Research Article

Preprocessing Signal for Speech Emotion Recognition

Bashar M. Nema, Ahmed A. Abdul-Kareem
Department of Computer Science, College of Science, Mustansiriyah University, IRAQ
*Correspondent email: bashar_sh77@uomustansiriyah.edu.iq

ArticleInfo

Received
10 Apr. 2017

Accepted
17 Oct. 2017

Abstract

In this paper the preprocessing signal for speech emotion recognition was introduced. The literature review on speech emotion recognition was presented. The discrimination between speech and music files was performed depend on a comparative between more than one statistical indicator such as mean, standard deviation, energy and silence interval. The preprocessing include silence removal, pre-emphasis, normalization and windowing so it is an important phase to get pure signal which is used in the next stage (feature extraction). The wave files (male, female) and the music file which are used in this paper have sample rate 48000; bit resolution 16 bit and mono channel. The wave files of this paper are taken from the Berlin dataset and RAVDESS dataset.

Keywords: Speech, Features extraction, Emotion, Vocal, LPC, MFCC.

الخلاصة

في هذا العمل تم تقديم المعالجة المسبقة للاشارة الكلامية للتعرف على العاطفة. تم عرض نتائج لدراسات سابقة للتعرف على عاطفة الكلام. عملية التمييز بين ملفات الكلام والملفات الموسيقية تمت بالاعتماد على المقارنة بين أكثر من مؤشر إحصائي واحد مثل المتوسط والانحراف المعياري والطاقة والفترة الصامتة. وتشمل المعالجة المسبقة إزالة فترات الصمت، والتركيز المسبق، والتطبيع، والنوافذ، لذلك فهي مرحلة هامة للحصول على إشارة نقية تستخدم في المرحلة التالية (استخراج الميزة). ملفات الموجة (ذكور وإناث) وملف الموسيقى التي استخدمت في هذه الورقة لها معدل عينة 48000؛ 16 بت وقناة أحادية. تم أخذ العينات الصوتية في هذه العمل من قاعدة البيانات برلين وقاعدة البيانات رافديس.

Introduction

Speech is one of the oldest human tools which are used for interaction among each other. Therefore it is one of the most natural ways to interact with the computers as well [1]. A typical speech signal consists of two main parts: one carries the speech information, and the other includes silent or noise sections. The verbal (informative) part of speech can be further divided into three categories: (a) The voice speech (b) unvoiced speech (c) silence. Voiced speech consists mainly of vowel sound. It is produced by forcing air through the glottis, proper adjustment of the tension of the vocal cords results in opening and closing of the cords, and a production of almost periodic pulses of air. These pulses excite the vocal tract. Psychoacoustics experiments show that this part holds most of the information of the speech and thus holds the keys for characterizing a speaker. *Unvoiced speech* parts are generated by forcing air through a constriction formed at a point in the vocal tract (usually toward the mouth end), thus producing turbulence. The last category is *Silence*, when there is no vibration of the vocal cords after the air is discharged from the lungs [2].

Speech recognition is the ability of a machine or program to identify words and phrases in spoken language and determine the emotions of the speaker such as normal, anger, happiness and sadness.

The founder of modern philosophy "René Descartes", identified six simple and primitive emotions wonder, love, hatred, desire, joy, and sadness. Other philosophers identified categories of emotions which include composed of some of these six or species of them [3].

As a result of the experiences and observations experienced by man over the centuries it became easy for him to distinguish emotions,

Copyright © 2017 Authors and Al-Mustansiriyah Journal of Science. This work is licensed under a Creative Commons Attribution-NonCommercial 4. 0 International Licenses.

such as when a person is angry, his tone raises, and his expression becomes stern. At the same time when a person is happy, he speaks in a musical tone, thus there is a look of glee on his face and the content of his speech is rather pleasant. Based on these observations, a person can quickly identify the state of the speaker whether he is happy, sad, energy or others states [4].

In this work, the preprocessing is applying to get a pure signal which is used to extract the features. After the features of this signal are extracted then it is used to distinguish the emotion. There are several methods for feature extraction of voice signal such as Liner Predictive Coding (LPC), Hidden Markov Model (HMM), Artificial Neural Network (ANN), and Mel-Frequency Cepstral Coefficient (MFCC) [5].

