# A PITCH SYNCHRONOUS FEATURE EXTRACTION METHOD FOR SPEAKER RECOGNTIION

*Samuel Kim, Thomas Eriksson†, Hong-Goo Kang, Dae Hee Youn*

Department of Electrical and Electronic Eng., Yonsei University, Korea
† Department of Signals and Systems. Chalmers University of Technology, Sweden
E-mail:*worshipersam@mcsp.yonsei.ac.kr*

**ABSTRACT**

This paper presents a novel feature extraction method to improve the performance of speaker identification systems. The proposed feature has a form of a typical conventional feature, mel frequency ceptral coefficients (MFCC), but a flexible segmentation to reduce spectral mismatch between training and testing processes. Specifically, the length and shift size of the analysis frame are determined by a pitch synchronous method, pitch synchronous MFCC (PSMFCC). To verify the performance of the new feature, we measure the cepstral distortion between training and testing and also perform closed set speaker identification tests. With text-independent and text-dependent experiments, the proposed algorithm provides 44.3 % and 26.7 % relative improvement respectively.

## 1. INTRODUCTION

The performance of speaker recognition systems has been improved significantly due to world-wide efforts by many researchers. State of the art speaker recognition systems typically use mel frequency cepstral coefficient (MFCC) and adopt Gaussian mixture models (GMM) for speaker modeling. In this paper, we focus on the feature extraction method. A dominant feature for speaker recognition, MFCC, has been shown to achieve fairly good performance not only in speaker recognition but also in speech recognition [1][2]. Since MFCC is obtained by spectral analysis, it is crucial to estimate spectral information with high accuracy. Conventionally, the window for spectral analysis has a fixed length of 20∼30 ms and a fiexd shift of 10∼15 ms under the assumption that the signal itself is quasi-stationary [2]. In the real speech signal analysis, however, there are some possibilities of causing spectral distortion due to mismatch of the position of the analysis frame and the non-stationarity in that frame. The drawbacks of conventional MFCC have been issued in both of speaker and speech recognition field [3][4].

In this paper, we propose a new method, called pitch synchronous mel frequency cepstral coefficient (PSMFCC), to overcome those drawbacks. The rationale behind the proposed scheme is that the distance of features between training and test could be minimized by using a flexible segmentation of the analysis frame. Retaining the consideration on human auditory characteristics on frequency domain, we apply the new pitch synchronous segmentation scheme to conventional features.

## 2. SPECTRAL DISTORTION IN FIXED FRAME ANALYSIS

Fig. 1 illustrates the conventional method of speech signal analysis. This analysis shows that exactly the same signal can give different results depending on the framing, causing the critical performance degradation. The signal, $x(n)$, in the figure is vowel-like artificial speech whose pitch period is 60 samples and assuming that there is a delay of *d* samples between training and test analysis frames, $s(n)$ and $t(n)$,

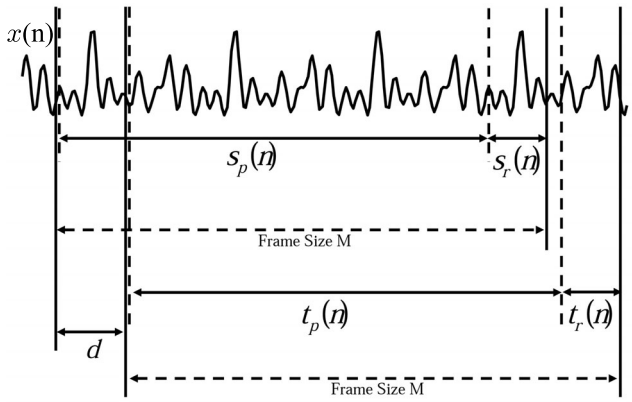


**Fig. 1**. Conventional signal segmentation method.

This work was supported by the Biometrics Engineering Research Center (KOSEF).

that is

$$s(n) = x(n)w(n)$$
$$t(n) = x(n+d)w(n) \quad (1)$$

where $w(n)$ is a window of length $M$.

