# Accuracy Improvement of MFCC Based Speech Recognition by Preventing DFT Leakage Using Pitch Segmentation

Sopon Wiriyarattanakul, Nawapak Eua-anant
*Department of Computer Engineering, Faculty of Engineering,*
*Khon Kaen University, Khon Kaen 40002, Thailand.*
*sopon_w@kkumail.com*

***Abstract*—Most MFCC based speech recognition algorithms employ frame segmentation to divide a signal into fixed-size frames as the first step prior to MFCC feature extraction. Commonly used fixed frame sizes, around 20-40 ms, do not usually fit into complete periods of speech signals. Consequently, in MFCC feature extraction, spectral leakage arises after Discrete Fourier Transform is applied to these fixed-size intervals resulting in smeared spectra and reduced speech recognition performance. In this paper, a pitch-based speech signal segmentation to reduce spectral leakage is proposed by utilizing a new technique of pitch detection based on Short-time Energy Waveform (SEW) to yield segmented speech intervals with complete periods. The proposed method utilizes local minima of SEW as markers for pitch segmentation. After segmenting speech signals into pitches, MFCC feature vectors are extracted and subsequently used as raw data for speech recognition using artificial neural networks. Speech recognition experiments using artificial neural networks, applied to collect Thai language speech signals from 40 speakers, were conducted. Empirical results indicate that speech recognition using speech signals segmented into pitches yields more accurate recognition results than those using speech signals segmented into a fixed frame.**

***Index Terms*—Short-time Energy Waveform (SEW); Pitch Segmentation; Spectral Leakage; Mel-Frequency Cepstral Coefficients (MFCC).**

## I. INTRODUCTION

Today, as the number of computers employed to perform human assistant tasks has been vastly increased, speech recognition has become more and more important because speech is a fundamental and convenient method for human-computer intercommunication. Besides improving a computer's hardware performance, a speech recognition accuracy can simply be increased by improving voice feature extraction [1-2] mapping voice signals into feature vectors with much smaller sizes while maintaining the important information essential for task recognition. Harmonic component analysis using Mel-Frequency Cepstral Coefficients (MFCC) is a well-known and widely used method of voice feature extraction due to its capability for representing voice data that, in nature, has high uncertainty.

In general, the first step of traditional MFCC based speech recognition starts with segmenting a voice signal into overlapped frames of fixed and equal size, normally around 20 - 40 milliseconds. In this method, it is assumed that, using an appropriate frame size, there are very few changes to statistic values in consecutive frames over a period of time. Therefore, the information extracted from segmented frames is suitable for speech recognition application. Next, the segmented frames are multiplied by smoothing windows [3] such as Hanning, Hamming Triangular, and Rectangular windows to produce convoluted spectra in the frequency domain. After that, subsequent speech segments are collected into the MFCC extraction process to obtain MFCC vectors for later use by speech recognition tools such as artificial neural networks (ANN) [4]-[6] or hidden Markov models (HMM) [7], [8]. To improve speech recognition performance, several techniques emphasize on the modification of learning models, such as increasing the number of nodes and layers, for instance, deep recurrent neural networks [9], and improving training algorithms, for instance, ANN with competitive learning algorithm [10] and a neural-fuzzy network trained by using an enhanced genetic algorithm [11]. In most papers, MFCC data extracted using a traditional method is assumed to be sufficient for speech representation and hence improvement of the MFCC extraction method is usually neglected. Nevertheless, in general, when a frame size does not match the period of a periodic signal, spectrum leakage, as well as undesired phase shift in consecutive frames, occurs [12]. A difficulty of speech signal processing arises from the nature that a speech signal is not an exact periodic signal, but a quasiperiodic signal which means that pitch lengths may vary slightly along a speech signal length. Therefore, fixed-frame segmentation used in a traditional MFCC extraction method, applied to a speech signal which is a quasiperiodic signal, can cause spectrum leakage which, subsequently, will reduce overall speech recognition performance. Popular methods for reducing spectral leakage employ multiplication between frames and a smoothing window to reduce the effects of the end points of a signal segment [13, 14]. However, all windowing methods cause unavoidable spectrum blurring as a result of window multiplication in the time domain corresponding to convolution in the frequency domain. To reduce frequency leakage and hence improve MFCC based speech recognition performance, this paper proposes a new pitch-based speech signal segmentation strategy that can effectively reduce the spectral leakage of speech signal segments. First, frequency leakage is explained in Section 2. Next, pitch detection and pitch-based speech segmentation methods based on short-time energy waveform (SEW), are introduced in Section 3. Frequency analysis of speech segmentation results are obtained using a new pitch-based segmentation algorithm and those obtained using traditional fixed frame segmentation are

given in Section 4. In Section 5, experimental results are compared between the speech recognition performances of ANNs using MFCC vectors extracted using pitch-based segments and fixed frame segments as inputs. Finally, Section 6 provides concluding remarks.