Theoretical Background

Literature Review

Many articles have been accomplished in field of speech emotion recognition. Some of them are present here briefly:

In [6], present an effective method for better and improvised performance which is Hidden Markov Model (HMM), Gaussian Mixture Model (GMM) and Support Vector Machine (SVM). There are two phases: training and testing. In the first step of training phase, the feature is extracted and the MFCC feature vectors are used for training the emotion recognition models (GMMs or HMMs). In the testing phase feature vectors are given .The SVM, GMM and HMM classifiers are used to evaluate and analyzed these experiments. In the last, they conclude that SVM is providing better results and the appropriate classification rate (86.6%) when compared with GMM and HMM.

In [7], used Mel Frequency Cepstral Coefficients (MFCC) for feature extraction from the speech signal and Support Vector Machine (SVM) for recognition of emotional states .English datasets with SVM Kernel functions are used for analysis of emotions. For using this analysis the machine is designed for detecting emotions in real time speech .The experimental results gives that using MFCC is reduce the frequency information of the speech signal into small number of coefficients which is easy and fast to compute, but the computation of SVM is very fast, effective and its accuracy is better in comparison to the other techniques.

In [8], proposed an effective method which is Mel Frequency Cepstral Coefficients (MFCC) and Vector Quantization (VQ). Accuracy obtained by using VQLBG algorithm was 72.5%. It was observed that take voice sample using high quality audio devices in a noise free environment can be improved the performance factor but degrades the computational efficiency. Hence an economical trade-off between code vectors and number of computation is required for optimized performance of VQLBG algorithm .Then they used another method for speaker identification known as GMM. Accuracy of 90.9% was obtained by using GMM algorithm for same data which clearly indicates its high efficiency.

In [9], present develops for speaker identification system. The system is developed into two phases: training and recognition. Each phase consists of multi stages. Speech files are used in wave form, which are achieved either directly using microphone or remotely by sending a recorded (.wav) file over network. In preprocessing of files used silence removal and framing and windowing the identification rate achieved is (94%) for 52 speakers these result is fulfillment by using MFCC and VQ.

Speech signal preprocessing

Before the extraction of the features of the signal, this signal is manipulate by using preprocessing. Preprocessing is mainly includes:

Silence removal: The speech signal usually include many parts of silence. The silence signal is not important because it is not contain information. There are several methods to remove these parts such as zero crossing rate (ZCR) and short time energy (STE). Zero-crossing rate is a measure of number of times in a given time interval such that the amplitude of the speech signals passes through a value of zero. Short time energy is a measure of energy of signal in time interval by using the following Equation [10]:

$$E_n = \sum_{m=-\infty}^{\infty} [x(m)w(n-m)]^2 \qquad (1)$$

Pre-emphasis: The pre-emphasis of the speech signal is the most important steps of preprocessing at high frequency. It's used to get comparable amplitude for all formant. To fulfill the assignment, the speech signal is passed through a high-pass filter (FIR). In simple form it can implement by using the following Equation [11] [12]:

$$x'(n) = x(n) - \propto x(n-1) \qquad (2)$$

α being the pre-emphasis parameter typically having a value between (0.9) and 1.

Normalization: It is a strategy for modifying the volume of sound to a standard level. Normalization is used the signal sequence divided by highest value of the signal to ensure that each sentence has a comparable volume level.

Windowing: It is done by using the following Equations [13] [14]:

$$y_1(n) = x_1(n)w(n), \quad 0 \le n \le N-1 \quad N-1 \qquad (3)$$

The most popular window is Hamming window. It has effect of smoothing the edges and reducing the effect of side lobe. The function of this window is defined as follows:

$$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \le n \le N-1 \qquad (4)$$

The window time is usually 25ms and overlaps every 10ms

Statistical Indicators

In the most recent decade the issue of discriminating between speech and music has gained interest. The solution is still open. The outcomes rely on upon many factors like corpus, features and test methods.

The distinction between speech wave and music wave can be performed using many statistical indicators such as [15] [16].

- **The Mean:** is referring to a central value, a discrete set of numbers specifically, the sum of the values divided by the number of values.
- **The Standard Deviation:** The standard deviation (SD additionally spoke to by the sigma σ or the Latin letter s) is a measure that is utilized to evaluate the measure of variety or scattering of a set of data values. A low standard deviation demonstrates that the information guides incline toward be near the mean (likewise called the normal esteem) of the set, while an elevated expectation deviation shows that the information focuses are spread out over a more extensive scope of qualities
- **The Energy:** Is the most basic feature of the speech signal. It is produced when sound waves move outward from a vibrating object or sound source.

