وزارة التعليم العالي والبحث العلمي
الجامعة التقنية الوسطى
الكلية التقنية الإدارية – بغداد

وقائع المؤتمر العلمي التخصصي الرابع

للكلية التقنية الإدارية – بغداد

للمدة من

28- 29 /11 / 2018

تحت شعار

الإبداع الإداري لتحقيق الرؤية المستقبلية

لمنظمات الأعمال

المجلد الثاني / رقم الايداع (642)

البحوث المنشورة محكمة

الفهرست
المجلد الثاني

Speech Recognition using Discrete Wavelet Transform and Neural Network

Rafeef Qasim Abbas Dr.Ismael Hadi Challob
Department of information Techniques
Technical College of management
Bagdad, Iraq

Abstract

This paper introduces a speech recognition algorithm based on Discrete Wavelet Transform (DWT). A big challenge of a real time speech recognition algorithm is the complex of calculation, so using a simple and fast algorithm is very important. Mel-Frequency Spectral Coefficients (MFCCs) is used for compression and feature extraction the DWT used to reinforcement the compression and feature extraction process. The classification step is performed by artificial neural network (ARNN) with a back propagation that compare the test speech with the speech that stored in enrollment database. The system is tested by two databases: the first one called Multiple microphone data recorded by a PDA-like mock-up (PDAm), and the 2nd one a real test database. Matlab programming tool has been used to performed this algorithm, and a recognition rate for the two databases are (97.5%) and (100%) respectively.

Keywords
MFCCs, DWT, ANN.SPEECH RECOGNITION

تمييز الصوت باستخدام التوزيع الموجي المتقطع والشبكات العصبية

المستخلص

في بحثنا هذا نقدم خوارزمية تمييز الصوت باستخدام التوزيع الموجي المتقطع (DWT). يتمثل تحدي تمييز الصوت في الزمن لحقيقي من خلال التعقيدات الحسابية. لذا فأن استخدام خوارزمية تتسم بالسرعة والسلاسة يعد امر غاية في الاهمية. تم في بحثنا استخدام طريقة معاملات الطيف الترددية (MFCCs) لاستخلاص الصفات المميزة للصوت وتقليل الخصائص، ثم استخدام (DWT). عملية التصنيف تضمنت استخدام الشبكات العصبية وبخوارزمية (Back Propagation) والتي تقارن

بين قاعدة بيانات الصوت الاختبارية مع قاعدة بيانات الصوت الفعلية. تم اختبار النظام باستخدام نوعين من قواعد بيانات الصوت (PDAm) وهي الافتراضية، والثانية هي قاعدة بيانات حقيقية مخزونة لغرض التصنيف.

تم استخدام ادارة ماتلاب البرمجية لتنفيذ الطريقة. كان معدل التمييز هو (97.5%) و (100%) على التوالي.

INTRODUCTION

Any person has features or fingerprints which distinguish him to other persons, face, iris, hand, speech, and others are examples of these features. Computer users, scientists, programmers, and software designers began to use certain devices to conduct precise experiments on the human voice to achieve the characteristics that characterize a person's voice Assigned to the rest of the votes.

Speech recognition became more popular due to the increased usage of digital-embedded systems like computers, mobile phones, cars, toys and other appliances. 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. The purpose of using the DWT is to benefit from its localization property in the time and frequency domains. ANN is nonlinear model that is easy to use and understand compared to statistical methods. ANN is non-parametric model while most of statistical methods are parametric model that need higher background of statistic. ANN with Back propagation (BP) learning algorithm is widely used in solving various classification and forecasting problems.

SPEECH RECOGNITION

Speech Recognition Algorithm

The speech signals are captured by the microphone. 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, (MFCCs) and DWT. The recognition is done by ANN with a one hidden layer of (800 neurons) and output layer determined by the number of person's speech[1].

