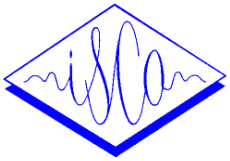

International Symposium on Chinese Spoken Language Processing (ISCSLP 2002)
Taipei, Taiwan
August 23-24, 2002

# COMPARISON OF MLLR AND CDCN FOR SPEECH RECOGNITION IN ADDITIVE NOISE BY EXPERIMENTS

*Guo-Hong Ding, Chengrong Li and Bo Xu*

National Laboratory of Pattern Recognition, Institute of Automation
Chinese Academy of Sciences, Beijing, 100080
{ghding, lich, xubo}@nlpr.ia.ac.cn

## ABSTRACT

This paper investigates the problem of robust speech recognition in additive noise. Model compensation and cepstral feature compensation techniques are evaluated and compared by experiments. The approaches considered here are MLLR and CDCN. Both approaches can be combined with CMN, which is a simple but efficient approach for robust speech recognition. Different combinations of CDCN and CMN are investigated in this paper. Noisy speech is simulated by adding different noise to clean speech with different SNR. Experiments are implemented on an isolated word recognition system. And the experimental results show that MLLR can give better performance in clean and light degraded environments, while CDCN can provide better in degraded conditions.

## 1. INTRODUCTION

In automatic speech recognition systems, noises cause mismatch between trained speech models and features of test speech and therefore often lead to severe degradation of recognition accuracy. Compensation for the mismatch can be divided into three categories with different domains: 1) in signal or spectral domain; 2) in cepstral feature domain; and 3) in model domain [1]. Over the past decades, state-of-the-art techniques for environment-robust speech recognition in the three categories are developed into: 1) speech enhancement in signal domain or spectral domain, such as speech enhancement based on HMM and spectral subtraction; 2) cepstral feature compensation in cepstral domain including CMN, CDCN [2], RATZ [7] and SPLICE; and 3) model compensation, such as MLLR [3,4] and PMC [5].

CMN is commonplace in most speech recognition systems. Because of its simplicity and efficiency, it has been fused in the standard frame of speech recognition. But for noisy speech recognition, CMN may destroy the standard environmental effect of convolutive and additive noise (2), which is the basis of many compensation algorithms, such as CDCN and PMC.

MLLR is a model-compensation algorithm. It was firstly used for speaker adaptation [3], and then it's introduced for environment adaptation [4]. Since it is a data-driven algorithm, it can be combined with CMN closely.

CDCN is a typical algorithm for robust speech recognition, which can gain good performance in convolutive and additive noises. Although it is based on the environmental effect, CDCN can be combined with CMN since the components of additive noise and filtering are estimated in the maximum likelihood criterion. In this paper, four combinations of CDCN and CMN are investigated. Experiments show the effectiveness of CMN in CDCN.

The remainder of this paper is organized as follows: The next section will give a brief review on CMN, MLLR and CDCN; in section three, several combinations of CMN and CDCN are presented; in section four, experiment evaluation is performed; and the conclusion is provided in section five.

## 2. REVIEW

In this section, CMN, MLLR and CDCN are breafly reviewed.

### 2.1 CMN

Cepstral Mean Normalization (CMN) was proposed by Atal in 1974, but it wasn't until the early 1990s that it became prevalent in most speech recognition systems [6].

For a set of $T$ cepstral vectors $\{x_0, x_1, ..., x_{T-1}\}$, its sample mean $\bar{x}$ is given by

$$\bar{x} = \frac{1}{T}\sum_{t=0}^{T-1} x_t$$

CMN consists of subtracting $\bar{x}$ from each vector $x_t$ to obtain the normalized cepstrum vector $\hat{x}_t$

$$\hat{x}_t = x_t - \bar{x} \qquad (1)$$

Since convolutive noise is represnted as the additive component in cepstral domain, CMN can reduce the effect of channel filtering greatly. Here, the effect of channel filtering includes not only the effect of convolutive environment interfering such as microphones, but also the effect of vocal tracts of different speakers.

### 2.2 MLLR

Maximum Likelihood Linear Regression (MLLR) is a data-driven approach, and is not based on prior knowledge. The technique adapts the speaker-independent acoustic model to match the noisy environment and the specific speaker using affine transformation.

Leggetter first proposed MLLR to adapt the speaker-independent model to match the specific speaker [3]. In [4], MLLR is first adopted for robust speech recognition in noisy environments. The basic idea is to adapt the model following the formulas below:

$$\hat{\mu} = A\mu + b, \quad \hat{\Sigma} = \Sigma$$

This work was supported by National Nature Science Foundation of China under grant number 69835003.

The affine transform matrix can be estimated using the EM algorithm in the ML criterion.

