Similarity in Arabic Continuous Speech Recognition

Khaled Nayef Abu Nab, Prof. Madya DR. Norita Binti Md Norwawi, DR. Waidah Binti Ismail

Abstract — In this paper, Dynamic time warping (DTW) algorithm is proposed to find the similarity between Arabic words based on 39 coefficients of MFCC features. The audio samples collected from different nationalities (Jordan (J), Malaysia (M), Pakistan (P), Nigeria (N),Iran(I) and Yemen(Y)) of continuous Arabic words. Then the MFCC coefficients are extracted from the database and DTW is introduced by Sakoe Chiba and it has been utilized as features matching techniques and recognition, where the voice signal itself tends to have different temporal rate. Training and testing phases are done using 39 coefficients MFCC features. The experimental results are provided using MFCC and Delta Delta Coefficients (DDMFCC). It is recommended that higher recognition rates can be accomplished using (DDMFCC) with DTW which is valuable for different time varying speech Arabic recognition words.

Keywords— Audio Speech Recognition, Continuous Speech ,DTW, MFCC.

1 Introduction

Arabic language is one of the six languages of the united nations and one of the more relevant spoken languages in the world. Statistics indicates that the Arabic language is the first language (mother-tongue) of 206 million native speakers [1]. In spite of its importance, there is a little research on Automatic Arabic Speech Recognition (ASR). ASR is a technique that uses the speech in order to communicate with machine and it automatically recognizes the spoken word of humans depending on the speech signal information [2]. According to [3] the speech recognition system is used to identify the attendance of words in a background of noise. The beginning and end point of a speech should be identified for further processing words. The main issue of speech recognition is the same word that is spoken by different speakers depending on speaking tone, style, region, speech patterns and gender. Furthermore, the noise and variant of speech signals over time are problems that occur in voice speech recognition.

In this paper, there are two phases to find the similarity between two speakers which are the feature extraction and classification (similarity) a wide range of techniques exist for parameterization the acoustic signal for the speech recognition module, such as Mel-Frequency Cepstrum Coefficients (MFCC), Linear Prediction Coding (LPC), and others. The MFCC at time is the most known and best popular used for feature extraction[4].The MFCC technique is proposed in this paper.

2 The Database

Sample data collected from ten people (males& females) in which they come from different countries (Malaysia, Jordan, Pakistan, Nigeria, Yemen and Iran). Since each person is asked to utter the Arabic sentence

(“ بِسۡمِ ٱللَّهِ ٱلرَّحۡمَٰنِ ٱلرَّحِيمِ "). All speech signals recorded in same conditions with almost same time (3 - 5) seconds .All the target group is non native Arabic speakers except Jordanians and Yemen .

3 Methodology

The signal similarity process contains two phases the feature extraction and recognition. Extracting most relevant features play important role for increasing the similarity between signals. Since MFCC technique mimics hearing human perception which can’t grasp frequencies more than 1KHZ. In other words MFCC use two filters which are spaced linearly above 1000HZ and low frequency below 1000 HZ. Each pitch is represented by Mel Frequency scale to capture the most relevant phonetic pattern of the speech. The general process of MFCC and DWT technique are shown in Fig. 1.

