

# An anti-noise MFCC extraction algorithm for speaker recognition

Wang Hong Pan Jin'gui

State Key Laboratory for Novel Software Technology,  
Nanjing University  
Nanjing, China  
whlogs@gmail.com

Wang Hong

Institute of Computer Application and Research,  
Changji University  
Changji, China

**Abstract**—In order to acquire satisfactory performance of speaker recognition system under noisy environment, an anti-noise Mel-scale frequency cepstrum coefficients (MFCC) extraction algorithm based on the general noisy speech model is proposed. The algorithm uses spectrum mean normalization (SMN) to suppress the additive noise, and uses cepstral mean normalization (CMN) to remove the effect of convolutional noise. Theoretical analyses show that the combination of SMN and CMN can inhibit additive and convolutional noise at the same time. To verify the performance of the new algorithm, we have conducted some speaker recognition tests by using this algorithm and the conventional MFCC approach, respectively. The additive white noise experiments and the additive factory noise experiments with the same convolutional noise component show that the proposed algorithm provides 10.5% and 9.6% relative improvement than the conventional MFCC approach, respectively.

**Index Terms**—Mel-scale Frequency Cepstral Coefficients, feature extraction, speaker recognition

## I. INTRODUCTION

As one of the most important research branches of the speech signal processing field and the biometric identification technology, speaker recognition has obtained great progress during the past half-century. Now, most speaker recognition systems have high performance in low noise and low distort environment, but due to the complex nature of the speech signal and the various noise which almost exists in every practical application environment, once the system applies in noisy environment, the performance is degraded dynamically. So, reducing the effect of noise is very crucial to the practicable speaker recognition system, and is one focus of current research work too.

Feature extraction is a very important key element in speaker recognition since it is the first step of the whole recognition process and it produces the parameters on which the recognition system is based. If the feature parameters used are not well extracted, the recognition performance is naturally limited. Mel-scale frequency cepstral coefficients (MFCC) are the most widely used feature parameters currently, and quite many improved approaches to produce

better recognition performance under noisy environments have been proposed, such as [1][2]. Furthermore, dynamic cepstral features such as delta and delta-delta cepstra have been shown to play an essential role in capturing the transitional characteristics of the speech signal. So, delta MFCC [3], delta-delta MFCC, and other related features such as delta cepstral energy (DCE) [4] and delta-delta cepstral energy (DDCE) are also has been introduced into the speaker recognition systems.

Meanwhile, more and more new anti-noise characteristic parameters have been developed. Reference [5] proposed a method to incorporate sub-band amplitude information with sub-band Mel-spectrum centroid (SMSC) to improve the robustness of a speaker identification system in stationary noises. Generally, the well extracted characteristic parameters can only suppress the effect of some simple or special noise. As a matter of convenience for research, it was usually assumed that the noise is stationary or slowly changing non-stationary noise. At the premise of this hypothesis, the speed of power spectrum change of noise is slower than that of the speech signal. So, as long as we develop a kind of filter which can filter the direct current (DC) sub-components and the slowly changing sub-components, such as the RelATive SpecTral (RASTA) algorithm [6] and the cepstral mean normalization (CMN) [7], the speaker recognition performance under noise environment is able to be improved. Currently, CMN has been widely used for its comparatively simplicity and effectiveness.

Aiming to acquire satisfactory performance of the speech recognition system under noisy environment, we proposed a general anti-noise MFCC extraction algorithm. The remainder of this paper is organized as follows. Section 2 studies the general anti-noise principle of feature mean normalization. Section 3 briefly introduces the general noisy speech model. Then, according to the simplified noisy speech model, an anti-noise MFCC extraction method using the Spectrum Mean Normalization (SMN) and CMN is proposed in section 4. Section 5 covers the experiments and discussions. Finally, a brief concluding remark is given in section 6.

## II. THE GENERAL ANTI-NOISE PRINCIPLE OF MEAN NORMALIZATION

Generally, what one speaker recognition system fundamentally deals with is not so much the speech signal as the observation sequence of the speech signal. The observation sequence contains the various characteristic parameters which extracted from the short speech frame, and it can be considered as a multidimensional signal  $y(m,k)$ , where  $m$  is the frame number of the feature, and each frame is with a duration of  $k$  samples of the speech signal. Let us assume that the observation sequence  $y(m,k)$  can be decomposed into two mutually independent components, as in

$$y(m,k) = x(m,k) + c(m,k), \quad (1)$$

where  $x(m,k)$  and  $c(m,k)$  represent observation sequences which have been assumed to come from the clean speech and the noise, respectively. With the further assumption that noise is stationary and is uncorrelated with the speech,  $c(m,k)$  becomes a constant which has no relation with  $m$ . hence (1) can be expressed as

