A NON-ITERATIVE MODEL-ADAPTIVE E-CMN/PMC APPROACH FOR SPEECH RECOGNITION IN CAR ENVIRONMENTS

M. Shozakai, S. Nakamura and K. Shikano
Graduate School of Information Science
Nara Institute of Science and Technology
Ikoma, Nara, 630-01 Japan
E-mail:{m-shozak | nakamura | shikano}@is.aist-nara.ac.jp

ABSTRACT

This paper investigates the Cepstrum Mean Normalization(CMN) which has been widely acknowledged useful for compensation of multiplicative distortions. However, the performance of usual CMN is limited because the normalization by a single cepstrum mean vector is not enough to compensate many factors of multiplicative distortion in real environments. To solve this problem, a new method E-CMN is proposed. The method estimates two cepstrum mean vectors, one for speech and the other for non-speech for each speaker and subtracts them from an input cepstrum. This method is capable of compensating various kinds of multiplicative distortion collectively to normalize input spectra. Furthermore, a new model-adaptive approach E-CMN/PMC, based on E-CMN and HMM composition, is proposed for environments with additive noise and multiplicative distortions. This method is simplified in a sense that it is possible to add speech models and an additive noise model without any iterative operations. Matching gains for all frequency bands of speech models to the noise model are uniquely estimated as a cepstrum mean vector for speech. The performance of E-CMN/PMC in adverse car environments is finally evaluated.

1. INTRODUCTION

Needs of speech interface for facilities like car navigation systems, and for mobile computing devices like Personal Digital Assistants, are stimulating research and development into speech recognition technologies in adverse environments. The drastic drop in performance that occurs in real environments is widely acknowledged to be due to multiplicative distortion and additive noise. For additive noise, speech-enhancement approaches such as spectrum subtraction and model-adaptive approaches such as HMM composition[7][8] have been proposed. This paper proposes E-CMN (Exact CMN) for compensation of multiplicative distortions, and proposes E-CMN/PMC for compensation of both multiplicative distortions and an additive noise. The E-CMN has two steps : an estimation step to calculate one cepstrum mean vector for speech frames for each speaker and another cepstrum mean vector for non-speech frames for each environment, and a normalization step to subtract cepstrum mean vectors from the input cepstrum vectors. The new non-iterative model-adaptive E-CMN/PMC approach is realized by combining E-CMN and PMC. We present results obtained from using E-CMN and E-CMN/PMC for speech recognition tasks in car environments.

2. MODELING MULTIPLICATIVE DISTORTION AND ADDITIVE NOISE

The long-term average of short-term spectra $S(\omega;t)$ of frequency ω at time t in the speech frame is called *speaker personality* and is defined as

$$H_{Person}(\omega) = \frac{1}{T} \cdot \sum_{t=1}^{T} S(\omega;t) \quad (1)$$

where T is a sufficiently large natural number. The *speaker personality* may be considered to represent a frequency characteristic which depends on the speaker's vocal tract and vocal cords. The normalized speech spectra is defined as

$$S^*(\omega;t) = S(\omega;t) / H_{Person}(\omega) \quad (2)$$

The short-time spectra $S(\omega;t)$ is interpreted as the generated output when the normalized speech spectra $S^*(\omega;t)$ passes through a time-invariant filter of gain $H_{Person}(\omega)$ which is a multiplicative distortion to $S^*(\omega;t)$. We may find three kinds of multiplicative distortion for $S^*(\omega;t)$ in addition to the $H_{Person}(\omega)$ in reality[1].

(1)*speaking style* $H_{Style(N)}(\omega)$: frequency characteristics peculiar to speaking styles(speed, loudness, Lombard effect etc.) which are affected by an additive noise, and

(2)*acoustical transmission characteristics* $H_{Trans}(\omega)$: spatial frequency characteristics from mouth to microphone, and

(3)*microphone characteristics* $H_{Mic}(\omega)$: frequency characteristics of microphone.