Proposed System

In our proposed system, the wave signal will be distinguish if it is music or speech. Then the music wave is neglected when it is appear and work on the speech wave. In the phase of discrimination a number of files will be used for example, five files for speech under the name (wf1, wf2, wf3, wf4, wf5) and five files for the music under the name (wf6, wf7,wf8, wf9, wf10). The processes of discrimination based on several statistical indicators which are energy, standard deviation and mean. Figure 1 shows the proposed emotion recognition system.

Copyright © 2017 Authors and Al-Mustansiriyah Journal of Science. This work is licensed under a Creative Commons Attribution-NonCommercial 4. 0 International Licenses.

Discriminating between Speech and Music signal

It is easy to note the difference, in the shape, between speech wave and music wave by looking at Figures 2 and 3.

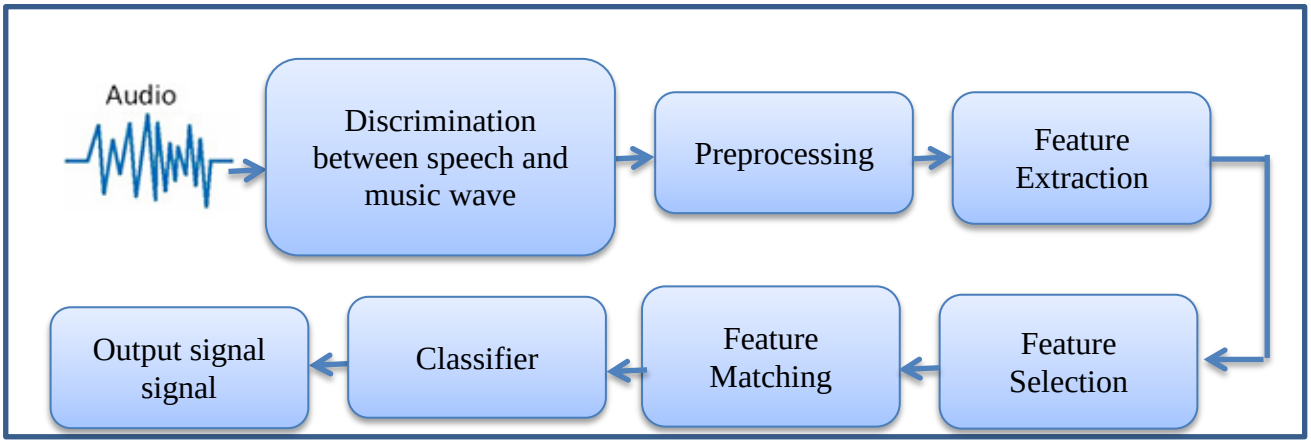


Figure 1: proposed Emotion Recognition System.

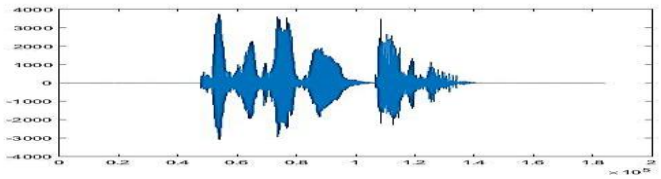
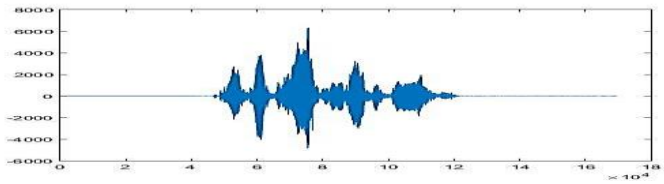
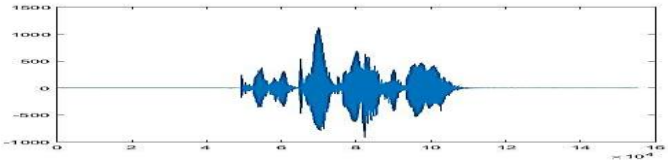
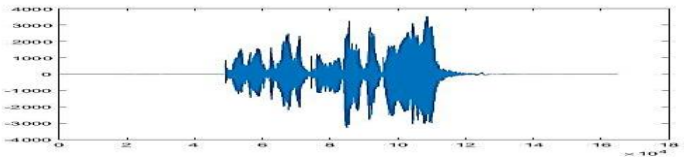
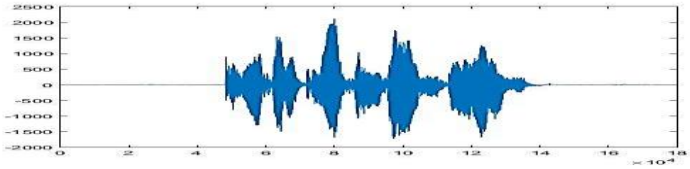
File name	Waveform	Type
Wf1.wav		**Speech**
Wf2.wav		**Speech**
Wf3.wav		**Speech**
Wf4.wav		**Speech**
Wf5.wav		**Speech**

