Speaker Recognition from Noisy Spoken Sentences

Fatima K. Faek
ELEC Dept. Engineering College, Salahaddin university, Kurdistan region , Iraq

Abdulbasit K. Al-Talabani
Computer Dept. Engineering College, Koya university, Kurdistan region , Iraq

ABSTRACT

In this paper a text independent speaker recognizer from controlled noisy speech signals has been investigated. A recorded data is used for 20 Kurdish speakers (10 males, and 10 females) . The feature used in this work is the MFCC, and k-NN is used as a classifier. The recognition performance from the noisy speech signals has been improved by a de-noising technique using wavelet transform. The result show that the de-noising technique could improve the performance of speaker recognizer by about 36%.

Keywords

Speaker recognition in noisy environment, MFCC features, k-NN classifier, de- noising signals in wavelet domain.

1. INTRODUCTION

Perfect speaker recognition from noisy speech signals is not easy due to some factors, for example , noise changes acoustic features of a speaker , making them different to those of a clean signal. Recently, many researchers have studied different techniques to reduce the effect of the noise[1]. Choosing feature extraction technique is very important for efficient speaker recognition. Unnecessary information of the speech must be eliminated, keeping only those features that are applicable for classification. Additionally, robust feature selection is helpful for better estimation of the model parameters, and less computational time, and appliances is needed.[2]

Mel-Frequency Cepstral Coefficients (MFCC), is used by researchers for speech and speaker recognition [**3**], since MFCC carry the vocal tract shape and length in addition to the frequency distribution that define sounds, , which are speaker specific features.

The aim of this work is to use a de-noising technique to decrease the effect of noise. MFCC feature is used , since MFCC are the most popular choice for any speaker recognition system, though the tradeoff of using MFCC is that the signal is supposed to be stationary in a given time interval ,therefore it is not suitable for analyzing the non-stationary signals (noisy signals) [4], thus a special de-noising technique is used here to avoid this problem.

The effect of different types of noises has been studied by many researchers.[5] , studies the problem of speaker recognition using speech signals corrupted with environmental noise. [6], describes how much the phase information is efficient for speaker recognition process in a noisy condition? Researchers like [1], [7], propose a method for enhancing the acting of speaker recognizer under limited data condition in a noisy environment, while , [8] studies an approach for speaker recognition in noisy condition using the multi-dimensional pulse signals .

[9] is a text independent study with the presence of noise, by using MFCC as feature and ANN as a classifier, [**4]** uses a technique for extracting the MFCC features by combining different types of AM-FM modulation/demodulation techniques for feature extraction . Different modifications are made by[10-13] to avoid the tradeoff of using MFCC with noisy signals.

2. THE DATA BASE

A recorded speech data of 20 Kurdish persons (10 males and 10 females) is used in this work. Two recorded speech statements are recorded for each person, one for training, and the other for recognizing, thus this work is a text independent speaker recognition.

3.THE SUGGESTED RECOGNIZER CHARECTERSTICS

3.1 Preprocessing (de-noising the noisy speech signal)

The de-noising technique used in this work is started by transforming the speech signal to the wavelet domain, thresholding the detail coefficients (outputs of high pass wavelet filter), and then reconstructing the speech signal .

Two types of thresholding are existed : hard and soft thresholding , hard thresholding can be described as the usual process of setting to zero the elements whose absolute values are lower than the threshold, this type is used for compressing signals .Soft thresholding is an extension of hard thresholding, first, elements of absolute values lower than the threshold are set to zero , and second, the nonzero coefficients are shrank towards 0, this technique is used for de-noising signals. Let t denote the threshold. The hard threshold signal is x if $|x| > t$, and is 0 if $|x|< t$. The soft threshold signal is $(|x| - t)$ if $|x| > t$ and is 0 if $|x| < t$. [14]

In this study the soft thresholding technique is used to de-noise the noisy speech signal.