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 1000 Hz and logarithmic spaced at frequency larger than 1000 Hz. The overall features extraction steps are described below in Figure (1) [2]

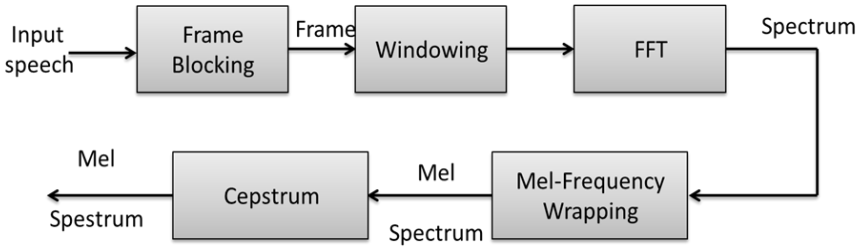

Fig 1: overall features extraction step[3]

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. The output equation of the pre-emphasis process is described as follows:[2]

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

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[2].

Step 3: Hamming windowing

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

$$W(n) = \begin{cases} 0.54 - 0.46\cos(\frac{2\pi n}{l-1}) & 0 \le n \le N \\ 0 & other \end{cases} \qquad (2)$$

Where:

N: Number of samples in each frame.

W(n) : Hamming window

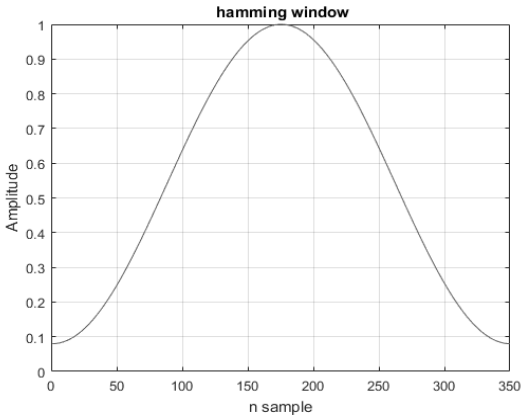

Fig 2: Hamming window [5]

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[4]. 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 Spectral analysis, this separation produce independent-speaker speech recognition. The equations below describe these statements.

$$y(f) = FFT[v(t) * g(t)] = V(F).G(F) \qquad (3)$$

Where:

V(t): vocal tracts signal.

g(t): vocal chords.

2.3 Mel-Frequency Spectral Coefficients (MFCCs)

To simulate the human perceptions, the warping from frequency in Hertz to Mel-scale is used[5]. The following equations describe the warping from frequency in Hertz to Mel-scale and vice versa.

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

$$F_H = 700\left(10^{\frac{F_{Mel}}{2595}} - 1\right) \qquad (5)$$

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.

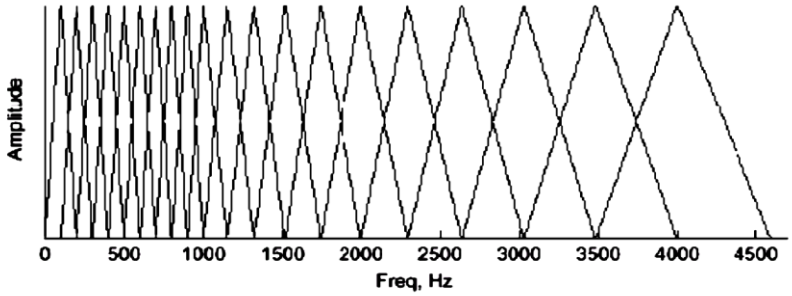

Fig 3: Triangular filter banks. [1]

DISCRETE WAVELET TRANSFORM

The Wavelet Transform (WT) is a technique for analyzing signals. It was developed as an alternative to the short time Fourier Transform (STFT) to overcome problems related to its frequency and time resolution properties. More specifically, unlike the STFT that provides uniform time resolution for all frequencies the DWT provides high time resolution and low frequency resolution for high frequencies and high frequency resolution and low time resolution for low frequencies. In that respect it is similar to the human ear which exhibits similar time-frequency resolution characteristics. The Discrete Wavelet Transform (DWT) is a special case of the WT that provides a compact representation of a signal in time and frequency that can be computed efficiently[6, 7].

