Hypothesis-Based Feature Combination of Multiple Speech Inputs for Robust Speech Recognition in Automotive Environments

Yasunari Obuchi and Nobuo Hataoka

Central Research Laboratory, Hitachi Ltd.
Kokubunji, Tokyo 185-8601, Japan
obuchi@rd.hitachi.co.jp, hataoka@crl.hitachi.co.jp

10.21437/Interspeech.2006-645

Abstract

In a microphone array system, feature combination in the MFCC domain can improve speech recognition accuracy. Multiple microphones provide different feature parameters such as MFCCs even if they have similar speech and noise signals, because of the phase difference and transmission characteristics. In this paper, we investigate how the recognition performance changes when we average multiple MFCC feature vectors. In addition, we extend Hypothesis-Based Feature Combination, which we formerly proposed for dual-microphone systems, to multi-input systems. Experimental results show that variance re-scaling is necessary when we combine multiple inputs with Cepstral Mean Normalization (CMN), in both MFCC average and HBFC. However, we can obtain better results without variance re-scaling if we use Mean and Variance Normalization (MVN) with MFCC average or HBFC. In the experiments using the database collected in a real automotive environment, HBFC-MVN reduced 22% of the recognition errors from the baseline single-microphone system.

Index Terms: robust speech recognition, feature combination, microphone array.

1. Introduction

Speech recognition accuracy in noisy environments can be improved by utilizing multiple speech inputs. A microphone array is used to collect multiple inputs, and there have been many studies of array signal processing to achieve better recognition performance. In this paper, we investigate various array processing techniques, aiming at robust speech recognition for practical user interface in automotive environments.

In most of the previous studies, speech (or speech feature) was enhanced in the time domain or the power spectral domain, and there were two assumptions related to these two domains. First, it was assumed that the interfering noises are directional and the observation process can be modeled as a linear combination of the delayed signals. Independent Component Analysis (ICA) is a typical solution for this model, which can separate multiple signals such as the voices from the driver and passenger seats using statistical independence of the signals. However, automotive applications are surrounded by various noises from outside of the vehicle, most of which should be treated as non-directional noise. In addition, inside of the vehicle is highly reverberant, and the independence requirement is not satisfied. The second assumption of the standard array processing is that the phase of the speech signal and the phase of the noise are not correlated and cancel out each other in the calculation of the power spectrum. Two-channel spectral subtraction [1] is a typical solution based on this assumption. However, as Droppo et al [2] pointed out, neglecting the cross term of the speech and noise signals is too naive an approximation even if they are uncorrelated. Moreover, it is impossible to maintain the completely equal frequency characteristics of the microphones. Therefore, the power spectra of the observed signals by the multiple microphones are not necessarily identical each other even if they share the same power spectra of the speech and noise. Hence we have various cepstral (MFCC) features corresponding to the multiple microphones, and it is worth combining them in the cepstral domain. 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 [3], we studied MFCC combination of the dual-microphone system, and proposed Hypothesis-Based Feature Combination. In this paper, we extend it to the multiple input system, where the straightforward extension may cause recognition accuracy degradation, and then attention must be paid to the variance scaling problem. We analyze the relationship between feature compensation and feature combination algorithms, and present how the recognition accuracy can be improved by the proposed method.

2. MFCC average and variance normalization

In this section, we start with a typical mixture model of speech and noise, which is described as

$$|Y_i(k)|^2 = |X(k)|^2 + |N(k)|^2 + |X(k)||N(k)|cos\theta_{ki} \quad (1)$$

where $X(k)$ and $N(k)$ are the k-th spectral component of the source and noise signals, $Y_i(k)$ is the k-th spectral component of the signal observed by the microphone i, and θ_{ki} is the phase difference of $X(k)$ and $N(k)$ observed by the microphone i. If we neglect the cross term of eq. (1), spectral subtraction:

$$|\hat{X}_i(k)|^2 = |Y_i(k)|^2 - |\hat{N}(k)|^2 \quad (2)$$

where $|\hat{X}_i(k)|^2$ and $|\hat{N}(k)|^2$ are the estimation of the source and noise signals, is a good approximation. However, eq. (1) suggests that we should have the same MFCC values for all microphones if they have the same values of $|X(k)|$ and $|N(k)|$ (far-field approximation). The experimental finding that the multiple microphones have different MFCC values leads us to the conclusion that we cannot neglect the cross term. Accordingly, we admit that the multiple

inputs provide different feature vectors, and we propose to combine them to obtain better MFCC sequence for speech recognition.