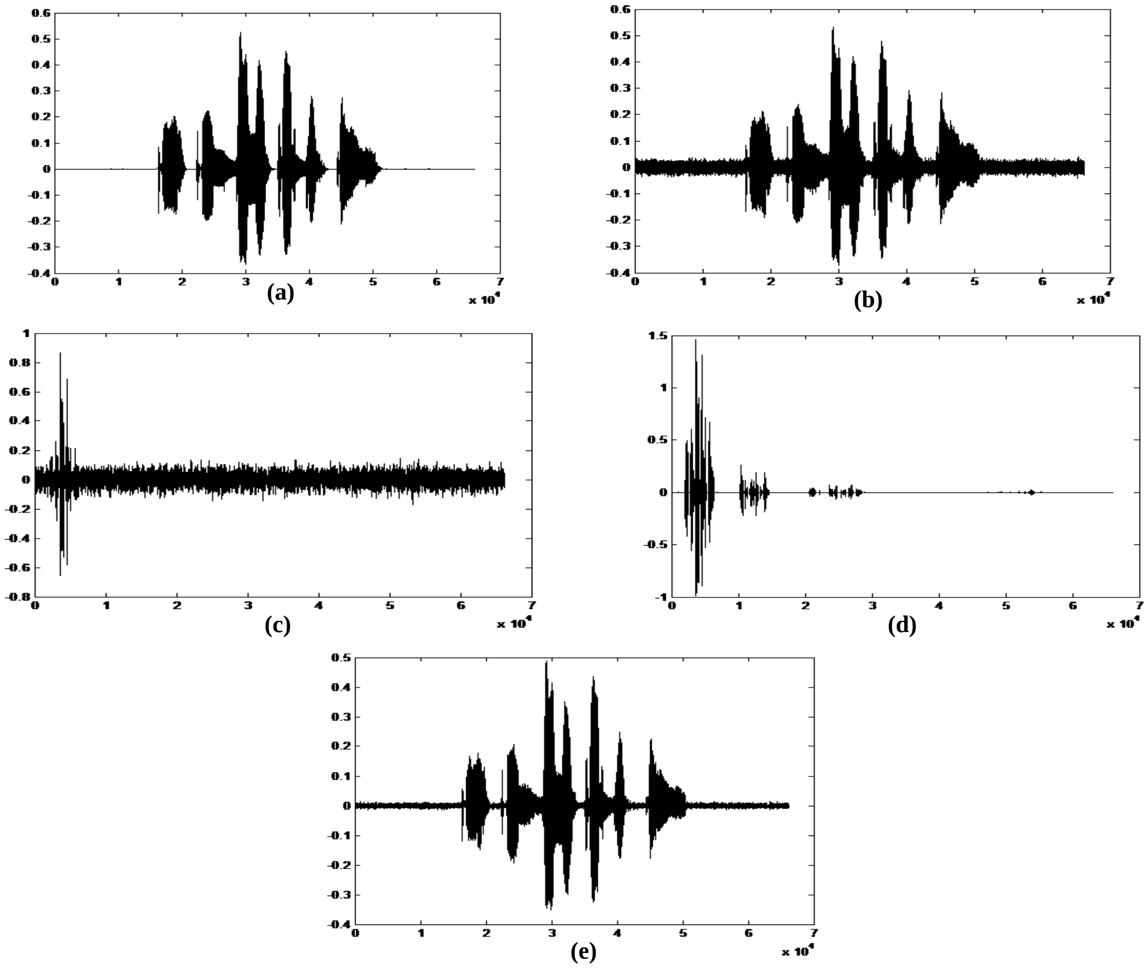


Figure 1: (a) The original speech signal . (b)The noisy speech signal(S/N ratio is 30dB). (c)The noisy wavelet coefficients. (d) The de-noised wavelet coefficients. (e) The de-noised speech signal .

An example of applying the described algorithm on a noisy speech signal is as shown figure (1):

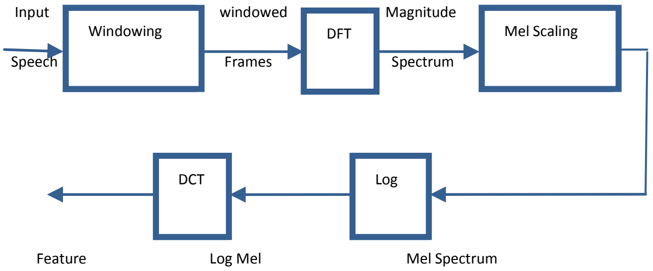
3.2 Feature extraction using MFCC

The MFCC parameter, was first used by Davis and Mermelstein [15], MFCC presents the energy distribution of the speech signal in the frequency domain. This method based on the Mel frequency scale and is related to human hearing , and it can be used as an anti-noise more than other parameters, like LPC. [16],[17]

In MFCC, feature extraction filters are spaced linearly at low frequencies and logarithmically at high frequencies, since most of the important human's perspective information are concentrated in the low frequency components of the speech signal. [4].

The technique of extracting MFCC features can be described by the block diagram shown in figure 2. [4]

Figure 2.The feature extraction technique using MFCC.



