IOSR Journal of Engineering (IOSRJEN)
ISSN: 2250-3021 Volume 2, Issue 7(July 2012), PP 11-14
www.iosrjen.org

Instantaneous Emotion Detection System using Vocalizations

M.JayaLakshmi[1], K.Maharajan[2], Dr.B.Paramasivan[3]
[1,2]Asst.Professor, [3]Professor & Head
Department of Computer Science & Engineering, National Engineering College
Kovilpatti, Thoothukudi Dist, Tamilnadu, India - 628 503

ABSTRACT: - *The importance of automatically recognizing emotions from human speech has grown immensely in human computer interaction application. This paper describes an experimental study on the detection of emotion from speech. The study utilizes a corpus containing emotional speech with 25 short utterances expressing three emotions: anger, happiness and neutral (unemotional) state, which were captured manually from recording. Emotions are so complex that most speech sentences cannot be precisely assigned into a particular emotion category; however, most emotional states nevertheless can be described as a mixture of multiple emotions. Based on this concept some of the acoustic features are extracted and the voice samples are trained using Locality Preserving Projection method and Kullback-Leibler Distance to recognize utterances within these three categories. Before detecting the emotion initially Kalman Denoise filtering technique is applied in order to remove noise. This paper is a model of "Emotion Detection System" in order to understand how emotional viewing could be coded and that has led us to the devising of emotional film structures. With this work both the negative and non-negative emotions can be easily detected.*

Keywords: - *Acoustic features, Kullback-Leibler , Kalman Denoise*

I. INTRODUCTION

As computers have become an integral part of our lives, the need has arisen for a more natural communication interface between humans and machines. To accomplish this goal, a computer would have to be able to perceive its present situation and respond differently depending on that perception. Part of this process involves understanding a user's emotional state. To make the Human-Computer Interaction (HCI) more natural, it would be more beneficial to give computers the ability to recognize situations the same way a human does. Researches concerning psychological and neurobiological analysis are also important and can improve the efficiency of intelligent human-machine interface. For example, psychological and linguistic researches show us, that emotions play significant role in decision-making process. In the field of HCI, speech is primary to the objectives of an emotion recognition system, as are facial expressions and gestures: It is considered a powerful mode to communicate intentions and emotions. This paper explores methods by which a computer can recognize human emotion in the speech signal. Emotion detection is rapidly gaining interests among researchers and industrial developers since it has a broad range of applications. In call center service, once the machine detects the possible angry or depressed emotion of customers, it transfers the calls to human representatives. The voicemail can be sorted according to the emotions or urgency level of messages of the callers. Automatic assessment of boredom or stress of vehicle drivers is highly valuable. One of the most basic capabilities for robots is the ability to decode and express human emotions. In Yi Lin Lin et al (2005), two classification methods, the Hidden Markov Model (HMM) and the Support Vector Machine (SVM), were studied to classify five emotional states: anger, happiness, sadness, surprise and a neutral state in which no distinct emotion is observed. The best feature vector with a dimension of five was determined from the 39 candidate instantaneous features before being input into the HMM classifier. That determination was made using the Sequential Forward Selection (SFS) method. For the SVM classifier, a novel feature vector that measures the difference between Mel scale sub-band energies was used. Classification experiments including gender dependent and gender independent cases were conducted on the Danish Emotional Speech (DES) database. In Schuller et al (2003), two classification methods, Gaussian Mixture Model (GMM) with global statistics and HMM with instantaneous features, were studied, but it was limited to features related to pitch and energy contour of the speech signal. As a background resulting from common behaviors of human being, we can take into account: joy, anger, sadness, disgust, fear, surprise and neutral state. Of course, for better understanding of mutual human relations it would be necessary to expand the list significantly, but for engineering purpose this background seems to be sufficient. Moreover, if our goal is focused on the correctness of emotion recognition, the number of emotions is inversely proportional to the recognition efficiency. The simplest analysis of only two emotions: negative and non-negative, can be sometimes very useful, especially for service improvement.