In [3], we showed that we can improve the speech recognition accuracy simply by averaging two MFCC sequences of the dual-microphone system. A straightforward extension of MFCC averaging to the multiple-input system is:

$$\mathbf{x}_{ave} = \frac{1}{N} \sum_{i=0}^{N} \mathbf{y}_i \tag{3}$$

where $\mathbf{y}_i = \{y_{itd} | 1 \leq t \leq T, 1 \leq d \leq D\}$ 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 equivalent to the geometric mean in the power spectral domain, a general equation

$$(\prod_{k=1}^{N} x_k)^{1/N} \leq \frac{1}{N} \sum_{k=1}^{N} x_k \quad \text{for any positive } \{x_k\} \tag{4}$$

indicates that the average of many MFCC values tends to become smaller than expected, especially if the observed MFCC values have largely different values. Our preliminary experiments showed that a simple MFCC average of many microphones tends to output the same hypothesis, which is also outputted if we multiply a small constant α (< 1.0) to all MFCC values. We will later examine if such degradation can be avoided by applying appropriate scaling of the MFCC values.

3. 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. Previously, we proposed Hypothesis-Based Feature Combination (HBFC) for dual-microphone systems, in which the speech model is used to synthesize an MFCC sequence from the recognition hypothesis.

In [3], we treated the two input signals evenly and compared the results obtained by decoding one signal or the other. However, in this paper, we assume that the microphone array is linear and the user is sitting near the center of the array, and the central microphone provides near-optimal signal. Therefore, we decode the signal obtained by the central microphone, and mix the synthesized feature from the hypothesis with the inputs of the other microphones. Figure 1 shows the schematic diagram of the Multi-input HBFC described in this paper.

The procedure to synthesize the MFCC sequence is as follows. First, the input feature of one channel is decoded in a standard manner to obtain a recognition hypothesis. Next, the HMM state sequence corresponding to the hypothesis is force-aligned with the input feature. For each pair of the HMM state and the input feature, the likelihood values are compared for each Gaussian mixture, and finally the most likely Gaussian is selected for each state. After that, the synthesized feature can be obtained simply by concatenating Gaussian means of all selected Gaussian mixtures.

After the synthesized feature $\mathbf{x}_{syn}$ was obtained, it is mixed with the average of the inputs of the other microphones, by linear

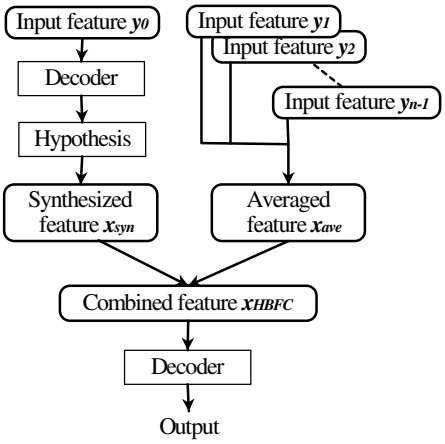

Figure 1: *Schematic diagram of Multi-input Hypothesis-Based Feature Combination.*

combination with a fixed weight parameter w as follows.

$$\begin{aligned} \mathbf{x}_{HBFC} &= w\mathbf{x}_{syn} + (1-w)\mathbf{x}_{ave} \\ &= w\mathbf{x}_{syn} + \frac{1-w}{N-1} \sum_{k=1}^{N-1} \mathbf{y}_k \end{aligned} \tag{5}$$

Here, we used the suffix 0 for the central microphone and suffix 1 to $N-1$ for the other microphones.