### 2.3 CDCN

Code-Dependent Cepstral Normalization (CDCN) [2] was proposed by Acero in 1990, which fulfills compensation in cepstral domain for noisy environments.

The effect of additive and convolutive noises in cepstral domain can be formulated as follows

$$y = x + q + r(x, n, q) \tag{2}$$

where

$$r(x, n, q) = DCT \ln(1 + e^{IDCT[n-q-x]})$$

In CDCN, a mixture gaussian model is trained for clean speech

$$p(x) = \sum_{k=0}^{K-1} c_k N(x; \mu_k, \Sigma_k)$$

Given $n$ and $q$, the noisy cepstra can be modeled using following probability density function

$$p(z) = \sum_{k=0}^{K-1} c_k N(z; \mu_k + q + r_k, \Sigma_k)$$

where $r_k = r(\mu_k, n, q)$. And $n$ and $q$ can be estimated iteratively using sufficient noisy utterances in the maximum likelihood criterion.

$$(\hat{n}, \hat{q}) = \arg\max_{n,q} p(Z \mid q, n) \tag{3}$$

Finally, given noisy cepstra, the MMSE estimate for clean speech cepstra has the form

$$\hat{x} = \sum_k p(k \mid y - q - r_k) \cdot (y - q - r_k) \tag{4}$$

In the equation, the correction vector $r_k = r(\mu_k, n, q)$ is no longer a function of $y$, so the cepstral normalization is codeword-dependent.

## 3. COMBINATION OF CMN AND CDCN

In this section, different combinations of CMN and CDCN are discussed. Before we discuss it, the effect of CMN for speech recognition in additive noise will be analyzed firstly.

### 3.1 CMN for speech recognition in additive noise

As pointed out last section, CMN is effective for the effect of convolutive noise. Here, we will analyze the capacity of CMN in additive noise.

Usually, the environmental effect can be formulated as

$$y(t) = x(t) * h(t) + n(t)$$

In spectral domain, it can be rewritten as

$$Y(\omega) = X(\omega)H(\omega) + N(\omega)$$

If no additive noise is present, environmental effect in log-spectral domain can be expressed as

$$y^c = x^c + h^c$$

So after CMN, the effect of $h^c$ can be eliminated.

If additive noise is present, the case is much complicated, just as illustrated in [5]. If we modify the effect of environmental effect in spectral domain as follows

$$Y(\omega) = (X(\omega) + N'(\omega))H(\omega)$$

where it must satisfy $N(\omega) = N'(\omega)H(\omega)$ and $N'(\omega)$ should be isolated-distributed. So in cepstral domain, we have

$$y^c = DCT \ln(X + N') + h^c \tag{5}$$

After CMN, the effect of convolutive noise can also be eliminated.

We have to admit that equation (5) is not strict in all cases, but this may be the reason that CMN is effective even in additive noise, since CMN can restrain the effect of convolutive noise.

### 3.2 Four combinations of CMN and CDCN

Figure 1 shows the four combinations of CMN and CDCN.

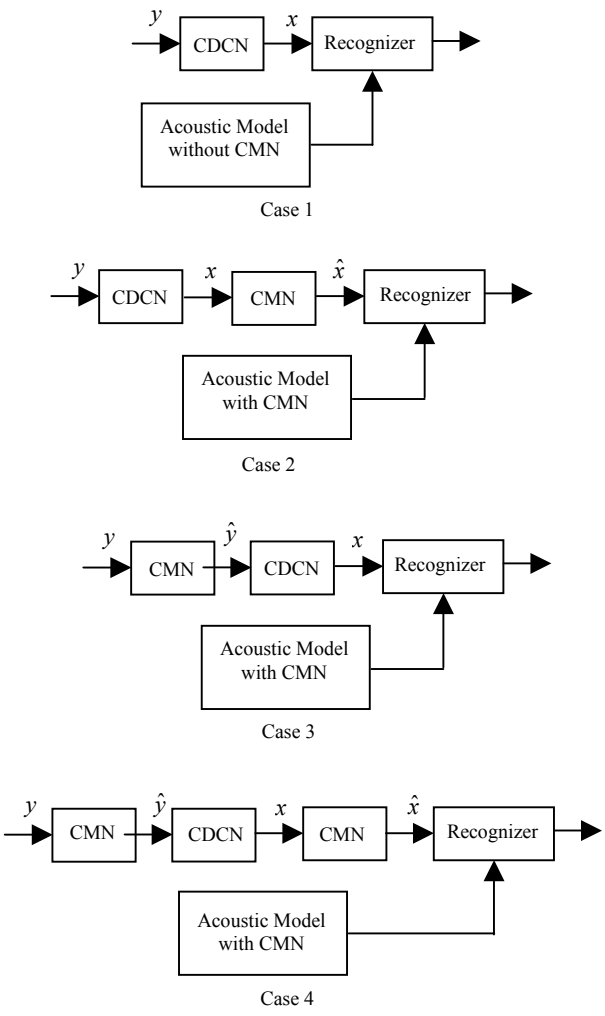


Figure 1: The four combinations of CDCN and CMN.

