Construction Method of Acoustic Models Dealing with Various Background Noises Based on Combination of HMMs

Motoyuki Suzuki, Yusuke Kato, Akinori Ito and Shozo Makino

Department of Electrical and Communication Engineering,
Graduate School of Engineering, Tohoku University, Sendai, Japan.
{moto, yusuke, aito, makino}@makino.ecei.tohoku.ac.jp

Abstract

Background noise is one of the biggest problem for speech recognition systems in real environments. In order to achieve high recognition performance for corrupted speech, we proposed a new construction method of HMMs dealing with various kinds of background noise. At first, each HMM dealing with a single noise is trained for each background noise, and then all Gaussian components of those HMMs are combined into a "multi-mixture HMM". From the experimental results, the multi-mixture HMM gave the highest recognition performance for any kind of noise and any variation of SNR.

Although the multi-mixture HMMs has high performance, it has a huge number of Gaussian components that makes the speech recognition slower. In order to solve the problem, we also proposed a reduction method of Gaussian components. It can decrease the number of Gaussian components with slight deterioration of recognition performance.

1. Introduction

One of the most important issues in speech recognition system is improvement of recognition performance for corrupted speech data involving background noise. Many methods have been proposed so far[1] (e.g. speech enhancement method using microphone array[2], spectral subtraction method[3], HMM decomposition method[4], and so on). Especially, several model-based methods are frequently used because they give high recognition performance and it is easy to use in an existing speech recognition system.

The conventional model-based methods are separated into the following three types.

- *HMM dealing with a single background noise*
  In this method, corrupted speech data is prepared by overlapping a single kind of noise on speech, and an HMM is trained using it. This method is easy to use, and the trained HMM gives high recognition performance for a "known" noise. However, it cannot achieve high recognition performance for "unknown" noises. In this paper, this type of trained HMM is called "single-noise HMMs".
- *HMM dealing with various background noises*
  This method is similar to the previous one, but various kinds of noise are used for training. At first, each corrupted speech data is prepared for each background noise, and then all of the data are used for training. This type of trained HMM is called "multi-condition HMMs", and it is robust for "known" noises. However, it is difficult to enroll a new "unknown" noise into an existing multi-condition HMM because the multi-condition HMM should be re-trained using all of the "known" and the new noise data.
- *HMM composition method*[5, 6]
  A phoneme HMM is trained using clean speech data, and a noise HMM is trained using noise data only. After that, an HMM dealing with the noise is constructed by composition of the phoneme and noise HMMs. If the noise HMM is trained using various kinds of noise data, the composed HMM has robustness for "known" noises. However, the composed HMMs gives lower recognition accuracy than the multi-condition HMMs because there are several approximate calculations in the composing process. In this paper, this type of HMM is called "composed HMMs".

We paid attention to the multi-condition HMMs because of high recognition performance and robustness for "known" background noises. However, the multi-condition HMMs has the problem as stated above. In order to solve the problem, we propose a new construction method of HMM dealing with various background noises.

2. HMM construction method by combining Gaussian components

The multi-condition HMMs gives high recognition performance for "known" noises (the known kind of noise and known SNR), however, it is difficult to enroll a new "unknown" noise into an existing HMM because the multi-condition HMM should be re-trained using all of the "known" and the new noise data. If the HMM can be constructed without re-training, it is easy to enroll a new noise. We propose a new construction method by combining Gaussian components.

2.1. Algorithm of the proposed method

The construction method is as follows:

1. *Training of each single-noise HMM for each "known" noise*
   For each "known" noise $(\varphi_1, \varphi_2, \cdots \varphi_E)$, a single-noise HMM $\mathcal{M}^{(\varphi_i)}$ is trained using the corrupted speech data involving the noise φ_i. Each single-noise HMM has the same number of states, and each state has a mixture Gaussian distribution with the same number of Gaussian components.
2. *Combining all Gaussian components of every single-noise HMMs into the unified HMM*
   The HMM $\mathcal{M}^{(\Phi)}$ is constructed by combining all Gaussian components of single-noise HMMs $\mathcal{M}^{(\varphi_i)}$, where Φ denotes a set of noises $(\Phi = \{\varphi_1, \varphi_2, \cdots \varphi_E\})$.