Different emotions can be presented in multidimensional space as a function of the features. Very simple example of such function is shown in Figure 1, where you have two-dimensional space describing by two axes: "Energy" and "Quality". If the number of space dimensions goes up, the situation becomes more and more complicate and different emotion spaces can overlap. In such a case increasing of the features can even decrease recognition efficiency.

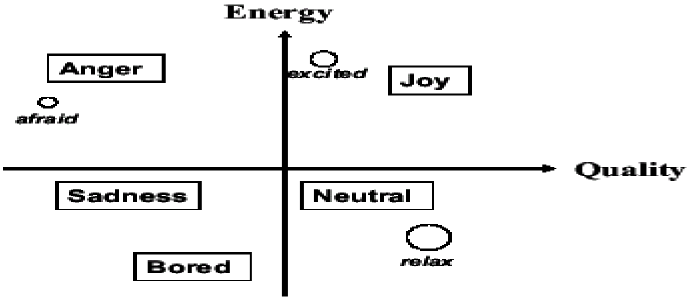

Figure 1: Emotions in two-dimensional space described by energy (prosody) and quality intensities

II. SPEECH DATA PREPARATION

Generally, there are three major categories of emotional speech samples. They are natural vocal expression, induced emotional expression, and simulated emotional expression. Natural vocal expression is recorded during naturally occurring emotional states of various sorts. Induced emotions are caused by using psychoactive drugs or some particular circumstances, such as in some kind of games or by using inducing words to get the speech sample of desired emotion. The third category of getting speech samples is the simulated emotional expression, that is, to ask actors to produce vocal expressions of certain emotions. In this way, the content and the emotions are given, and the process can be controlled to get more typical expressions. In literature, the most preferred way of getting emotional speech samples is the third one. The database used in this paper has been collected with use of dynamic AKG-1000s MK-II microphone in an acoustically isolated room. The phrases were all collected in English language and are throughout acted emotions. This has been widely discussed due to the fact that acted emotions are not spontaneous and tend to be exaggerated. The two speakers were in average 25 years old and female. The speech is expressed in three emotional states: anger, happiness and neutral. The utterances vary in length and spoken content throughout the corpus meeting the challenge of ensuring greater independence.

III. Feature Extraction

The proper choice of speech features has very significant influence for an efficiency of emotion recognition. In this experiment, the main goal for emotion detection is to build a system which does not require a speech recognizer or a speech understanding component and all the information which system can use comes from the acoustic level. It is well known that speech is a short time stationary signal, and we think the emotion is expressed with some kind of trajectory pattern. For example, energy is a measure of emotion but it is the fluctuation/trajectory of energy, not the energy itself makes the speech sound emotional. The same argument is true for the pitch. Pitch and Energy are very useful in the emotion detection but, pitch is hard to extract reliably; the energy measure is sensitive to the channel distortion and environment noise. Besides the popular feature – pitch in emotion detection, the Zero Crossing Rate (ZCR), and short time energy measurement is included. In the case of fundamental frequency/pitch calculation, two basic methods are available: autocorrelation and cepstrum method. The complex values of cepstrum $C(T)$ can be obtained using the following equation:
$C(T) = F^{-1} [\log (F(x(t)))]$ (1) where F is a Fourier transform and $x(t)$ represents speech signal. In this transform the convolution of glottis excitation and vocal tract is converted, first to the product after Fourier transform, separated them finally as the sum. ZCR is defined as the number of times the audio sequence changes signs in the processing window. The ZCR is applied to help capturing the end-point of audio (pause, voiced speech, and unvoiced speech). Average Zero-Crossing Rate is defined as $z_n = \sum_{m=-\inf}^{m=\inf} |\mathrm{sgn}[x(m)] - \mathrm{sgn}[x(m-1)]| w(n-m)$ w(n) = 1/2N, 0 <= n <= N-1. This measure could allow the discrimination between voiced and unvoiced regions of speech, or between speech and silence. Unvoiced speech has in general, higher zero-crossing rate. The signals in the graphs are normalized. Short-Time energy is a simple short-time speech measurement. It is defined as $E_n = \sum_{m=-\inf}^{m=\inf} [x(m)w(n-m)]2$ .The processing window for each frame is 0.025 second. So there are 100 frames in every second of audio.

