ROBUST DISTANT SPEECH RECOGNITION BY COMBINING POSITION-DEPENDENT CMN WITH CONVENTIONAL CMN

Longbiao Wang, Norihide Kitaoka, Seiichi Nakagawa

Department of Information and Computer Sciences, Toyohashi University of Technology, Japan
{wang, kitaoka, nakagawa}@slp.ics.tut.ac.jp

ABSTRACT

We proposed an environmentally robust speech recognition method based on Position-Dependent Cepstral Mean Normalization (PDCMN) to compensate for channel distortion depending on speaker position. PDCMN can efficiently compensate for the channel transmission characteristics while it cannot normalize speaker variation because position-dependent cepstral mean does not contain speaker characteristics. Conventional CMN can compensate for the speaker variation while it cannot obtain good recognition performance for short utterances. In this paper, we propose a robust distant speech recognition by combining position-dependent CMN with the conventional CMN to address the above problems. The position-dependent cepstral mean is linearly combined with conventional cepstral mean with following two types of processing. The first method is to use a fixed weighting coefficient over whole test data to obtain the combinational CMN, which is called *fixed-weight combinational CMN*. The second method is to calculate the output probability of multiple features compensated by a variable weighting coefficient at each frame, and a single decoder using these output probabilities is used to perform speech recognition, which is called *variable-weight combinational CMN*. We conducted the experiments of our proposed method using small vocabulary (100 words) distant isolated word recognition in a real environment. The proposed *variable-weight combinational CMN* method achieved a relative error reduction rate of 56.3% from conventional CMN and 22.2% from PDCMN, respectively.

Index Terms— Robust speech recognition, distant-talking environments, position-dependent CMN, conventional CMN, multiple microphone processing

1. INTRODUCTION

Automatic speech recognition (ASR) systems are known to perform reasonably well when the speech signals are captured by a close-talking microphone. However, there are many environments where the use of close-talking microphone is undesirable for reasons of safety or convenience. Hands-free speech communication [2, 3] has been more and more popular in some special environments such as an office or a cabin of a car. Unfortunately, in a distant environment, channel distortion may drastically degrade speech recognition performance. This is mostly caused by the mismatch between the practical environment and the training environment.

Compensating an input feature is the main way to reduce a mismatch. Cepstral Mean Normalization (CMN) has been used to reduce channel distortion as a simple and effective way of normalizing the feature space [4]. CMN reduces errors caused by the mismatch between test and training conditions, and it is also very simple to implement. Thus, it has been adopted in many current systems. However, the system should wait until the end of speech to activate the recognition procedure when adopting the conventional CMN [4]. The other problem is that the accurate cepstral mean can not be estimated especially when the utterance is short. However, the recognition of short utterances such as commands, city names etc. is very important in many applications. In [1], we proposed a robust speech recognition method using a new real-time CMN based on speaker position, which we call Position-Dependent CMN (PDCMN).

PDCMN can indeed efficiently compensate for the channel transmission characteristics depending on speaker position, but it cannot normalize the speaker variation because position-dependent cepstral mean does not contain speaker characteristics. On the contrary, the conventional CMN can compensate for the speaker variation. It, however, cannot obtain good recognition performance for short utterances.

In this paper, we propose a robust distant speech recognition by combining position-dependent CMN with conventional CMN to address the above problems. The *a priori* estimated position-dependent cepstral mean is linearly combined with utterance-wise cepstral mean with following two types of combination method. The first method is to use a fixed weighting coefficient over whole test data to obtain the combinational CMN, which is called *fixed-weight combinational CMN*. However, the optimal weight seems to depend on the the speaker position and the length of the utterance to be recognized. Thus, a fixed weighting coefficient could not obtain the optimal result. A variable weighting coefficient may obtain a better performance. A single input feature compensated by the combinational cepstral means with different weighting coefficients generates multiple input features. Thus, the problem turns to how to obtain the optimal performance given the multiple input features. Voting on the different hypotheses generated from the multiple input features has been studied in [1, 5]. In [6], a new algorithm to select a suitable channel for speech recognition using the output of the speech recognizer was proposed. All above methods used the output hypotheses generated by multiple decoders to estimate the final result.