If we assume that speech and noise are additive in the linear spectrum domain, the observed spectra $O(\omega;t)$ is modeled as

$$O(\omega;t) = H^*(\omega) \cdot S^*(\omega;t) + \tilde{N}(\omega;t) \quad (3)$$

$$H^*(\omega) = H_{Mic}(\omega) \cdot H_{Trans}(\omega) \cdot H_{Style(N)}(\omega) \cdot H_{Person}(\omega) \quad (4)$$

$$\tilde{N}(\omega;t) = H_{Mic}(\omega) \cdot N(\omega;t) \quad (5)$$

where $N(\omega;t)$ is an environmental additive noise.

3. E-CMN

The CMN[2] has been widely used for compensating

Table 1 Classification of CMN.

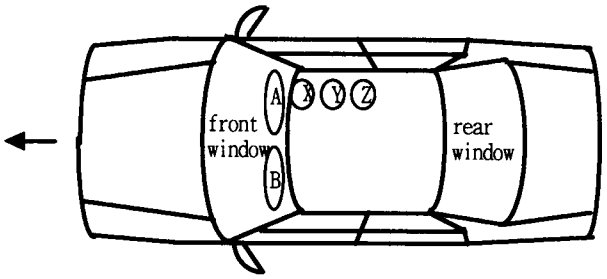
cepstrum mean for speech/non-speech	utterance-based	speaker-based
common	Type 1	Type 3
separate	Type 2	Type 4

Table 2 Combinations of microphone and seat positions.

Combination	Mic. Position A : sun-visor at driver's seat B : sun-visor at seat beside driver	Seat Position X : front Y : intermediate Z : rear	Distance (cm)
1	A	X	26
2	A	Y	34
3	A	Z	42
4	B	X	66
5	B	Y	71
6	B	Z	76

microphone characteristics. Recently, it has been suggested that calculating cepstrum mean vectors separately for speech and non-speech gives better performance than calculating one common cepstrum mean vector[3][4]. Equ.(3) justifies this conclusion, because the multiplicative distortion in speech frames and the multiplicative distortion in non-speech frames are different. At the same time, equs. (3), (4) suggest that CMN can be interpreted as a method to normalize absolute speech spectra by the product of the four kinds of multiplicative distortion. It should be noted that *speaker personality* cannot be isolated from the product of multiplicative distortions. Furthermore, the *speaker personality* estimated from short utterances may vary depending on the phoneme balance. For these reasons, a new CMN(Type 4) method: calculate two cepstrum mean vectors -- one for speech frames in sufficiently-long utterances, and the other for non-speech frames -- for each speaker separately seems to give better performance. We classify four variations of CMN in Table 1. *Utterance-based* means that the cepstrum is calculated utterance by utterance. *Speaker-based* means that the cepstrum mean is calculated from sufficient lengths of each speaker's speech.

The recognition task is speaker-independent 520 Japanese words with 54 context-independent tied-mixture HMM models which are derived from a speech database(ATR Database C set) for 40 speakers. The acoustic analysis uses 8kHz sampling, 32ms frame length and 20ms frame shift. The parameters are 10 MFCC(Mel-Frequency Cepstrum Coefficient)s, 10 Delta MFCCs and Delta energy. The number of shared Gaussian distributions are 256, 256 and 64 respectively. Six impulse responses, from dummy head(Head and Torso Simulator TYPE4128 by B&K Inc.)'s mouth to omni-directional microphone, measured in a car cabin by TSP(Time Stretched Pulse) method[5] are convoluted with the evaluation data(2 males and 2 females) as $H_{Trans}(\omega)$. Impulse responses are measured for 6 combinations of Table 2 in a car environment shown in Fig.1. Fig.2 shows those measured impulse responses



Microphone
A : sun-visor at driver's seat
B : sun-visor at seat beside driver

Driver's Seat
X : front
Y : intermediate
Z : rear

Fig. 1 Positions of microphone and driver's seat.