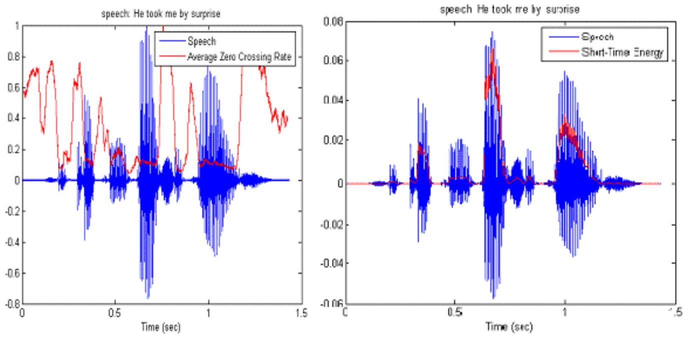

Figure 2: Audio feature representation of sample emotional speech

IV. Feature reduction by LPP

To improve classification performance, we need to reduce the dimension of the features. In this work, a new linear dimensionality reduction algorithm, called Locality Preserving Projections (LPP) is proposed. It builds a graph incorporating neighborhood information of the data set. Using the notion of the Laplacian of the graph, we can compute a transformation matrix which maps the data points to a subspace. This linear transformation optimally preserves local neighborhood information in a certain sense. The Locality Preserving Projection may be simply applied to any new data point to locate it in the reduced representation space. LPP should be seen as an alternative to Principal Component Analysis (PCA) – a classical linear technique that projects the data along the directions of maximal variance. When the high dimensional data lies on a low dimensional manifold embedded in the ambient space, the Locality Preserving Projections are obtained by finding the optimal linear approximations to the eigenfunctions of the Laplace Beltrami operator on the manifold. As a result, LPP shares many of the data representation properties of nonlinear techniques such as Laplacian Eigenmaps or Locally Linear Embedding. Given a set of high-dimensional data $X = \{x_1, x_2 ... x_m\} \in R^N$, the basis problem of dimensionality reduction is to find out a transform matrix A, that could map the input data X into a low dimensional space $R^l, l < N,$ i.e. $y_i = A^T x_i$, and $y_i \in R^l, i = 1,2,...m$, is the transformation result. The LPP algorithm is summarized in the following three steps 1: Constructing the adjacency graph: Create a graph G with n nodes. The ith node denotes the data xi. Put an edge between the ith node and the jth node if xi or xj is among k nearest neighbors of each other. Step 2: Computing the weights: The weights evaluate the local structure of the data space. In our work, if the ith node and the jth node are connected, put $w_{ij} = \exp(-\left|x_i - x_j\right|^2 / t)$, else $w_{ij} = 0$. So W is a sparse and symmetric n x n matrix with the weight w_{ij}.

Step 3: Eigenmaps: The optimal projection preserving the locality could be solved by minimizing the following objective function $\min \sum_{i,j} (y_i - y_j)^2 w_{ij}$ (2). This minimization problem is to ensure that if xi and xj are close, then yi and yj are close as well. And this problem can be solved through computing the eigenvalue λ and the eigenvector μ by $XLX^T \mu = \lambda XDX^T \mu$, (3) where D is a diagonal matrix and Let the solutions of $D_{ij} = \sum_j w_{ij} = \sum_j w_{ji}, L = D - W.$ (3) be the column vectors $\mu_0, \mu_1, ... \mu_{L-1}$, and their corresponding eigenvalues $\lambda_0, \lambda_{1,}, ... \lambda_{l-1}$, are sorted by $\lambda_0 \leq \lambda_1 \leq ... \leq \lambda_{l-1}$. The final linear dimensionality reduction mapping is as follows: $x_i \rightarrow y_i = A^T x_i$, (4) and $A = (\mu_0, \mu_1 .... \mu_{l-1})$. Then xi is reduced to a l dimensional vector yi which preserves the local relationship with its neighbors. The obtained projections are actually the optimal linear approximation to the eigenfunctions of the Laplacian Beltrami operator on the manifold.

