A noise-robust pitch synchronous feature extraction algorithm for speaker recognition systems

Samuel Kim, Sungwan Yoon, Thomas Eriksson†, Hong-Goo Kang and Dae Hee Youn

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

Abstract

A noise-robust pitch synchronous feature extraction algorithm for speaker recognition systems is proposed in this paper. Since the pitch synchronous algorithms utilize pitch information, which is meaningful only for periodic segments, we propose a new scheme to deal with non-periodic ones such as unvoiced and noise-corrupted. The experimental results show that the proposed algorithm outperforms the conventional algorithm using fixed length of analysis window in actual identification tasks even in low SNR noisy environments.

1. Introduction

In state of art speaker recognition systems, mel frequency cepstral coefficient (MFCC) has been shown to achieve fairly good performance. Conventionally, MFCC features are extracted from the spectral analysis of 20∼30 ms long speech frames with an overlap of 10∼15ms [1, 2]. The length of the analysis window and the size of overlap is usually fixed for each system. The drawbacks of a fixed length analysis window have been issued by many researchers [2, 3, 4, 5]. Most efforts have focused on pitch synchronous algorithms to endow the analysis window with flexibility because pitch denotes the very fundamental period of speech signal.

Zilca *et al* introduced the *depitch* algorithm to remove pitch information from the speech signal [3]. Since it had been considered negligible for MFCC to include pitch information in it, he extracted one pitch period of signal in an analysis window to generate a feature vector. Experimental results showed that the *depitch* algorithm helped to alleviate the problem of "goat speakers" although the depitched feature vector by itself could not improve the performance. In [4], a pitch synchronous cepstrum (PSC) was proposed along with a channel compensation method called formant broadened CMS (FBCMS). Our previous work introduced the need of pitch synchronous analysis to minimize cepstral distance between training and test sets [5]. The proposed pitch synchronous MFCC (PSMFCC) which generates a feature vector from each pitch period outperformed the conventional MFCC.

In this paper, we propose a new feature extraction algorithm called pitch synchronous mel frequency cepstral coefficient (PSMFCC). Even though the pitch synchronous algorithm itself is not new, we devise the novel scheme to deal with the unvoiced region in which there are no, or meaningless pitch information. We also perform the identification tasks in noisy environments which were known to be difficult to get accurate pitch information, as well as in clean circumstance. In Section 2, we start with a detailed algorithm description, followed by experimental results and discussions in Section 3.

2. Pitch Synchronous Feature Extraction

Fig. 1 shows the basic idea of extracting a pitch synchronous waveform. To generate pitch synchronous feature vectors, we need to extract a pitch synchronous waveform first. When the speech signal enters the system, the pitch contour of the signal is estimated and the signal is segmented in a pitch synchronous way. Based on the pitch information, we segment the speech signal with flexibility, and the pitch synchronous waveforms are fed to the feature extraction algorithm to generate feature vectors. The rest of the procedures are identical to the conventional MFCC algorithm [1, 2].

Since the pitch synchronous feature extraction process is based on the pitch information of the speech signal, it is very crucial to have the accurate pitch estimation. We adopt a pitch estimation algorithm from the enhanced variable rate codec (EVRC), which is one of the well known speech coders [8]. This algorithm is based on finding the value of p that maximizes the reliability coefficient for the speech signal in the corresponding analysis window. The reliability coefficient is defined as

$$\beta = \max\left\{0, \min\left\{\frac{\sum_{i=0}^{L-p-1} r(i)\, r(i+p)}{\sqrt{\sum_{i=0}^{L-p-1} r(i)^2 \sum_{i=0}^{L-p-1} r(i+p)^2}}, 1.0\right\}\right\}, \quad (1)$$

where L is the length of the analysis window, and $r(\cdot)$ represents the residual signal which is an output signal from linear prediction analysis. When p is close to the true pitch period or an integer multiple of it, the value of β will be close to 1.0. Otherwise, in case of unvoiced speech, it tends to be lower. Thus, to find the true pitch period, we search for the p that maxi-

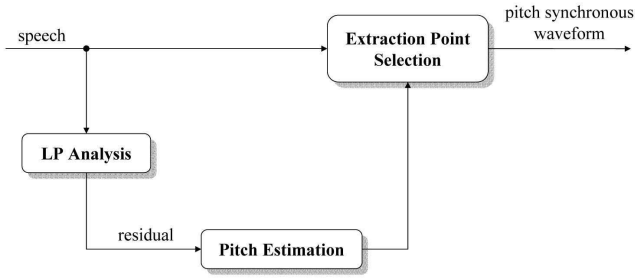


Figure 1: A basic diagram of extracting pitch synchronous waveform.