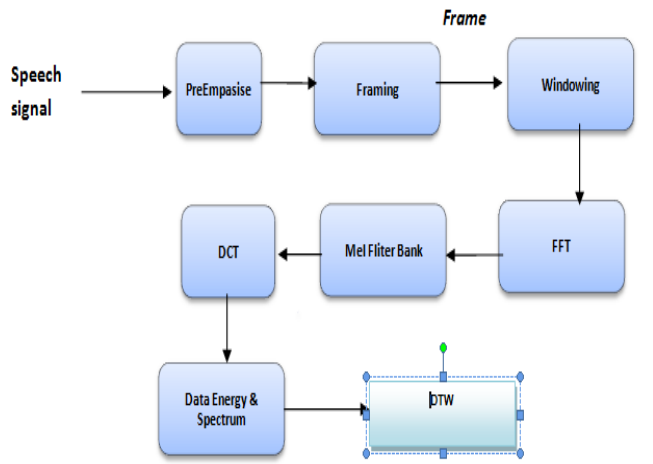

Fig.1. Similarity Block Diagram

3.1 Feature extraction

MFCC feature yields better for speech recognition than LPCC and LPC techniques according experiments in the previous study [5], [6]and[7]. In order to obtain the feature vector for recognition process (similarity) the following steps should be done :

1. Pre-Emphasis.
2. Hamming Windowing.
3. Power Spectrum by computing FFT.
4. Melfilter bank.
5. Discrete cosine transforms DCT.
6. MFCC.
7. Or (Delta MfCC).
8. Or (Delta Delta MFCC).

1. Pre-Emphasis

First order (FIR) filter is used to emphasize higher frequencies that increase energy of speech signal at higher frequency .

FIR filter equation is F [S] = X[S] -0.95 X[S].

2. Hamming Windowing

In automatic speech recognition, the most known window is the hamming window[8]. Hamming window is form of window through keeping the next block in feature extraction and integrates all the frequency lines that are closest to avoid anomalous discontinuities in the signal segment and distortion in the underlying Spectrum[9]. The Hamming window as shown in Fig.2 is represented in the Eq. (1). If the window is set as:

$W(s) , 0 \quad s \leq F-1$

$S [s] = X (s) * W (s)$ (1)

$\leq$

where F= Number of samples for each frame.

S[s] = The output signal.
X (s) = The input signal.
W(s) = Hamming window.
Then the windowing signal result is shown in vious Eq (1).

Frame

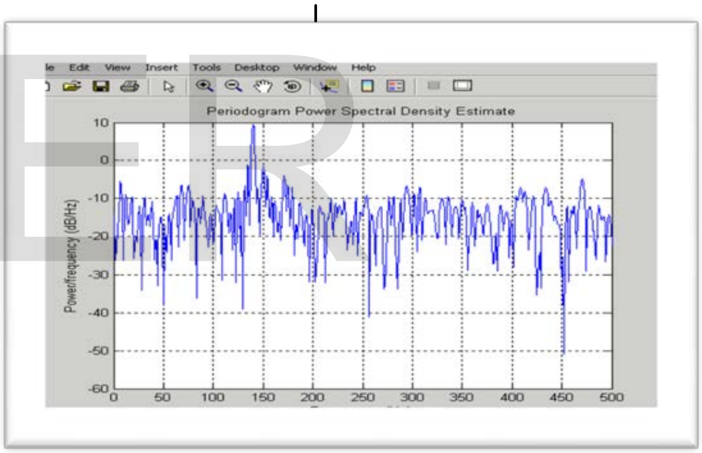
le Edit View Insert Tools Desktop Window Help
Periodogram Power Spectral Density Estimate
Power/frequency (dB/Hz)
10
0
-10
-20
-30
-40
-50
-60
0
50
100
150
200
250
300
350
400
450
500

Fig. 2. Hamming Window

3. Fast Fourier Transform (FFT)

The main function of fast fourier transform is to convert each frame after hamming window applied from time domain, which is usually defined as the convolution of the glottal pulse & the vocal tract impulse response, into the frequency domain[10].