## II. SPECTRAL LEAKAGE

Let $x(n)$ be a discrete-time sampled signal of length $N$ points obtained using a sampling frequency $f_s$, and $X(k)$ be the discrete Fourier transform (DFT) $x(n)$. Then, the corresponding frequency of the $k^{th}$ point of $X(k)$ is equal to $kf_s/N$, and a frequency resolution of $X(k)$, i.e., the difference between corresponding frequencies of consecutive points of $X(k)$, is given by $\Delta f = f_s/N$. This means that for $x(n)$ obtained by sampling a continuous-time sinusoidal signal $x(t)$ with frequency, $f_0 = kf_0/N$ where $k, N$ are integers, $k < N/2$, $X(k)$ will contain only a single frequency component at the kth point, i.e., all energy of $x(n)$ will be confined in the kth point of $X(k)$ corresponding to frequency $f_0$, and thus there is no frequency leakage. In other words, when the condition $kT_0 = NT_s$ is met, i.e., the total sampling time, $NT_s$, is equal to a $k$-multiple of a period, $T_0 = 1/f_0$ ,there is no frequency leakage in $X(k)$ as shown in Figure 1, where $f_0$ =10 Hz, $T_0$ =0.1 Sec, Hz , $f_s$ =1,000 samples/Sec, $T_s$ =0.001 Sec, $N$ =300 points and $k$ =3 periods.

In the case where the condition $kT_0 = NT_s$ cannot be met, frequency leakage will occur as a result of no corresponding frequency component $kf_s/N$ of $X(k)$ exactly matching $f_0$ causing the energy at $f_0$ to spread to nearby frequency components of $X(k)$ as illustrated in Figure 2, where $f_0$ =10 Hz, $T_0$ =0.1 Sec, Hz , $f_s$ =1,000 samples/Sec, $T_s$ =0.001 Sec, and $N$ =350 points. In order to explain the cause of frequency leakage, a signal $x(n)$ of length $N$ points can be obtained by multiplying an infinite-length sampled version of $x(t)$ with a rectangular window of length $N$ points. Since, in the discrete-time Fourier transform (DTFT), window multiplication in the time domain is equivalent to convolution in the frequency domain between the DTFT of a window and that of a signal. As a result, DTFT of $x(n)$ is obtained by convolution between the DTFT of a rectangular window [3] and that of an infinite-length sampled version of $x(t)$ resulting in an undesired, broadened spectrum and ripples. Later, $X(k)$, the DFT of $x(n)$, is achieved by sampling the DTFT of $x(n)$ at corresponding frequencies $kf_s/N, k = 0,\ldots,N-1$ when $kT_0 = NT_s$, the kth point of $X(k)$ with the corresponding frequency $kf_s/N$ will take a sample at the center of the main lobe of the DTFT of $x(n)$, while other points of $X(k)$ will take samples at zero crossings of the DTFT of $x(n)$ resulting in a DFT with no frequency leakage. On the other hand, when $kT_0 \neq NT_s$, points of $X(k)$ that are located neither at the center of a main lobe, nor zero crossings but on lobes of the DTFT of $x(n)$ results in frequency leakage in DFT.

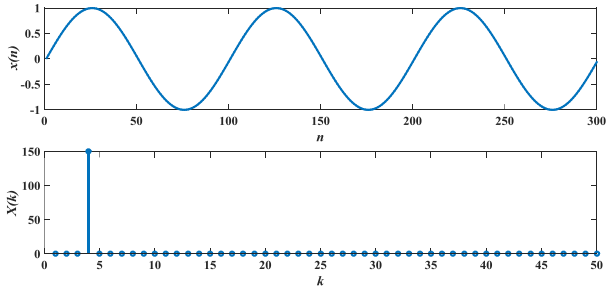


Figure 1: $x(n)$ obtained by sampling $x(t) = \sin(20\pi t)$ at $f_s$ =1,000 samples/Sec for 300 points and the corresponding DFT Magnitude $|X(k)|$ without frequency leakage

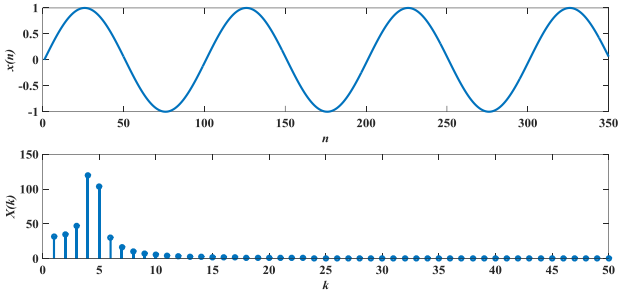


Figure 2: $x(n)$ obtained by sampling $x(t) = \sin(20\pi t)$ at $f_s$ =1,000samples/Sec for 350 points and the corresponding DFT Magnitude $|X(k)|$ with frequency leakage