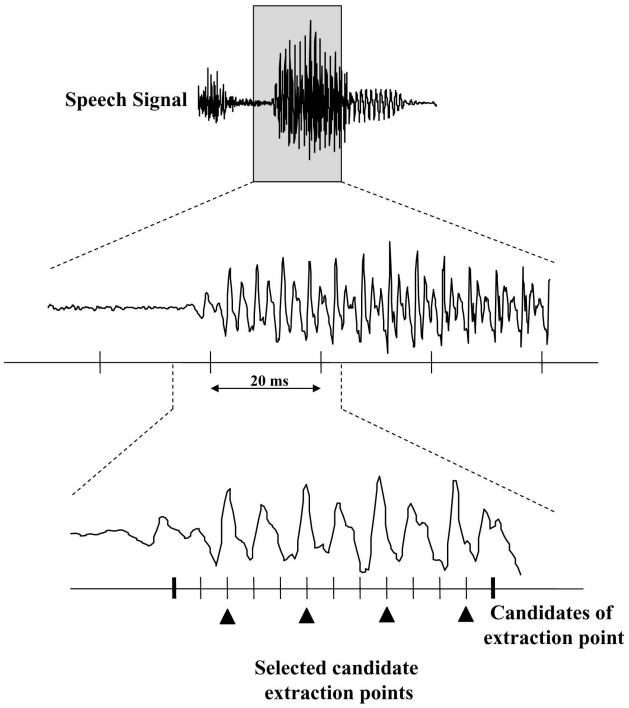


Figure 2: Candidates of extraction point and selected candidate extraction points.

mizes β. Since we need pitch information at every sample to extract a pitch synchronous waveform, we adopt a linear interpolation scheme to accomplish the extraction with a reasonable complexity.

2.1. Extraction Point Selection

Assuming that one pitch period of speech signal is an impulse response of the vocal tract, we shift the analysis frame according to the pitch period to produce feature vectors from every pitch period. An additional point that we have to consider in this algorithm is to find the extraction point of the pitch synchronous waveform. Fig. 2 illustrates the idea of selecting the extraction points. Since we have found that around 1ms interval of predefined candidates of extraction point is enough to have good quality as well as low complexity, we set 12 predefined candidates of the extraction point in a frame with an uniform interval [9]. The processor will find how many pitch periods are in the frame and which candidates are selected depending on the pitch information.

After the processor decides the extraction point, an additional procedure depicted in Fig. 3 obtains the offset value δ by searching the minimal energy point at the boundary of one pitch long speech signal whose center is at the selected candidate extraction point, i.e.

$$\delta = \arg\min_{\delta} \sum_{k=-\frac{\varepsilon}{2}}^{\frac{\varepsilon}{2}} s^2\left(n+\delta-\frac{p}{2}+k\right) + s^2\left(n+\delta+\frac{p}{2}+k\right) \tag{2}$$

where $s(\cdot)$, n, p and ε represent speech signal, time index of corresponding selected extraction point, pitch period and the

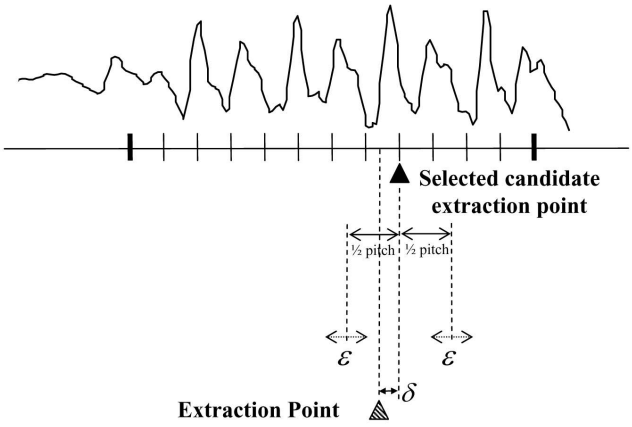


Figure 3: Adjusting the extraction point by obtaining an offset value.