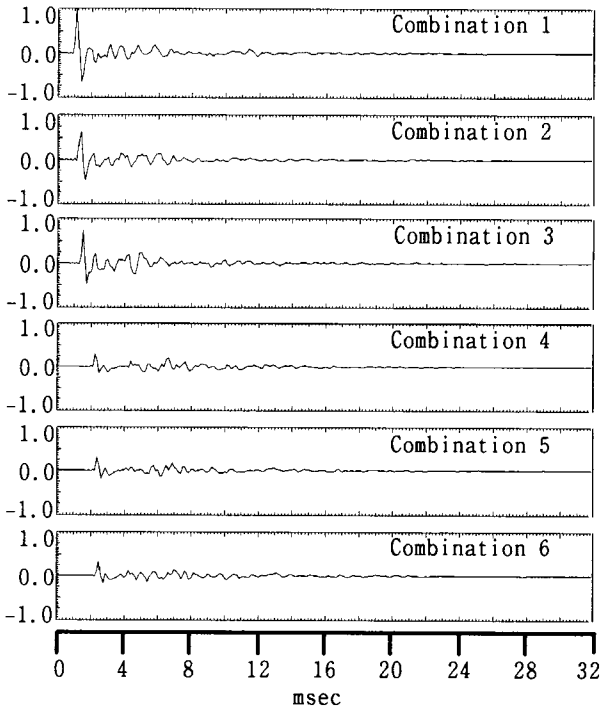


Fig. 2 Measured Impulse Responses.

which are normalized so that maximum amplitude for combination 1 is equal to 1.0. No additive noise was added to the evaluation data in this simulation. The speech and non-speech frame are detected by enhanced voice activity detection algorithm based on [6].

The recognition performances of no CMN, CMN(Type 1) and CMN(Type 2) are shown in Table 3. The recognition performances for CMN(Type 3) and CMN(Type 4) are shown in Table 4 and Table 5 respectively. For CMN(Type 3) and CMN(Type 4), the cepstrum means were averaged over 520 words, 50 words, 10 words and 5 words.

We found that:
(1)CMN(Type 4) was the most effective. The cepstrum mean as a product of various multiplicative distortions can be estimated

Table 3 Recognition performance for no CMN, CMN(Type 1) and CMN(Type 2).

Combination	no CMN	CMN Type 1	CMN Type 2
1	80.1	90.8	88.4
2	75.9	90.2	88.0
3	72.3	89.7	87.1
4	82.9	87.9	83.9
5	83.2	87.7	82.1
6	80.5	87.3	81.9
average	79.2	88.9	85.2

Table 4 Recognition Performances for CMN(Type 3).

Combination	5 w	10 w	50 w	520 w
1	92.6	92.4	93.0	93.3
2	92.3	92.2	92.8	92.9
3	91.0	91.5	91.8	91.7
4	89.8	89.5	90.5	90.6
5	89.3	89.5	89.4	89.6
6	88.2	88.3	88.3	88.5
average	90.5	90.6	91.0	91.1

Table 5 Recognition Performances for CMN(Type 4).

Combination	5 w	10 w	50 w	520 w
1	92.3	93.3	93.0	93.1
2	92.9	93.4	93.9	93.8
3	92.0	92.9	93.2	93.0
4	90.4	91.4	91.2	91.4
5	89.6	90.1	90.3	90.3
6	89.2	89.6	89.4	89.8
average	91.1	91.8	91.8	91.9

accurately with around 10 words.

(2)With no additive noise, high recognition performance was obtained regardless of variation in seat and microphone positions.

(3)CMN(Type 3) gives slightly worse performance than CMN(Type 4) in this evaluation task, where all speech data used in this experiment have equally 250ms non-speech intervals at the beginning of and at the end of speech intervals.

(4)The poor performance for CMN(Type 2) is given because the cepstrum mean calculated from short word utterance with unbalanced phoneme distribution varies a lot utterance by utterance.

We rename CMN(Type 4) as E-CMN and summarize the algorithm as follows:

(*Estimation Step*) : Two cepstrum mean vectors are calculated for each speaker. One, obtained from speech frames of sufficiently-long utterances, is speaker-dependent. The other, obtained from non-speech frames, is environment-dependent.

(*Normalization Step*) : The speaker-dependent cepstrum mean for speech is subtracted from the input cepstrum vector in speech frames. The environment-dependent cepstrum mean for non-speech is subtracted from the input cepstrum vector in non-speech frames.