Table 1: Experimental conditions

Training data	8,000 sentences uttered by 1,938 speakers
Test data	200 sentences uttered by 28 speakers
#components in a state	
single-noise HMM	16
multi-mixture HMM	192 (= 16 × 12)
multi-condition HMM	16, 32, 64, 128
composed HMM	128, 256

The output probability density distribution $p_s^{(\Phi)}(x)$ at the s-th state in $\mathcal{M}^{(\Phi)}$ is given by Eq.(1).

$$p_s^{(\Phi)}(x) = \frac{1}{E} \sum_{\varphi_i \in \Phi} \sum_{m=1}^{M} w_{sm}^{(\varphi_i)} \mathcal{N}_{sm}^{(\varphi_i)}(x) \qquad (1)$$

where, $\mathcal{N}_{sm}^{(\varphi_i)}(x)$ denotes the m-th Gaussian component at the s-th state in $\mathcal{M}^{(\varphi_i)}$, and $w_{sm}^{(\varphi_i)}$ denotes the mixture weight.

If a new noise $\acute{\varphi}$ is given to "known" noises, the following three steps are only required.

1. Train a new single-noise HMM $\mathcal{M}^{(\acute{\varphi})}$
2. Add $\acute{\varphi}$ into Φ
3. Re-calculate $p_s^{(\Phi)}(x)$ using Eq.(1).

In this paper, the HMM $\mathcal{M}^{(\Phi)}$ is called "multi-mixture HMMs".

The another method to combine all single-noise HMMs $\mathcal{M}^{(\varphi_i)}$ into unified HMM $\mathcal{M}^{(\Phi)}$ is parallel connection of all HMMs. However, preliminary experiments showed that this type of HMMs gave lower recognition performance than the multi-mixture HMMs.

2.2. Speech recognition experiments

In order to confirm the effectiveness of the multi-mixture HMMs, several speech recognition experiments were carried out. We compared the recognition performance of the multi-mixture HMMs, the composed HMMs[5, 6], the multi-condition HMMs and "noise-matched HMMs". The noise-matched HMMs means a single-noise HMM dealing with the noise of the test data.

Sixteen kinds of background noise[7] (Exhibition hall, Railway stations, Crowded street and so on) were used for the experiments. Twelve kinds of noise were used as "known" noises, and the other four kinds of noise were used as "unknown" noises. Another experimental conditions are shown in Table 1.

2.2.1. Recognition performance for fixed SNR

At first, we compared the performance of each method in a fixed SNR condition. SNR was set to 10dB. Experimental results are shown in Table 2. From these results, the multi-mixture HMM showed the highest accuracy for "known" and "unknown" conditions. In the "known" condition, the multi-condition HMMs with 64 or 128 Gaussian components gave lower accuracy than that with 16 or 32 Gaussian components because these HMMs were over-trained due to the limit of the amount of training data.

Table 2: Recognition accuracy (SNR = 10dB)

HMM	#Gaussian components	kind of noise	
		"known"	"unknown"
multi-mixture	192	70.5%	69.0%
composed	128	58.4%	56.6%
	256	58.9%	57.3%
multi-condition	16	66.9%	65.5%
	32	66.9%	64.9%
	64	66.3%	64.0%
	128	65.1%	62.8%
noise-matched	16	67.9%	—

Although the noise-matched HMM is expected to give the highest accuracy of all, it showed lower accuracy than the multi-mixture HMM. As several triphones appeared few times in the training data, these triphones could not acquire the characteristics of the noise sufficiently. On the other hand, the multi-mixture HMM could acquire the characteristics of the noise because many variations of noise were included by combining various single-noise HMMs.

