Real-Time Arabic Speech Recognition

Zaid Y. Mohammed
Computer Engineering Department
University of Mosul
Iraq

AbdulSattar M. Khidhir, Ph.D
Mosul Technical InstituteFoundation of Technical Education
Iraq

ABSTRACT

Speech recognition system needs to perform a high complex calculation and short time to complete it. This is a big challenge for the real-time systems. However, using a simple and fast algorithm may do this task for the slow systems. Thus, the main objective of this paper is to design and implement a Real-Time Arabic Speech Recognition system using MATLAB environment. It is capable of accurately identifying some letters while remaining simple and fast. It uses the Mel-Frequency Cepstral Coefficients (MFCCs) as a feature extraction and Euclidean distance to compare the test sound and the database. A recognition rate of 89.6% has been reached.

General Terms

Speech Recognition.

Keywords

Feature extraction, *Mel-Frequency Cepstral Coefficients* (*MFCCs), Feature match*.

1. INTRODUCTION

Speech recognition became more popular due to the increased usage of digital-embedded systems like computers, mobile phones, cars, toys and other appliances [1]. These systems have to understand the Arabic language because, it is the second language in the world.The main idea is to convert voice signal to text by the computer in real-time manner, There are many algorithms to do such conversions. These algorithms depend on how the voice signals are processed and how the features are extracted, how can the speech recognition system recognize and identify these features and how fast these algorithms to be suitable for real-time systems. The (MFCCs) is very good algorithm for speech recognition application which is based on human hearing perceptions, it is used in this paper for features extraction[2], due to the simplicity the Euclidean distance is used for features match and identification.The main aim of this paper is to design and implement independent-speaker Arabic speech recognition system. The description of how the signals are processed and how the features are extracted is explained in section 2, the results and the discussions are presented in section 3 and finally the conclusions are briefly described in section 4.

2. SPEECH RECOGNITION

2.1 Speech Recognition Algorithm

The microphone captures the speech signals. Those signals are sampled and converted to digital form by the analog to digital converter (A/D) at frequency 11025 Hertz. The features are extracted from these signals by applying some steps including Pre-emphasis, Framing, Windowing and (MFCCs).The recognition is done by calculating the minimum Euclidean Distance measured between the (MFCCs) and database to determine which letter is pronounced [3].

2.2 Feature extraction

The recognition rate depends absolutely on the features that are extracted from the input speech signals, better feature extraction for better recognition rate with minimum error rate. The MFCCs algorithm is chosen because, it is less sensitive to the speaker-depend variations that appear in the speech signals, it is based on human hearing perceptions, which is a linear spaced at frequency less than 1000Hertz and logarithmic spaced at frequency larger than 1000Hertz.The overall features extraction steps are described below (see Figure 1) [4].

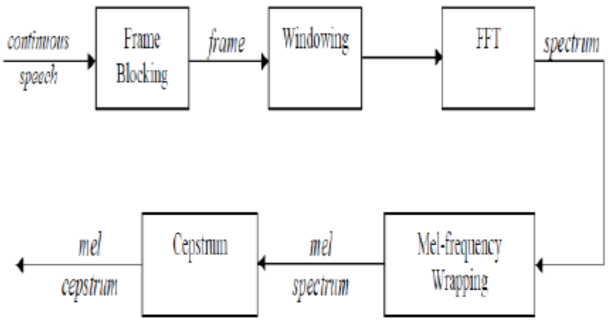

Fig 1: overall features extraction steps.

Step 1: Pre–emphasis (spectrum normalization)
The pre-emphasis process is to straighten the spectrum of the speech signals, it is simply, first order high pass filter to attenuate the high energy of the low frequency band [5]. The output equation of the pre-emphasis process is described as follows:

$$y(n) = x(n) - 0.95x(n-1) \qquad (1)$$

Step 2: Framing
The speech signal is assumed to be stationary signal if it is divided into frames[6], these frames determine the system complexity and efficiency, small frame size should be processed in small period of time and produce redundancy data, whereas signal stationary may be violated for large frame size, the frame size typically about 10-20 ms with 50% overlapping[7].