In summary, in order to prevent frequency leakage in DFT, generally, for a periodic signal containing harmonics with a fundamental frequency $f_0$ and fundamental period $T_0 = 1/f_0$ the condition $kT_0 = NT_s$ must be retained, i.e., the total sampling time, $NT_s$, must be equal to a $k$-multiple of $T_0$. Therefore, with a traditional fixed frame speech segmentation method, widely used as the first step of MFCC extraction, there is a high possibility that a frame size does not match the period of the speech signal causing spectral leakage in DFT. Windowing techniques [3, 14] extensively used to reduce frequency leakage do not actually solve the frequency leakage problem, instead mitigate the problem by smearing the spectrum. In addition, although, the problem of frequency leakage in harmonic analysis has been thoroughly investigated [3], since the introduction of MFCC for speech processing [15-17], the effect of frequency leakage on speech recognition performance has never been addressed. In contrast, since a speech signal is a quasiperiodic signal, segmentation by complete periods or pitches of the signal is more suitable to reduce frequency leakage. In this paper, a new approach which effectively reduces the problem of spectral leakage by utilizing pitch detection in speech signal segmentation process is developed.

## III. PITCH BASED SPEECH SEGMENTATION

In speech processing, pitch refers to quasi-repetitive patterns occurring in speech waveforms as illustrated in Figure 3. Pitch detection is used for locating pitches or estimating the fundamental frequency of a periodic signal. Traditionally, techniques based on detecting the local maxima of an autocorrelation function (ACF) [18], widely used for pitch detection, are prone to noise. In addition, real

time speech processing involving pitch detection, to be accomplished in a short time, cannot be done using computationally exhaustive ACF. YIN [19] algorithm is an improved ACF-based pitch detection method for addressing the problem of false peaks due to sub harmonic components. Nevertheless, in the YIN method, there is still a limitation of pitch detection in the transient stages of a speech signal at the beginning and end of voices [20]. In this paper, a new technique for pitch detection based on Short-time Energy Waveform (SEW), defined in [21], which is able to handle the problem of sub harmonic peaks, as well as detecting pitches in transient intervals of a speech signal, is introduced.

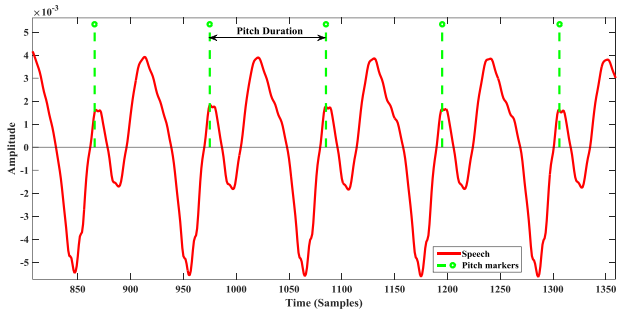


Figure 3: A speech signal, pitches, and pitch duration

A short-time energy waveform (SEW) is defined as the energy of windowed signal segments as a function of time:

$$E(n) = \sqrt{\sum_{i=0}^{W} x^2(n-i)} \qquad (1)$$

where W is the window size.

In Equation 1, from the view point of frequency domain analysis, summation acts as low pass filtering while power-of-2 operation results in frequency subtraction between constituent frequency components of $x(n)$. As a result, in SEW, a fundamental frequency component can be emphasized by the power-of-2 operation, and high frequency components are suppressed by the summation operation. The advantage of using SEW for detecting fundamental frequency is that it is computationally simple and can work with most periodic signals. The cases where SEW can probably not be used for detecting fundamental frequency components are trivial cases where $x(n)$ is a single frequency sinusoid or cases where $x(n)$ contains no consecutive harmonic components.

Figure 4 demonstrates a speech signal sampled using a sampling frequency of 22500 samples/sec and its corresponding power spectrum, SEW obtained using Equation 1 with W=300 points and the power spectrum of SEW. Obviously, a peak corresponding to a fundamental frequency of the signal can clearly be seen in Figure 4d. To calculate SEW, a window size, W, must be chosen appropriately since, in Equation 1, summation acts similarly to a moving average filter with its bandwidth inversely proportional to the window size. Therefore, if W is too small, complex peaks due to high frequency components will occur in the SEW while if W is too large, the SEW will be over smoothed as a result of a low pass filter with a narrow bandwidth. Therefore, to choose an appropriate value of W, W is initialized using a large number, in this paper, 300 points. Then, the fundamental frequency $f_0$ of a speech signal is estimated by determining a frequency associated with the largest peak, excluding the peak corresponding to a DC component, of a power spectrum of SEW. Later, W is chosen to be a proportion of a fundamental period $T_0$, i.e., $W = cT_0/T_s$ where $0 < c \leq 1$, in this paper, $W = 0.5T_0/T_s$. Later, in order to obtain an appropriate SEW, the SEW is recomputed using a new value of W.

