Robust Feature Combination for Speech Recognition using Linear Microphone Array in a Car

Yasunari Obuchi[1] *and Nobuo Hataoka*[2]

[1] Central Research Laboratory, Hitachi Ltd, Kokubunji, Tokyo 185-8601, Japan
[2] Department of Electronics and Intelligent Systems, Graduate School of Electronics, Tohoku Institute of Technology, Sendai, Miyagi 982-8577, Japan

ABSTRACT

When speech recognition is performed in a car environment, there are two important robustness issues that should be taken into account. The first robustness is related to the noisy acoustic condition, and it has been one of the most popular research topics of in-vehicle speech recognition. In contrast, the second robustness, which is related to unstable calibration of the audio input, has not attracted much attention. Consequently, the performance of speech recognition would degrade greatly in a real application if the input device such as a microphone array is badly calibrated. We propose robust feature combination in the MFCC domain using speech inputs from a linear microphone array. It realizes robust (from both the noise and calibration viewpoints) and practical speech recognition applications in car environments. Even a simple MFCC averaging approach is effective, and a new algorithm, Hypothesis-Based Feature Combination (HBFC), improves the performance. We also extend cepstral variance normalization as variance re-scaling, which makes the feature combination approach more robust. The advantages of the proposed algorithms are confirmed by the experiments using the data recorded in a moving car.

1. INTRODUCTION

There have been a lot of studies to realize robust speech recognition in noisy environments such as in cars and in public spaces (stations, stores, airports, etc.). Speech recognition using a microphone array is one of the successful approaches to realize such robustness. In most cases, microphone array techniques are implemented in the time domain or in the spectral domain to enhance the input signal to obtain better recognition performance. It is because those techniques are mainly focusing on the phase difference between the target signal and interfering noises. If the noise is directional, the phase difference can be measured clearly, and those typical microphone array approaches would work effectively. However, they are less effective for non-directional noises. In car environments, the speech recognition system is surrounded by various noise sources, and the directional noise assumption does not hold. The array processing algorithm should treat non-directional noises effectively.

The second problem is the robustness in the cases when calibration of the microphones and audio systems are not maintained well. When we use a microphone array in real applications, it is hard to maintain the stability of microphone characteristics, and it is the reason why there is a large discrepancy between the performance in the laboratory and in the real field. Hence the assumption that the power spectra of multiple inputs are identical does not hold, and we have various cepstral (MFCC) features corresponding to the multiple microphones. It is then reasonable to expect that combining them in the cepstral domain may improve the speech recognition performance. In fact, the Gaussian statistical nature of the cepstral features of speech suggests the isotropic nature of the cepstrum space, and approves the effectiveness of feature combination in the cepstral domain. Moreover, since the speech can be modeled precisely in the cepstral domain using hidden Markov models (HMMs), we can take advantage of the prior knowledge about speech if we work in the cepstral domain.

In [1], we studied MFCC combination of the dual-microphone system, and proposed Hypothesis-Based Feature Combination (HBFC). In [2], the concept of feature combination in the cepstral domain was extended to a linear microphone system, and a problem of cepstral variance normalization was raised. In this paper, these issues are investigated in detail, and an approach to solve the cepstral variance normalization problem will be presented.

2. MFCC AVERAGE AND VARIANCE RE-SCALING

In [1], we showed that we can improve the speech recognition accuracy simply by averaging two MFCC sequences of the dual-microphone systems. Naturally, it can be extended to a multiple-input system as:

$$\mathbf{x}_{ave} = \frac{1}{N}\sum_{i=0}^{N}\mathbf{y}_i \qquad (1)$$

where $\mathbf{y}_i = \{y_{itd} \mid 1<=t<=T,\ 1<=d<=D\}$ is the MFCC feature vector made from the observed signal by the *i*-th microphone, and $\mathbf{x}_{ave}$ is the corresponding combined feature vector. *N* is the number of microphones, *T* is the number of time frames, and *D* is the dimension of MFCC used.

However, this simple averaging does not work well, especially in the case of large *N*. Taking into account that the arithmetic mean in the MFCC domain is almost equivalent to the geometric mean in the power spectral domain, it tends to take a smaller value. In particular, such a change occurs if the observed MFCC values have largely different values. Figure 1 shows the results of our preliminary experiments, in which we calculated framewise ratios of absolute values of two feature vectors obtained by single input and by averaging seven inputs and made a histogram. It is shown that the feature vector became smaller by averaging in more than 65% frames. The mean of the ratio was 0.95.