4. E-CMN/PMC

Various model-adaptive approaches for multiplicative distortion and additive noise were investigated. The most typical one is an HMM composition method such as NOVO[7] and PMC[8] or their derivatives. The HMM composition method uses equ. (6) to adapt clean HMM models to adverse environments with the estimated multiplicative distortion and the estimated additive noise model.

$$O(\omega;t) = H(\omega)\cdot S(\omega;t) + \tilde{N}(\omega;t) \tag{6}$$

We assume that the additive noise model can be estimated in advance. Then, the multiplicative distortion is estimated by ML estimation using

$$\hat{H} = \arg\max_{H} \left[p(O_T|H,M_S,M_N) \right] \tag{7}$$

where O_T, M_S, M_N are the observed linear spectra, the clean speech spectra model and the additive noise spectra model respectively.

To solve equ.(7), a steepest descent method[9] and EM algorithm[4][10] were investigated. This paper proposes a non-iterative model-adaptive E-CMN/PMC method based on spectra normalization by E-CMN. By making three assumptions: (1)M_S, M_N can be modeled as one state/one Gaussian distribution HMM, (2)the multiplicative distortion is independent on variances of M_S, M_N and (3)order-dependent optimization of equ.(7) is feasible, we can estimate the multiplicative distortion by

$$H(\omega) = \frac{\bar{O}(\omega) - \bar{N}(\omega)}{\bar{S}(\omega)} \tag{8}$$

where $\bar{O}(\omega)$, $\bar{N}(\omega)$, $\bar{S}(\omega)$ are long-term averages of the observed spectra, the mean vector of single distribution in M_N, and the mean vector of single distribution in M_S respectively. Equ.(6) and (8) lead to as follows:

$$O(\omega;t) = H^*(\omega)\cdot S^*(\omega;t) + \tilde{N}(\omega;t) \tag{9}$$

$$H^*(\omega) = \bar{O}(\omega) - \bar{N}(\omega) \tag{10}$$

$$S^*(\omega;t) = S(\omega;t) / \bar{S}(\omega) \tag{11}$$

$S^*(\omega;t)$ is equal to the normalized spectra in equ.(2). The best normalization for multiplicative distortion was realized by E-CMN as stated before. Equ. (9) suggests that once the HMM models are trained from the normalized cepstrum converted from normalized spectrum by E-CMN, we can adapt the HMM models to any adverse environments using the estimated multiplicative distortion $H^*(\omega)$ and the estimated additive noise M_N. Fig.3 briefly describes the algorithm of the E-CMN/PMC method. We note here that the multiplicative distortion is obtained as the cepstrum mean vector for speech frames by E-CMN(*Estimation Step*). The advantages of this E-CMN/PMC method over other algorithms[4][9][10] are as follows:

(1)More accurate estimation of multiplicative distortion from around 10 words is possible by E-CMN(*Estimation Step*).

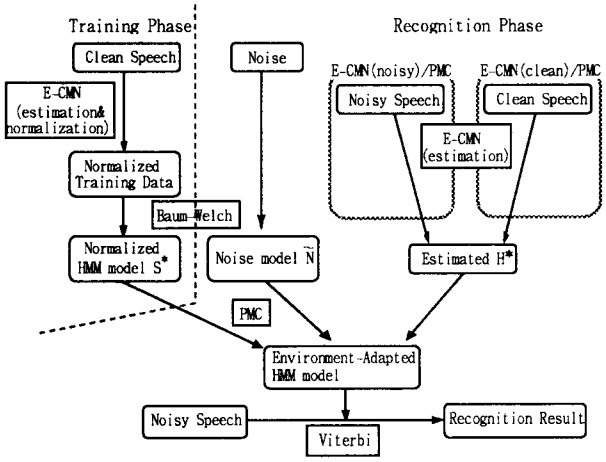


Fig. 3 E-CMN/PMC method.

(2)No iterative operations are required to adapt the HMM models, which are derived from training with normalized speech. The matching gains(multiplicative distortions) for all frequency bands of HMM models to additive noise model are uniquely estimated as speaker-dependent cepstrum mean vector by E-CMN(*Estimation Step*).

We investigate two variations of E-CMN/PMC.
(1)E-CMN(clean)/PMC : The cepstrum mean is calculated from 10 words without additive noise. Accurate estimation of multiplicative distortion is possible.
(2)E-CMN(noisy)/PMC : The cepstrum mean is calculated from 10 words with additive noise. No additive noise cancellation is done. The estimation of multiplicative distortion is contaminated by additive noise.