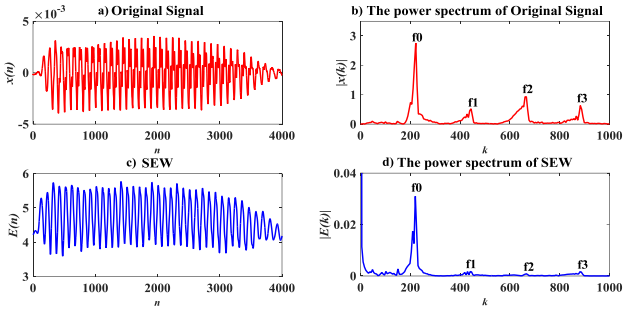


Figure 4: A speech signal, SEW and corresponding power spectra

For pitch analysis, as seen in Figures 4a and 4c, local extrema of SEW inherently synchronize with pitches. Thus, detection of pitches can be done by locating pitch marker features, such as local minima or local maxima of SEW. In this paper, to detect pitch markers, a simple local minimum searching algorithm, using a search window of size M points performed on SEW is used. However, determining the local minima of SEW can be cumbersome or it can even skip some local minima if M is too large. On the other hand, if M is too small, spurious local minima may be detected. In this paper, M is chosen to be equal to W, the window size for computing SEW. Figure 5 illustrates an enlarged graph of the speech signal in Figure 4a plotted along with its corresponding SEW and pitch detection results.

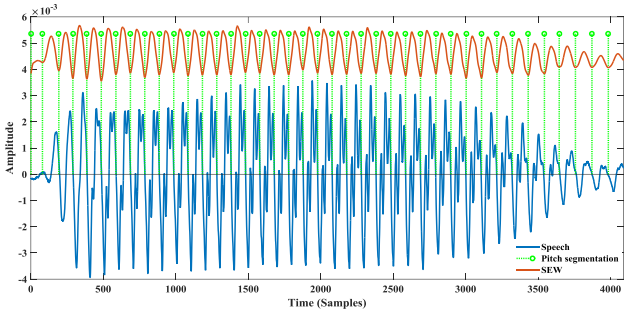


Figure 5: An original speech signal, SEW and detected pitch markers based on SEW

## IV. Frequency Analysis of Speech Segments

After all pitches are detected, prior to DFT computation and MFCC extraction, speech segmentation can be done using locations of detected pitch markers. To obtain high resolution DFTs, the length of speech segments must be large enough so that each segment may cover several pitches. Figure 6(a) illustrates 36 speech segments of a length of approximately 5 pitch duration of the speech signal in Figure 5, obtained using the proposed pitch based segmentation in comparison with those in Figure 6(b), obtained using traditional fixed frame segmentation with a frame size of 22.3ms (490 samples). To prevent abrupt changes of speech segment data, all segments are multiplied with the Hamming window, and segments obtained using the proposed algorithm are set to overlap adjacent segments by about 4 pitches, while those obtained

using a fixed frame segmentation are set to overlap adjacent segments by about 17.6ms (384 points). As can be seen, the speech segments in Figure 6(a) are less chaotic than those in Figure 6b.

To demonstrate the effect of frequency leakage, Figures 7(a) and (b) show the power spectra of speech segments obtained using the proposed pitch based method and a fixed frame method, respectively. In Figure 7(a), since the lengths of the speech segments obtained using the pitch based method are slightly varied, a frequency variable of each power spectrum graph is scaled to be in the same frequency range. As seen, in both Figures 7(a) and (b), the peaks of power spectra are slightly shifted, but the peaks of the power spectra of the pitch based segments in Figure 7(a) are narrower than those of the fixed frame segments in Figure 7(b). Harmonic analysis results in terms of the means and variances of detected fundamental and harmonic frequencies are shown in Table 1. It is clear that the fundamental and harmonic frequencies obtained from the power spectra of pitch based segments have less variation than those obtained from the power spectra of fixed frame segments, which means that spectral leakage can be reduced by using pitch based segmentation.

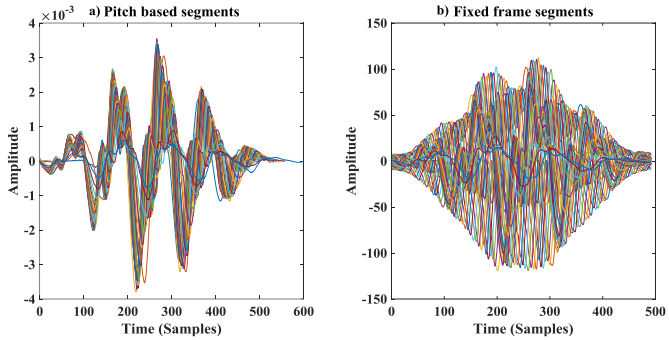


Figure 6: Speech segments, multiplied by the Hamming window, obtained using a) pitch-based segmentation with 5 pitch duration and b) fixed-frame segmentation with a frame size of 22.3ms.