4. Experimental Results

4.1. Database and setup

We carried out several sets of experiments to evaluate various implementations of MFCC average and HBFC. 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. The variance of the estimated SNR over the microphones was smaller than the estimation error.

For the recognition experiments, we prepared our original decoder and two acoustic models. Each acoustic model was trained using phonetically balanced sentences recorded in a quiet room, spoken by 160 speakers (80 male and 80 female), 16 hours in total. The first and second acoustic models were trained with utterance-based CMN and MVN respectively. Our acoustic model is triphone-based (based on 34 Japanese phonemes), made of 1614 states with 6 mixtures each, and compressed by subvector quantization with 128 codewords in each of 9 subspaces. As the feature

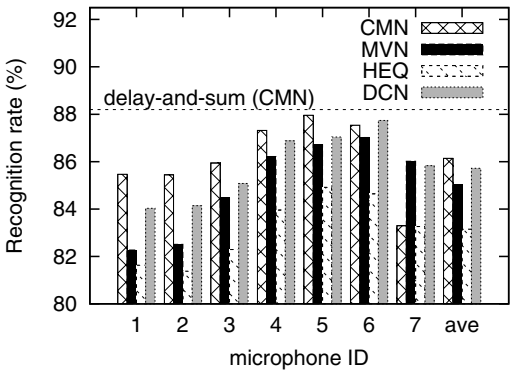

Figure 2: *Baseline performance of each single microphone.*

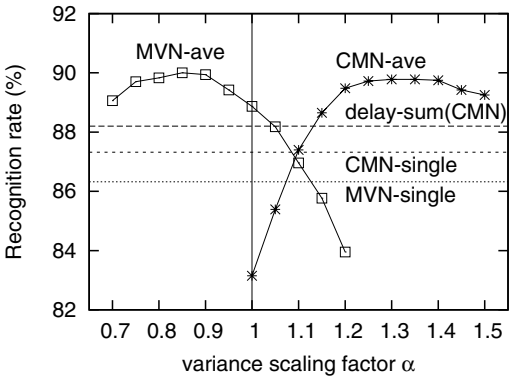

Figure 3: *Experimental results of MFCC average. CMN-single and MVN-single are the recognition rates of microphone #4 without variance re-scaling*

vector, 13 MFCCs including C0 and their first and second-order time derivatives were extracted every 10ms from the 16kHz sampled wave data.

Figure 2 shows the results of the baseline experiments. Since the task is IWR, the recognition rate was simply defined as the ratio of the correctly recognized utterances to the total utterances. In the baseline experiments, we tried four feature normalization algorithms: CMN, MVN, Histogram Equalization (HEQ) [4], and Delta-Cepstrum Normalization (DCN) [5]. In all cases, normalization was applied on the whole utterance basis (off-line). Among these four normalization algorithms, CMN achieved the best recognition rate, 87.32%, using the central microphone. DCN and MVN were close to CMN, and HEQ was less accurate. It was unexpected that microphone #5 and #6 achieved better recognition rate than the central microphone (#4), but the difference was rather small. The average of all microphones had the same tendency of the order of CMN, DCN, MVN, and HEQ, but the order was not kept in some specific microphones (#6 and #7).

We also tried a simple delay-and-sum beamformer and ICA [6]. In the delay-and-sum beamformer experiments, since the position of the speaker is almost fixed, we calculated the delay between microphones geometrically. The original 16kHz sampled data of 7 microphones were first upsampled to 64kHz, the fixed delays (5 pts for #1 and #7, 1 pt for #2 and #6, and nothing for #3, #4, #5) were applied, and then all data were added and downsampled to 16kHz. Using this beamformer, we obtained 88.20% recognition rate, which is slightly better than the baseline result. On the other hand, ICA could not improve the recognition rate at all, either applied in the time domain or in the frequency domain, and either using two (#3 and #5) or all microphones

4.2. MFCC average

Next, we averaged 7 MFCC feature vectors at each frame after feature normalization by CMN or MVN, and decoded them. The results were 83.15% (CMN) and 88.87% (MVN). Since the feature vectors after CMN have larger variety over microphones, their variance tends to become small by taking the average. It was proved by the fact that 6.63% of the utterances were recognized as the POI "Tourist Hotel" though only 0.55% of the utterances were that word and only 1.30% were recognized as that word in the baseline experiment.