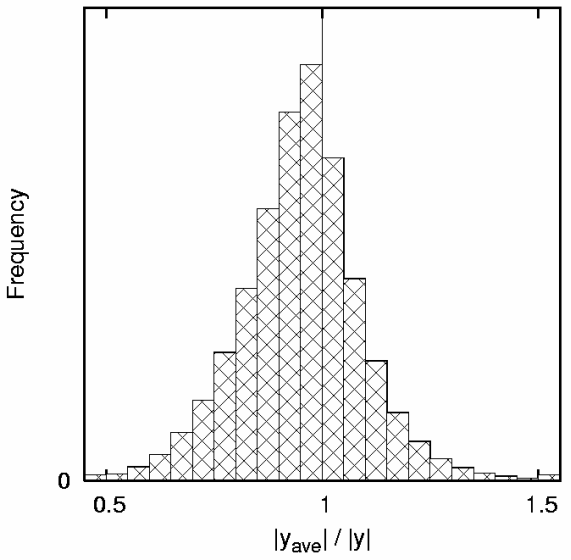

Fig.1 Comparison of single input feature and averaged feature.

A simple solution to this problem is to multiply a fixed normalization factor to all the MFCC values

$$\mathbf{z}_{ave} = \alpha\mathbf{x}_{ave} \qquad (2)$$

If we use Cepstral Mean Normalization (CMN) [3], the cepstral mean does not change by eq. (2), and eq. (2) can be interpreted as re-scaling of the cepstral variance.

3. GMM-BASED VARIANCE NORMALIZATION

As shown in Fig. 1, the cepstral variance of the averaged feature vector tends to be smaller than the original one. It can be compensated by eq. (2), but the optimal value of the scaling factor alpha is not obvious from eq. (2) itself. Therefore, we use Gaussian Mixture Models (GMMs) to estimate the effectiveness of a specific value of alpha.

It is natural to expect that the GMM score of the original feature vector is higher than that of the corrupted feature vector. We carried out a set of preliminary experiments, in which we added all frame-level GMM scores to obtain an utterance-level GMM score, using various values of alpha. Contrary to our expectation, it was revealed that the GMM score takes the maximum with a very small value of alpha such as 0.2, whether the original or averaged feature vector was re-scaled. The results indicate that either too large or too small GMM score means highly corrupted feature vectors. Therefore, our criteria must not be the absolute value of the GMM score, but the relative value of the GMM score compared with the single input feature vector.

Equation (3) is the definition of so-called GMM-based variance normalization, proposed in this paper as the conclusion of the above discussion.

$$\alpha_{opt} = \arg\max_i(|S(\alpha_i\mathbf{x}_{ave}) - S(\mathbf{y}_0)|) \qquad (3)$$

Here, the optimal value of alpha is chosen from a finite set of candidate values. $S(\mathbf{x})$ is the GMM score of the feature vector $\mathbf{x}$.

4. HYPOTHESIS-BASED FEATURE COMBINATION OF MULTIPLE INPUTS

As the huge success of the speech recognition research indicates, the speech signal can be modeled precisely in the cepstral (MFCC) domain. Working in the MFCC domain has an advantage that we can use the prior knowledge about the speech model in a framework of the feature combination. Figure 2 shows the schematic diagram of Hypothesis-Based Feature Combination (HBFC) applied to more than two microphone inputs. On

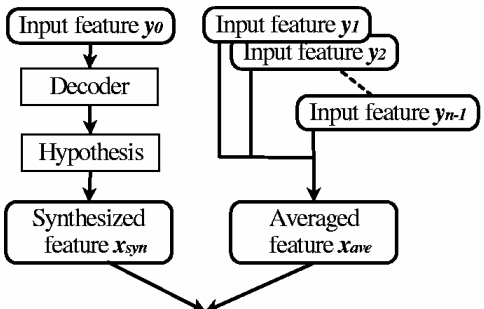

Fig. 2. Schematic diagram of Multiple-input Hypothesis-Based Feature Combination.

the left hand side, one channel (typically the central microphone) was chosen to be used in the first decoding process, and the obtained speech recognition hypothesis is used to synthesize the feature vector using the forced-alignment result and HMM. Feature synthesis is a simple procedure in which the mean vectors of the HMM state sequence (result of forced-alignment) are simply concatenated. On the right hand side, feature vectors of all the other channels are averaged. Finally, the outputs of two separate processes are combined by taking a simple weighted average.

$$\begin{aligned} \mathbf{x}_{HBFC} &= w\mathbf{x}_{syn} + (1-w)\mathbf{x}_{ave} \\ &= w\mathbf{x}_{syn} + \frac{1-w}{N-1}\sum_{k=1}^{N-1}\mathbf{y}_k \end{aligned} \quad (4)$$

where $\boldsymbol{x}_{syn}$ is the output of the left hand side of Fig. 2, $\boldsymbol{x}_{ave}$ is the output of the right hand side of Fig. 2, N is the number of inputs, and w is the weight parameter. In our experiments, N=7 and w=0.1 are used.