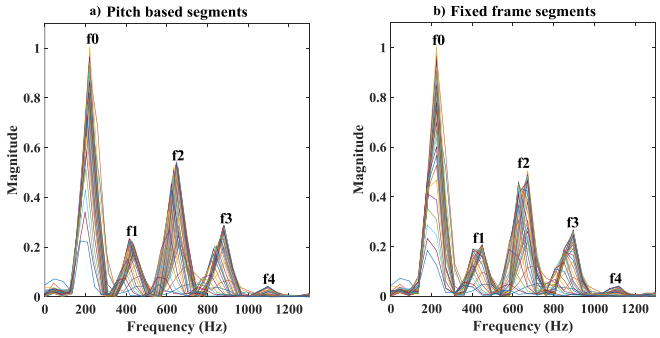


Figure 7: Power spectra of speech segments in Figure 6: a) Pitch based segments and b) fixed frame segments.

Table 1
Means and variances of $f_0$, $f_1$, $f_2$ and $f_3$ estimated from power spectra of pitch based segments and fixed frame segments.

| | | $f_0$ | $f_1$ | $f_2$ | $f_3$ |
|---|---|---|---|---|---|
| Mean | Pitch frame | 213.167 | 429.752 | 641.823 | 859.986 |
| | Fix frame | 217.860 | 431.986 | 651.092 | 870.198 |
| Variance | Pitch frame | 52.174 | 431.785 | 1022.564 | 4333.263 |
| | Fix frame | 247.085 | 706.187 | 1547.874 | 5552.261 |

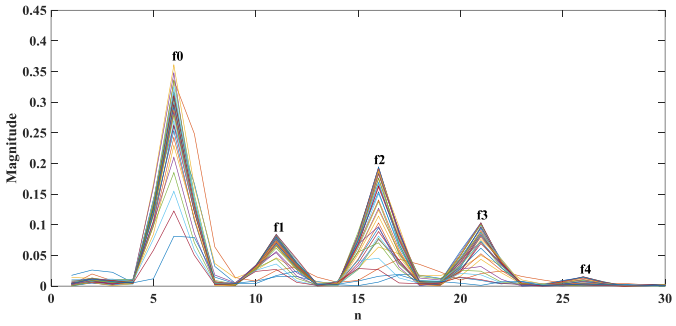


Figure 8: The power spectra of the pitch based speech segments plotted according to DFT index.

Table 2
Means and variances of DFT indices of power spectrum peaks corresponding to $f_0$, $f_1$, $f_2$ and $f_3$ of pitch based speech segments

| | $f_0$ | $f_1$ | $f_2$ | $f_3$ |
|---|---|---|---|---|
| Mean | 6 | 11.083 | 16.055 | 21.166 |
| Variance | 0 | 0.078 | 0.111 | 1.228 |

Furthermore, when the power spectra of pitch based speech segments are plotted according to a DFT index, k, instead of an actual frequency variable, f, as shown in Figure 8, the peaks of all power spectra are, interestingly, located almost at the same indices. Table 2 shows harmonic analysis results in terms of means and variances of indices of power spectrum peaks corresponding to fundamental and harmonic frequencies of pitch based speech segments. As seen, variances of indices of fundamental and harmonic frequencies are remarkably very low and the mean of the index of each frequency is separated from those of adjacent frequencies by almost the same value. Consequently, since power spectra plotted with respect to the DFT index provide the best harmonic analysis results, in this paper, superior quality MFCC are extracted from the original DFT of pitch based speech segments without frequency normalization.

MFCC [22-24], the features most commonly used for speech recognition, are discrete cosine transform coefficients of cepstrum of short-time energies of a signal in multiple frequency bands defined by a logarithmically spaced mel-frequency scale filter bank. MFCC are commonly computed by first segmenting a signal into short, overlapped, fixed-size frames of approximately 20-40 ms in length. Next, the power spectrum of each frame is computed using the fast Fourier transform (FFT) and then mapped into overlapped log-scale frequency slots defined by a mel-frequency scale filter bank. Finally, a set of log outputs from the filters is converted into discrete cosine transform coefficients, namely MFCC. In traditional MFCC computation, fixed-frame speech segmentation is performed before MFCC is computed. In this paper, to reduce frequency leakage during MFCC extraction, fixed-frame speech segmentation is replaced by the proposed pitch-based speech segmentation before calculating MFCC. The overall MFCC extraction process is shown in Figure 9.