2.2.2. Recognition performance for various SNRs

We also carried out the recognition experiments for various SNR conditions. In these experiments, "known" kind of noises were only used, and SNR was set to 5, 10, 15 and 20dB. The multi-mixture HMM was constructed from 48 single-noise HMMs (12 noise variations × 4 SNR variations) and each state has 768 (= 16 × 12 × 4) Gaussian components. We also examined a "multi-path HMMs"[8]. It is constructed by connecting several HMMs in parallel, and each HMM is a multi-condition HMM with a fixed SNR. In these experiments, the multi-path HMM has four paths.

Figure 1 shows the experimental results. From these results, the multi-mixture HMM gave the highest accuracy for all SNRs. It means that the multi-mixture HMMs is robust for "known" SNRs.

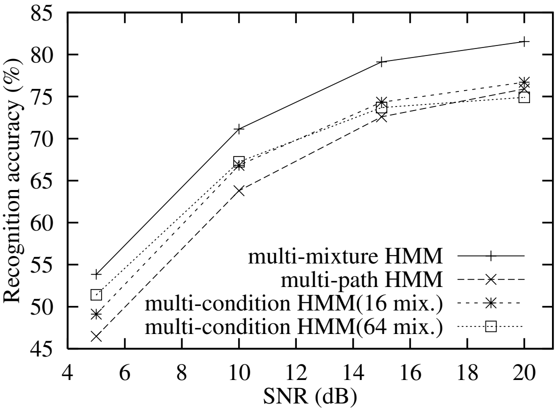

Figure 1: Recognition accuracy (SNR = 5, 10, 15, 20dB)

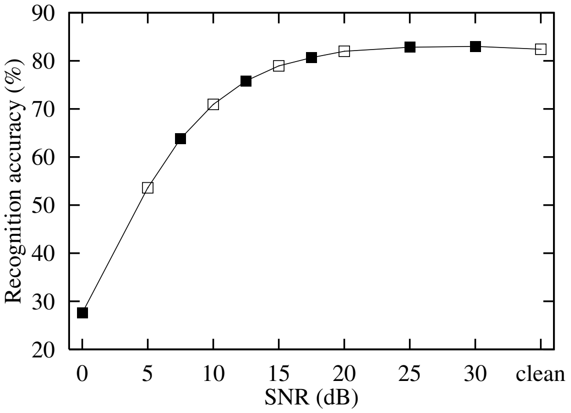

Figure 2: Recognition accuracy for "unknown" SNR

2.2.3. Recognition Performance for "unknown" SNRs

In order to confirm the robustness for variation of SNRs, we checked the performance for "unknown" SNRs. The multi-mixture HMM was constructed by combining five HMMs, which corresponds to SNR = 5, 10, 15, 20 and ∞ respectively. SNR of the test data were set to 0, 5, 7.5, 10, 12.5, 15, 17.5, 20, 25, 30 and ∞.

Figure 2 shows the recognition accuracy. The multi-mixture HMM gave the same recognition performance for "unknown" SNRs as that for "known" SNRs. This result says that the multi-mixture HMMs is robust for any variation of SNR.

3. Reduction method of a number of Gaussian components in multi-mixture HMMs

The multi-mixture HMMs gives high recognition performance, however, it has a huge number of Gaussian components. For example, the multi-mixture HMM used in Section 2.2.2 in total has about 4.8 million Gaussian components. A huge number of Gaussian components causes a huge calculation time, memory, storage, and so on. In this section, we propose a reduction method of a number of Gaussian components keeping high recognition performance.

3.1. Algorithm of the reduction method

The reduction method is as follows:

1. For each state, distance between two Gaussian components is calculated for all pairs of Gaussian components in the state. Definition of the distance is described in Section 3.2.
2. The pair of Gaussian components with the smallest distance is merged into a single Gaussian component. Let w_i be a mixture weight of a Gaussian component $\mathcal{N}_i$, and let $\vec{\mu}_i = (\mu_i^{(1)}, \mu_i^{(2)}, \cdots \mu_i^{(V)})$ be a mean vector, and $\sigma_i^{(v)}$ be a v-th diagonal element of the covariance matrix Σ_i of $\mathcal{N}_i$. Note that Σ_i is a diagonal matrix. The mixture weight $\hat{w}$, mean vector $\hat{\vec{\mu}}$ and v-th diagonal element $\hat{\sigma}^{(v)}$ of covariance matrix $\hat{\Sigma}$ of merged Gaussian component $\hat{\mathcal{N}}$ are given by Eq.(2), (3) and (4).