To avoid such effect, we applied re-scaling of the MFCC parameters.

$$\mathbf{z}_{ave} = \alpha \mathbf{x}_{ave} \tag{6}$$

Since the cepstral mean was already subtracted, it is equivalent to re-scaling of the variance. Figure 3 shows how the recognition rate changes when we sweep the scaling factor α.

The average of 7 MFCC feature vectors after CMN provides better recognition rates if we use $\alpha > 1.0$. The best value of 89.78% was obtained with $\alpha = 1.30$ and $\alpha = 1.35$. Contrastingly, the average of 7 MFCC feature vectors after MVN has the peak near $\alpha = 1.0$, which provided the higher recognition rate of MVN-ave without variance re-scaling. The best value of MVN-ave was 90.00% with $\alpha = 0.85$.

Since it is impossible to know the optimal value of α without knowing the correct words (like supervised adaptation), it is an advantage of MVN-ave that we can obtain near-optimal recognition rate with $\alpha = 1.0$.

We also checked how the microphones with lower accuracy contribute to feature combination. As shown in Fig. 2, some microphones have lower input quality, and it is not certain if adding those microphones improves the recognition rate. Table 1 shows the experimental results with various microphone numbers and weights. The top row shows the recognition rate obtained by microphone #4 alone. The second row shows the results obtained by adding microphone #3 and #5, and the third row shows the results of all 7 microphones. These results confirmed that even the lower quality inputs contributed to improve the recognition rate by MFCC average. Next, we applied biased weights so that the central microphone has the larger weight. However, the best recogni-

Table 1: *Experimental results with various microphone numbers and weights*

# of mic.	weights	Recognition rate (%)	
		CMN (α=1.3)	MVN (α=1.0)
1		79.81	86.22
3	equal	89.01	87.60
5	equal	89.72	88.34
7	equal	89.78	88.87
7	1:2:3:4:3:2:1	89.67	88.59
7	1:4:9:16:9:4:1	89.59	88.20

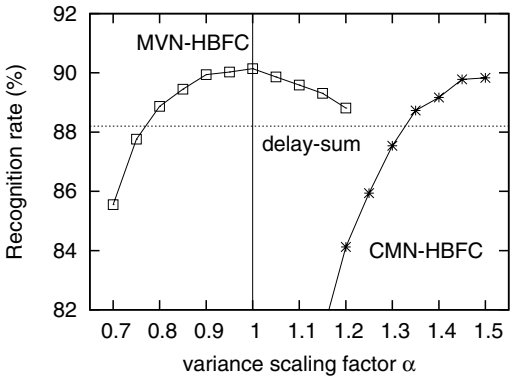

Figure 4: *Experimental result of Multi-input HBFC.*

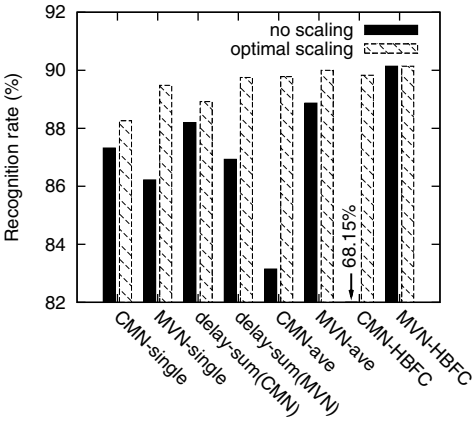

Figure 5: *Summary of the experiments.*

tion rate was obtained when we set the equal weights for all microphones.