The first step is to use hamming window for windowing the signal [4], this window is given by equation (1):

$$w[n] = 0.54 - 0.46\,cos\left(\frac{2\pi n}{L}\right);\ 0 \le n \le L - 1 \ldots..(1)$$

The second step is to compute the spectral information from the windowed signal using DFT, which gives the information about the energy rate at each frequency band. The frequency output are re-arranged according to Human hearing sensitivity to the frequency bands (Male scale) . The relation of the frequency Mel scale is linear below 1000 Hz , and logarithmic above 1000 Hz [4] .The Mel frequency *m* can be computed from equation (2):

$$Mel(f\,) = 1127\ln\left(1+\left(\frac{f}{700}\right)\right)\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2)$$

A bank of filters will collect the energy from each frequency band, with 10 linearly spaced filters below 1000 Hz, and 10 logarithmically spread filters above 1000 Hz [4],[15]

Finally (20) MFCC are calculated by taking discrete cosine transform of the logarithm of the power spectrum [4]. Most researchers used 12 MFCC, while, in the final part of this work only 5 MFCC are used.

3.3 Feature Selection

In this study robust MFCC features are selected, the five highest MFCC coefficients are selected after the de-noising process, since their frequencies are the highest, for they are affected by the de-noising technique , these features are to be fed to the k-NN classifier for the recognition process.

3.4 The k-NN Classifier

The k-Nearest-Neighbors (k-NN) is a simple arbitrary classifier. This classifier is highly applicable in many cases [18] [19], k-NN approach can be described by the following classification example: If a recorded data x is to be classified, its *k* nearest neighbors are referred to, and this forms a neighborhood of x . Simply this classifier classifies each set of the data in sample into one of the groups in training.[19]

4. RESULTS AND DISCUSSION

This work consists of three parts :

1-Speaker recognition using 12 MFCC features, and k-NN classifier from clean speech signals (before adding white Gaussian noise to these signals). Both utterances of each person is used here, one for training the classifier and the other for recognizing the speaker. The result of the recognition is 100%.

2-Speaker recognition using 12 MFCC features, and k-NN classifier from noisy speech signals that are corrupted by white Gaussian noise ,and with five values of S/N ratios (10dB,20dB,30dB,40dB,and 50 dB). One of the spoken sentences is used for training and the other one is corrupted by noise with five S/N ratios , these noisy signals are used for recognizing the speaker. Results of this part are shown in table (1) ,the graph is shown in figure (3):

The drawback of MFCC features with noisy signals [4], is the reason behind obtaining these results.

Table 1. The results of part two of this work.

S/N Ratio (dB)	10	20	30	40	50
Recognition Rate	20%	30%	30%	40%	45%

3-Speaker recognition using the 5 highest MFCC features, and k-NN classifier from de-noised speech signals, for five S/N ratios. One of the spoken sentences is used for training and the other one is corrupted by noise with the previous five S/N ratios , and then de-noised in the wavelet domain , the de-noised signals are used for recognizing the speaker. The results of this part are shown in table(2), the graph is shown in figure (3):

The results show the robustness of using the described de-noising technique with the MFCC features, in addition to the process of feature selection.

Table 2. The results of part three of this work.

S/N Ratio (dB)	10	20	30	40	50
Recognition Rate	45%	55%	70%	80%	95%

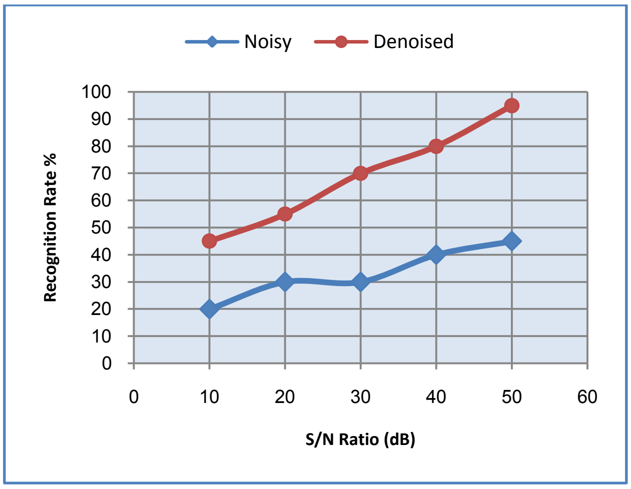


Figure 3.The results of part two and three of this work.