$$\begin{aligned}\hat{w} &= w_i + w_j &(2)\\ \hat{\mu}^{(v)} &= \frac{1}{\hat{w}}\left(w_i\mu_i^{(v)} + w_j\mu_j^{(v)}\right) &(3)\\ \hat{\sigma}^{(v)} &= \frac{1}{\hat{w}}\left\{w_i\left(\sigma_i^{(v)} + \left(\mu_i^{(v)}\right)^2\right) + \right.\\ &\left. w_j\left(\sigma_j^{(v)} + \left(\mu_j^{(v)}\right)^2\right)\right\} - \left(\hat{\mu}^{(v)}\right)^2 &(4)\end{aligned}$$

3. Re-calculate the distance between $\hat{\mathcal{N}}$ and all Gaussian components in the same state.
4. Finish the merging if the total number of Gaussian components reaches to the pre-defined number. Otherwise, go to Step 2.

3.2. Definition of the distance between two Gaussian components

We defined four distance measures between two Gaussian components.

- *Bhattacharyya distance*
  Bhattacharyya distance is one of the most popular distances between two probability density distributions. It is defined as Eq. (5).

$$D_1 = -\log \int \sqrt{\mathcal{N}_i\mathcal{N}_j}dx \qquad (5)$$

- *Sum of mixture weights*
  It is assumed that Gaussian component with small mixture weight is not important for the output probability density distribution.

$$D_2 = w_i + w_j \qquad (6)$$

- *Combination of previous two definitions*
  The previous two definitions are combined with normalized parameter α.

$$D_3 = D_1 \times D_2^{\alpha} \qquad (7)$$

- *Difference of two likelihoods between before and after merging*
  Two likelihoods are calculated for the output probability density distribution before and after merging of $\mathcal{N}_i$ and $\mathcal{N}_j$. The difference between two likelihoods is defined as the distance between two Gaussian components.

$$D_4 = \int \left\{\sum_k w_k\mathcal{N}_k - \left(\hat{\mathcal{N}} + \sum_{k\neq i,j} w_k\mathcal{N}_k\right)\right\}^2 dx \qquad (8)$$

3.3. Speech recognition experiments

In order to confirm the effectiveness of the reduction method, several experiments were carried out. The training and test data were the same as Table 1.

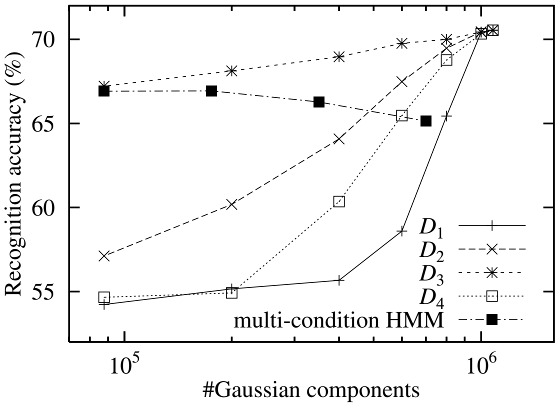

Figure 3: Recognition accuracy for fixed SNR

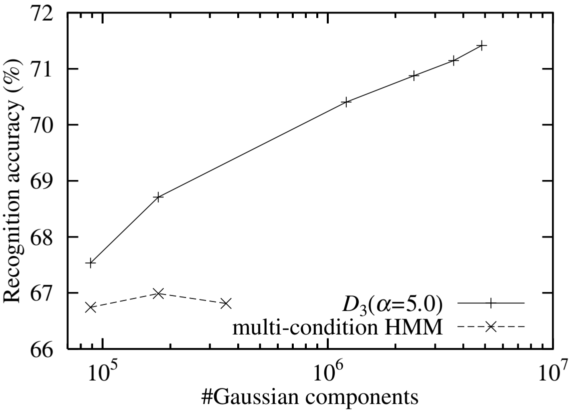

Figure 4: Recognition accuracy for variable SNR

3.3.1. Comparison of each distance with a fixed SNR

The multi-mixture HMM constructed in Section 2.2.1 were used in these experiments. SNR was fixed to 10dB, and the multi-mixture HMM in total had 1,079,028 Gaussian components. The parameter α used in the distance D_3 was set to 5.0 according to the results of preliminary experiments.