The FFT represented as in Eq. (2)[11].

$$Y (w) = FFT [h (t) * X (t)] = H (w) * X (w) \qquad (2)$$

4. Mel-frequency

Human Hearing perception is not equally to all bands of frequencies. The perception is poor at higher frequency, when it is more than 1000 Hz. Thus the nature of human hearing for frequency is non-linear. Mel filter bank as shown in Fig.3 is series of filter of the triangular shape and decrease linearly to zero at centre frequency of two adjacent filters [12]. The sum of filtered spectral components is the output for each filter according to the Eq. (3) which is used to compute the Mel for frequency f in Hz:

$$F\ (Mel) = [2595 * \log 10[1+ f\ /700] \qquad (3)$$

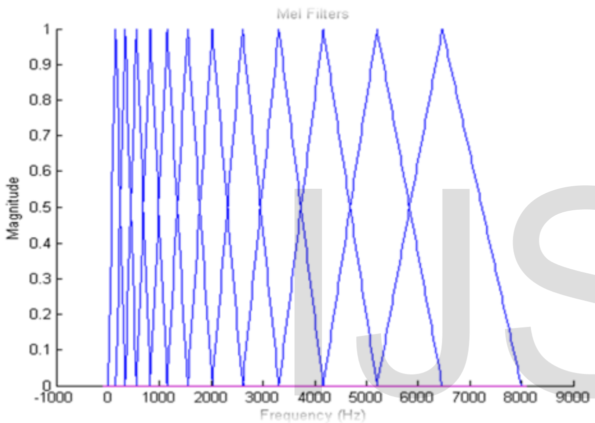

Fig. 3. Mel-Frequency Banks

5. Discrete cosine transform (DCT)

The Mel spectrum coefficients are real numbers; in order to convert the log Mel spectrum into time domain DCT is used. The result called Mel Frequency Cepstrum Coefficient (MFCC). The acoustic vectors (set of coefficient) result from conversion. Therefore, each input utterances is converted into a series of acoustic vector [3].

6. Delta energy and delta spectrum

The first two coefficients of MFCC are discarded, since they are changed between different utterances of the same speech. There is a need to add features related to the variation in cepstral features over time, acceleration and delta coefficients are found from the MFCC to increase the dimensions of coefficients for each frame, therefore an increase of accuracy occurs. The 39 delta delta coefficients are added in the experiment. The energy is represented using the following Eq. (4):

$$\text{Energy} = \sum_{n=0}^{N-1} x^2(n) \qquad .(4)$$

3.2 The Similarity

There are many techniques which are used for feature-matching in the speech recognition domain such as Dynamic Time Warping (DTW), Vector Quantization, and Hidden Markov Modeling (HMM). DTW technique is proposed for features matching.

3.3 Feature matching (DTW)

DTW technique is based on Dynamic Programming; DTW is used for measuring the similarity between two time sequences which are different in speed and/or time. Furthermore, this technique is used to find the optimal alignment between two times sequences, if one time sequence may be warped (non-linearly) by shrinking it along its time axis. This warping between two time sequences can then be used to find corresponding regions that can determine the similarity between two time sequences. To do alignment between two sequences using DTW, an N by M matrix is built, where the (ith, jth) element of the matrix contains the distance d (ai, bj) between the two points ai and bj. Then, the absolute distance between the values of two sequences is calculated by using the euclidean distance equation as in Eq. (5).

$$d\ (ai,bj) = (ai - bj)2 \qquad (5)$$

Each matrix element (i , j) identifies the alignment between two points (ai and bj). Then, the distance is accumulated and measured by "Eq. (6)".

$$DS\ (i, j) =$$

$$\text{Minimum}[DS(i-1, j-1),DS(i-1, j),DS(i, j -1)]+ d(i,) \quad (6)$$

4 Results

The techniques (Delta Delta MFCC & DTW) are applied on the input voice signals of different and same speakers. The results have been obtained by comparing different speakers are shown in the Table 1, 2, 3, 4, 5, 6, 8 and 9. The results by comparing same speakers are shown in a Table 10. The following example shows the similarity between two speakers using 39 coefficients of MFCC as shown in Fig 4.