5. EXPERIMENTAL RESULTS

5.1. Database and setup

We carried out several sets of experiments to evaluate various implementations of feature combination in the MFCC domain. The evaluation data was recorded in a real car which was running on urban roads. The database is made of 3620 utterances in total, uttered by 18 speakers (11 male and 7 female). The task is 152 Japanese POI (points of interest) isolated word recognition (IWR) to input the destination to the navigation system. The speaker sat in the passenger seat, and was prompted each time to speak by a beep. Each utterance was roughly endpointed by a fixed time-window from the beep position, so the utterance contains relatively high-percentage of the non-speech segment. The utterances were recorded by a microphone array, which is made of seven linearly located microphones. These microphones were numbered from 1 to 7 in the direction from the driver's side to the window side, and placed at intervals of 10cm, 5cm, 5cm, 5cm, 5cm, and 10cm. The average SNR of all recorded data was estimated as -3,4dB, but most of the noise exists in lower frequency range, and the estimated SNR increased to 10.0dB after applying a bandpass filter with a 400Hz-5500Hz pass band. More details of the database can be found in [4].

For the recognition experiments, we prepared our original decoder and acoustic models. The acoustic models were made of Japanese triphones (3 states/model, 6 Gaussian mixtures/state) and trained using 16 hours of clean speech. All speech data were sampled by 16kHz, converted to a 13 dimension MFCC feature vector every 10ms, and either CMN or Cepstral Mean and Variance Normalization (MVN) [5] was applied.

5.2. MFCC average

Figure 3 shows the recognition results obtained by the averaged and re-scaled MFCC features. The recognition rates of single input (CMN:87.43% and MVN:86.32%) and a delay-and-sum beamformer [6] (CMN:88.20%) are shown by the horizontal lines for comparison.

When we look at the data points on the line of alpha=1.0, we can confirm the recognition rates without

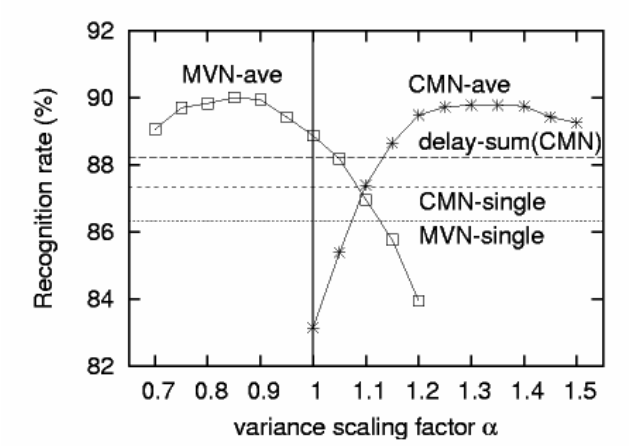

Fig. 3. Experimental results of MFCC average.

variance re-scaling, which are 83.15% (CMN) and 88.87% (MVN). Since the feature vectors after CMN have larger variety over microphones than MVN, the variance of their average tends to be smaller, and it results in the lower recognition rate. However, if we can use the optimal value of alpha (1.30 for CMN and 0.85 for MVN), the recognition rate would be much higher than the standard delay-and-sum beamformer.

Next, we tried GMM-based variance normalization. The results obtained with CMN are shown in Fig. 4. In this figure, "GMM score" means the sum of the utterance-level relative GMM scores:

$$S^{rel}(\alpha) = \sum_{j=1}^{N} |S(\alpha \mathbf{x}_{ave,j}) - S(\mathbf{y}_{0,j})| \qquad (5)$$

where subscript j represents each utterance and N is the number of test utterances. It is clear that the curve of the recognition rate of the average and the GMM score are almost symmetric. The GMM score takes the smallest value with alpha=1.25, with which the recognition rate is 89.72%. If we apply GMM-based variance normalization utterance by utterance, the recognition rate becomes 90.00%, which is shown as "Rate (optimized)" in the figure.