V. EMOTION DETECTION CLASSIFIER

During the last years Support Vector Machines (SVM's)[7] have become extremely successful discriminative approaches to pattern classification and regression problems. Excellent results have been reported in applying SVM's in multiple domains. However, the application of SVM's to data sets where each element has variable length remains problematic. Furthermore, for those data sets where the elements are represented by large sequences of vectors, such as speech the direct application of SVM's to the original vector space is typically unsuccessful. The Kullback-Leibler distance is a measure of distance between probability distributions and it is widely used in information theory to estimate the information extracted from a correlation matrix by correlation filtering procedures. We are also able to compute the expected value of the Kullback-Leibler distance between two distinct samples of the correlation matrix obtained from the same random source. This is a powerful and accurate tool able to characterize the information and stability of sample, model and filtered correlation matrices and it is a useful quantitative indicator for the relative amount of information and the relative stability of correlation matrices of multivariate data. Let P and M be two probability measures associated with the measurable space. The Kullback-Leibler Distance (KLD) of P with respect to M is given by

$$D(P|\,|M) = \sum P(A)\log\frac{P(A)}{M(A)} \ .$$

VI. Expected Result

Emotion Detection process involves three steps. Feature extraction, Feature reduction and Classification. When a voice sample is given as input in order to recognize the emotion, it goes through the training process. In this procedure, all but one data is used for training. The selection criterion is the correct classification rates by KLD. The best recognition rate of 100% will be obtained when only emotional speech from female subjects is considered. For male subjects the expected correct classification rate will be nearly 90%. In the confusion matrices shown in Table 1, the columns show the emotions that the speakers tried to induce, and the rows will be the output recognized emotions.

Stimulus	Recognized Emotions (%)		
	Neutral	Angry	Happy
Neutral	100	0	0
Angry	0	100	0
Happy	0	0	100

Table 1: Confusion matrix using Kullback Leibler Distance

VII. CONCLUSION

The results of the emotion recognition system are rewarding, but sometimes we can observe mistakes in recognition process. However, such recognition makes difficulties also for human evaluation. On the other hand, a quality of the program depends on the training processes. Unfortunately, it is difficult to obtain a proper base of voice examples for different emotions. However, it is discovered, that the program could recognize two states: positive and negative emotions with almost 100% precision.

References

[1]. Ben Gold and Nelson Morgan (2006) ‘Speech and Audi Audio Signal Processing’, Berkeley, pp. 302-348

[2]. Chul Min Lee, Narayanan S. (2005) ‘Toward Detecting Emotions in Spoken Dialogs’, IEEE Trans. Speech and Audio Processing, vol. 13, no 2, pp. 293-303

[3]. Cowie R, Douglas-Cowie E, Tsapatsoulis N. (2001) ‘Emotion recognition in human-computer interaction’, IEEE Signal Processing magazine, vol. 18, no. 1, pp. 32-80

[4]. Laurence Devillers, Lori Lamel and Ioana Vasilescu (2003) ‘Emotion Detection in Task-Oriented Spoken Dialogs’, ICME Proceedings, vol.3, page(s): III- 549-52

[5]. Schuller B, Rigoll G, and Lang M. (2003) ‘Hidden Markov model-based speech emotion recognition’, Proceedings of the IEEE ICASSP Conference, Hong Kong, pp. 1-4.

[6]. Ververidis D. and Kotropoulos C. (2004) ‘Automatic speech classification to five emotional states based on gender information’, Proceedings of the EUSIPCO2004 Conference, Austria, pp. 341-344.

[7]. Yi-Lin Lin and Gang Wei (2005) ‘Speech Emotion Recognition Based on HMM and SVM’ Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, Guangzhou, pp. 18-21

[8]. Zygmunt Ciota (2006) ‘Feature Extraction of Spoken Dialogs for Emotion Detection’ ICSP Proceedings , vol.1