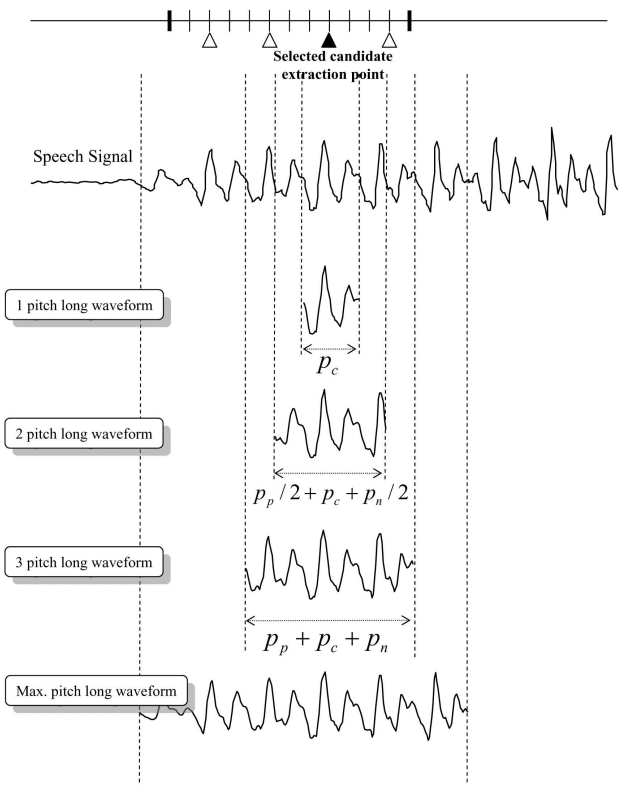


Figure 4: Pitch synchronous waveforms with various analysis window length.

window size of calculating energy at the boundary of one pitch long speech signal, respectively. With the obtained offset value δ, we can finely adjust the position of the extraction point to extract a pitch synchronous waveform. The reason of this additional procedure is simply to minimize possible discontinuities.

2.2. Analysis Window Length

To decide the length of analysis window, we present several cases in this paper; integer multiples of the pitch period, maxi-

mum pitch period and even conventional fixed length. Fig. 4 depicts the extracted pitch synchronous waveforms depending on the analysis window length, where p_c, p_p and p_n represent the pitch period at the current, previous and next candidate extraction points, respectively. *Max. pitch long waveform* represents that the pitch synchronous analysis window includes as many pitch periods as possible with the constraint of not exceeding the FFT size (256 samples in this tasks). We also use the conventional fixed analysis window length, which is hereafter denoted as *fixed*. Since we use different window length instead of one pitch period p, the offset value δ in (2) could differ in each case. Note that successive pitch synchronous waveforms can be overlapped, when the waveforms contain more than one pitch period.

2.3. Unvoiced Region

We have discussed the pitch synchronous feature extraction algorithm assuming that the speech signal is periodic, at least it has pitch periods. This is, however, obviously not true in unvoiced region. Even if we could find the pitch period in an unvoiced region that maximizes (1), it would not be meaningful. As we described earlier, the reliability coefficient β will be the measurement of how much the signal is periodic; the value of β would be small and the pitch period fluctuating in unvoiced speech. Therefore, the decision of either voiced or unvoiced will be simply made whether the reliability coefficient β is less than the threshold or not. We set the threshold to be 0.75 in this paper.

There are three possible schemes to deal with an unvoiced region; one is totally relying on the pitch period and applying the pitch synchronous scheme, another is retaining the conventional fixed analysis window length and overlap size, and the other is omitting the feature vector obtained from the unvoiced region.

The prefix PS, which means *pitch synchronous*, is added to the name of the feature vector if it is done with pitch synchronous scheme. The subscription $_{\beta}$ or $_{-\beta}$ will be attached when the feature extraction scheme depends on the value of β; $_{\beta}$ indicates that the conventional scheme is retained in unvoiced speech region, and $_{-\beta}$ indicates that feature vectors in unvoiced speech region are omitted.

3. Experimental Results

3.1. Setup

We use the YOHO database for the text independent experiments. It was designed for the speaker recognition systems that use limited vocabulary, text-independent input signals. The particular vocabulary employed in this database consists of two-digit numbers, spoken continuously in sets of three (e.g. "36-45-89"). 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 recognition. The sampling rate of the speech signals is 8 kHz and stored in 12-bit resolution. More details about the database are given in [6].