Originally, CDCN is implemented for cepstral features without CMN, since it is founded on the standard environmental effect described as equation (2). Case 1 and case 2 follow the rule. In case 1, no CMN is used; and in case 2, CDCN is performed with original cepstra, and then CMN is implemented.

Since CMN is effective for additive noise, CDCN can also be used to normalize the feature after CMN. A simple explanation for this is that according to equation (5), CMN can eliminate part of the convolutive noise, and the rudimental parts of filtering and additive noises can be compensated by CDCN. In case 3, CDCN is performed with cepstra that are compensated firstly by CMN; and in case 4, CDCN is performed after CMN, and then, the compensated cepstra are filtered again by CMN.

Here, acoustic models with or without CMN are used in different cases for corresponding conditions in some sense.

## 4. EXPERIMENT EVALUATION

In this section, several experiments are performed, including those with and without CMN, those with the combination of MLLR and CMN and those with different combinations of CMN and CDCN.

### 4.1 Recognition engine and database

The experiments are implemented on an isolated speech recognition system. Each acoustic signal is sampled at 8kHz. 20ms speech frames are windowed and analyzed with FFT every 10ms. 20 mel filters are applied, and after logarithm and DCT, 12-d cepstra are obtained.

We record isolated words from 33 persons (16 female, 17 male), 108 words per speaker. Utterances of 10 male and 10 female are used to train the model and utterances of the rest 13 persons are used for testing. The acoustic model is single gaussian, which is built using whole words as modelling unit.

The noise data used here are noisex-92, which are resampled and added to clean speech to generate noisy data with different SNR. In this paper, additive white noise and additive babble noise are used.

### 4.2 Experiments for the effect of CMN in noisy environments

Table 1 and table 2 give the experimental results of recognition accuracy with or without CMN in different SNR with additive white and babble noise.

| | clean | 20dB | 15dB | 10dB | 5dB |
|---|---|---|---|---|---|
| No CMN | 93.16 | 42.95 | 21.44 | 9.54 | 5.20 |
| CMN | 96.58 | 79.13 | 61.11 | 36.25 | 11.25 |

Table 1: Recognition accuracy(percentage accuracy) as a function of the test-data SNR with additive white noise. Experimental results with and without CMN are dispalyed.

| | clean | 20dB | 15dB | 10dB | 5dB |
|---|---|---|---|---|---|
| No CMN | 93.16 | 73.43 | 47.65 | 22.93 | 8.55 |
| CMN | 96.58 | 89.67 | 79.20 | 58.62 | 32.27 |

Table 2: Recognition accuracy with and without CMN as a function of the test-data SNR with additive babble noise.

The experimental results show that the recognition system with CMN, in clean, light degraded and heavy degraded conditions, can gain much better performance than that without CMN, which testifies the effectiveness of CMN for robust speech recognition. Furthermore, in clean environements, both can gain perfect performance, but with environments getting worse, both get worse.

### 4.3 Experiments for combination of CMN and MLLR

MLLR is a data-driven approach, and it can be combined with CMN closely. Table 3 and table 4 give the experimental results of combination of CMN and MLLR in different conditions with additive white and babble noise.

| | clean | 20dB | 15dB | 10dB | 5dB |
|---|---|---|---|---|---|
| CMN+ MLLR | 98.24 | 85.82 | 69.71 | 41.67 | 16.35 |

Table 3: Recognition accuracy as a function of different SNR with additive white noise for combination of MLLR and CMN.

| | clean | 20dB | 15dB | 10dB | 5dB |
|---|---|---|---|---|---|
| CMN+ MLLR | 98.24 | 93.43 | 86.62 | 67.07 | 38.78 |

Table 4: Recognition accuracy as a function of different SNR with additive babble noise for combination of MLLR and CMN.

The experimental results show that the combination of MLLR and CMN can provide better performance, and it increases the recognition accuracy by nearly 5%. It achieves perfect in clean and light degraded conditions, but with environemts getting worse, the performance gets worse.