The spectral distance is given by

$$D(S,T) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log|S(\omega)| - \log|T(\omega)|\, d\omega \quad (2)$$

where

$$S(\omega) = \sum_{n=0}^{M-1} s(n)^{-j2\pi\omega\frac{n}{M}}$$

$$T(\omega) = \sum_{n=0}^{M-1} t(n)\, e^{-j2\pi\omega\frac{n}{M}}$$

Now, we decompose the signal in the frame into two parts, one is the purely periodic portion of the signal, $s_p(n)$ and $t_p(n)$, and the other is the remaining part of signal, $s_r(n)$ and $t_r(n)$.

$$s(n) = \begin{cases} s_p(n), & for\ 0 \le n < NP \\ s_r(n), & for\ NP \le n < M \end{cases} \quad (3)$$

$$t(n) = \begin{cases} t_p(n), & for\ 0 \le n < NP \\ t_r(n), & for\ NP \le n < M \end{cases} \quad (4)$$

where $N$ is a number of pitch period in one frame and $P$ is the pitch period. Using DFT properties, and $s_p$ and $t_p$ can be also represented as

$$s_p(n) = \frac{1}{NP}\sum_{k=0}^{NP-1} \{X(k) * W(k)\} e^{j2\pi n\frac{k}{NP}} \quad (5)$$

$$t_p(n) = \frac{1}{NP}\sum_{k=0}^{NP-1} \{X(k) * W(k)\} e^{j2\pi n\frac{k+d}{NP}} \quad (6)$$

where $X(k)$ and $W(k)$ represent the fourier transform of $x(n)$ and $w(n)$ respectively.

From above, we can rewrite $S(\omega)$ and $T(\omega)$ as

$$S(\omega) = \sum_{n=0}^{NP-1} \frac{1}{NP}\sum_{k=0}^{NP-1} \{X(k) * W(k)\} e^{j2\pi n\frac{k}{NP}} e^{-j2\pi\omega\frac{n}{M}} + \sum_{n=NP}^{M-1} s_r(n) e^{-j2\pi\omega\frac{n}{M}} \quad (7)$$

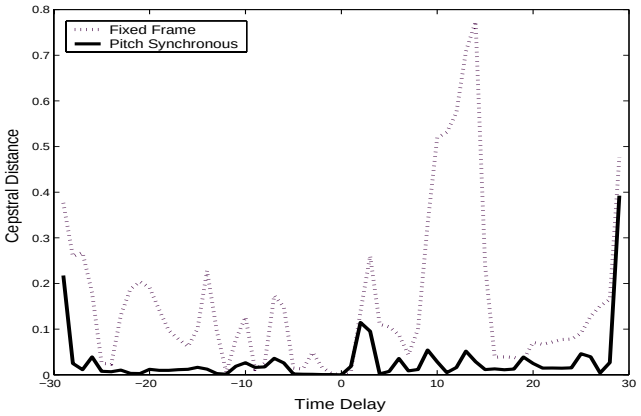


**Fig. 2**. Cepstral distance versus time delay

$$T(\omega) = \sum_{n=0}^{NP-1} \frac{1}{NP}\sum_{k=0}^{NP-1} \{X(k) * W(k)\} e^{j2\pi n\frac{k+d}{NP}} e^{-j2\pi\omega\frac{n}{M}} + \sum_{n=NP}^{M-1} t_r(n) e^{-j2\pi\omega\frac{n}{M}} \quad (8)$$

Even though *S(ω)* and *T(ω)* are derived from the same signal, they differ due to the delay *d*. Trivially, the distance *D(S,T)=0*, when *d=0*. In case of *M=NP* , the last term of *S(ω)* and *T(ω)* will be eliminated. This fact is very attractive because the mismatch of the position of the frame is inevitable to some degree. It indicates that by removing the redundant signal, which is outside of periodic components, the spectral distortion could be minimized.

## 3. A PROPOSED FEATURE, PSMFCC

In this section, we explain the details of the proposed method. Fig. 2 shows the cepstral distortion obtained by varying the number of samples of analysis frame delay in Fig. 1 with following criterion [2].

$$D_{C^{(d)}} = \sum_{i=1}^{N_C} |C_{\text{reference}}(i) - C^{(d)}_{\text{delayed}}(i)|^2 \quad (9)$$

where $C$ is a feature vector, $N_C$ is the feature order, $d$ is the number of delayed samples whose range is $-P/2 \le d \le P/2$.