4.3. Multi-input HBFC

In the final set of experiments, we tried HBFC with similar variance re-scaling. In eq. (5), the weight parameter w was set to 0.1 according to the results reported in [3]. Since the experiments in the previous work and in this paper do not share anything (data, decoder, task, language, etc.), adopting this value would be more persuasive if we obtain good results.

In the experiments without variance re-scaling, the recognition rate of CMN-HBFC degraded terribly to 68.15%. In that experiment, 15.5% of the utterances were recognized as "Tourist Hotel," which is suggesting the necessity of variance re-scaling. Contrastingly, MVN-HBFC provided additional improvements, and the recognition rate was 90.14%. It means 22% relative error reduction from the baseline (CMN-single). Then we applied variance re-scaling to HBFC, and the results are shown in Fig. 4. CMN-HBFC has a more drastic curve, and we can obtain satisfactory results if we set around $\alpha = 1.5$. In contrast, degradation caused by α is small in MVN-HBFC, and the optimal recognition rate was obtained without variance re-scaling, which is slightly better (3.0% relative error reduction) than the optimal case of CMN-HBFC.

5. 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 in the MFCC domain tends to make the variance of MFCC features small, so it is beneficial to introduce variance re-scaling. However, it is an ideal case to know the optimal value of the scaling factor, so the comparison should be interpreted as mere reference.

Figure 5 is the summary of the experiments using various methods. When we compare the results obtained without variance re-scaling, MVN-HBFC provides significant improvement. The relative error reduction is 22% from CMN-single, 28% from MVN-single, 16% from the delay-and-sum beamformer with CMN, and 11% from MVN-ave. On the other hand, CMN-ave and CMN-HBFC degraded the accuracy. If we assume that we have a chance to know the optimal value of the variance scaling factor α, the difference between various methods becomes small. However, MVN-HBFC achieved the best recognition rate even under such assumption, which has approved the robustness of the proposed method.

In [7], we discussed how the weight parameter w of HBFC should be optimized. In this paper, we added another issue related to optimization of the scaling factor α. So far we used reasonable setting $w = 0.1$ and $\alpha = 1.0$ in MVN-HBFC and obtained good results. However, if we go forward and try to extend the applicability of the proposed method, these parametric optimization issues would be an important future work.

6. 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. This work was supported by New Energy and Industrial Technology Development Organization (NEDO), Japan.

7. References

[1] H.-Y. Kim, F. Asano, Y. Suzuki, and T. Sone, "Speech Enhancement Based on Short-time Spectral Amplitude Estimation with Two-channel Beamformer," *IEICE Trans.*, Vol.E79-A, No.12, pp.2151-2158, 1996

[2] J. Droppo, A. Acero, and L. Deng, "A Nonlinear Observation Model for Removing Noise from Corrupted Speech Log Mel-spectral Energies," *Proc. ICSLP*, Denver, CO, USA, 2002

[3] Y. Obuchi, "Hypothesis-Based Feature Combination for Dual-Microphone Speech Recognition," *Proc. HSCMA*, Piscataway, NJ, USA, 2005

[4] A. de la Torre, et al., "Non-linear Transformation of the Feature Space for Robust Speech Recognition," *Proc. ICASSP*, Orlando, FL, USA, 2002

[5] Y. Obuchi and R. M. Stern, "Normalization of Time-derivative Parameters Using Histogram Equalization," *Proc. EUROSPEECH*, Geneva, Switzerland, 2003

[6] N. Murata, S. Ikeda, and A. Ziehe, "An Approach to Blind Source Separation Based on Temporal Structure of Speech Signals," *BSIS Technical Report*, 00-6, 2000

[7] Y. Obuchi, "Mixture Weight Optimization for Dual-microphone MFCC Combination," *Proc. ASRU*, San Juan, Puerto Rico, 2005