Figure 5 shows the equivalent results obtained with MVN. It is not as symmetric as in Fig. 4, but there is a similar tendency, and the recognition rate of the GMM-based variance normalization is 88.26%, which is still better than the delay-and-sum beamformer.

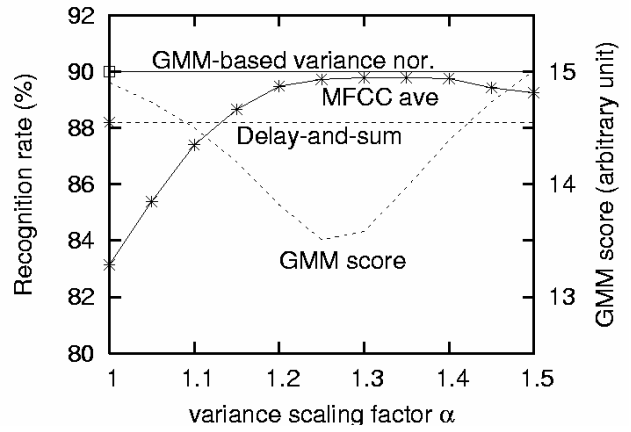

Fig. 5. Comparison of GMM scores, recognition rates for averaged features, and recognition rates of GMM-based variance normalization. All feature vectors were normalized by CMN.

5.3. Hypothesis-Based Feature Combination

In the final set of experiments, we tried HBFC with variance re-scaling. Figure 7 shows the results obtained with HBFC and CMN. The symmetric nature of the GMM score and the recognition rate is preserved well. With a fixed scaling factor, the highest recognition rate of 89.93% was obtained with alpha=1.5, and the recognition rate of GMM-based variance normalization is still higher than it (90.00%).

Figure 8 shows the results obtained with HBFC and MVN. In this case, the lowest GMM score was obtained with the alpha even lower than 0.7, and the two curves

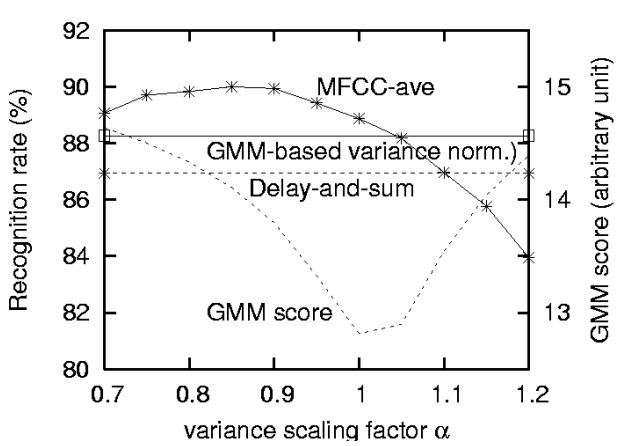

Fig. 6. Comparison of GMM scores, recognition rates for averaged features, and recognition rates of GMM-based variance normalization. All feature vectors were normalized by MVN.

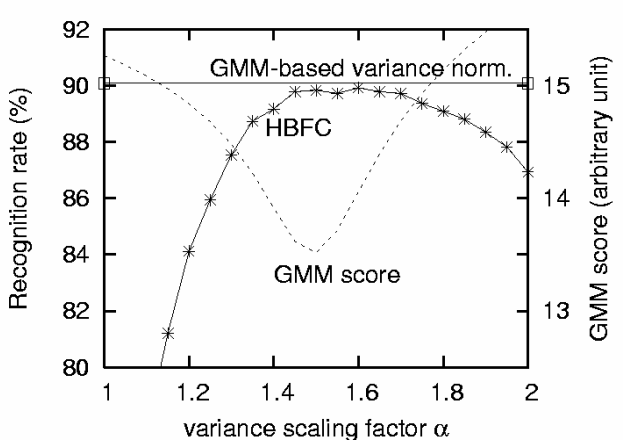

Fig. 7. Comparison of GMM scores, recognition rates for HBFC, and recognition rates of GMM-based variance normalization. All feature vectors were normalized by CMN