Figure 2: The original form of speech wave

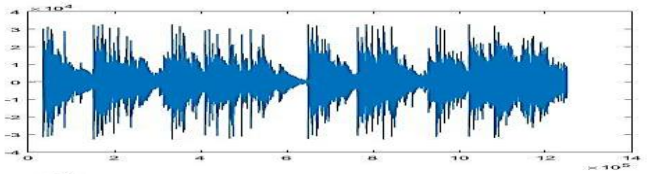
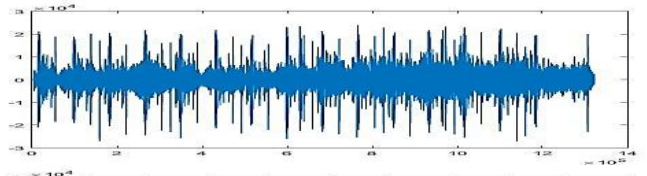
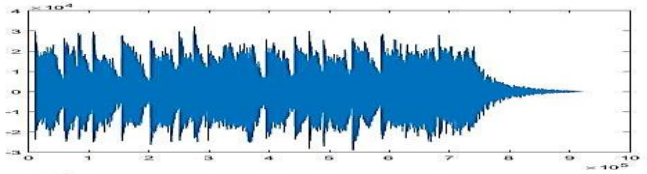
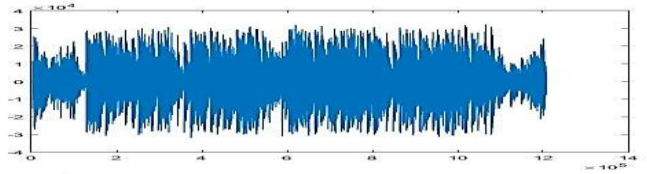
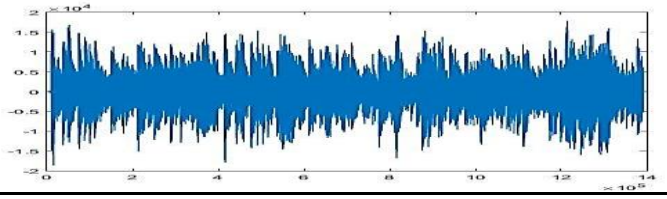
File name	Waveform	type
Wf6.wav		**Music**
Wf7.wav		**Music**
Wf8.wav		**Music**
Wf9.wav		**Music**
Wf10.wav		**Music**

Figure 3: The original form of music wave.

The following Tables 1 and 2 explain the results of using the statistical indicators to discriminating between the speech wave and music wave**:**

Table 1: Statistical indicators result for speech files.

Wave file name	Mean	Standard Deviation	Energy	Silence interval	Proposed System Result
Wf1	0.0043	0.1150	1272	50105	Speech
Wf2	8.3333	0.0714	490	48585	Speech
Wf3	-0.0016	0.0921	815	58196	Speech
Wf4	-1.5625	0.0948	863	50915	Speech
Wf5	-7.7083	0.0686	452	52000	Speech
Average	3.7500	0.0884	778.4	51960.2	

Table 2: Statistical indicators result for music files

Wave file name	Mean	Standard Deviation	Energy	Silence interval	Proposed System Result
Wf6	0.0015	0.1299	1619	5207	Music
Wf7	-1.6667	0.1046	1050	3116	Music
Wf8	0.0083	0.1701	2785	10292	Music
Wf9	-9.8958	0.1202	1387	1867	Music
Wf10	0.0077	0.1526	2240	5668	Music
Average	-2.309	0.1355	18162.2	5230	

Copyright © 2017 Authors and Al-Mustansiriyah Journal of Science. This work is licensed under a Creative Commons Attribution-NonCommercial 4. 0 International Licenses.