Gaussian mixture models with diagonal covariance matrices are adopted to estimate the pdfs of feature vectors; 32 mixtures for 12 dimensional feature vectors are used. Only one enrollment session is chosen for training, while we use whole recognition sessions for test. Cepstral mean removal (CMR) is also used to compensate possible channel differences. Speaker identification tasks use each of the whole test utterance with removing the silent regions at the beginning and the end of each utterance. We use MFCC with 20ms long analysis window and 50% overlap as a baseline.

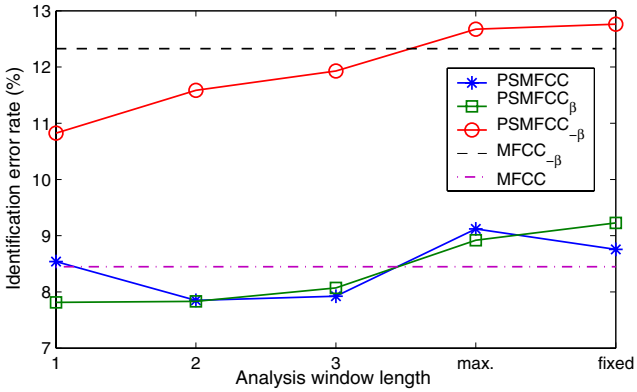


Figure 5: Identification error rate versus various schemes of feature extraction.

3.2. Identification Tasks

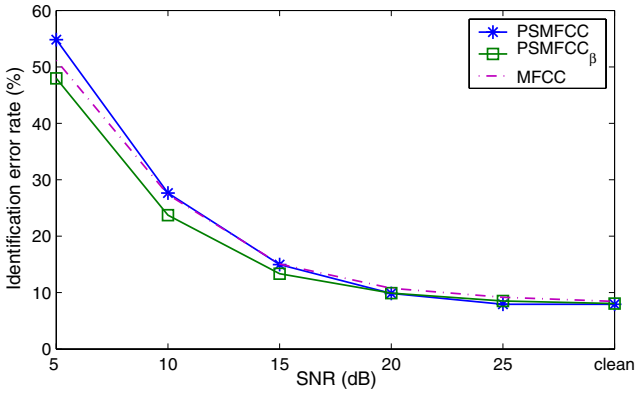
We perform identification tasks based on the general scenario that certain phrases or sentences by clients are used for identification. Fig. 5 shows the results of the experiments. Firstly, omitting feature vectors in unvoiced speech region causes severe performance degradations. It turns out that the feature vectors from unvoiced speech region cannot be simply ignored even if it is widely believed that the feature vectors from unvoiced speech region are less important than those from voiced speech region. The number of feature vectors from voiced speech region only would be insufficient to make a right decision.

Secondly, using 2 or 3 pitch period long analysis window for both PSMFCC and PSMFCC$_{\beta}$ can improve the performance, even better than the conventional MFCC. It indicates that the feature vectors from pitch synchronous algorithm are more speaker-dependent than those from the conventional one. One might ask if the performance improvement only due to the fact that the number of feature vectors are greater. Having more feature vectors, however, does not always guarantee better performance. In case of *fixed*, even though it produces approximately 2 times more feature vectors than the conventional algorithm while they use the same analysis window, the performance of pitch synchronous algorithm is rather worse than others. This indicates that the proposed pitch synchronous feature extraction algorithm is very powerful for the given limited length of speech signal.

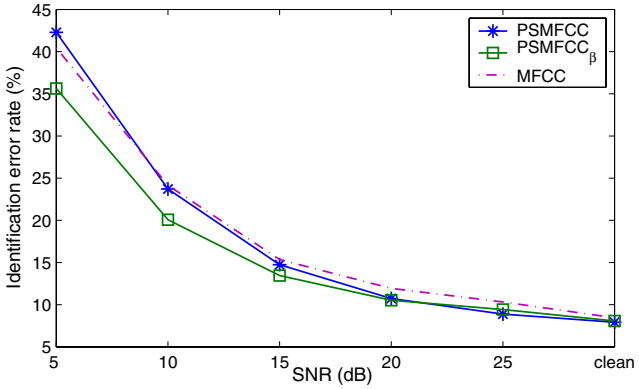
3.3. Robustness against noise

Pitch synchronous feature extraction algorithms require a pitch estimation process, which is known to be very vulnerable in noisy environments. In this paper, we perform the identification tasks with the speaker models trained in clean circumstance and various noise-corrupted test data without any preprocessing for speech enhancement. This is called a *mismatched* condition. Three types of noise such as babble, car, and street noises with 5dB, 10dB, 15dB, 20dB, and 25dB SNR is used to evaluate the performance of the systems. The noise sources are taken from the NOISEX-92 database.

