Voice Recognition Based on Adaptive MFCC and Deep Learning

Hyan-Soo Bae
Robotics and Control System Lab
Yeungnam University
Gyeongsan, Korea
bhs8017@naver.com

Ho-Jin Lee
Robotics and Control System Lab
Yeungnam University
Gyeongsan, Korea
huruhuru91@gmail.com

Suk-Gyu Lee
Robotics and Control System Lab
Yeungnam University
Gyeongsan, Korea
sglee@ynu.ac.kr

Abstract—**In this paper, we propose an enhanced voice recognition method using Adaptive MFCC and Deep Learning. To improve the voice recognition rate, it is important to extract the audio data from original signal. However the existing Algorithms which is used to remove the noise of particular band deteriorate the audio signal. Differently from the existing MFCC, the proposed filter is built up compactly in the data density area to reduce data loss, and impose the weighted value to the data area. As a result, it prevents the data loss which results in improving the recognition rate. In addition the Deep Learning makes it possible to use the Voice recognition without DB. Therefore, it can be effectively used for electronic devices with small memory.**

Keywords— Voice recognition, MFCC, Deep Learning, Noise, Filter

I. Introduction

The Voice recognition technology to transform the voice signal into text is applied to mobile communication, home application, navigation equipment and robot. The voice recognition methods generally are based on comparison between ‘the audio signal to be recognized’ and ‘the audio signal to be set’. In ideal case, embedded system can be used for this purpose. However, the noised voice signal results in fatal error in voice recognition. The speech enhancement method is proposed to eliminate the noise from the audio signal itself,[1][2][3][4]. To compensate for the damaged model for the effects of noise, ‘The model compensation’ method is effective. [5][6], In addtion, The inherently robust speech feature method extracts the feature vector for the robust noise, [7][8], etc. ‘The speech enhancement’ method is Spectral Subtraction[1], MMSE(Minimum Mean Square Error), Wiener filtering[2], Adaptive noise Cancelling[3], Microphone Array[4], etc. The model compensation method deals with is HMM decomposition[5], PMC(Parallel Model Compensation) [6]. The inherently robust speech feature method inclues MFCC(Mel-Frequency Cepstral Coefficient), PLP (Perceptual Linear Prediction)[7], SMC(Short-time Modified Coberence)[8], and Cepstral Compensation[9]. Recently, the MFCC method is relatively popular than other methods. Basically the MFCC calculates the Logarithm energy of the filter bank configured in consideration of human auditory characteristics using DCT(Discrete Cosine Transform). Even though the recognition performance is decreased in the signals with low signal to noise ratio, this method shows good performance. Noise removal filter that is proposed in this paper extracts the higher frequency band over certain decibel by DFT of audio signal. With a different weighting to each extracted frequency, it is distinguished from any other band. In addition, the Filterbank uses the triangle band-pass filter. The MFCC is common that the middle value of the n-th bandpass filter is the first value of (n+1) th filter. However, since in proposed algorithm, the last value of the n-th filter is the first value of the (n+1)th filter, the filter is not overlapped to remove the noise, and extract the cepstrum for voice recognition. By applying the cepstrum value obtained through the DCT to Deep Neural Network of the Neural Network Algorithm, we can build a speech recognizer which can be used in the Embedded system with small memory capacity, and improves the recognition rate.

II. Adaptive MFCC

The average amplitudes of typical noise is much smaller than the average amplitudes of the input signal. Such a small noise is eliminated using ‘Smooth’ in the input signal.

$$\mathrm{s(n)} = \begin{cases} s(n) & s(n) \geq \ value \\ 0 & s(n) \leq value \end{cases} \tag{2-1}$$

The signal without small size noise using a ‘Smooth’ is applied to the Adaptive MFCC Algorithm which converts the inputted voice data to the Mel-scale Algorithm. We can recognize the voice even if the noise is mixed in the sound by using this principle. This method shows excellent performance even in noisy environments due to the processing of non-linear frequency scale which is log scale. The equation 2-2 converts the inputted voice signal to the Mel-scale which is robust to noise.

$$f_{mel} \ = 2595 \times \log_{10}\left(1 + \frac{f}{700}\right) \tag{2-2}$$

This work was supported by BK21+ (Brain Korea 21 Plus) Program funded by the Ministry of Education, Science and Technology.

Where f is the frequency.

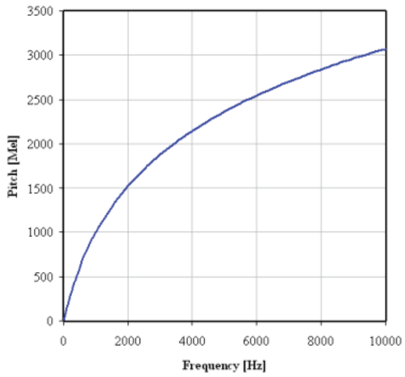


Fig. 1. The relation between Mel-scale and linear frequency

The Figure 1 shows the relation between Mel-scale and linear frequency.