The statistical indicators of above tables show that the music signal has higher energy values than the speech signal and when this value is increases, the mean value and standard deviation value are also increased and vice versa. Also, the silence intervals in speech files are greater than silence intervals in music files. Based on these results, the distinctions between music and speech waves were done.

The process of selecting the statistical indicators mentioned above has helped us in the process of deciding on the files if they are music or speech. The following Figures 4 and 5 explain the statistical indicators form for each file.

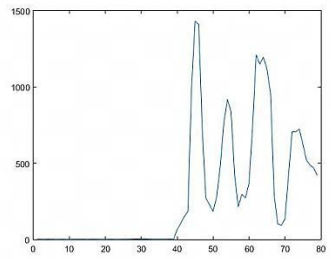
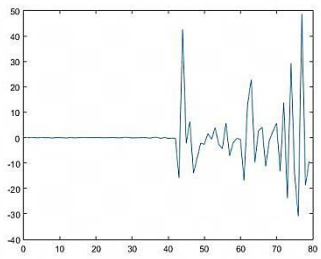
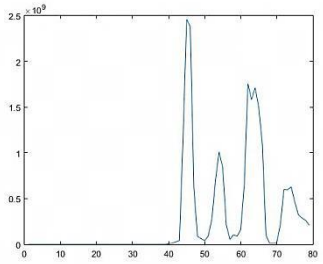
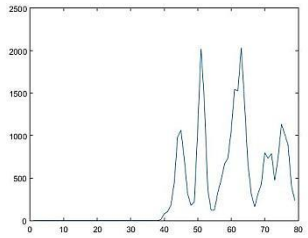
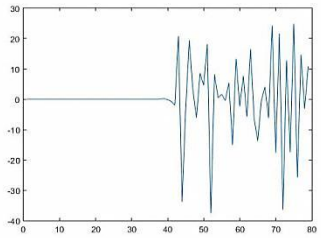
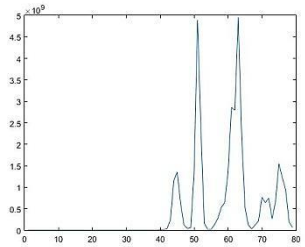
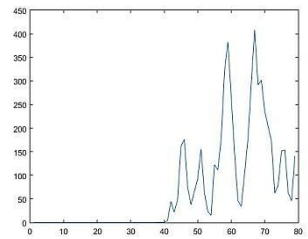
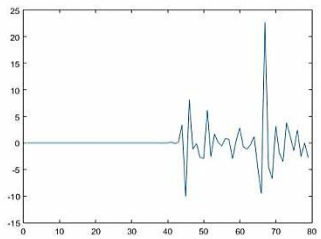
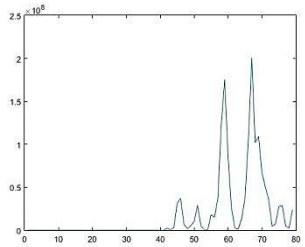
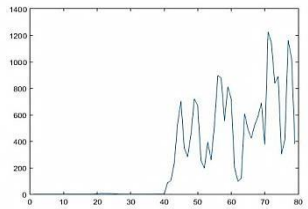
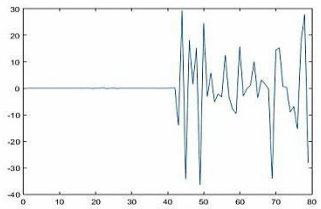
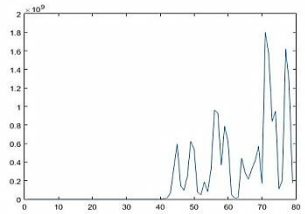
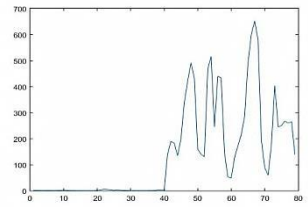
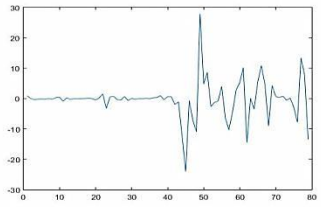
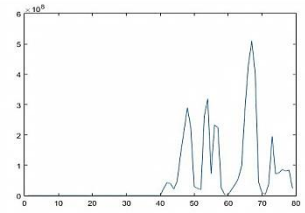
Speech wave name	Mean	Standard Deviation	Energy
Wf1.wav			
Wf2.wav			
Wf3.wav			
Wf4.wav			
Wf5.wav			

Figure 4: Statistical indicators form of speech files.

