Emotion Recognition in Speech Using Neural Network

Fatin B. Sofia, Sahar K. Ahmed &Abdul-basit K. Faeq
Mosul University

Received **Accepted**
08/01/2007 **09/05/2007**

الخلاصة

نمت فكرة تمييز الحالة النفسية للمتكلم بعد تقنيات تمييز الصوت . ان الهدف الأساسي لهذه الدراسة هو تسهيل عمل واجهة كلامية بين الإنسان والكومبيوتر والتي تساعد على التوسع بدراسة الأصوات بالحاسوب.

هذه الدراسة تمت باستخدام أربع متكلمين ينطقون جمل وبحالات نفسية مختلفة والمتكلمين هم من فئة الممثلين الجيدين وقد استخدمت الحالات النفسية التالية: سعادة، غضب، حزن، خوف، ووضع طبيعي (بدون أي عاطفية) . تم استخلاص الخواص المهمة من الإشارة الكلامية في الزمن الحقيقي ومتجه الخواص تضمن 12 برامتر من معامل الترميز الخطي وبرامتر يمثل طبقة الصوت وآخر للقوة و 3 برامترات للترددات الأساسية للصوت . وأما عملية التمييز فقد تمت باستخدام تقنيات الشبكات العصبية الذكائية (شبكة الانتشار العكسي) . تم استخدام عشرة شبكات منفصلة وكل منها تحقق تمييز لانتماء حالة معينة أو لا لأنه تم اخذ جميع الاحتمالات للحالات النفسية الخمسة المذكورة وبهذه الطريقة حصلنا على نسبة تمييز عالية .

هذا النظام المصمم تمكن من تمييز كل الحالات بنسبة 75% ماعدا حالة الخوف كانت نسبة تمييزها 50% .

ABSTRACT

Emotion recognition in speech studies has grown after the growing of speech recognition technologies. The aim of such studies is to make language interfaces in human-computer interaction applications more wide in use and to make it efficient. And it may help the studiers of the human sound areas.

This study deals with four spoken sentences of sixteen short utterance expressing five emotions: a happiness, anger, sadness, fear, and normal (unemotional) state. The feature extraction techniques are used to capture the most important information of the signal, this process is applied in the time domain. The extracted feature vector contain 12 LPC (linear predictive coding) parameters, the pitch, the power, and the three first formant frequencies. While in the second part of the system (the emotion recognizer) the artificial neural network technology is used, the

designing network is with feedforward backpropagation. This design used ten separated nets. each one make a decision whether the spoken sentence is belong to one of only tow emotions. So the computational time is increased with this system, while such a system has better performance with respect to other systems.

The designed system could recognize all emotions with ratio 75% except the fear one which recognized in ratio of 50%.

1. Introduction

The importance of automatically recognizing emotions from human speech has grown with the increasing role of spoken language interfaces in human-computer interaction applications [1].

Emotion recognition in speech has many potential application. One current use is in interactive movies [5]. Emotion understanding can allow characters in an interactive movie to react to the emotions conveyed by a participant's utterances.

Another possible use of emotion recognition is as an aid to speech understanding. It is possible that by recognizing the emotion in speech, one be able to 'subtract' them form the speech and improve the performance of speech understanding system.

Finally, emotion recognition for speech could serve as a kind of 'emotional translator'. Emotions are often portrayed differently in different cultures and languages. [6] for example, one type of intonation which indicates admiration in Kurdish can indicate disbelief in English.

2. Emotion Recognition Systems

The emotion recognition system consists of two main blocks as shown in figure (1): feature extraction, and emotion recognizer. A typical interactive spoken dialog system comprises elements enabling dialog management between human and machines, and the emotion recognition system can help the machine manage the interaction in a natural and effective manner [1]. The focus of this work is in the design of an emotion recognizer that will work in conjunction with such a system. In the next sections the practically steps will be discussed.