The dashed line depicts the cepstral distance when we use a fixed length of analysis frame for feature extraction. The solid line depicts the distance using the proposed method, with only periodic signal obtained by removing the redundant signal. The distance of the pitch synchronous analysis is very low and stable even when it has long deviation, but

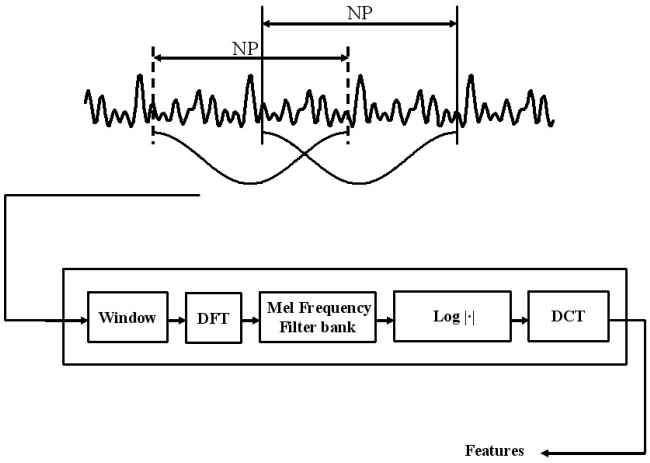


**Fig. 3**. A basic diagram for the proposed method

that of the fixed length analysis is much higher and unstable. This figure implicates the advantage of the pitch synchronous feature extraction method.

Fig.3 shows the basic idea of how we extract the proposed feature PSMFCC. When the speech signal comes into the system, the pitch contour of the signal is estimated and the signal is segmented in a pitch synchronous way. We use a very reliable pitch extraction algorithm described in [5]. Based on the pitch information, we segment the speech signal with flexibility.

For the length of the analysis frame, as we have shown in the previous section, we choose integer multiples of the pitch period for the length of analysis frame to minimize the spectral distortion. Concerning one pitch period of speech signal is an impulse response of vocal tract [3], we shift the analysis frame according to the pitch period in this paper. To eliminate possible distortion due to the discontinuity, we also find the minimum energy points through a full-search algorithm before feature extraction.

The following steps are basically identical with the conventional MFCC procedure [2]. Finally, we extract 12 dimensional PSMFCC and no delta cepstral coefficient is used.

## 4. EXPERIMENT

### 4.1. Speaker Models

We use the Gaussian mixture model (GMM) to represent the characteristics of each speaker [1][6]. For a *D*-dimensional feature vector denoted as $\underline{\mathrm{x}}$, the mixture density for speaker $s$ is defined as

$$p\left(\underline{\mathrm{x}}|\lambda_{(s)}\right) = \sum_{i=1}^{M} p_i^{(s)} b_i^{(s)}(\underline{\mathrm{x}}) \qquad (10)$$

The density is a weighted sum of *M* component Gaussian densities, $b_i^{(s)}(\underline{\mathrm{x}})$ , each parameterized by a mean vector, $\mu_i^{(s)}$ , and covariance matrix, $\Lambda_i^{(s)}$ ;

$$b_i^{(s)}(\underline{\mathrm{x}}) = \frac{\exp\{-\frac{1}{2}(\underline{\mathrm{x}}-\mu_i)^T(\Lambda_i^{(s)})^{-1}(\underline{\mathrm{x}}-\mu_i)\}}{(2\pi)^{D/2}\left|\Lambda_i^{(s)}\right|^{1/2}} \qquad (11)$$

The mixture weights, $p_i^{(s)}$, is determined to satisfy the constrain $\sum_{i=1}^{M} p_i^{(s)} = 1$ . We use the expectation maximization (EM) algorithm for the mixtures to get maximum likelihood [1][6]. Diagonal covariance matrices are used for the models. For an initialization process, we assign the speaker model parameters by choosing random samples [6].

### 4.2. The Closed-set Speaker Identification

#### *4.2.1. Text-independent Experiment*

We use the YOHO DB for the text-independent experiment. The sampling rate is 8 kHz and stored in 12-bit resolution. There are 138 speakers (32 female, 106 male); for each speaker there are 4 sessions of 24 utterances for enrollment and 10 sessions of 4 utterances for verification. More details are given in [7]. Although the YOHO DB was designed for speaker verification, we use it for the speaker identification task in this paper [1]. Speaker models for each speaker are modeled by a 64 component GMM from enrollment session 1 through 4 (average of 6 minutes). Identification test is done with 10 verification sessions consisting of four utterances for each speaker (average of 15 seconds). This setup is based on the previous work in [1]. To verify the advantage of the proposed method, we perform similar experiments with reduced enrollment time to test the algorithms under more difficult conditions. We choose the 24 utterances in the first session as train data (average of 1.5 minutes) for 16 component GMM speaker models, and same test data as in the previous experiment.