In our previous study [7], the combination of multiple input streams at frame level using a single decoder was proposed. A more robust performance with a lower computational cost was achieved in compared with the method proposed in [1]. In this paper, we extend this method to the combinational CMN. The second method of the combinational CMN is to calculate the output probability of each input feature at frame level, and a single decoder using these output probabilities is used to perform speech recognition, which is called *variable-weight combinational CMN*. It is very easy to implement in both isolated word recognition systems and continuous speech recognition systems. Furthermore, the proposed combinational CMN is also integrated with the multiple microphone-array processing proposed in [7] which uses multiple microphone-arrays to obtain more robust spatial filtering in a real environment.

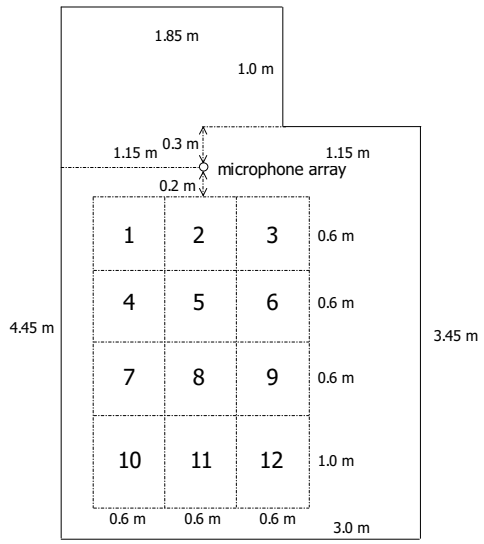

Fig. 1. Room configuration (room size: (W)3 m x (L)3.45 m x (H)2.6 m)

2. CONVENTIONAL CMN AND PDCMN

Convolutional noise (channel distortion) can be compensated by CMN in the cepstral domain as:

$$\tilde{C}_t = C_t - \Delta C, \qquad (t = 0, ..., T). \tag{1}$$

where $\tilde{C}_t$ and C_t are compensated and original cepstrums at time frame t, respectively.

In conventional CMN, the compensation parameter ΔC is approximated by:

$$\Delta C \approx \bar{C}_t - \bar{C}_{train}, \tag{2}$$

where $\bar{C}_t$ and $\bar{C}_{train}$ are cepstral means of utterances to be recognized and those to be used to train the speaker-independent acoustical models, respectively. This method stands for the assumption that the test utterance is phonetically well balanced, but the assumption may not be correct especially when the utterance is short.

We proposed an environmentally robust channel compensation method called Position-Dependent CMN (PDCMN) [1]. The new compensation parameter for PDCMN is defined by:

$$\Delta C = \bar{C}_{position} - \bar{C}_{train}, \tag{3}$$

where $\bar{C}_{position}$ is the cepstral mean of utterances affected by the transmission characteristics between a certain position and the microphone.

In our experiments in Section 5, we divide the room into 12 areas as Fig. 1 and measure the $\bar{C}_{position}$ corresponding to each area. The system estimates the speaker position in a 3-D space based on microphone arrays [8]. The system adopts the compensation parameter (that is, $\bar{C}_{position}$) corresponding to the estimated position and compensates the distortion using Equations (1) and (3) [1] and performs speech recognition.

[1] ΔC derived from Equation (2) includes the individual difference of articulation and thus compensates the speaker variability. ΔC derived from Equation (3), on the other side, should be speaker-independent for speaker-independent recognition system.