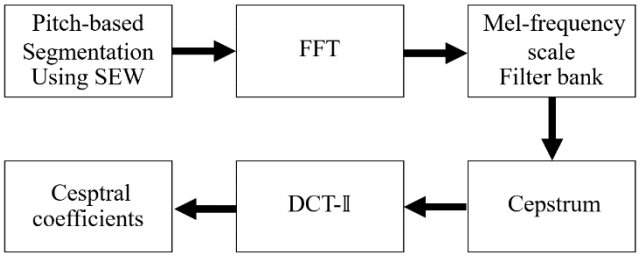


Figure 9: Block diagram of the MFCC Extraction Process

## V. EXPERIMENTS

In this paper, multilayer perceptron networks were used as platforms for evaluating the speech recognition performance. Two types of input data were used in comparison: MFCC vectors of speech segments obtained using the proposed pitch based speech segmentation and those obtained using the fixed frame speech segmentation method. The speech data sampled at 22050 samples/s were gathered from 40 speakers; 20 males and 20 females aged between 20 – 30 years, recorded in a clear environment. The training data set consisted of 11 pronunciations of single numbers, from 0 to 10 in Thai language, repeatedly spoken 10 times by 15 male and 15 female speakers (3300 words in total). The training set was then divided into 10 subsets, each containing 330 words obtained from each speaking time. For a blind test, the test data set consists of the same words as those in the training data set, repeatedly spoken 10 times by 5 male and 5 female speakers who were excluded as speakers in the training data set (1100 words in total). All neural networks comprise of 2 hidden layers with 25 and 20 nodes and an output layer with 11 output nodes while the numbers of input layer nodes were chosen by the size of the MFCC input vectors, a factor affecting upon speech recognition performance.

In the first experiment, to find the appropriate size of MFCC vectors, MFCC vectors of various sizes with 4, 8, 12, 16, 20, and 24 elements were tested. For each size of MFCC vectors, 10 models of neural networks, each trained using one training subset, were implemented as illustrated in Figure 10. In total, 120 neural network models were implemented: 60 models used MFCC vectors obtained using the proposed pitch based speech segmentation as input data and the other 60 models used MFCC vectors obtained using the fixed frame speech segmentation as input data. For each neural network model, the remaining training data subsets, excluding those used in training the network, and the blind test data set were used to evaluate the speech recognition performance of the network for a closed speaker group test and an outside speaker group test respectively, as shown in Figure 11. Average speech recognition results obtained from the first experiment are shown in Table 3. Remarkably, in all cases, the networks using MFCC vectors obtained using the proposed pitch based speech segmentation as input data outperformed those using MFCC vectors obtained using the traditional fixed frame speech segmentation as input data. According to the results in Table 3, the size of MFCC vectors that provided the best recognition results was 16.

In the second experiment, to investigate the effect of noise on the recognition performance, the neural networks providing the best results in the first experiment with MFCC input vector of size 16 elements were tested under noisy input conditions where all testing speech signals were added by the addition of white Gaussian noises with signal to noise ratios (SNR) set to 25, 20, 15 and 10 dB. As SNR decreased, Table 4 shows the declining recognition performances received in the second experiment. Similar to the results in the first experiment, the networks using MFCC vectors obtained using the proposed pitch based speech segmentation as input data still outperformed those using MFCC vectors obtained using the traditional fixed frame speech segmentation as input data.

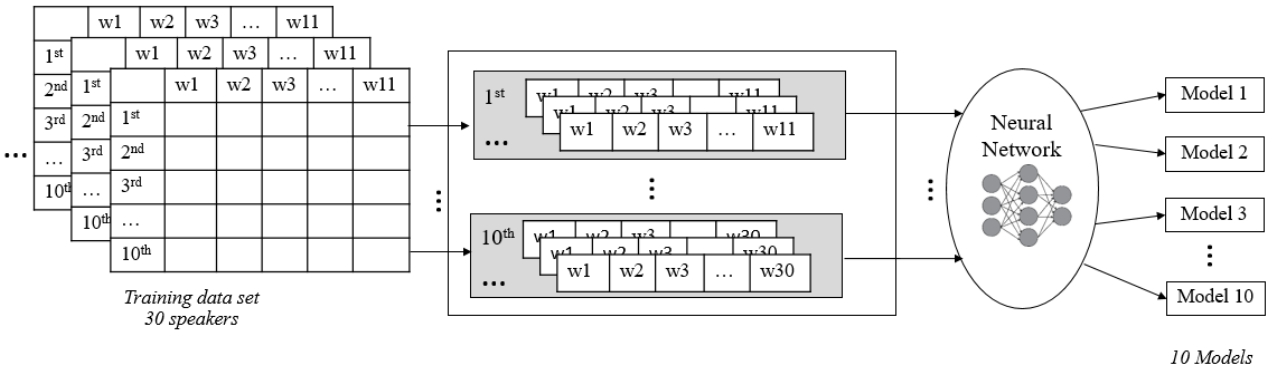


Figure 10: The training data set, training data subsets, and neural network models