Step 3: Hamming windowing
Framing process produce discontinuity frames, the Hamming window (Figure 2) is used to lessen these discontinuities as much as possible [8]. The Hamming window can be described below:

$$w(n) = 0.54 - 0.46 \cos\left({}^{2\pi n}/_{N - 1}\right) \qquad (2)$$

Where:
N: Number of samples in each frame.
$w(n)$: Hamming window.

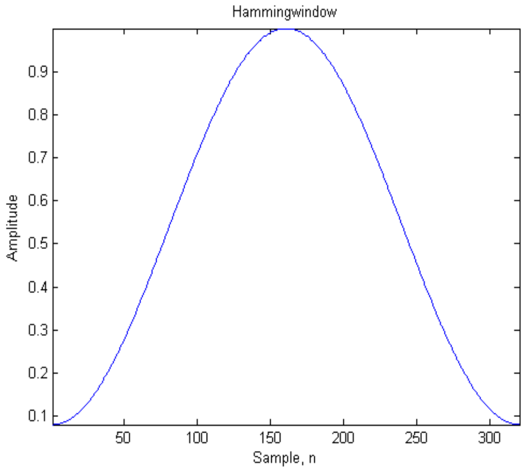

Fig 2: Hamming window.

Step 4: Fast Fourier Transform
The frequency domain tells more information about the speech signal than time domain does. Therefore, the Fast Fourier Transform is used to transform the signal from time domain to frequency domain [9]. The convolution process in time domain between the vocal chords and the resonance vocal tract can be converted to multiplication in frequency domain so as to be separated by Cepstral analysis, this separation produce independent-speaker speech recognition. The equations below describe these statements [10]:

$$\mathrm{Y(f)} = \mathrm{FFT}[\mathrm{v(t)} * \mathrm{g(t)}] = \mathrm{V(f)}.\,\mathrm{G(f)} \qquad (3)$$

Where:
$v(t)$: Vocal tracts signal.
$g(t)$: Vocal chords.

2.3 Mel-Frequency Cepstral Coefficients (MFCCs)

To simulate the human perceptions, the warping from frequency in Hertz to Mel-scale is used [11]. The following equations describe the warping from frequency in Hertz to Mel-scale and vice versa.

$$\mathrm{F_{Mel}} = 2595\log_{10}(1 + \frac{\mathrm{F_{Hz}}}{700}) \qquad (4)$$

$$\mathrm{F_{H}} = 700(10^{\frac{\mathrm{F_{Mel}}}{2595}} - 1) \qquad (5)$$

The warping can be done using triangular filter banks (see Figure 3) that are linear spaced below 1000 Hertz and logarithmic spaced above 1000 Hertz, frequencies below 1000H contain more information than other frequencies, therefore more triangular filter banks are used to capture these information [12].

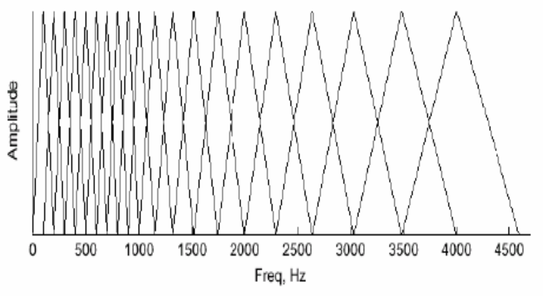

Fig 3: Triangular filter banks.

3. FEATURE MATCH

After the (MFCCs) processing, the result is 20 dimensions feature vector, this feature vector will be compared with the reference module (database).The Euclidean Distance is used to compute the distance between feature vector of the unknown pronounced letter and all the letters that are stored in the database[13,14]. Equation below is used to compute the Euclidean Distance as follow:

The Euclidean Distance between unknown feature vector F(x1,x2,….,x20) and the database feature vector D(y1,y2,….,y20) is:

$$\mathrm{E.D} = \sqrt{(\mathrm{y}_1 - \mathrm{x}_1)^2 + (\mathrm{y}_2 - \mathrm{x}_2)^2 + \cdots. + (\mathrm{y}_{20} - \mathrm{x}_{20})^2}$$

$$= \sqrt{\sum_{\mathrm{i}=1}^{20} (\mathrm{y_i} - \mathrm{x_i})^2} \qquad (6)$$

After computing all the Euclidean Distances between the feature vector of the unknown letter and feature vectors that are stored in the database and represents all the letters, the letter with lowest Euclidean Distance is chosen to be the pronounced letter.