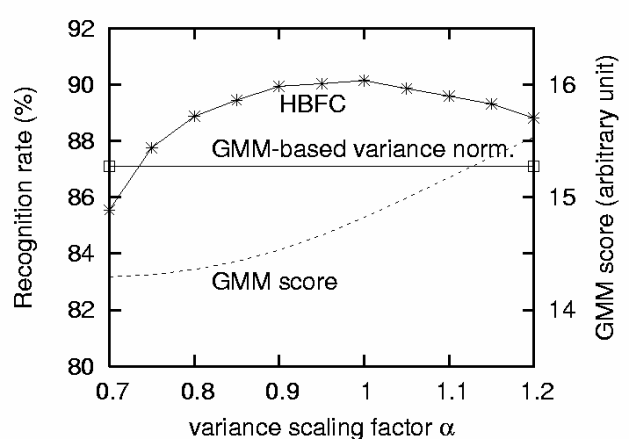

Fig. 8. Comparison of GMM scores, recognition rates for HBFC, and recognition rates of GMM-based variance normalization. All feature vectors were normalized by MVN

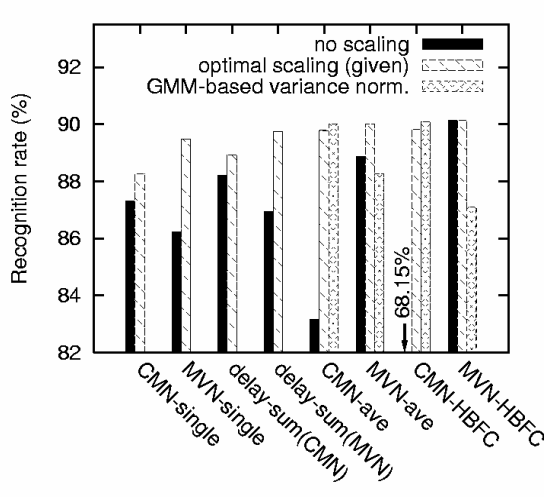

Fig. 9. Comparison of various algorithms evaluated in this paper. The highest recognition rate in all is 90.14% with MVN-HBFC (optimal scaling and no scaling coincided).

(GMM score and HBFC recognition rate) are not symmetric. Consequently, the recognition rate of GMM-based variance normalization is lower than the HBFC recognition rates near alpha=1.0. However, it is worth mentioning that the recognition rate of GMM-based variance normalization is higher than that of HBFC with alpha=0.7, where the GMM score is smallest (in this figure). The degradation of GMM-based variance normalization from the highest recognition rate of HBFC is 3.07 points.

6. CONCLUSIONS

In this paper, we have shown how speech recognition accuracy can be improved by various ways of feature combination in the MFCC domain. Simple averaging of MFCC features tends to lower the recognition rate, especially when only the cepstral means are normalized (CMN). However, such degradation can be explained by the fact that the averaged MFCC features tend to be smaller than the original MFCC feature, and it can be compensated by introducing variance re-scaling.

Next, we proposed a new algorithm to estimate the optimal value of the scaling factor, using the GMM score of the re-scaled feature vector and the single input feature vector. Experimental results have shown that the proposed algorithm worked quite well if the feature vectors were normalized by CMN.

Figure 9 shows the comparison of the algorithms evaluated in this paper. As mentioned in [2], MVN-HBFC gives the highest recognition rate without variance re-scaling, but the problem with CMN-ave and CMN-HBFC were solved by the proposed algorithm.

7. ACKNOWLEDGMENTS

The authors are thankful to Prof. Sadaoki Furui of Tokyo Institute of Technology and Prof. Tetsunori Kobayashi of Waseda University for their valuable comments. A part of this work was supported by New Energy and Industrial Technology Development Organization (NEDO), Japan.

8. REFERENCES

[1] Y. Obuchi, "Hypothesis-based feature combination for dual-microphone speech recognition," *Proc. HSCMA*, Piscataway, NJ, USA, 2005.

[2] Y. Obuchi and N. Hataoka, "Hypothesis-based feature combination of multiple speech inputs for robust speech recognition in automotive environments," *Proc. Interspeech2006-ICSLP*, Pittsburgh, PA, USA, 2006.

[3] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," Journal of Acoustical Society of America, Vol.55, No.6, pp.1304-1312, 1974.

[4] Y. Obuchi and N. Hataoka, "Development and evaluation of speech database in automotive environments for practical speech recognition systems," *Proc. Interspeech2006-ICSLP*, Pittsburgh, PA, USA, 2006.

[5] J. P. Openshaw and J. S. Mason, "On the limitations of cepstral features in noise," *Proc. ICASSP*, Adelaide, Australia, 1994.

[6] W. Kellermann, "A self steering digital microphone array," *Proc. ICASSP*, Toronto, Canada, 1991.