recognition with Rapid Environment Adaptation by Spectrum Equalization", Proc. of ICSLP, Vol. 3, S18.10, pp. 1023-1026, 1994.

[3] A.Acero and R.M.Stern: "Environmental Robustness in Automatic Speech Recognition", ICASSP90, Vol. 2, S15b.11, pp. 849-852, 1990.

[4] A.Acero: "Acoustical and Environmental Robustness in Automatic Speech Recognition", KLUWER ACADEMIC PUBLISHERS, 1993.

[5] K.Shinoda, K.Iso, and T.Watanabe: "Speaker Adaptation For Demi-Syllable Based Continuous Density HMM", ICASSP91, S13.7, pp. 857-860, 1991.

[6] M. Berouti, R. Schwarz, and J. Makhoul: "Enhancement of Speech Corrupted by Acoustic Noise", Proc. of ICASSP, pp. 208-211, 1979.

[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] 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*).

[9] K. Takagi, H. Hattori, and T. Watanabe: "Speech Recognition Using Input Environment Adaptation and Speaker Adaptation", Proc. of ASJ Autumn Meeting, 2-8-18, pp. 73-74, 1994(*in Japanese*).