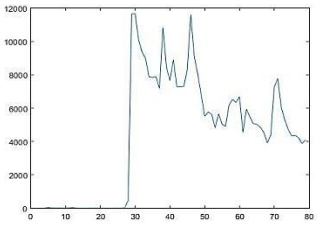
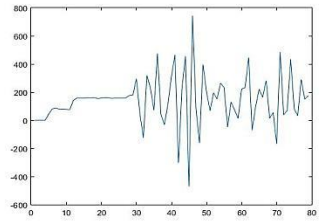
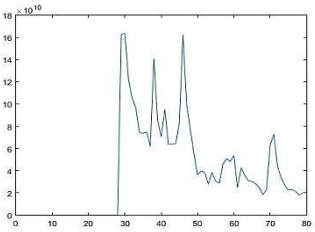
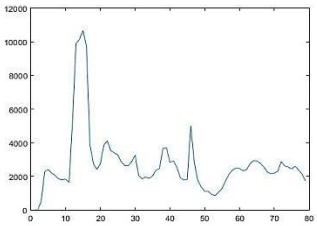
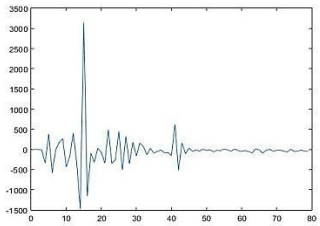
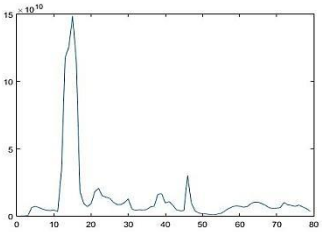
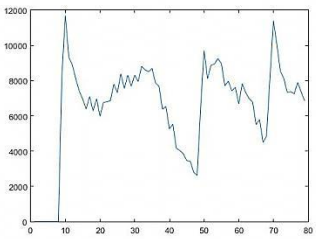
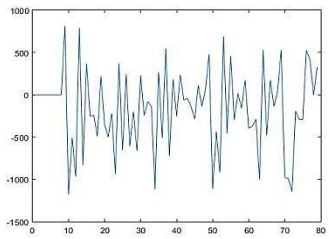
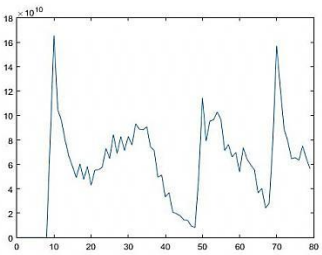
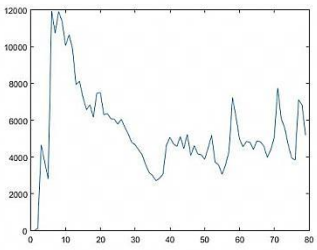
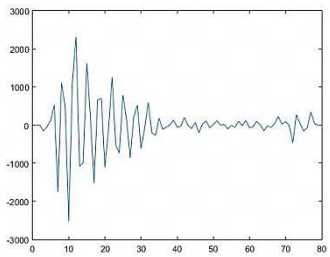
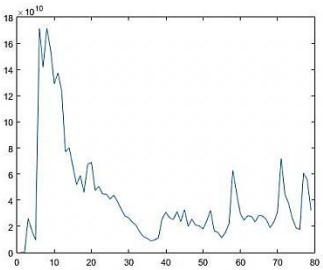
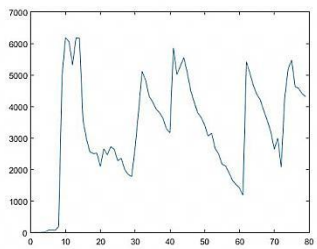
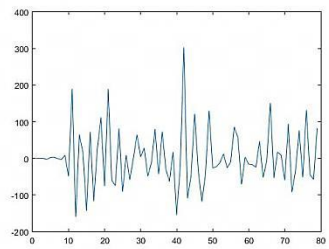
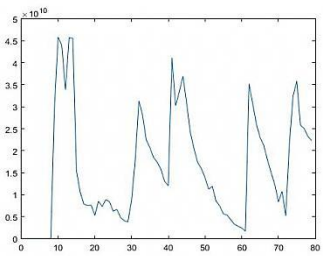
Music wave name	Mean	Standard Deviation	Energy
Wf6.wav			
Wf7.wav			
Wf8.wav			
Wf9.wav			
Wf10.wav			