5. CONCLUSIONS

Adding white Gaussian noise with different ratios affect the recognition rate negatively especially using MFCC features, since MFCC based on DFT which is not suitable with non-stationary signals (noisy signals). The suggested de-noising technique is not able to improve the performance of the speaker recognizer using 12 MFCC features, while using high MFCC (five highest) , improves the performance by about 36%. The reason behind this might be the fact that the de-noising technique used affect the high frequency components of the speech signal, which means that the enhancement process has been done on the high frequency components of the noisy speech signal.

6. REFERENCES

[1] N.McLaughlin, J. Ming, and D. Crookes. 2011.Speaker recognition in Noisy Condition With limited Training Data,19th European Signal Processing Conference , Barcelona , Spain .

[2] M. Zamalloa, G. Bordel, L. J. Rodriguez, M. Penagarikano.2006.Feature Selection Based on Genetic Algorithms for Speaker Recognition,IEEE Odyssey - The Speaker and Language Recognition Workshop.

[3] A. Srinivasan. 2012. Speaker Identification and Verification using Vector Quantization and Mel Frequency Cepstral Coefficients, Research Journal of Applied Sciences, Engineering and Technology 4(1): 33-40.

[4] V.Tiwari ,G. Ganga , J. Singhai , and M. Azad .2011. Wavelet Based Noise Robust Features for Speaker Recognition ,An International Journal (SPIJ), Volume (5) : Issue (2) .

[5] Ji Ming, Timothy J. Hazen, James R. Glass, and Douglas A. Reynolds. 2007. Robust Speaker Recognition in Noisy Conditions, Transaction on Audio, Speech, and Language Processing , vol. 15, no. 5, 1711-1723.

[6] L.Wang, K. Minami, K. Yamamoto, S. Nakagawa. 2010. Speaker Identification By Combining MFCC and Phase Information in Noisy Environments , IEEE 4502-4505 ICASSP.

[7] P.Krishnamoorthy , H.S. Jayanna , S.R.M. Prasanna .2011.Speaker recognition under limited data condition by noise addition ,Expert Systems with Applications 38 13487–13490.

[8] T.Azetsu, M. Abuku, N. Suetake and E. Uchin. 2012.Speaker Identification in Noisy Environment with Use of the Precise Model of the Human Auditory System ,Proceeding of the international multi Conference of Engineers and Computer Scientists Vol1, IMECS, Hong Kong .

[9] H. Seddik, A. Rahmouni and M. Sayadi, C. Esstt, AV. Taha Hussein .2010.Text Independent Speaker Recognition Using The Mel Frequency Cepctral Coefficients And A Neural Network Classifier, 1008, Tunis, Tunisia , Ecole National Des Scince Information, Manouba, IEEE.

[10] R Sarikaya et.al. 1998.Wavelet packet transform features with application to speaker identification, in proceedings of the IEEE Nordic signal processing symposium .

[11] R. Sarikaya and J. H. L. Hansen.2000.High resolution speech feature parameterization formonophone-based stressed speech recognition, IEEE signal processing letters vol7(7),pp. 182-185.

[12] O. Farooq and S. Datta .2001.Mel Filter- Like Admissible Wavelet packet Structure for Speech Recognition , IEEE Signal Processing letters, Vol.8 , No. 7,pp 196-198 .

[13] P. Maragos, J. F. Kaiser and T. F. Quatieri .1993.Energy Separation in Signal Modulations with Application to Speech Analysis, IEEE transactions on signal processing, vol. 41, no. 10, pp. 3024-3051

[14] D.L. Donoho, 1995.De-noising by soft-thresholding, IEEE, Trans. on Inf. Theory, 41, 3, pp. 613-627.

[15] S. Davis and P. Mermelstein. 1980.Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, IEEE Transaction Acoustics Speech and Signal Processing, vol. 4, pp. 375-366.

[16] J. Wu and Z. G. Cao,2005.Improved MFCC-Based Feature for Robust Speaker Identification, TSINGHUA Science and Technology, vol. 10, pp. 158–161.

[17] Wang Yutai, Li Bo, Jiang Xiaoqing, Liu Feng, Wang Lihao.2011.Speaker Recognition Based on Dynamic MFCC Parameters, IEEE .

[18] D. Hand, H. Mannila, P. Smyth.2001. Principles of Data Mining.The MIT Press.

[19] G. Guo, H. Wang ,K.Greer ,D. Bell , and Y. Bi .2001.KNN Model-Based Approach in Classification ,This work was partly supported by the European Commission project ICONS, project no. IST-2001-32429.