The recognition task is the same as the previous one. The impulse response of Combination 1 in Table 2 is convoluted with evaluation data(2 males and 2 females) as a multiplicative distortion. Noise recorded in a car cabin was added to the evaluation data with SNR 29dB, 22dB, 15dB and 8dB. The recognition performance using only E-CMN, only PMC are shown in Fig.4. The recognition performances using E-CMN/PMC are shown in Fig.5. The recognition performance for no adaptation is also shown in Fig.4 and Fig.5. These results show that
(1)E-CMN outperforms PMC at higher SNR, and
(2)E-CMN(noisy)/PMC has worse performance than E-CMN(clean)/PMC at lower SNR.

5. CONCLUSION

We have proposed an E-CMN consisting of two steps, an estimation step to calculate each speaker's cepstrum mean vectors for speech frames and non-speech frames separately, and a normalization step to subtract these vectors from the input cepstrum. Moreover, a new model-adaptive E-CMN/PMC approach is proposed and evaluated for recognition task in car environments.

6. ACKNOWLEDGMENT

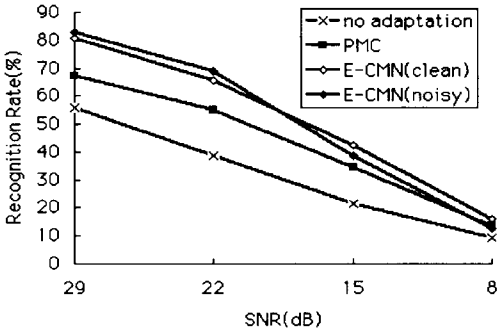


Fig. 4 Comparison of E-CMN method and PMC method.

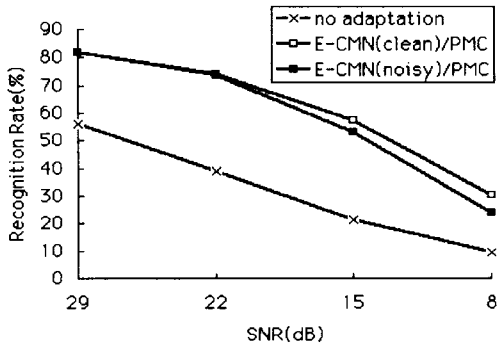


Fig. 5 Performance of E-CMN/PMC.

We would like to thank many individuals in Speech and Acoustics Laboratory of Nara Institute of Science and Technology for useful discussions and suggestions.

7. REFERENCES

[1] A. Acero, "*Acoustical and Environmental Robustness in Automatic Speech Recognition*", Kluwer Academic Publishers, 1992.
[2] S. Furui, "*Cepstral Analysis Technique for Automatic Speaker Verification*", IEEE Trans. ASSP-29, pp.254-272.
[3] X. Huang, A. Acero, A. Alleva, M. Y. Hwang, L. Jiang and M. Mahajan, "*Microsoft Windows Highly Intelligent Speech Recognizer:Wisper*", Proc. ICASSP, Detroit, 1995.
[4] A. Sanker and C. H. Lee, "*Robust Speech Recognition Based on Stochastic Matching*", Proc. ICASSP, Detroit, 1995.
[5] N. Aoshima, "*Computer-generated pulse signal applied for sound measurement*", J. Acoustic. Soc. Am. 69, 1484-1488, 1981.
[6] Recommendation GSM 06.32.
[7] F. Martin, K. Shikano and Y. Minami, "*Recognition of Noisy Speech by Composition of Hidden Markov Models*", Proc. Eurospeech, pp.1031-1034, 1993.
[8] M. Gales and S. Young, "*Cepstrum Parameter Compensation for HMM Recognition*", Speech Communication, vol.12, no.3, pp.231-239, 1993.
[9] Y. Minami and S. Furui, "*A Maximum Likelihood Procedure for a Universal Adaptation Method Based on HMM Composition*", Proc. ICASSP, pp.129-132, 1995.
[10] Y. Minami and S. Furui, "*Adaptation Method Based on HMM Composition and EM Algorithm*", Proc. ICASSP, pp.327-330, 1996.