The DWT is defined by the following equation:

$$W(j,k) = \sum_{j}\sum_{k} x(k)2^{-j/2}\,\psi(2^{-j}n - k) \qquad (6)$$

Where ψ(t) is a time function with finite energy and fast decay called the mother wavelet. The DWT analysis can be performed using a fast, pyramidal algorithm related to multi rate filter banks. As a multi rate filter bank the DWT can be viewed as a constant Q filter bank with octave spacing between the centers of the filters. Each sub band contains half the samples of the neighboring higher frequency sub band . In the pyramidal algorithm the signal is analyzed at different frequency bands with different resolution by decomposing the signal into a coarse approximation and detail information. The coarse approximation is then further decomposed using the same wavelet decomposition step. This is achieved by successive high pass and low pass filtering of the time domain signal and is defined by the following equations:

(7)

(8)

$$Y_{high}[k] = \sum_n x[n]g[2k-n] \qquad (7)$$

$$Y_{low}[k] = \sum_n x[n]h[2k-n] \qquad (8)$$

where y [k], y [k] high low are the outputs of the high pass (g) and low pass (h) filters, respectively after subsampling by 2. Because of the down

sampling the number of resulting wavelet coefficients is exactly the same as the number of input points. A variety of different wavelet families have been proposed in the literature. In our implementation, the 4 coefficient wavelet family (DAUB2) proposed by Daubechies is used[8, 9].

FEATURE EXTRACTION & CLASSIFICATION

After the MFCCs coefficients are extracted they fed to DWT then The extracted wavelet coefficients provide a compact representation that shows the energy distribution of the signal in time and frequency. In order to further reduce the dimensionality of the extracted feature vectors, statistics over the set of the wavelet coefficients are used. That way the statistical characteristics of the "texture" or the "music surface" of the piece can be represented. For example the distribution of energy in time and frequency for music is different from that of speech.

The classifier that used for classification phase is the ANN with a back propagation. It consists of three layers: input, hidden, and output. The input layer determined by the size of input feature vector which extracted by the feature extraction steps. The second layer (hidden) set to (800) nodes by experimentally. The last layer (output) determined by the number of persons that are identified. Gradient descent with adaptive back propagation algorithm is used. The ANN is iterated 500 times. For more generalization the ANN is trained for 5 times and the average result is taken.

RESULTS & DISCUSSION

The speech recognition system presented in this paper is implemented and work under MATLAB environment. Two databases are tested: PDAm which consist of speech of a number of persons (men, women). Each one record a different phrases in English language. Ten different persons are chosen for testing. Six samples from each one used for training and four speech for testing. The average recognition rate calculated due to the following equations :

$$p_k = \frac{No.\ of\ correct\ recognition\ test\ speech}{total\ number\ of\ tested\ speech} \qquad (9)$$

$$Average\ rate = \sum_{i=1}^{u} {}^{p_i}/_{u} \qquad (10)$$

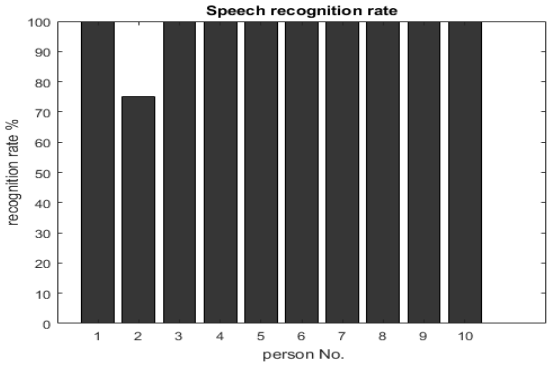

Fig 4: Average rate of tested PDAm datab [researcher]

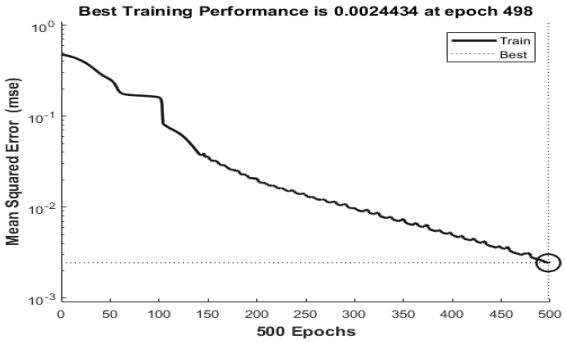

Fig 5: Training performance of tested PD[researcher]

The second tested database is a real test database The sound signal is recorded from six persons, two females and four males. Six samples for each person are chose for training and four speeches for testing. The recorder speech was colors and phrases in Arabic language. These signals pass through the system components and after pre-processing and feature extraction and the classification phase the recognition rate calculated due to equations (9 &10). Perfect recognition rate (100%) achieved in this database as shown below.