### 4.4 Experiments for combination of CMN and CDCN

In section 3, four combinations of CMN and CDCN are illustrated. Experiments of all the four cases are implemented and the results are given in table 5 and table 6.

| | clean | 20dB | 15dB | 10dB | 5dB |
|---|---|---|---|---|---|
| Case 1 | 75.21 | 55.41 | 25.36 | 13.68 | 6.48 |
| Case 2 | 93.66 | 78.35 | 59.76 | 35.90 | 13.53 |
| Case 3 | 52.35 | 77.99 | 67.52 | 46.44 | 21.79 |
| Case 4 | 79.34 | 82.05 | 74.15 | 56.27 | 33.05 |

Table 5: Recognition accuracy as a function of different SNR with additive white noise. The four cases are described in figure 1.

| | clean | 20dB | 15dB | 10dB | 5dB |
|---|---|---|---|---|---|
| Case 1 | 75.21 | 72.79 | 57.27 | 32.76 | 8.69 |
| Case 2 | 93.66 | 88.89 | 77.99 | 58.19 | 32.05 |
| Case 3 | 52.35 | 83.62 | 84.62 | 74.36 | 45.94 |
| Case 4 | 79.34 | 91.81 | 87.68 | 76.28 | 49.57 |

Table 6: Recognition accuracy as a function of different SNR with additive babble noise.

Table 5 and table 6 show the experimental results of the four cases in different conditions with additive white and babble noise. It is showed that all cases, except case 2, provide worse performance in clean conditions, and in other conditions, the performance is increased by degree from case 1 to case 4 and case 4 can provide the best

performance. These testify the effectiveness of CMN even in the use of CDCN.

4.5 Discussion

Experimental results show that CMN is effective both in clean and in degraded conditions, and combination of MLLR and CMN can gain better performance in clean and light degraded condition, and the combinations of CDCN and CMN, especially case 4, can provide perfect performance in degraded conditions.

It's noted that cepstra are compensated in feature domain in CDCN, so the features can be differentiated to contain dynamic information, which is another advantage of CDCN over prior knowledge-based model-compensation algorithms, such as PMC. Below we will give some expeimental reults.

Since case 4 of the combinations of CMN and CDCN can give best performance in additive noise, we adjust the contour of case 4 in figure 1 and differentiate the features after compensation. So the feature used is 12-d cepstra and their 1st-order and 2nd-order difference. The experimental results are listed in table 7 and table 8.

| | clean | 20dB | 15dB | 10dB | 5dB |
|---|---|---|---|---|---|
| Diff of Case 4 | 72.58 | 89.74 | 87.18 | 76.71 | 50.28 |

Table 7: Recognition accuracy as a function of different SNR with additive white noise. The compensation processing follows case 4 of figure 1, and the feature is 36-d, which consists of the cepstra and their 1st and 2nd differences.

| | clean | 20dB | 15dB | 10dB | 5dB |
|---|---|---|---|---|---|
| Diff of Case 4 | 72.58 | 93.38 | 91.74 | 84.97 | 61.54 |

Table 8: Recognition accuracy as a function of different SNR with additive babble noise.

Table 7 and table 8 display experimental results in the case of difference after cepstra compensation by case 4 in figure 1 in additive white and babble noises. The results show that the compensation aglorithm with differentiation can provide much better performance in the degraded condition. And the results in lighted degraded conditions are desirable.

## 5. CONCLUSION

In this paper, several cases of MLLR and CDCN with the combination of CMN are discussed and experiments are implemented. Experimental results show that CMN is effective both in clean and degraded conditions, and combination of MLLR and CMN can obtain good performance in clean and light degraded conditions, and combination of CDCN and CMN, especially with case 4, can provide perfect performance in degraded conditions.

## 6. REFERENCES


[1] A. Sankar, C. -H. Lee, A Maximum-Likelihood Approach to Stochastic Matching for Robust Speech Recognition, *IEEE Trans. Speech Audio Proc*., 4(3): 190-201, 1996.

[2] A. Acero, R. M. Stern, Environmental Robustness in Automatic Speech Recognition, In *Proc. ICASSP*, 849-852, 1990.

[3] C. J. Leggetter, P. C. Woodland, Speaker Adaptation of Continuous Density HMMs Using Linear Regression. In *Proc. ICSLP- 1994*, pp: 451-454.

[4] P. C. Woodland, M. J. F. Gales and D. Pye, "Improving environmental robustness in large vocabulary speech recognition", In *Proc. ICASSP*, 1: 65-68, 1996.

[5] M. J. F. Gales, S. J. Young, PMC for Speech Recognition in Additive and Convolutive Noise, *Technical Report CUED/F-INFENG/ TR135*, 1993.

[6] X. D. Huang, A. Acero, H. Hon, and S. Meredith, *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*, Prentice Hall, 2001.

[7] P. J. Moreno, B. Raj, E. Gouvea, R. M. Stern, Multivariate-Gaussian-Based Cepstral Normalization for Robust Speech Recognition, In *Proc. ICASSP*, 1: 141 -144, 1995.