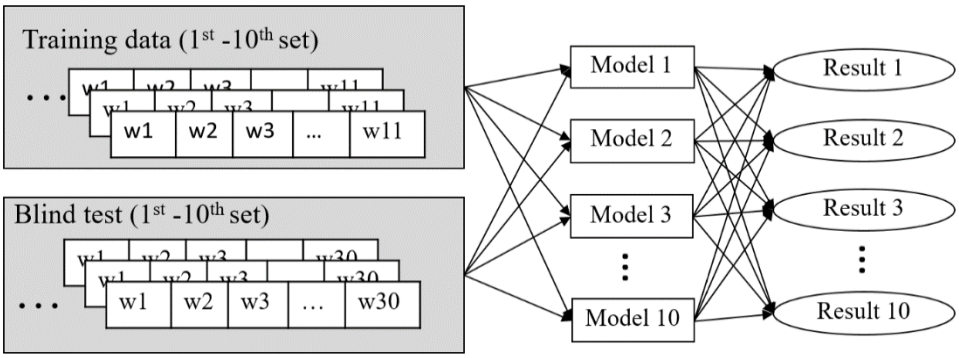


Figure 11: Data sets for evaluating the speech recognition performance

Table 3
Speech recognition performances of ANNs using MFCC vectors obtained from 2 speech segmentation schemes as input data.

| Size of MFCC Vectors | Recognition rates (%) | | | |
|---|---|---|---|---|
| | Fixed frame | | Pitch Based | |
| | Closed speakers | Blind test | Closed speakers | Blind test |
| 4 | 75.29 | 64.05 | 79.85 | 67.99 |
| 8 | 78.42 | 68.82 | 81.81 | 79.25 |
| 12 | 82.56 | 74.84 | 87.04 | 85.12 |
| 16 | 85.26 | 81.15 | 90.48 | 87.15 |
| 20 | 87.15 | 75.48 | 89.80 | 85.45 |
| 24 | 80.61 | 74.18 | 87.05 | 82.76 |

Table 4
Speech recognition performances of the best ANNs with MFCC input vectors of size 16 elements under simulated noisy input conditions

| SNR (dB) | Recognition rates (%) | | | |
|---|---|---|---|---|
| | Fixed frame | | Pitch based | |
| | Closed speakers | Blind test | Closed speakers | Blind test |
| No noise | 85.26 | 81.15 | 90.48 | 87.15 |
| 25 | 77.92 | 74.55 | 89.52 | 86.26 |
| 20 | 71.84 | 70.16 | 88.30 | 82.25 |
| 15 | 68.61 | 63.28 | 78.07 | 72.54 |
| 10 | 52.45 | 47.36 | 64.49 | 59.20 |

## VI. CONCLUSIONS

In this paper, the problem of frequency leakage affecting the speech recognition performance of MFCC-based speech recognition methods, due to fixed frame speech segmentation has been addressed. In the frequency domain analysis, the spectral leakage arises when the length of speech segments does not match complete periods of the speech signal. The solution to reducing frequency leakage can be obtained by using the proposed pitch based speech segmentation method in utilizing the local minima of the corresponding short-time energy waveform as pitch marker features for pitch segmentation. Experiments using multilayer perceptron networks, with MFCC vectors used as input data, as platforms to evaluate the speech recognition performances were conducted. It is found that, compared to the results obtained from the networks using MFCC vectors obtained using the traditional fixed frame speech segmentation as input data, the networks using MFCC vectors obtained using the proposed pitch based speech segmentation as input data provided results with higher accuracy in all cases of both clear sound and noisy conditions due to lower spectral leakage occurring during the MFCC extraction process.

## ACKNOWLEDGEMENTS

This research received financial support from Uttaradit Rajabhat University and the Computer Engineering Research and Development Group, Department of Computer Engineering, Faculty of Engineering, Khon Kaen University.

## REFERENCES

[1] Y. Zouhir and K. Ouni, “Feature Extraction Method for Improving Speech Recognition in Noisy Environments,” Journal of Computer Science, vol. 12, no. 2, pp. 56–61, Mar. 2016.

[2] J. Chaloupka, P. Červa, J. Silovský, J. Žd’ánský, and J. Nouza, “Modification of the speech feature extraction module for the improvement of the system for automatic lectures transcription,” in Proceedings ELMAR-2012, 2012, pp. 223–226.

[3] F. J. Harris, “On the use of windows for harmonic analysis with the discrete Fourier transform,” Proceedings of the IEEE, vol. 66, no. 1, pp. 51–83, Jan. 1978.