Speech data used for testing emotion recognition can be grouped under three categories depending on the way the speech signal was captured. The first method uses actors to record utterances, where each utterance is spoken with multiple feigned emotions. The actors are usually given the time to imagine themselves into a specific situation before speaking. The second method called Wizard-Of-Oz (WOZ) uses a program that interacts with the actor and drives him into a specific emotion situation and then records his responses. The third method, which is hard to obtain, is actual real-world recording of utterances that

express emotions[7]. The method that has been used in this work is the first method which explained in the following figure.

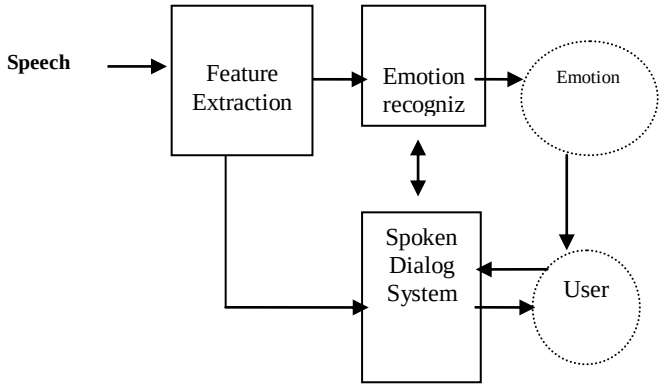

Figure (1): Emotion recognition system

3. Feature Extraction

The speech signal must be divided into frames in order to obtain more stability. So the speech signal is divided in this work in to frames of length (512) value for each frame. For each pair of frames the second frame is started in the middle of the first one, that is to avoiding the loosing of general features near the end of the first frame and the start of the second one[4]. Figure (2) show this step upon the previous speech signal.

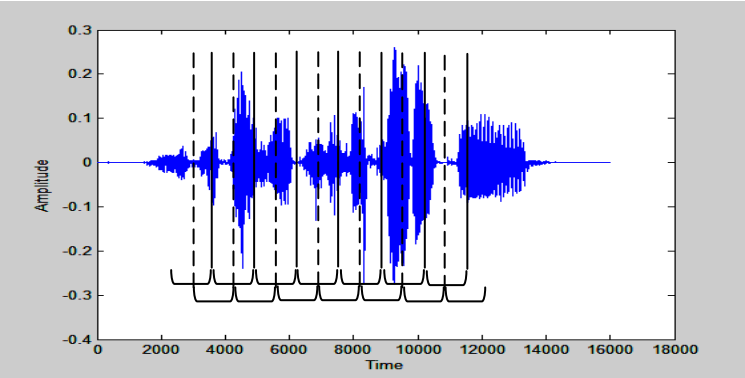

Figure (2): The sampling process upon the spoken Kurdish language sentence signal

The main aim of applying features extraction algorithms and representing the speech signal in a robust way is to minimize amount of input data for processing through extracting minimum numbers of features, and to distill the information into a more concentrated and manageable form. This can be done by avoiding redundant data and keeping the required information for the next processing operation. [3]

feature extraction algorithms decreases the difficulty of the processing. As a result greater emphasis has been done by many researchers to develop efficient feature extraction algorithms that can be used in different applications.[8]

The feature extraction in this dissertation based on time domain algorithms and the features that used in the designed system are (LPC, pitch, power, and the formant).

3.1 LPC Parameters Calculation:

The LPC, functions find the coefficients of a digital rational transfer function that approximates a given time-domain impulse response [2]. The autocorrelation is the used method in the time domain to estimate the most features. Calculated LPC parameters can be done with applying the autocorrelation function and then by Durbin's method the parameters is calculated. And so they form the first (12* number of frames) feature vector parameters.