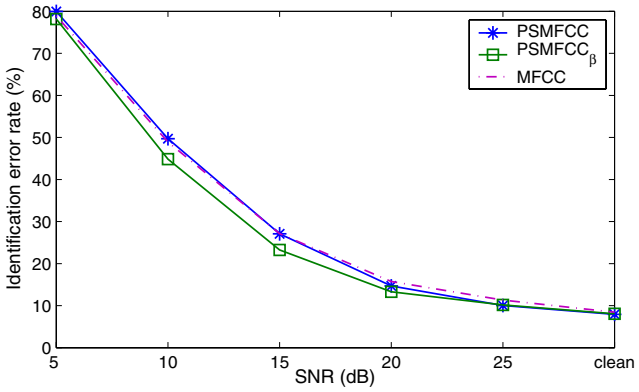
Fig. 6 shows the identification error rate for various types of noise and SNR. For convenience, we only include the results



(a) Babble noise



(b) Car noise



(c) Street noise

Figure 6: Identification error rate versus SNR for various noise-corrupted test data.

of 3 pitch period long analysis window cases of PSMFCC and PSMFCC$_\beta$ with the conventional MFCC. The performance of PSMFCC in low SNRs is worse than that of the conventional MFCC. This shows that it is very crucial to have accurate pitch information. On the other hand, PSMFCC$_\beta$ constantly outperforms the others even in very low SNR environments. This confirms that applying the pitch synchronous scheme only to voiced speech region and retaining the conventional scheme elsewhere can improve the performance. This is reasonable because that the badly corrupted speech region where the pitch information could be incorrect will be considered as a normal unvoiced region, which is still useful for applying the method of conventional fixed analysis.

4. Conclusions

This paper proposed a new feature extraction method called PSMFCC, which was based on pitch synchronous flexible analysis window scheme. The practical issues such as the length of the analysis window and schemes to deal with unvoiced speech region were discussed. Text-independent speaker identification tasks were done for GMM-based speaker recognition systems.

The results of speaker identification tasks showed that the proposed algorithm outperformed the conventional one with the appropriately chosen analysis window length. The robustness against noise was also verified. In noisy environment where the pitch estimation easily fails, a pitch synchronous algorithm with retaining the conventional scheme for unvoiced region, i.e. PSMFCC$_\beta$ improved the performance. We conclude that the proposed pitch synchronous algorithm is very powerful in both clean and noisy environments.

5. Acknowledgements

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

6. References

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

[2] T. F. Quatieri, *Discrete time speech signal processing*, Prentice Hall, 2002.

[3] R. D. Zilca, J. Navratil, and G. N. Ramaswamy, "The role of fundamental frequency in speaker recognition," *Proc. Internat. Conf. Acoust. Speech Signal Process.*, Vol. II, pp. 81–84, 2003.

[4] Y. J. Kim, and J. H. Chung, "Pitch synchronous cepstrum for robust speaker recognition over telephone channels," *Electroincs letters*, Vol. 40, No. 3, pp. 207–209, 2004.

[5] S. Kim, T. Eriksson, H.-G. Kang, and D. H. Youn, "A pitch synchronous feature extraction method for speaker recognition," *Proc. Internat. Conf. Acoust. Speech Signal Process.*, Vol. I, p.p. 405–408, 2004.

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

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

[8] Telecommunication Industry Association, TIA/EIA/PN-3292, *EIA/TIA Interim Standard, Enhanced Variable Rate Codec (EVRC)*, Mar. 1996.

[9] W. B. Kleijn and J. Haagen, "Waveform interpolation for coding and synthesis," in *Speech Coding and Synthesis*, pp. 175–208, Elsevier, 1995.