$$y(m,k) = x(m,k) + c(k). \quad (2)$$

Let us consider the average of  $y(m,k)$  with respect to the  $m$  frame, which is defined as

$$E[x(m,k)]_m \approx \frac{1}{N} \sum_{m=1}^N x(m,k), \quad (3)$$

where  $N$  is the total frame number of the given speech signal. By subtracting  $E[y(m,k)]_m$  from  $y(m,k)$ , we get

$$\begin{aligned} \hat{y}(m,k) &= y(m,k) - E[y(m,k)]_m \\ &= \{x(m,k) + c(k)\} - \{E[x(m,k)]_m + c(k)\}, \quad (4) \\ &= x(m,k) - E[x(m,k)]_m \\ &= \hat{x}(m,k) \end{aligned}$$

where  $\hat{y}(m,k)$  denotes the mean normalized feature of  $y(m,k)$ . As we can see from (4), the mean normalized feature of the noisy observation sequence is equal with that of the clean speech and become no related with the noise. Thereby, it can be concluded that any feature which satisfies (2) can be mean normalized to improve its robustness.

## III. THE GENERAL NOISY SPEECH MODEL

Considering all the potential interference factors over the whole speech signal transmission channel, Hansen & Arsian [8] proposed a general noisy speech generation and transmission model, as in

$$y(t) = \left( (x(t)|_{Lombard} + n_1(t)) * h_r(t) + n_2(t) \right) * h_c(t) + n_3(t), \quad (5)$$

Where  $x(t)$  is the clean speech signal,  $n_1(t)$ ,  $n_2(t)$ ,  $n_3(t)$  denotes the additive environmental noises which injected into the signal at the different spot of the speech signal channel., such as ambient noise, recording equipment, and transmission channels,  $h_r(t)$  and  $h_c(t)$  are transfer functions of the recording equipment and the transport channel, respectively, and the subscription *Lombard* represents the Lombard Effect.

Equation (5) is too complicated to deal with, but, the noise generally can be approximated as the equivalent additive noise and the equivalent convolutional noise. If we further take the assumption that noise is uncorrelated with the speech signal, then we can get a simplified noisy speech model as

$$y(t) = x(t) * h(t) + n(t). \quad (6)$$

## IV. ANTI-NOISE MFCC EXTRACTION ALGORITHM WITH SPECTRUM MEAN NORMALIZATION

From the very basic signal processing knowledge, we know that, when the noise is not correlated with the signal and is stationary, the simplest feature which satisfies (2) is autocorrelation or power spectrum of the signal, and the characteristics of the power spectrum are the foundation of most commonly used speaker features including the MFCC. so we firstly focus our investigation on the conventional MFCC extraction approach step by step. From then on, according to the general noisy speech model, an anti-noise MFCC extraction algorithm for speaker recognition is theoretically proposed by using the spectrum mean normalization and the cepstral mean normalization conjunctively.

Given that the speech signal has been sliced into  $M$  frames, and each frame contains  $N$  samples, we can conduct a frame based noisy speech model from (6) as

$$y(m,n) = x(m,n) * h(m,n) + w(m,n) \quad 0 \leq m \leq M-1, 0 \leq n \leq N-1, \quad (7)$$

Where  $y(m,n)$  denotes the noisy speech signal,  $x(m,n)$  denotes the clean speech signal,  $h(m,n)$  represents the convolutional noise, and  $w(m,n)$  is the additive noise. Hereon, we not only assume that the additive noise is stationary and is uncorrelated with the speech, but also assume the power spectrum of the convolutional noise is stationary or changes considerably slow. We know that the second assumption means  $h(m,n)$  is linear and shift invariant, thus, (7) could be simplified as

$$y(m,n) = x(m,n) * h(n) + w(n) \quad 0 \leq m \leq M-1, 0 \leq n \leq N-1. \quad (8)$$

By expressing (8) in the power spectrum we get

$$P_y(m,k) = P_x(m,n) |H(n)|^2 + P_w(n) \quad 0 \leq m \leq M-1, 0 \leq n \leq N-1, \quad (9)$$

where  $P$  denotes power spectrum of the signal which is marked as its subscription, and  $H(n)$  is the transfer function of  $h(n)$ . Now applying the same mean normalization to (9) as illustrated in (4) lead to the relation

$$\begin{aligned}\hat{P}_y(m,n) &= P_y(m,n) - E[P_x(m,n)]_m \\ &= P_x(m,n) |H(n)|^2 - E[P_x(m,n) |H(n)|^2]_m \quad (10) \\ &= \{P_x(m,n) - E[P_x(m,n)]_m\} |H(n)|^2 \\ &= \hat{P}_x(m,n) |H(n)|^2\end{aligned}$$

We can see the effect of the additive noise has been eliminated in (10).