3.2 Power Calculated:

The autocorrelation function at ($m=0$) gives the power of the speech signal. Using a created function (power feature) that include a simple MATLAB code which calculate the summation of the squire of value of the signal amplitude, the power for each frame are computed and added to the feature vector. And so the power can be estimated from the equation :

$$\mathbf{E}_n=\sum_{m=1}^{M} x_l^2(m) \qquad \text{……..(1)}$$

Where E_n represent power and $x_i(m)$ represent samples of speech signal .

3.3 The Formant Frequency Calculation

By sorting the LP filter coefficient the formant frequencies can be obtained. The process of calculating the formant frequencies is started with calculating the root of the polynomial function that it's coefficient are number of LPC parameters equal to 2+sample rate/1000. Choosing this number of parameters is depending on the rule of thumb for formant estimation. In the next step we ignoring all the negative imaginary and real part of these roots. And then sort this roots after converting it to Hertz by multiplying the roots with the sample rate of the signal. the sorted roots gives the formant frequencies (Fi), started with F1,F2,… and so on.

4. Neural Network Part

The used network is the back propagation algorithm- multilayer feed forward [9]. In this work the recognizer emotion system is treated with five emotions (natural, happy, anger, sad and fear). And so the input to the network is a feature vector such that each element of the feature vector is a vector of five elements.

For the previous example the input layer has 1071 neurons. Each input is weighted with an appropriate w. The sum of the weighted inputs and the bias forms the input to the transfer function f. Neurons may use any differentiable transfer function f to generate their output.

The single hidden layer in this work has 600 neurons with a log sigmoid transfer function, while the output layer has only (1) neuron which is a vector with length =5.

In order to make the neural network more efficient, the extracted feature vector of only two emotions will form the input vector, so the output will also consist of one neuron with only two elements; this implies that the number of nets will be increased to ten nets (pair of emotions will be compared in each net). Finally the average of all nets results will be computed which will give the final recognition ratio of each emotion.

5. Results

5.1 Human Ability of Recognition (Subjective Test)

Before testing the designed recognizer system, a human ability of recognition is tested in order to estimate the actors' efficiency to utter the sentence. This can be done by hearing spoken sentences from a human and the hearing ability will test the spoken sentence and decide the emotion that the sentence carries it. The process is done with four persons that they hear four spoken sentences with five different emotions. The table (1) shows the results in this testing process.

Table(1): The results of human ability of recognition**.**

Emotions	Anger	Happy	Natural	Sad	Fear	Successful ratio
Anger	**19**	1	0	0	0	95%
Happy	0	**19**	1	0	0	95%
Natural	0	1	**15**	0	0	94%
Sad	0	0	2	**16**	2	80%
Fear	0	0	2	3	**15**	75%

The table above shows that 10% of spoken sentences with sad emotion are recognized as natural , 10% of spoken sentences with the same emotion are recognized as fear. 15% of spoken sentences with fear emotion are recognized as a sad. 10% of spoken sentences with sad emotion are recognized as natural , 10% of spoken sentences with the

same emotion are recognized as natural these results calculated by equation .

$$\textbf{performance} = \frac{\textbf{Total succeeded number of test}}{\textbf{Total number of test}} * \textbf{100}$$

5.2 The Results With Different LPC Parameter Numbers

To show the ability of the emotional recognizer system we take different numbers of LPC parameters in feature extraction stage, the result is computed in the following Tables (2, 3 and 4) which is show the ability of the emotional recognizer system used with number of LPC parameters equal 10, 12, and 14 parameters.

Table (2):The results of emotion recognizer with (10)LPC parameters

Emotions	Anger	Happy	Natural	Sad	Fear	Successful ratio
Anger	**3**	0	1	0	0	75%
Happy	0	**1**	2	0	1	25%
Natural	0	0	**2**	1	1	50%
Sad	0	0	1	**2**	1	50%
Fear	0	0	1	0	**3**	75%

Table(3): The results of emotion recognizer with (12)LPC parameters

Emotions	Anger	Happy	Natural	Sad	Fear	Successful ratio
Anger	**3**	0	1	0	0	75%
Happy	0	**3**	1	0	0	75%
Natural	0	0	**3**	1	0	75%
Sad	0	0	1	**3**	0	75%
Fear	0	1	1	0	**2**	50%