Since the voice signal is continuous input. it is hard to process at once. it is divided in scale of the frame in 25ms ~ 30ms using the ‘Window’. In this case, for the continuity of the start and the end, the Hamming Window as in the equation (2-3) is used.

$$w(n) = 0.54 - 0.46\, cos\left(\frac{2\pi n}{N-1}\right) \quad 0 \le n \le N \quad (2\text{-}3)$$

where ‘N’ is the length of Window function. ‘Figure 2 ’ shows the characteristic curve of Hamming window.

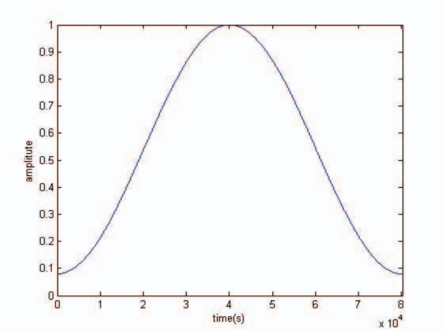


Fig. 2. Characteristic curve of Hamming window

In the Adaptive MFCC method, the frames which is divided by the Hamming window is applied in the Filterbank which consists of Triangle bandpass and filters out the noise. In this case, the intervals of the frequency band become wider in the higher frequency band because Mel-scale is applied. In the Figure 3 which shows the characteristic curve of the Triangle Bandpass Filterbank in the existing MFCC, we can see the overlap between filter.

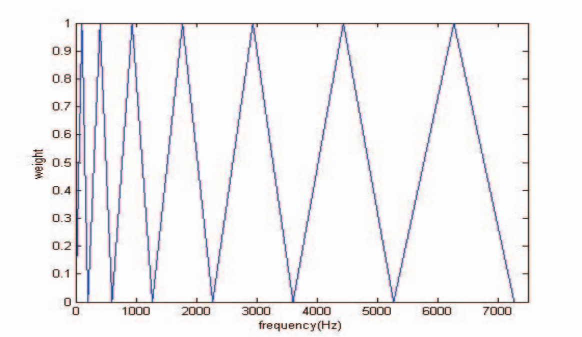


Fig. 3. Characteristic curve of Filterbank

However, in the Adaptive MFCC, if the pass band of the n-th filter was 500 ~ 600 Hz, the pass band of the (n+1)th filter should be designed 600 ~ 750Hz so as not to overlap the each of filters. With this process the noise is eliminated effectively.

Figure 3 shows the characteristic of the modified Filterbank without overlapped filter.

$$y(k) = w(k) \sum_{n=1}^{N} x(n) cos \frac{\pi(2n-1)(k-1)}{2N}, k = 1, \dots, N \quad (2\text{-}4)$$

$$w(k) = \begin{cases} \frac{1}{\sqrt{N}} & k = 1 \\ \sqrt{\frac{2}{N}} & 2 \le k \le N \end{cases} \quad (2\text{-}5)$$

In equation 2-4 and 2-5, N, k, and x(n) describe the number of filters, the degree of feature vector, and the n-th filter value, respectively. The result obtained by the DCT filtered signal was learned by using the DNN(Deep Neural Network) based on the Back propagation. The hidden layer is updated using equation (2-6).

$$\nabla w_{ij}(t+1) = \nabla w_{ij}(t) + \eta \frac{\partial C}{\partial w_{ij}} \quad (2\text{-}6)$$

Using the learning, updated value of the hidden layer calculates the value of output layer to calculate the result of the speech recognizer.

III. Experiment

To prove the efficiency of the proposed algorithm, the result is obtained under irregular noise environments using MATLAB. In the experiment, the Senheiser e835s was used in dynamic MIC system. Especially, the DNN algorithm was used for efficient speech recognition. The DNN with 16 hidden layer for a given number of inputs and 13th filter were use.

A. Removing noise using Adaptive MFCC

In this experiments, the 6 words such as ‘Hello’, ‘Turn On’, ‘Turn Off’, ‘Up’, ‘Down’, and ‘Good bye’ are recoded and used as data. The frame was divided by 30ms by using Hamming Window. The Figure 4 indicates the voice signal of ‘Hello’ recorded in the general laboratory environment.

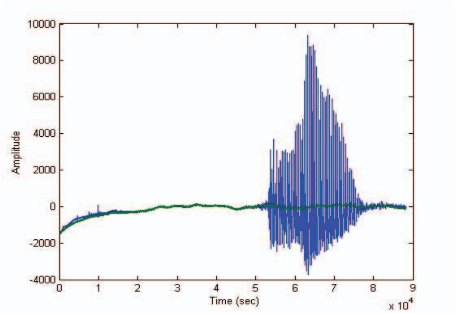


Fig. 4. Voice data for ‘Hello’

In this experiments, after the voice signal of ‘Hello’ is converted to the Mel-scale which is robust to noise, the time domain was converted to the Frequency domain based on DFT. The noise was eliminated using the Filterbank which consists of the Triangle bandpass filter. Figure 5 shows the audio signal after the recorded audio signal is converted into Mel-scale and the converted data using DFT.