Figure 3 shows the experimental results. Each line denotes the recognition accuracy for each distance and multi-condition HMMs which has various numbers of Gaussian components (16, 32, 64, 128 in a state). These results shows the distance D_3 with $\alpha = 5.0$ gave the highest accuracy of all. The recognition performance of the multi-mixture HMM which has 600,000 Gaussian components (it is about 55% compared with the original multi-mixture HMM) was decreased only 1 point compared with that of the original HMM. The reduction method with D_3 can reduce the number of Gaussian components effectively.

3.3.2. Experimental results for various SNRs

We also checked the performance of the multi-mixture HMMs which supports various kinds of noise and various SNRs. The multi-mixture HMM constructed in Section 2.2.2 were used in these experiments. SNR of the training and the test data was set to 5, 10, 15, 20dB, and the multi-mixture HMM in total had 4,836,034 Gaussian components. We compared the performance with the multi-condition HMMs. Each state of multi-condition HMMs has 16, 32 and 64 Gaussian components.

The test data was the same as Section 2.2.2, and Fig. 4 shows mean of recognition accuracy for four SNRs. The recognition performance of the multi-mixture HMM which has 1,209,024 Gaussian components (it is 25% compared with the original HMM) was decreased only 1 point compared with that of the original HMM. Moreover, the multi-mixture HMM showed higher performance than the multi-condition HMM with the same number of Gaussian components.

4. Conclusion

In this paper, we propose a new construction method of HMMs dealing with various kinds of background noise. At first, each HMM dealing with a single noise is trained for each background noise, and then those HMMs are combined into the multi-mixture HMM. From the experimental results, the multi-mixture HMMs gave the highest recognition performance for "known" and "unknown" noises.

In order to reduce the number of Gaussian components of the multi-mixture HMMs, we proposed a reduction method of Gaussian components. It can decrease the number of Gaussian components with slight deterioration of recognition performance.

5. References

[1] S. V. Vaseghi and B. P. Milner, "Noise compensation methods for hidden Markov model speech recognition in adverse environments", IEEE Trans. Speech and Audio Processing, vol. 5, no. 1, pp. 11–21, 1997.

[2] J. Adcock, Y. Gotoh, D. Mashao, and H. Silverman, "Microphone-array speech recognition via incremental MAP training", Proc. ICASSP'96, pp. 897–900, 1996.

[3] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction", IEEE Trans. Acoustics, Speech and Signal Processing, vol. ASSP-27, no. 2, pp. 113–120, 1979.

[4] A. P. Varga and R. K. Moore, "Hidden Markov model decomposition of speech and noise", Proc. ICASSP'99, pp. 845–848, 1999.

[5] M. J. F. Gales and S. J. Young, "HMM recognition in noise using paralled model combination", Proc. EUROSPEECH, pp. 837–840, 1993.

[6] F. Martin, K. Shikano, and Y. Minami, "Recognition of noisy speech by composition of hidden Markov models", Proc. EUROSPEECH, pp. 1031–1034, 1993.

[7] S. Itahashi, "Creating speech copora for speech science and technology", IEICE Transactions, vol. E74-A, no. 7, pp. 1906–1910, July 1991.

[8] M. Ida and S. Nakamura, "Rapid environment adaptation method based on HMM composition with prior noise GMM and multi-SNR models for noisy speech recognition", The Transactions of the Institute of Electronics, Information and Communication Engineerings, vol. J86-D-II, no. 2, pp. 195–203, Feb. 2003.