Table(4): The results of emotion recognizer with (14)LPC parameters

Emotions	Anger	Happy	Natural	Sad	Fear	Successful ratio
Anger	**2**	0	0	1	1	50%
Happy	0	**0**	0	0	4	0%
Natural	0	0	**0**	0	4	0%
Sad	0	0	0	**0**	4	0%
Fear	1	0	0	0	**3**	75%

The chart in the figure (3) shows that the higher ratio of speech recognizer system is the sounds that it’s extracted features contain (12) LPC parameters.

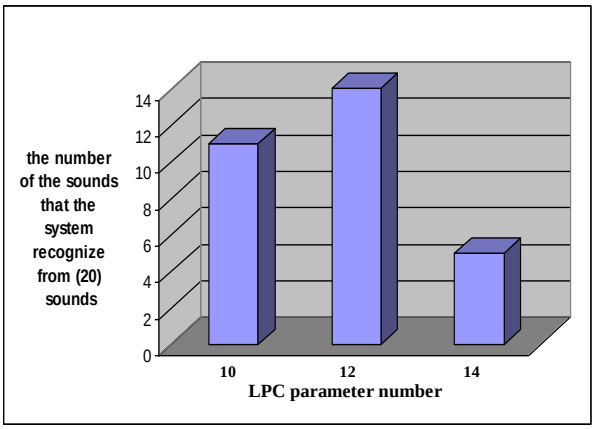

Figure (3) : The ratio of speech recognizer system with different LPC parameters number

5.3 The Results Without Formants Value in the Feature Vectors

when we take the feature vector contain only the features (LPC, pitch and power). The emotion recognizer system give the result shown in Table (5)

Table(5): The results of emotion recognizer without formants feature

Emotions	Anger	Happy	Natural	Sad	Fear	Successful ratio
Anger	**2**	0	1	0	1	50%
Happy	0	**0**	1	0	3	0%
Natural	0	0	**0**	1	3	0%
Sad	0	0	0	**1**	3	25%
Fear	1	0	1	0	**2**	50%

5.6 The Results Without Pitch Value in the Feature Vectors

when the feature extracted vector contain only the features (LPC, formants (1,2,3) and power). The emotion recognizer obtained the results shown in Table (6)

Table(6): The results of emotion recognizer without pitch value feature

Emotions	Anger	Happy	Natural	Sad	Fear	Successful ratio
Anger	**3**	0	1	0	0	75%
Happy	1	**2**	0	1	0	50%
Natural	1	1	**2**	0	0	50%
Sad	1	0	0	**2**	1	50%
Fear	0	3	0	0	**1**	25%

5.7 The Results With Different Numbers of Neurons In The Hidden Layer

The previous results show that the best result will obtained with the feature vector that contained the features (12 LPC parameters , pitch, power, and the formants (1,2 and 3)). with the above conditions different numbers of neurons in the hidden layer that the network contained gives different results. Tables (7and 8) shows this results.

Table(7): The results of emotion recognizer with 107 neurons in the hidden layer

Emotions	Anger	Happy	Natural	Sad	Fear	Successful Ratio
Anger	**1**	1	2	0	0	25%
Happy	0	**3**	0	1	0	75%
Natural	0	2	**2**	1	0	50%
Sad	0	2	1	**1**	0	25%
Fear	0	3	1	0	**0**	0%

Table(8): The results of emotion recognizer with 500 neurons in the hidden layer

Emotions	Anger	Happy	Natural	Sad	Fear	Successful ratio
Anger	**3**	0	1	0	0	75%
Happy	0	**3**	1	0	0	75%
Natural	0	0	**3**	1	0	75%
Sad	0	0	1	**3**	0	75%
Fear	0	1	1	0	**2**	50%

6. Discussion Of Results

Table (1) shows that the recognition meets difficulty with human ability of hearing, that is the natural, sad, and fear emotions are mixed one with another in human testing ability with ratio from 10-15%.