From the point of view of the conventional MFCC extraction approach, our next step is to apply the Mel-scale frequency filter bank to (10). Actually, our Mel-scale frequency filter bank can be viewed as to do certain weighting process to  $\hat{P}_y(m,n)$ . So if we assume  $H(n)$  is constant or almost constant within every critical frequency band, then the output of the Mel-scale frequency filter bank is

$$A(m,d) = w(n,d) \hat{P}_x(m,n) |H(n)|^2, \quad (11)$$

where  $w(n,d)$  is the weighting coefficients which is independent of the  $m$ , and  $d$  denotes the No.  $d$  filter in the Mel-scale frequency filter bank.

Taking logarithm on both sides of (11), we obtain

$$\log[A(m,d)] = \log[w(n,d) \hat{P}_x(m,n)] + \log[|H(n)|^2] \quad (12)$$

Taking discrete cosine transform (DCT) on both sides of (12), yielding

$$\begin{aligned}c_{\hat{y}}(m,d) &= DCT\{\log[A(m,d)]\} \\ &= DCT\{\log[w(n,d) \hat{P}_x(m,n)]\} + DCT\{\log[|H(n)|^2]\}, \quad (13) \\ &= c_{\hat{x}}(m,d) + c_h(d)\end{aligned}$$

where  $c_{\hat{x}}(m,d)$  and  $c_{\hat{y}}(m,d)$  are named as spectrum mean normalized MFCC (SMN-MFCC) coefficients with respect to the imaginary clean speech and to the original noisy speech, respectively, and  $c_h(d)$  is the SMN-MFCC coefficient with respect to the convolutional noise.

Now applying the same mean normalization to (13) as illustrated in (4) and (9) again, lead to the relation

$$\begin{aligned}\hat{c}_{\hat{y}}(m,d) &= c_{\hat{y}}(m,d) - E[c_{\hat{y}}(m,d)]_m \\ &= \{c_{\hat{x}}(m,d) + c_h(d)\} - \{E[c_{\hat{x}}(m,d)]_m + c_h(d)\}, \quad (14) \\ &= c_{\hat{x}}(m,d) - E[c_{\hat{x}}(m,d)]_m \\ &= \hat{c}_{\hat{x}}(m,d)\end{aligned}$$

where  $\hat{c}_{\hat{x}}(m,d)$  and  $\hat{c}_{\hat{y}}(m,d)$  are named as cepstral mean normalized SMN-MFCC (CMN-SMN-MFCC) coefficients with respect to the imaginary clean speech and to the original noisy speech, respectively. That  $\hat{c}_{\hat{y}}(m,d)$  equals  $\hat{c}_{\hat{x}}(m,d)$  means the CMN-SMN-MFCC coefficient is noisy robust.

Consequently, we get a new MFCC extraction algorithm which using SMN to suppress the additive noise while using CMN to suppress the convolutional noise at the same time. By now, we can figure out the new extraction approach wholly by the block diagram Fig. 1.

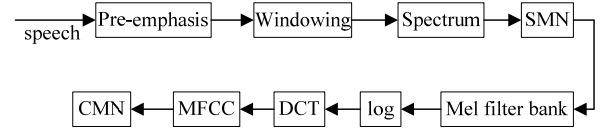


Figure 1. MFCC extraction algorithm using SMN and CMN jointly

## V. EXPERIMENTS

### A. Experimental setup

Speaker recognition experiments have been conducted to test the performance of the proposed algorithm. In the experiments, we use real clean speech recordings with sampling rate of 8 kHz, and the speech analysis frame rate is set to 256 samples with 80 samples skip rate. Of the recording speech, silence and low-energy speech parts are removed using a general energy detection technique, and the frames that have higher energy than the pre-defined threshold are selected to concatenate the experimental speech signal.

To obtain the noisy speeches for recognizing, the clean speech first feeds into a 10-tap FIR filter, which is used to simulate the convolutional noise and its amplitude response is shown in Fig. 2, then the filter output are mixed to the additive noise according to the given Signal-to-noise ratio(SNR). Here we select two kind of additive noises, one is White Gaussian Noise (WGN) and the other is Factory Noise[9].

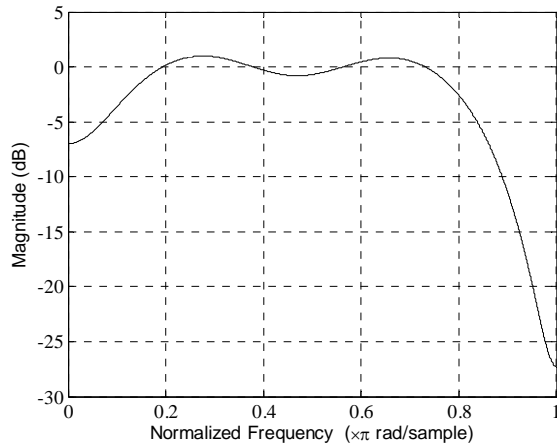


Figure 2. Amplitude response of the 10-tap FIR channel filter used to simulate the conventional noise

As for feature extraction, the clean speech and the noisy speech are all pre-emphasized with the filter  $(1-0.95z^{-1})$ , and there are 18 filters in the Mel-scale frequency filter bank. As for recognizing, Gaussian mixture model (GMM) with 32 Gaussian components is selected as speaker model. And there are 10 male speakers and each has one 10-second worth of speech for training his GMM model, and each has 10 one-second worth of speech for recognizing test.