[4] S. Shabani and Y. Norouzi, “Speech recognition using Principal Components Analysis and Neural Networks,” in 2016 IEEE 8th International Conference on Intelligent Systems (IS), 2016, pp. 90–95.

[5] W. Hu, M. Fu, and W. Pan, “Primi speech recognition based on deep neural network,” in 2016 IEEE 8th International Conference on Intelligent Systems (IS), 2016, pp. 667–671.

[6] D. Malewadi and G. Ghule, "Development of Speech recognition technique for Marathi numerals using MFCC & LFZI algorithm," *2016 International Conference on Computing Communication Control and automation (ICCUBEA)*, Pune, India, 2016, pp. 1-6.

[7] T. Kinjo and K. Funaki, “On HMM Speech Recognition Based on Complex Speech Analysis,” in IECON 2006 - 32nd Annual Conference on IEEE Industrial Electronics, 2006, pp. 3477–3480.

[8] L. Yuan, “An improved HMM speech recognition model,” in 2008 International Conference on Audio, Language and Image Processing, 2008, pp. 1311–1315.

[9] A. Graves, A. r Mohamed, and G. Hinton, “Speech recognition with deep recurrent neural networks,” in 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013, pp. 6645–6649.

[10] X. Wang, J. Tian, and M. Wang, “Parameter influence on speech recognition rate of modified RBF neural network,” in 2010 International Conference on Intelligent Control and Information Processing, 2010, pp. 76–78.

[11] K. F. Leung, F. H. F. Leung, H. K. Lam, and P. K. S. Tam, “Recognition of speech commands using a modified neural fuzzy network and an improved GA,” in The 12th IEEE International Conference on Fuzzy Systems, 2003. FUZZY ’03, 2003, vol. 1, pp. 190–195 vol.1.

[12] A. Oppenheim, S. Willsky, and S. Nawab, Signals And Systems, second edition ed. PHI Learning, 2009.

[13] M. Sahidullah and G. Saha, “A Novel Windowing Technique for Efficient Computation of MFCC for Speaker Recognition,” IEEE Signal Processing Letters, vol. 20, no. 2, pp. 149–152, Feb. 2013.

[14] I. Reljin, B. Reljin, V. Papic, and P. Kostic, “New window functions generated by means of time convolution-spectral leakage error,” in Electrotechnical Conference, 1998. MELECON 98., 9th Mediterranean, 1998, vol. 2, pp. 878–881 vol.2.

[15] Z. Qawaqneh, A. A. Mallouh, and B. D. Barkana, “Deep neural network framework and transformed MFCCs for speaker’s age and

gender classification," Knowledge-Based Systems, vol. 115, pp. 5–14, Jan. 2017.

[16] G. Zhai, J. Chen, C. Li, and G. Wang, "Pattern recognition approach to identify loose particle material based on modified MFCC and HMMs," Neurocomputing, vol. 155, pp. 135–145, May 2015.

[17] P. Barua, K. Ahmad, A. A. S. Khan and M. Sanaullah, "Neural network based recognition of speech using MFCC features," *2014 International Conference on Informatics, Electronics & Vision (ICIEV)*, Dhaka, 2014, pp. 1-6.

[18] K. Kolhatkar, M. Kolte and J. Lele, "Implementation of pitch detection algorithms for pathological voices," *2016 International Conference on Inventive Computation Technologies(ICICT)*, Coimbatore, 2016, pp. 1-5.

[19] A. de Cheveigneb and H. Kawahara, "Yin, a fundamental frequency ´ estimator for speech and music," J. Acoust. Soc. Am (2002).

[20] B. Kumaraswamy and P. G. Poonacha, "Improved pitch detection using fourier approximation method," in 2015 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), 2015, pp. 522–527.

[21] T. T. Swee, S. H. S. Salleh, and M. R. Jamaludin, "Speech pitch detection using short-time energy," in 2010 International Conference on Computer and Communication Engineering (ICCCE), 2010, pp. 1–6.

[22] T. Ganchev, N. Fakotakis, and G. Kokkinakis, "Comparative evaluation of various MFCC implementations on the speaker verification task," in 10th International Conference on Speech and Computer (SPECOM 2005), Vol. 1, 2005, pp. 191–194.

[23] V. Tyagi and C. Wellekens, On desensitizing the Mel-Cepstrum to spurious spectral components for Robust Speech Recognition, in Acoustics, Speech, and Signal Processing, Proceedings. (ICASSP '05). IEEE International Conference on, vol. 1, 2005, pp. 529–

[24] K. Gupta and D. Gupta, "An analysis on LPC, RASTA and MFCC techniques in Automatic Speech recognition system," *2016 6th International Conference - Cloud System and Big Data Engineering (Confluence)*, Noida, 2016, pp. 493-497.