Figure 5: Statistical indicators form of music files.

Preprocessing

For the proposed emotion recognition system, the preprocessing methods are used in the first stage to enhance the feature extraction process. This method (preprocessing) includes:

a. Silence removal: It is the first preprocess applied to detect and delete silence frame from speech wave. Silence removal is fulfillment by compute the maximum value for each frame and compared this value with a certain threshold value. If maximum value for the one frame is larger than threshold, then added as speech frame, otherwise removed it as silence frame.

b. Pre-emphasis: It has effect of reinforce high frequencies. In this work, the pre-

Copyright © 2017 Authors and Al-Mustansiriyah Journal of Science. This work is licensed under a Creative Commons Attribution-NonCommercial 4. 0 International Licenses.

emphasis will be computed by using the equation 2 with α equal to 0.95.

c. Normalization: It is third stage of preprocessing. In this stage, calculate the maximum value of signal and then divided the whole signal sequence by this maximum.

d. Windowing: For reducing stopping part from speech wave it is useful to use the windowing function. It is calculate by using equation 3.One of the types of the windowing function is Hamming window which is used in this paper and it is defined by equation 4.

(a) (b)

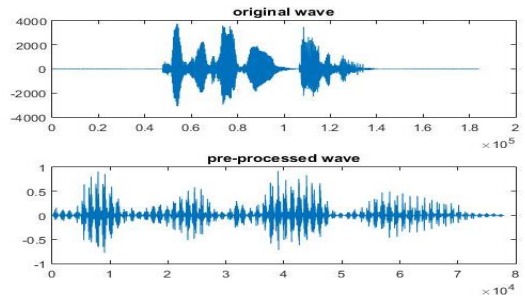


Figure 6: wf1 speech waveform before and after preprocessing.

(a) (b)

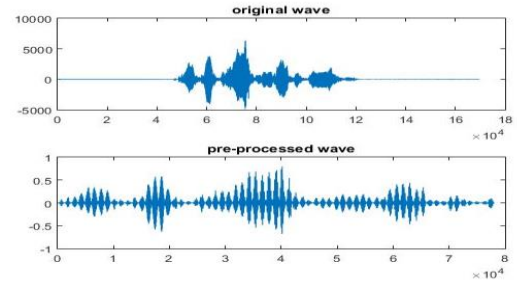


Figure 7: wf2 speech waveform before and after preprocessing.

(a) (b)

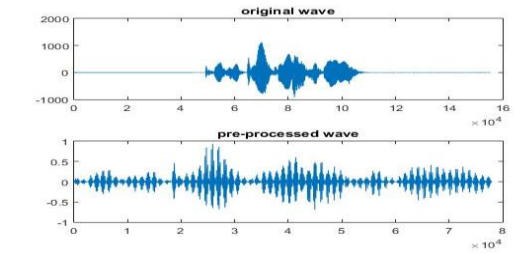


Figure 8: wf3 speech waveform before and after preprocessing.

(a) (b)

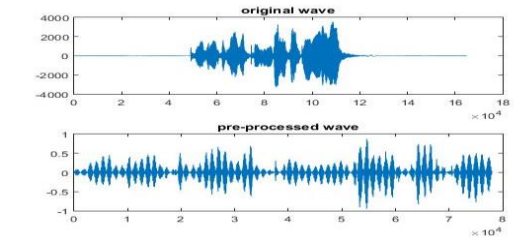


Figure 9: wf4 speech waveform before and after preprocessing.