#### *4.2.2. Text-dependent Experiment*

We use the Soongsil University DB for the text-dependent experiment. Contents are fixed for every speaker. The sampling rate for the speech files is 16 kHz. There are 195 speakers (97 female, 98 male); for each speaker, there are 3 sessions of 5 utterances each, and each session was recorded in every another week. We use 5 utterances of the first two sessions (average 20 seconds) for training and the rest of 5 utterances for testing (average 2 seconds each). Each speaker model is trained using 10 components of GMM of which is decided experimentally.

## 5. SIMULATION RESULTS

Table 1 shows the error rate of text-independent experiment with full enrollment database. The performance of the MFCC feature is comparable with previous reports [1][7]. The proposed PSMFCC feature provides overall 33.3 % relative improvement.

To verify the superiority of the proposed method, we perform similar experiments but by reducing enrollment data. Table 2 shows the error rate of text-independent experiment with only one session of enrollment database. The proposed PSMFCC feature provides overall 44.3 % (61.9 % for female, 17.9 % for male) improvement. This confirms that the proposed feature, PSMFCC, is very efficient even in the small database experiments.

Table 3 shows the error rate of text-dependent experiment. The proposed PSMFCC provides overall 26.7 % (28.0 % for female, 25.0 % for male) relative improvement. Previous experiments for another database in English demonstrated that performance was worse for the female speakers [1][3][7]. Similar results are achieved also with Korean speakers.

**Table 1**. Error Rate of Text-Independent Experiments to the Full Set Training.

| Error Rate (%) | MFCC | PSMFCC | Relative Improvement |
|---|---|---|---|
| **Female** | 0.94 | 0.31 | 66.7 |
| **Male** | 0.28 | 0.28 | 0.00 |
| **Overall** | 0.44 | 0.29 | 33.3 |

**Table 2**. Error Rate of Text-Independent Experiments to the Subset Training.

| Error Rate (%) | MFCC | PSMFCC | Relative Improvement |
|---|---|---|---|
| **Female** | 13.8 | 5.02 | 61.9 |
| **Male** | 2.64 | 2.17 | 17.9 |
| **Overall** | 5.08 | 2.83 | 44.3 |

**Table 3**. Error Rate of the Text-Dependent Experiments.

| Error Rate (%) | MFCC | PSMFCC | Relative Improvement |
|---|---|---|---|
| **Female** | 5.00 | 3.60 | 28.0 |
| **Male** | 4.08 | 3.06 | 25.0 |
| **Overall** | 4.55 | 3.33 | 26.7 |

The results confirm that the proposed PSMFCC provides the improvement of speaker identification performance in both text-dependent and text-independent experiments. Note that both experiments show higher improvements for the female speakers compared to the male speakers. It implicates that a pitch synchronous segmentation method can provide more detailed and accurate features for the female speakers whose pitch periods fluctuate with a large variation.

## 6. CONCLUSION

This paper proposed a new feature extraction method called PSMFCC, which was based on pitch synchronous spectral analysis. We verified the performance of the new feature by closed set speaker identification experiments; text-dependent and text-independent. In both experiments, the performance of speaker identification was significantly improved compared to the conventional MFCC feature, especially for female speakers. The robustness of the proposed feature in adverse conditions such as noisy or channel distorted environments should be verified in the future.

## 7. REFERENCES


[1] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture models", *Speech Communication*, vol. 17, pp. 91-108, 1995.

[2] L. R. Rabiner, B.H. Juang, "*Fundamentals of speech recognition*", *Prentice Hall* , 1993.

[3] Ran D. Zilca, Jiri Navratil, and Ganesh N. Ramaswamy, "Depitch and the role of fundamental frequency in speaker recognition", *Proc. Internat. Conf. Acoust. Speech Signal Process.*, pp. 2-81, 2003.

[4] K. Aikawa, K. Ishizuka, "Noise-robust speech recognition using a new spectral estimation method, 'PHASOR'", *Proc. Internat. Conf. Acoust. Speech Signal Process.* vol. I, pp. 397-400, 2002.

[5] S. Ono and K. Ozawa, "2.4 kbps pitch prediction multi-pulse speech coding", *Proc. Internat. Conf. Acoust. Speech Signal Process.*, (New York), pp. 175-178, 1988.

[6] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture models", *IEEE Trans. Speech and Audio Processing*, vol. 3, pp. 72-83, 1995.

[7] J. P. Campbell, "Testing with the YOHO CD-ROM voice verification corpus", *Proc. Internat. Conf. Acoust. Speech Signal Process.*, pp. 341-344, May 1995.