4. RESULTS AND DISCUSSION

The speech recognition system presented in this paper is implemented and work under MATLAB environment. The sound signal is recorded from four persons, two females and two males, these signals pass through the system components and after pre-processing and feature extraction the comparator phase will start, the comparator will calculate and find the minimum Euclidean distance to investigate the pronounced letter. The software was running under windows 7 using PC core i3/2.1G, 4G RAM. The system results are described in table 1.

Table 1. The results.

	File length[sec]	Processing time[sec]	Recognition rate
Person1	**86**	**2.660100**	**90.08%**
Person2	**95**	**2.747884**	**88.4%**
Person3	**104**	**2.850499**	**89.4%**
Person4	**142**	**3.849604**	**90.86%**
average	**106.75**	**3.027021**	**89.685**

5. CONCLUSIONS

In this paper we have designed an Arabic speech recognition system in Matlab environment which is preprocess the signal with Pre-emphasis, Framing, Windowing. We use the Mel-Frequency Cepstral Coefficients (MFCCs) algorithm to extract feature from the speech signals, we depend on the Euclidean Distance to compare the test sound and the database, the recognition rate was between 88.4% - 90.86%. Our future work is to use Mel-Frequency Cepstral Coefficients (MFCCs) and the Euclidean Distance in the implementation of speech recognition system based on Field Programmable Gate Arrays (FPGAs).

6. REFERENCES

[1] S. J. Melnikoff, S. F. Quigley and M. J. Russell, Implementing a simple continuous speech recognition system on an FPGA, Proc. of the 10th Annual IEEE Symposium on Field-Programmable Custom Computing Machines, pp.275-276, 2002.

[2] J. R. Deller, J. H. L. Hansen and J. G. Proakis, "Discrete-Time Processing of Speech Signals", IEEE Press, 2000, 56-63 380-385

[3] S. Sakti, K. Markov, S. Nakamura, W. Minker, "Incorporating Knowledge Sources into Statistical Speech Recognition", Springer, 2009, page 39-40.

[4] Lawrence rabinar,biing-hwangjuang, "fundamental of speech recognition ",prentice hall,1993.

[5] L. Deng and D. O'Shaughnessy, Speech Processing A Dynamic and Optimization-Oriented Approach, Marcel Dekker, New York, 2003.

[6] W. C. Chu, Speech Coding Algorithms, John Wiley and Sons, Wiley-IEEE, 2003.

[7] L. Muda, M. Begam, I. Elamvazuthi, " Voice Recognition Algorithms using MelFrequency Cepstral Coefficient (MFCC) andDynamic Time Warping (DTW) Techniques", Journal of Computing, Volume 2, Issue 3, March 2010, ISSN 2151-9617

[8] X. Huang, A. Acero and H. Wuenon, Spoken Language Processing: A Guide to Theory, Algorithm and System Development, Pearson, 2005.

[9] R. Vergin, "An algorithm for robust signal modeling in speech recognition," Proceedings of the 1998 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '98), Vol. 2, pp. 969-972, May, 1998.

[10] B. P. Lathi, "Modern Digital and Analog Communication Systems", California state universtiy,1998

[11] Wei Han, Cheong-Fat Chan, Chiu-Sing Choy and Kong-Pang Pun "An efficient MFCC extraction method in speech recognition" Department of Electronic Engineering" The Chinese University of Hong Kong, Hong, IEEE – ISCAS, 2006.

[12] Steven B. Davis, and Paul Mermelstein, Comparison of parametricrepresentations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. ASSP, 1980.

[13] R. S. Kurcan, "Isolated Word Recognition From in-ear Microphone Data Using Hidden Markov Models (HMM)",M. Sc. Thisis,2006,178 pp.

[14] Stephen E. Levinson, "Mathematical Models for Speech Technology", John Wiley&Sons,ltd, University of Illinois at Urbana-Champaign, USA,2005.