Using (12) LPC parameters in feature extraction stage gives the best results in this work. Typically the number of the LPC parameters that used are from 8 to 16 parameters. Since 12 is the medium of this range, usually the best results are obtained with this number of LPC parameter.

With using 10 and 12 LPC parameters, and also using the extracted feature vector without pitch the best results are obtained, in this stages there is no fear spoken sentence recognized as an anger and conversely. This is explicated by the scientific agreement that fear and anger are primary emotions.

While in using 14 LPC parameters, the only truly recognized sounds are the fear and the anger sounds. Also the other spoken sentences are recognized as a fear or an anger sounds.

Ignoring the pitch feature in the extracted feature has an effect on the result, that's the successful result ratio in this stage is generally 50%. while without ignoring this feature the successful result ratio is 75%.

The formant frequencies has a large effect on the results, that's the successful result ratio is decreased without using formant frequencies in the extracted features from 75% to 25%. This is mean that the formants is more effective than the pitch, that is each vowel has different formant frequencies and bandwidths, and the emotions is consist of vowel sounds more than the constants.

The number of neurons equal to 500 in the hidden layer is the most useful number that can be used. This number is approximately is 46% of the neurons number in the input layer that is equal to 1071 neurons. A number of neurons in the hidden layer is chosen depending on the problem, for example in letter recognition the number of neurons that chosen is approximately equal 10% of the number of neurons in the hidden layer. But generally the researches in sound recognition use number of neurons equal 50% of the number of neurons in the hidden layer.

7. Conclusion

After applying the steps of the emotion speech recognizer and using the discussed algorithms of feature extraction, with the neural network that designed for this system. The following points can be noted:

1- The best number of LPC parameters that must be used is (12) parameters.
2- With using number of LPC parameters equal (14) parameter, the secondary emotions have very little chance to recognize.
3- The primary emotions have better capacity of recognition.
4- The formant frequencies have a better capacity of representing the speech signal features than the pitch.
5- Number of neurons equal to 50% of the hidden layer neurons number is the best number that is chosen.
6- Difficulty of acting the uttered sentences, and this lead to mixture of the emotions.
7- The difficulty of expressing the emotions that used to train the network, has it's negative effects in the process.
8- The difficulty of formants frequency estimation makes the feature vector not to represent the speech signal typically.

References

[1] Chul Min Lee, “Emotion Recognition for Spoken Language”, University of Southern California, 2004

[2] Jackson, L.B., “Digital Filters and Signal Processing, Second Edition”, Kluwer Academic Publishers, 1989. pp. 255-257.

[3] Kivimaki, J., “Very Low Bit Rate Speech Coding Using Speech Recognition, Analysis and Synthesis”, M.Sc. Thesis, Department of information Technology, Tampere university of technology, 2000.

[4] Lawrence R., “Fundamentals of Speech Recognition”, Prentice Hall PTR, 1993 pp.[25-32 40] .

[5] Nakatsu, R., Tosa, N., Ochi, T., “Construction of an Interactive Movie System for Multiperson Participation.” Proc Int Conf on Multimedia Computing and Systems 1998; 228-232.

[6] Perry R. Cook, "Identification of Control Parameters in an Articulatory Vocal Tract Model”, with applications to the synthesis of singing", 1990, Ph.D Dissertation, CCRMA .

[7] Sherif Y, “Recognition of Emotions in Interactive Voice Response Systems” , HP Laboratories Palo Alto, July, 2003.

[8] Theiler, J. “Two Realization of a General Feature Extraction Framework”, NIS-2, Los Alamos National Lab, pp 875-887.

[9] علام زكي عيسى، "الشبكات العصبية البنية الهندسية الخوارزميات، التطبيقات "، شعاع للنشر والعلوم، 2000 .