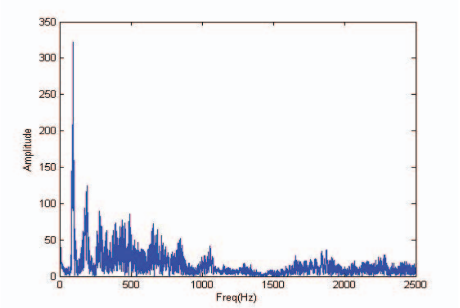
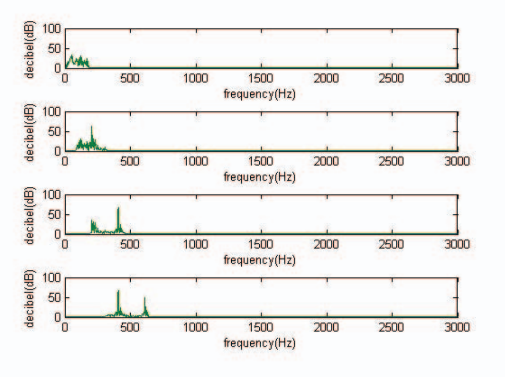


Fig. 5. DFT conversion of the original signal

The noise in the signal that is converted by DFT is eliminated using the Mel-scale Filterbank. Figure 9 shows the signal which passed through the each Filterbank.



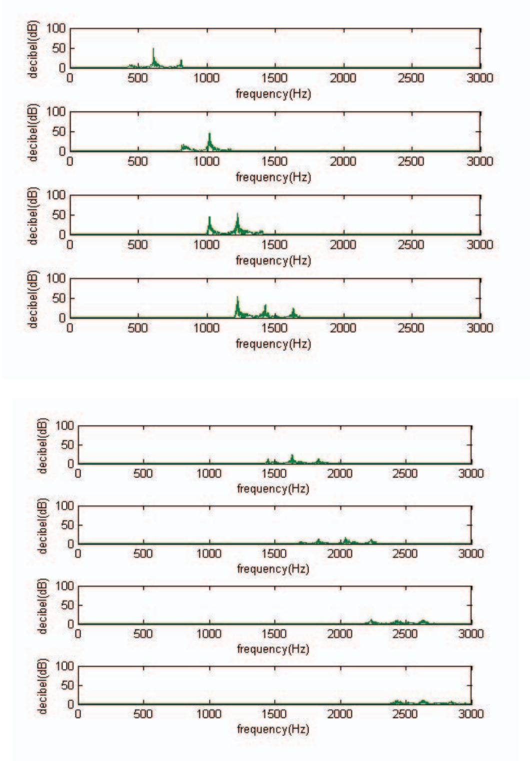


Fig. 6. DFT signal passed through Filterbank

Table 1. Table Type Styles

68Hz	**73** Hz	**77** Hz	**95** Hz
101 Hz	**106** Hz	**131** Hz	**132** Hz
133 Hz	**178** Hz	**256** Hz	**421** Hz
422 Hz	**423** Hz	**479** Hz	**613** Hz
622 Hz	**645** Hz	**673** Hz	**742** Hz
803 Hz	**804** Hz	**805** Hz	**835** Hz
836 Hz	**852** Hz	**853** Hz	**869** Hz
870 Hz	**884** Hz	**911** Hz	**912** Hz
944 Hz	**948** Hz	**980** Hz	**981** Hz
1058 Hz	**1059** Hz	**1067** Hz	**1272** Hz
1275 Hz	**1289** Hz	**1462** Hz	**1543** Hz
1546 Hz	**1693** Hz	**1695** Hz	**1847** Hz
1849 Hz	**1852** Hz		

As shown in Table 1, the bigger weight is imposed to the higher frequency range because it has higher probability to contain more audio data. In order to calculate the each cepstrum using the signal obtained through the Filterbank, the DCT transform was applied. Figure 7 and Figure 8, show the voice signal passed through Filterbank and DFT conversion of filtered signal.