3. COMBINING PDCMN WITH CONVENTIONAL CMN

3.1. Fixed-weight combinational CMN

To compensate the channel distortion and speaker characteristics simultaneously, position-dependent cepstral mean is linearly combined with the conventional cepstral mean. The new compensation parameter ΔC for combinational CMN is defined by:

$$\Delta C = \lambda \bar{C}_{position} + (1 - \lambda)\bar{C}_t - \bar{C}_{train}, \tag{4}$$

where λ denotes a weighting coefficient. When using a fixed λ to the whole test data, we call this method *fixed-weight combinational CMN*.

3.2. Variable-weight combinational CMN

In Section 3.1, a fixed weighting coefficient λ is used to combine PDCMN with the conventional CMN. The effect of the channel distortion (that is, position-dependent cepstral mean) depends on speaker position, and the confidence of estimated speaker characteristics (that is, the conventional cepstral mean) depends on the length of the utterance. Therefore, the weighting coefficient λ should be adjusted depending on the speaker position and the length of the utterance. Given a set of variable weights λ_s, an automatic decision algorithm of the optimal weighting coefficient λ is required. A single input feature compensated by the combinational cepstral means with different weighting coefficients generates multiple input features. Thus, the problem turns to how to obtain the optimal performance given the multiple input features.

In our previous study [7], an optimal input decision algorithm which calculates the output probability of each input stream at frame level and selects the input with maximum probability as the optimal input was proposed. We extend and modify this algorithm to the so-called *variable-weight combinational CMN*.

For multiple inputs, a conventional Viterbi algorithm [9] is used for each input stream, k, and the probability $\alpha(t, j, k)$ of the most likely state sequence at time *t* which has generated the observation sequence $O_k(1) \cdots O_k(t)$ (until time t) of *k-th* input ($1 \le k \le K$) and ends in state j is defined by:

$$\begin{aligned} \alpha(t, j, k) &= \max_{1 \le i \le S} \{\alpha(t-1, i, k) a_{ij} b_j(O_k(t))\}, \\ O_k(t) &= \tilde{C}_t - (\lambda_k \bar{C}_{position} + (1 - \lambda_k)\bar{C}_t - \bar{C}_{train}). \end{aligned} \tag{5}$$

where $a_{ij} = P(s_t = j | s_{t-1} = i)$ is the transition probability from state i to state j, $1 \le i, j \le S$, $2 \le t \le T$; and $b_j(O_k(t))$ is the output probability for an observation sequence $O_k(t)$ at state j. λ_k is the *k-th* weighting coefficient. In this conventional multiple-decoder method, the Viterbi algorithm is performed for each input stream independently, so K (the number of input streams) times computational complexity is required. Thus, both the calculation of output probability and the rest of the processing cost such as finding a best path (state sequence), and so forth, are K times that of a single input.

In order to use a single decoder for multiple inputs, we modify the Equation (5) as follows:

$$\alpha(t, j) = \max_{1 \le i \le S} \{\alpha(t-1, i) a_{ij} \max_k b_j(O_k(t))\}. \tag{6}$$

This method is called *single decoder processing*. In Equation (6), the maximum output probability of all K inputs at time t and state j is used. So only one best state sequence for all K inputs using the maximum output probability of all K inputs is obtained. This means that extra $K-1$ times the calculation of only the output probability is required compared to that of a single input. Furthermore, the derivatives of the K input cepstrums compensated by different

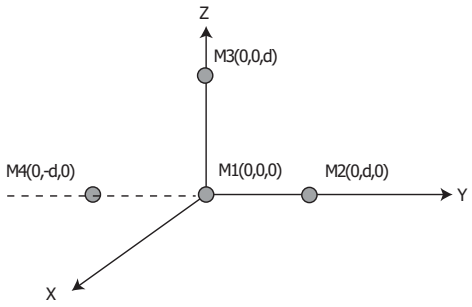

Fig. 2. Microphones' setup (d = 20 cm)

combinational cepstral means have same values. Thus, the calculation depending only on the derivatives can be shared with the input streams.