Table-1: Comparison Between Different Speakers

1	YMali1 and JGirl	1.37310721623412E+03
2	YMali1 and PMALE	1.45402140009952E+03
3	YMali1 AND MMale1	1.70614435632676E+03
4	YMali1 and Ngirl	1.07473688593493E+03
5	YMali1 and JMale	1.14804760926416E+03
6	YMali1 and MGirl	1.59818648727200E+03
7	YMali1 and JGirl2	1.31846709389293E+03
8	YMali1 and MGirl2	1.25562492161377E+03

Table-2: Comparison Between Different speakers

1	JGIR1 and PMALE =	1.38581140552803E+03
2	JGIR1 and MMale1=	1.46377865570291E+03
3	JGIR1 and Ngirl=	1.27788768105951E+03
4	JGIR1 and JMale. =	1.42754540287978E+03
5	JGIR1 and MGirl	1.86473576726604E+03
6	JGIR1 and JGirl2	1.34503949375431E+03
7	JGIR1 and MGirl2	1.68968765518772E+03

1	Pmale and MMale1=	1.03725251369831E+03
2	Pmale and Ngirl=	1.09136653222776E+03
3	Pmale and JMale. =	1.21628731861672E+03
4	Pmale and MGirl	2.00137676896414E+03
5	Pmale and JGirl2	1.50604980595396E+03
6	Pmale and MGirl2	1.85541662055786E+03

Table-3: Comparison Between Different Speakers

Table-4 : Comparison Between Different Speakers

1	MMale1 and Ngirl=	1.20763085970045E+03
2	MMale1 and JMale. =	1.33524777444977E+03
3	MMale1 and MGirl1	2.05189981980785E+03
4	MMale1 and JGirl2	1.73513153807371E+03
5	MMale1 and MGirl2	1.85368035294308E+03

Table-5 : Comparison Between Different Speakers

1	Ngirl and JMale. =	8.86938039213884E+02
2	Ngirl and MGirl1	1.50045410921242E+03
3	Ngirl and JGirl2	1.21212147295591E+03
4	Ngirl and MGirl2	1.15069043564533E+03

Table-6: Comparison Between Different Speakers

1	Jmale and Mgirl 1	1.49942556856828E+03
2	Jmale and JGirl2	1.31695623807601E+03
3	Jmale and MGir2	1.24629240086720E+03

Table-7: Comparison Between Different Speakers

1	MGirl 1 and JGirl2	1.50671511389671E+03
2	MGirl 1 and MGirl2	1.26820106804908E+03

Table-8: Comparison Between Different Speakers

1	JGirl2 and MGirl2	1.36353874966478E+03

Table-9: Comparison Between Different Speakers

1	IMALE and IYMali1	0.0
2	Jgirl1and Jgirl1	0.0
3	PMale and PMale	0.0
4	Mmale and Mmale	0.0
5	Ngirl and Ngirl	0.0
6	JMale and JMale.	0.0
7	Mgirl1 and Mgirl1	0.0
8	Jgirl2 and Jgirl2	0.0
9	mgirl2 and mgirl2	0.0

Table-10: Comparison Between Same Speakers

	Speakers	Similarity Cost
1	Imale and Ymali1	1.29289545959301E+03
2	Imale and Jgirl1	1.32692424756723E+03
3	Imale and Pmale	1.00089701997836E+03
4	Imale and Mmale	1.27705246680712E+03
5	Imale and Ngirl	1.17737858376310E+03
6	Imale and Jmale.	1.23322326001636E+03
7	Imale and Mgirl1	1.89164412577178E+03
8	Imale and Jgirl2	1.53031538787592E+03
9	Imale and Mgirl2	1.78647055689049E+03