### B. Experimental Results

For the clean speech, the WGN stained noisy speech and the Factory noise stained noisy speech with 24, 18, 12, 16, 0 SNR, respectively, the proposed CMN-SMN-MFCCs and the conventional MFCCs are both extracted to the GMM model, and the recognition rates are shown in Table 1 and Table 2.

TABLE I. RECOGNITION RATE UNDER WGN AND CONVOLUTIONAL NOISE CONDITION(%)

Speaker feature	SNR(dB)					
	clean	24	18	12	6	0
MFCC	77	75	63	45	23	13
SMN-CMN-MFCC	74	78	70	52	34	19

TABLE II. RECOGNITION RATE UNDER FACTORY NOISE AND CONVOLUTIONAL NOISE CONDITION(%)

Speaker feature	SNR(dB)					
	clean	24	18	12	6	0
MFCC	77	72	59	44	19	11
SMN-CMN-MFCC	76	76	64	50	29	14

The experimental results show that the proposed algorithm provides 10.5% and 9.6% relative improvement than the conventional approach, respectively, while its performance in clean speech is not significantly affected.

## VI. CONCLUSIONS

A major deficiency in state-of-the-art automatic speaker recognition (ASR) systems is the lack of robustness in additive and convolutional noise. Aiming to improve the performance of speaker recognition in noise conditions, in this paper, we first have investigated the general anti-noise principle of the mean normalization for the observation sequence of the speech signal. Then, according to the simplified noisy speech model, we propose a new anti-noise MFCC extraction algorithm which employs SMN to suppress the additive stationary noise while using CMN to suppress the convolutional stationary noise at the same time. Finally, the CMN-SMN-MFCC have been examined and have been compared with the common MFCC in ASR experiments, and the experimental results demonstrate the effectiveness and robustness of the algorithm in noisy conditions, while its performance in clean speech is not significantly affected.

## ACKNOWLEDGMENT

This work was supported mainly by the Department of Education of Xinjiang Uygur Autonomous Region of China, Initial Research Funds for Young College Teachers, (XJEDU2006S34), 2006 and by a Grant-in-Aid of Changji University, China.

## REFERENCES

- [1] BAI Jun-mei, ZHANG Shi-lei, ZHANG Shu-wu and XU Bo, "Robust Speaker Recognition in Noisy Environment," Journal of Chinese Information Processing. Vol. 20, pp. 91-97, February 2006.
- [2] Z. H. Chen, Y. F. Liao and Y. T. Juang, "Prosodic modeling and Eigen-Prosody Analysis for Robust Speaker Recognition," Proc. ICASSP 2005. PA. USA. vol. 1, pp. 185-188, March 2005.
- [3] L. R. Rabiner and B. H. Juang, Fundamentals of Speech Recognition. Prentice-Hall, NJ, 1993.
- [4] Nosratighods, M. Ambikairajah, E. and Epps, J., "Speaker Verification Using A Novel Set of Dynamic Features," Pattern Recognition, 2006. ICPR 2006. 18th International Conference on. Hong Kong, China, vol. 4, pp. 266-269, September 2006.
- [5] DENG Jing, ZHENG Fang, LIU Jian and WU Wenhui, "Using subband Mel-spectrum centroid and Gaussian mixture correlation for robust speaker identification," Acta Acustica. Vol. 5, pp.471-475, 2006.
- [6] H. Hermansky, N. Morgan. "RASTA processing of speech signal," IEEE Trans. On speech and Audio Processing. vol. 4, pp. 578-589, February. 1994.
- [7] F.H. Liu, A. Acero, R. Stern. "Efficient joint compensation of speech for the effects of additive noise and linear filtering," Proc. Of IEEE ICASP. pp. 257-260. January 1992.
- [8] J.H.L Hansen, L.M. Arslan. Robust feature-estimation and objective quality assessment for noisy speech recognition using credit card corpus[J]. IEEE Trans. On Speech and Audio Processing, 1995, 3(3): 169-184.
- [9] A. P. Varga, H. J. M. Steeneken, M. Tomlinson, D. Jones. "The NOISEX-92 study on the effect of additive noise on automatic speech recognition," Documentation included in the NOISEX-92 CD-ROMS, 1992