4. MULTIPLE MICROPHONE-ARRAY PROCESSING

Many microphone array-based speech recognition systems have successfully used delay-and-sum processing to improve recognition performance because of its spatial filtering ability and simplicity, so it remains the method of choice for many array-based speech recognition systems [3, 10]. Beamforming can suppress reverberation for the speech source of interest. Beams with different properties would be formed by the array structure, sensor spacing and sensor quality [10]. The multiple microphone-array processing using the *single decoder processing* described in Section 3.2 could obtain a more robust performance than a single beamforming [7]. We also integrate the combinational CMN with multiple microphone-array processing.

In our research, the four microphones are set as shown in Fig. 2. Array 1 (microphone 1, 2, 3), array 2 (microphone 1, 2, 4), array 3 (microphone 1, 3, 4), array 4 (microphone 2, 3, 4) and array 5 (microphone 1, 2, 3, 4) are used as individual arrays, and thus we can obtain 5 channel input streams using delay-and-sum beamforming. These streams are used as inputs of the *single decoder processing* to obtain the final result. We call this method *multiple microphone-array processing*. By combining the combinational CMN with *multiple microphone-array processing*, the maximum output probability at state j for an observation sequence $O_{c,k}(t)$ of input stream from the *c-th* array with *k-th* weighting coefficient at time t is $\hat{b}_j(O_{\hat{c},\hat{k}}(t)) = \max_{c,k} b_j(O_{c,k}(t))$.

5. EXPERIMENTS

5.1. Experimental setup

We performed the experiments in a room measuring 3.45 m $\times$ 3 m $\times$ 2.6 m without additive noise. The room was divided into the 12 (3×4) rectangular areas shown in Figure 1, where the area size is 60 cm $\times$ 60 cm. We measured the transmission characteristics (that is, the cepstral means of utterances recorded *a priori*) from the center of each area. In our experiments, the room was set up as the seminar room with a whiteboard beside the left wall, one table and some chairs in the center of the room, one TV and some other tables, and so forth. The reverberation time of the seminar room was about 150 ms. In our past study [8], we revealed that the speaker position could be estimated with estimation errors of 20–25 cm by the 4 T-shaped microphone system shown as Fig. 2. In the present study, therefore, we assumed that the position area was accurately estimated, and we purely evaluated only our proposed speech recognition methods.

Twenty male speakers uttered 200 isolated Japanese words to a close-microphone. The average time of all utterances was about 0.6 second. For the utterances of each speaker, the first 100 words were used as test data and the rest for estimation of cepstral mean $\bar{C}_{position}$ in Equation (3). The same compensation parameters ($\bar{C}_{position}$) were used for all 20 speakers (that is, speaker-independent). All the utterances were emitted from a loudspeaker located in the center of each area and recorded for test and estimation of $\bar{C}_{position}$ to simulate the utterances spoken at various positions. The sampling frequency was 12 kHz. The frame length was 21.3 ms, and the frame shift was 8 ms with a 256 point Hamming window. Then, 116 Japanese speaker-independent syllable-based HMMs (strictly speaking, mora-unit HMMs [11]) were trained using 27992 utterances read by 175 male speakers (JNAS corpus). Each continuous-density HMM had 5 states, 4 with pdfs of output probability. Each pdf consisted of 4 Gaussians with full-covariance matrices, which correspond to about 32 Gaussions with diagonal covariance matrices. The feature space was comprised of 10 MFCCs. First- and second-order derivatives of the cepstrums plus the first- and second-order derivatives of the power component were also included.

Table 1. Baseline recognition result (%)

W/o CMN	Conv. CMN	PICMN	PDCMN
91.4	**93.6**	96.1	**96.4**

5.2. Experimental results

5.2.1. Baseline result

We conducted the speech recognition experiment of isolated words emitted by a loudspeaker in a distant environment. The recognition results of a conventional delay-and-sum beamforming (array 5) are shown in Table 1.