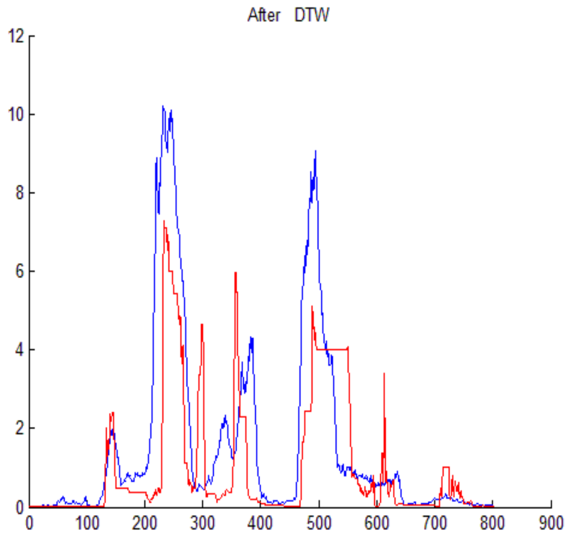

Fig .4. The similarity between imale and ymale by using 39-cofficients of MFCC.

5 CONCLUSION

This paper has examined two phases used for audio recognition system which are important to improve the performance and accuracy in speech recognition. The first phase provides the detailed information to extract Delta Delta MFCC coefficients from the audio signals. The second phase presents the similarity between two speakers using DTW algorithm. The MFCC with DTW techniques have been applied on the same speakers signals as well as they have been applied on different speakers speech signals. It has been concluded that if the speech signal similarity cost equals zero, it indicates the same word and signal. In contrast, if the speech signal similarity cost does not equal zero, it indicates different speakers .

7 REFERENCES

[1] R. G. Gordon Jr, “Ethnologue, languages of the world,” http//www. Ethnol. com/, 2005.

[2] L. Muda, M. Begam, and I. Elamvazuthi, “Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques,” arXiv Prepr. arXiv1003.4083, 2010.

[3] N. T. Hai, N. Van Thuyen, T. T. Mai, and V. Van Toi, “MFCC-DTW algorithm for speech recognition in an intelligent wheelchair,” in 5th International Conference on Biomedical Engineering in Vietnam, 2015, pp. 417–421.

[4] S. D. Dhingra, G. Nijhawan, and P. Pandit, “Isolated speech recognition using MFCC and DTW,” Int. J. Adv. Res. Electr. Electron. Instrum. Eng., vol. 2, no. 8, pp. 4085–4092, 2013.

[5] S. Pathak and A. Kulkarni, “Recognizing emotions from speech,” in Electronics Computer Technology (ICECT), 2011 3rd International Conference on, 2011, vol. 4, pp. 107–109.

[6] D. Ververidis and C. Kotropoulos, “Emotional speech recognition: Resources, features, and methods,” Speech Commun., vol. 48, no. 9, pp. 1162–1181, 2006.

[7] Z. Xiao, E. Dellandrea, W. Dou, and L. Chen, “Features extraction and selection for emotional speech classification,” in Advanced Video and Signal Based Surveillance, 2005. AVSS 2005. IEEE Conference on, 2005, pp. 411–416.

[8] J. R. Deller Jr, J. G. Proakis, and J. H. Hansen, Discrete time processing of speech signals. Prentice Hall PTR, 1993.

[9] B. Gold, N. Morgan, and D. Ellis, Speech and audio signal processing: processing and perception of speech and music. John Wiley & Sons, 2011.

[10] Z. Ali, A. W. Abbas, T. M. Thasleema, B. Uddin, T. Raaz, and S. A. R. Abid, “Database development and automatic speech recognition of isolated Pashto spoken digits using MFCC and K-NN,” Int. J. Speech Technol., vol. 18, no. 2, pp. 271–275, 2015.

[11] A. BALA, “Voice Command Recognition System Based on Mfcc and Dtw,” Int. J. Eng. Sci. Technol., vol. 2 (12), no. 9, pp. 7335–7342, 2010.

[12] S. M. Azam, Z. A. Mansoor, M. S. Mughal, and S. Mohsin, “Urdu spoken digits recognition using classified MFCC and backpropgation neural network,” in Computer Graphics, Imaging and Visualisation, 2007. CGIV’07, 2007, pp. 414–418.