(a) (b)

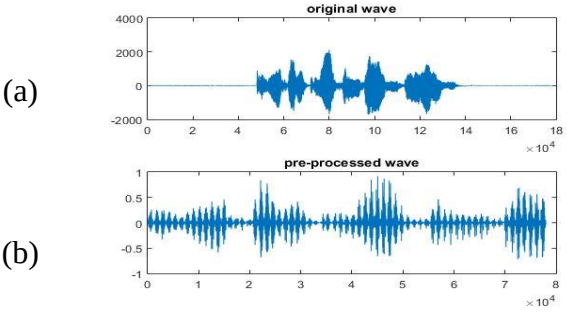


Figure 10: wf5 speech waveform before and after preprocessing.

Conclusions

The working in specialized scientific research in speech is considered a major challenge and was evident in most branches of audio work. We find this more complex and difficult in the process of access to the characteristics of sound in terms of voice and other. In this paper, the most important thing is the access to the audio characteristics which have the ability to distinguish emotions, so the first phase is the preprocessing. After the process of the preprocessing on the signal we noticed the following important points:

In order to implement speech emotion recognition system, it is necessary to perform the preprocessing signal according to the previously mentioned steps without providing or delaying.

Noticed that the new files part (b) of the Figures (6, 7, 8, 9, and 10) after preprocessing is the slimmer comparative with the original files in part (a). In spite of the new files are slimmer than original files, but it is still retain important information. In the last part of the preprocessing the pure signal is obtained and it is used in the next step (feature extraction).

References

[1] M. S. Unluturk, K. Oguz, and C. Atay, "A Comparison of Neural Networks for Real-Time Emotion Recognition from Speech Signals."

[2] A. S. Utane and S. Nalbalwar, "Emotion recognition through Speech," *International Journal of Applied Information Systems (IJAIS)*, pp. 5-8, 2013.

[3] S. P. Robbins, *Organizational Behavior, 13/E*: Pearson Education India, 2009.

[4] K. S. Rao, T. P. Kumar, K. Anusha, B. Leela, I. Bhavana, and S. Gowtham, "Emotion recognition from speech," *International Journal of Computer Science and Information Technologies,* vol. 3, pp. 3603-3607, 2012.

[5] D.-I. Kim and B.-C. Kim, "Speech recognition using hidden markov models in embedded platform," *Indian Journal of Science and Technology,* vol. 8, 2015.

[6] O. A. B. Sujatha, "Speech Emotion Recognition Using HMM, GMM and SVM Models," *International Journal of Professional Engineering Studies,* vol. 4, pp. 311-318, 2016.

[7] A. C. S. R. D. Shah, "Speech Emotion Recognition Based on SVM Using MATLAB," *International Journal of Innovative Research in Computer and Communication Engineering,* vol. 4, pp. 2916-2921, 2016.

[8] J. Martinez, H. Perez, E. Escamilla, and M. M. Suzuki, "Speaker recognition using Mel frequency Cepstral Coefficients (MFCC) and Vector quantization (VQ) techniques," in *Electrical Communications and Computers (CONIELECOMP), 2012 22nd International Conference on*, 2012, pp. 248-251.

[9] Y. A. Mohammed, "Speaker Identification Using MFCC and VQ," Master, Computer Science Department, Mustansiriyah University, College of Science, 2016.

[10] T. F. Quatieri, *Discrete-time speech signal processing: principles and practice*: Pearson Education India, 2006.

[11] D. O'Shaughnessy, "Interacting with computers by voice: automatic speech recognition and synthesis," *Proceedings of the IEEE,* vol. 91, pp. 1272-1305, 2003.

[12] A. Peinado and J. Segura, *Speech recognition over digital channels: Robustness and Standards*: John Wiley & Sons, 2006.

[13] H. Beigi, *Fundamentals of speaker recognition*: Springer Science and Business Media, 2011.

[14] V. Radha, C. Vimala, and M. Krishnaveni, "Isolated word recognition system using Back propagation network for Tamil Spoken Language," *Trends in Computer Science, Engineering and Information Technology,* pp. 254-264, 2011.

[15] S. Deviant, *The Practically Cheating Statistics Handbook*: Lulu. com, 2011.

[16] B. M. Nema, "Hybrid Secure Conversation System," *Research Journal of Applied Sciences,* vol. 11, pp. 1039-1044, 2016.

Copyright © 2017 Authors and Al-Mustansiriyah Journal of Science. This work is licensed under a Creative Commons Attribution-NonCommercial 4. 0 International Licenses.