In Table 1, PDCMN is compared with recognition without CMN, conventional CMN, and PICMN (Position-Independent CMN). PICMN means the method by which the averaged compensation parameters over 12 areas were used. Without CMN, the recognition rate was not good according to the distance between the sound source and the microphone. Conventional CMN could not obtain enough improvement because the average duration of all utterances was too short (about 0.6 second). By compensating the transmission characteristics using the compensation parameters measured *a priori*, both PICMN and PDCMN effectively improved the performance of speech recognition from W/o CMN and conventional CMN. In a distant environment, the reflection may be very strong and may be very different depending on the given areas, so the difference of transmission characteristics in each area should be very large. In other words, obstacles caused complex reflection patterns depending on the speaker positions. The conventional CMN and PDCMN using the beamforming (array 5) were used as baseline in the following sections.

5.2.2. Results for the fixed-weight combinational CMN

The recognition results for *fixed-weight combinational CMN* were shown in Table 2. The range of weighting coefficients was from 0.3 to 0.9 with a step of 0.1. In Table 2, the optimal weighting coefficient of each speaker position was different for each other. The best average performance was obtained when weighting coefficient $\lambda = 0.5$. Since the combinational CMN compensated the channel distortion and speaker characteristics simultaneously, it achieved a better performance than the Conv. CMN and the PDCMN. The result of the combinational CMN achieved a relative error reduction

Table 3. Comparison of recognition accuracy of individual CMNs with combinational CMN (%)

	single Mic.		single array				multiple arrays			
			individual CMN		combinational CMN		individual CMN		combinational CMN	
	Conv. CMN	PD-CMN	Conv. CMN	PD-CMN	fixed-weight	variable-weight	Conv. CMN	PD-CMN	fixed-weight	variable-weight
Recognition rate	92.5	95.4	93.6	96.4	96.9	97.2	93.9	96.9	97.3	97.4
Relative computational cost	1.0	1.0	1.0	1.0	1.0	1.26	1.0	3.58	3.58	4.88

Table 2. Recognition results for fixed-weight combinational CMN (%)

	weights λ						
area	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1	97.0	97.3	97.7	97.0	97.3	97.6	**97.8**
2	97.6	98.0	97.9	98.2	**98.4**	98.2	98.0
3	97.1	97.6	97.7	97.3	**97.8**	97.6	97.1
4	96.2	**96.8**	96.7	96.4	96.0	95.9	95.6
5	97.4	97.5	98.0	97.9	**98.2**	97.4	97.5
6	97.2	**97.7**	**97.7**	97.4	**97.7**	97.6	97.4
7	96.3	96.4	**96.9**	96.8	96.7	**96.9**	96.6
8	95.5	95.9	**96.2**	95.8	95.9	95.7	**96.2**
9	96.5	97.1	**97.2**	96.8	96.6	97.0	96.9
10	94.8	95.1	**95.4**	95.1	94.5	**95.4**	95.2
11	95.2	95.7	95.8	**96.1**	95.8	95.9	95.5
12	94.8	94.5	**95.2**	95.1	95.0	**95.2**	94.8
Ave.	96.3	96.6	**96.9**	96.7	96.7	96.7	96.5

rate of 51.6% from the Conv. CMN and 13.9% from the PDCMN, respectively.

5.2.3. Results for the variable-weight combinational CMN

We also conducted the experiment for the *variable-weight combinational CMN*. The recognition results of the *variable-weight combinational CMN* were compared with those of individual CMNs in Table 3. K was set as 3, that is, λ_1, λ_2, and λ_3 were set as $0.4, 0.5$, and 0.6, respectively. The *variable-weight combinational CMN* selected the optimal weighting coefficient at each frame in an utterance, so it worked better than the *fixed-weight combinational CMN* (relative error reduction rate of 9.7%).