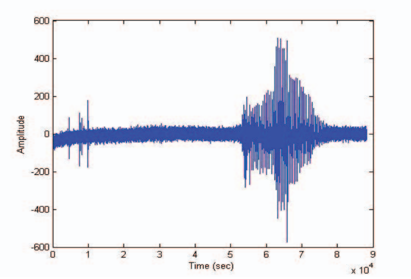


Fig. 7. Voice signal passed through Filterbank

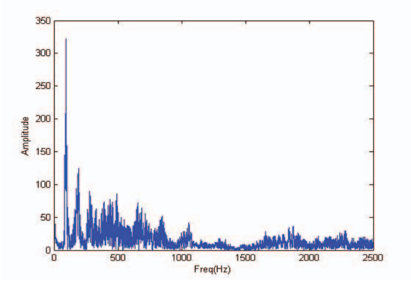


Fig. 8. DFT conversion of filtered signal

In this process, the 13 order MFCC with lower DCT coefficients can be obtained from the coefficients by 20 Filterbanks. Figure 9 shows the capstrum obtained by Adaptive MFCC.

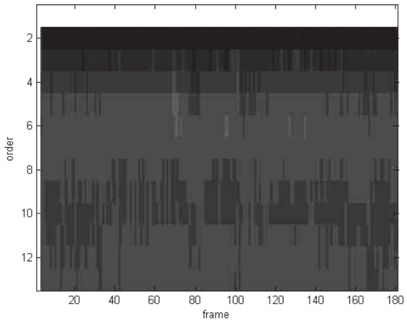


Fig. 9. Capstrum by Adaptive MFCC

B. Recognition of the MFCC and Adaptive MFCC

The analyzer learned 10000 learning data by using the same DNN algorithm with 6 words with the noise of the random frequency of 0.1dB, 10dB, 50dB.

Table 2. Recognition of the MFCC and Adaptive MFCC

	MFCC	Adaptive MFCC
Hello	96.7%	98%
Turn On	96.2%	96%
Turn Off	96.5%	97.2%
Up	96.4%	97.6%
Down	96.3%	98%
Good bye	96%	97.4%

Table 2 shows 96% recognition rate of the MFCC and about 96~98% of Adaptive MFCC.

IV. Conclusion

In voice recognition method, comparing the input audio signal through Mic or specific input device with the data of existing Database is the most widely used. In this method, to remove noise from the obtained signal is one of the most important parts.

In this paper, the enhanced MFCC effectively removes the noise for robust voice recognition. The recognized voice through the Deep Learning, the results are compared in terms of the recognition rate. Especially, to improve the recognition rate, the noise is eliminated without data loss by using ‘Smooth’ and ‘Adaptive filter’. The conventional MFCC which has the disadvantage in removing the specific band noise, has inherently low recognition rates. However, the Adaptive MFCC removes noise in all bands without damaging the audio data by removing the small size noise through the ‘Smooth’ in advance. For robust voice recognition in all frequency range, the average and the variance of about 50 frequency band with bigger magnitude are calculated. The adaptive filter which is robust to in all frequency range is designed by imposing the weighting factor to each filter.

A more advanced adaptive filter is under going to remove white noise expecially for home appliances**.**

REFERENCES

[1] **Boll S, " A spectral subtraction algorithm of acoustic noise in speech", IEEE International Conference on ICASSP '79, vol 4, pp.200-203, 1979.**

[2] **Song Young-Chul, "Effective Noise Suppression in Edge Region Using Modified wiener Filter," The transactions of the Korean Institute of Electrical Engineers D/D 2003, vol 52, no. 3, pp.173-180, 2003**

[3] Widrow. B. et all, "Adaptive Noise Cancelling, Principles and Applications", Proc. Of IEEE 63(12), pp.1692-1716, 1975.

[4] Han Chul-Hee, Kang Hong-Goo, et all, "A Mocrophone Array Beamformer for the Performance Enhancement of Speech Recognizer in Car", The journal of the acoustical society of Korea, vol. 24, no. 7, pp. 423-430, 2005

[5] Miki Kazuhiko, Nishiura Takanobu, et all, "Speech recognition based on HMM decomposition and composition method with a microphone array in noisy reverberant environments," Electronics & communications in Japan Part2 Electronic, vol. 85, no. 9, pp. 13-22

[6] Wang F-M, Kabal P, et all, "Frequency domain adaptive postfiltering for enhancement of noisy speech," Speech communication, vol. 12, no. 1, pp. 41-56, 1993

[7] Shin Won-Ho, Yang Tae-Young, et all, "Speech Recognition Using Noise Robust Features and Spectral Subtraction," the journal of the acoustical society of Korea, vol. 19, no.2, pp. 38-43, 1969

[8] Nitsch B. H, "A Frequency-selective stepfactor control for an adaptive filter algorithm working in the frequency domain," Signal processing the official publication of the European Association for Signal Processing, vol. 80, no. 9, pp. 1733-1745, 2000

[9] Liu Q- G, Champagne. B, Ho D.K.C, "Simple design of oversampled uniform DFT filter banks with applications to subband acoustic echo cancellation," Signal processing the official publication if the European Association for Signal Processing, vol. 80, no.5, pp.831,-847, 2000