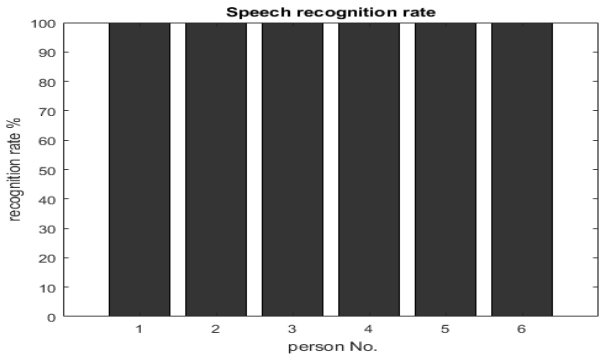

Fig 6: Average rate of tested real test[researcher]

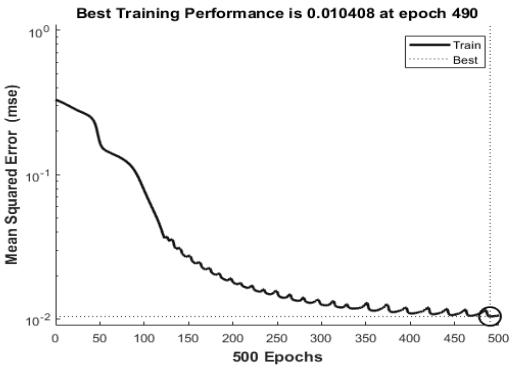

Fig 7: Training performance of tested real test database[researcher]

CONCLUSIONS AND FUTURE WORK

This paper presents an algorithm of speech recognition system in which:

1- Two databases are tested by this algorithm. The performance of the system affected by the type and size of the database.

2- DWT and MFCC in feature extraction better than using MFCC alone.

3- ANN used as a classifier give better performance than traditional classifier.

This suggested system may be improved in future if :

1- Proposing an open system for an unlimited number of speakers and for any number of words.
2- DWT replaced by Contourlet Transform (CT) which is a best tool in representation and capturing smooth transition and geometry of edges in two-dimensional data.
3- ANN can combinational with HMM (Hidden Markov Model) to give best results.

REFRENCES

[1] Mohammed, Z.Y, and Khidhir, A.S.M., 'Real-Time Arabic Speech Recognition' International Journal of Computer Applications , 81, (4) , 2013
[2] Hagino, T, Hiryu, S, Fujioka, S., Riquimaroux, H., and Watanabe, Y.: 'Adaptive SONAR sounds by echolocating bats', in Editor (Ed.)^(Eds.): 'Book Adaptive SONAR sounds by echolocating bats' (IEEE, 2007, edn.), pp. 647-651
[3] Rabiner, L.R. and Juang, B.-H, 'Fundamentals of speech recognition' (PTR Prentice Hall Englewood Cliffs, 1993.
[4] Perez-Meana, H.: 'Advances in Audio and Speech Signal Processing: Technologies and Applications: Technologies and Applications' (Igi Global, 2007. 2007)
[5] Oppenheim, A.V, and Schafer, R.W, 'Discrete-time signal processing' (Pearson Education, 2014.
[6] Tawfiq, Z.R.: 'Voice Based Authentication Using Artificial Neural Network ', 2012
[7] Nengheng, Z.: 'Speaker recognition using complementary information from vocal source and vocal tract', PhD Thesis, The Chinese University of Hong Kong, 2005
[8] George Tzanetakis, G.E., Perry Cook: 'Audio Analysis using the Discrete Wavelet Transform'
[9] Burrus, C.S., Gopinath, R.A., Guo, H., Odegard, J.E., and Selesnick, I.W.: 'Introduction to wavelets and wavelet transforms: a primer' (Prentice hall New Jersey, 1998. 1998)
[10] Antonini, M., Barlaud, M., Mathieu, P., and Daubechies, I.: 'Image coding using wavelet transform', IEEE Transactions on image processing, 1992, 1, (2), pp. 205-220