The multiple microphone-array processing described in Section 4 was also conducted and incorporated into the combinational CMN. The result of PDCMN using multiple arrays achieved 0.5% improvement over that using a single array. By integrating the combinational CMN with multiple arrays, more improvement was achieved. The proposed *variable-weight combinational CMN* using multiple arrays achieved a relative error reduction of 59.4% from the conventional CMN using a single array and 27.8% from PDCMN using a single array, respectively.

We also compared the computational costs among the methods in Table 3. As described in Section 3.2, the computational cost of *variable-weight combinational CMN* was less than that of multiple arrays because the calculation of the output probability for derivatives of the input cepstrums was same for different weighting coefficients in *variable-weight combinational CMN*. Using a single array, the proposed *variable-weight combinational CMN* achieved a relative error reduction of 56.3% from the conventional CMN and 22.2% from PDCMN at only 1.26 times the computational cost.

6. CONCLUSION

In this paper, we proposed a robust distant speech recognition by combining position-dependent CMN with the conventional CMN. Two combination methods were proposed. The first was to use a fixed weighting coefficient over whole test data to obtain the combinational CMN, which was called *fixed-weight combinational CMN*. The second was to use the maximum output probability among those from multiple features compensated by variable weighting coefficients at each frame, and a single decoder used this output probability, which was called *variable-weight combinational CMN*. The proposed combinational CMNs were also integrated with the *multiple microphone-array processing*. The experiment was conducted on small vocabulary distant isolated word recognition in a realistic environment. The proposed *variable-weight combinational CMN* using multiple arrays achieved a relative error reduction of 59.4% from the conventional CMN using single array and 27.8% from PDCMN using a single array, respectively.

7. REFERENCES

[1] L. Wang, N. Kitaoka and S. Nakagawa, "Robust distant speech recognition based on position dependent CMN using a novel multiple microphone processing technique," Proc. of EUROSPEECH-2005, pp. 2661-2664, 2005.

[2] B. H. Juang, F. K. Soong, "Hands-free telecommunications," Proceedings of the Workshop on Hands-Free Speech Communication (HSC-2001), pp. 5–10, 2001.

[3] M.L. Seltzer, B. Raj, R.M. Stern, "Likelihood-maximizing beamforming for robust hands-free speech recognition," IEEE Trans. Speech, and Audio Processing, vol. 12, no. 5, pp. 489–498, September 2004.

[4] S. Furui, "Cepstral analysis technique for automatic speaker verification," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 29, no. 2, pp. 254–272, 1981.

[5] J. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)," IEEE ASRU Workshop, pp. 347-352, 1997.

[6] Y. Obuchi, "Mixture weight optimization for dual-microphone MFCC combination," IEEE ASRU Workshop, pp. 325-330, 2005.

[7] L. Wang, N. Kitaoka and S. Nakagawa, "Robust distant speech recognition by combining multiple microphone-array processing with position-dependent CMN," EURASIP Journal on Applied Signal Processing, vol. 2006, Article ID 95491, pp. 1-11, 2006.

[8] L. Wang, N. Kitaoka and S. Nakagawa, "Distant speech recognition based on position dependent cepstral mean normalization," Proceedings of the 6th IASTED International Conference on Signal and Image Processing (SIP-2004), pp. 249–254, 2004.

[9] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," IEEE Trans. Information Theory, 13(2), pp. 260–269, 1967.

[10] J. Flanagan, J. Johnston, R. Zahn and G. Elko, "Computer-steered microphone arrays for sound transduction in large rooms," Journal of the Acoustical Society of America, vol. 78, pp. 1508–1518, June 1985.

[11] S. Nakagawa, K. Hanai, K. Yamamoto, and N. Minematsu, "Comparison of syllable-based HMMs and triphone-based HMMs in Japanese speech recognition," Proc. International Workshop on Automatic Speech Recognition and Understanding, pp. 393–396, 1999.