Multi-Feature Based Speech Emotion Recognition

Feature combination for speech emotion recognition

Sarveda S1, Nirmala L2
1PG Scholar, 2Faculty
1,2 ECE Department & REVA Institute of Technology and Management, Bangalore, Karnataka, India,

Abstract—This paper deals with the use of multi features for speech emotion recognition. We analyze the tradeoff among the different combination of features for the efficient recognition rate. The method of multi-feature combination is to make recognition rate more accurate but by increasing the computational complexity. In this paper we use Mel Frequency Cepstrum Coefficients (MFCC) which is an efficient feature of Gaussian mixture model, Linear Predictive Coding (LPC), and Discrete Wavelet Transform (DWT) which is trained to Neural Networks using SAVEE emotion base database. Our recognition rate of 85% is better than other method of using HMM with MFCC, Delta MFCC, and speech energy.

Key words: Speech recognition, Neural networks.

I. INTRODUCTION

Speech is the most basic and main communication tool in human-to-human interaction. Emotion can make it’s meaning more complex and the listeners can react differently according to what kind of emotion the speaker transmit, e.g., consoling a sad one with soft words. Speech signal contains rich information including: 1) speech content; 2) speaker identified[2] and recognized by the voice signal; 3) speaker’s emotions[1]. From the signal processing point of view, speech signal includes the linguistic information, speaker’s tone and emotion. There are several applications for automatic machine recognition of the type of emotion expressed in a given speech. For example, it may be desirable to include information of emotions to other party in conventional video teleconferencing and web-based teaching for added effects. In distance teaching, if a student does not understand what the teacher is saying, it may be detected from expression of emotions on his face and in his speech. In this paper, we first discuss which kind of features to extract from the speech. Next, we use these features to recognize emotions contained in the given speech. Later, we demonstrate that our approach can be used effectively.

Referring the previous work of other researchers, Atassi, Esposito and Smekal [9] showed that high-level features performed well in terms of speech emotion classification, and recognition rates for six emotions are more than 80%. They used a lot of features: the Mel Frequency Cepstrum Coefficients (MFCC), pitch, the Perceptual Linear Predictive (PLP), the sub band based cepstral coefficients (SBC), the wavelet decomposition (WADE), the MELBS, the HFBs and the LFBs. Similarly, Emerich and Lupu used the MFCC and the Discrete Wavelet Transform (DWT) to identify seven kinds of emotions, and achieved 95.42% accuracy rate. Iliou and Anagnostopoulos[10] also achieved very high accuracy (94.3%). They used twenty-five features in their work including the pitch, the MFCC, the energy and the formant. The major drawback of the above three works is that the number of features used is so large that their time complexity becomes too high. Other researchers used a large number of features to identify a small number of emotions, and acquired good results. [11] Pan and Shen used seven features, which are the energy, the pitch, the linear predictive spectrum coding (LPCC), the MFCC and the mel-energy spectrum dynamic coefficients (MEDC), to identify three emotions. Their method achieved up to 95.1% accuracy rate. However, the speaker-dependent approach is not practicable in many applications, since it mostly works with a very large number of possible users (speakers). To our knowledge, for speaker independent applications, the best classification accuracy was 81% [20] and was obtained on the Berlin Database of Emotional Speech (BDES) using a two-step classification approach and a unique set of spectral, prosodic and voice features selected through the Sequential Floating Forward Selection (SFFS) algorithm. The Majority of work addressing the emotion recognition has focused on the classical approach to the task, that is, the input speech signal is usually assigned to one class “emotion” according to a certain classification criteria. This approach has several limitations; since the output of such systems is determined within predefined emotions whereas it is proved that human emotional states are characterized by a high level of variability. Moreover, it is almost impossible to build a speech database that covers all possible human emotional states. Our literature survey shows that most speech emotion recognition methods use spectral and prosodic features. Using different set of features also lead to quite different emotion recognition rate.

In our study, we select the following features i) MFCC, this feature has been widely used in the related work because Mel frequency is proposed according to the characteristics of the human auditory and the MFCC include language characters richer than other speech features. ii) Linear predictive coding (LPC), and iii) DWT.

II. METHODOLOGY

2.1 Gaussian Mixture Model

A Gaussian Mixture Model (GMM)[2] is a parametric probability density function represented as a weighted sum of Gaussian component densities. GMMs are commonly used as a parametric model of the probability distribution of continuous measurements or features in a biometric system,
such as vocal-tract related spectral features in a speaker recognition system.

A Gaussian mixture model is a weighted sum of $M$ component Gaussian densities as given by the equation

$$p(x|\lambda) = \sum_{i=1}^{M} w_i b_i(x)$$  \hspace{1cm} (1)

Where $x$ is a D-dimensional vector $w_i$, $i=1,\ldots,M$, are the mixture weights, and $b_i(x)$

$$b_i(x) = \frac{1}{\sqrt{(2\pi)^D}} \exp\left(-\frac{1}{2}(x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i)\right)$$  \hspace{1cm} (2)

with mean vector $\mu_i$ and covariance matrix $\Sigma_i$.

The complete Gaussian mixture model is parameterized by the mean vectors, covariance matrices and mixture weights from all component densities. These parameters are collectively represented by the notation

$$\lambda = \{w_i, \mu_i, \Sigma_i\} \hspace{1cm} i = 1, \ldots, M$$  \hspace{1cm} (3)

2.2 Feature Extraction

The first step in any automatic speech recognition system is to extract features i.e. identify the components of the audio signal that are good for identifying the linguistic content and discarding all the other stuff which carries information like background noise, emotion etc. Feature extraction of the voices is the important part of speech emotion recognition. In this paper, we will evaluate the emotion of the voice based on the MFCC, LPC and DWT.

A. Mel frequency cepstrum coefficients (MFCC)

Mel Frequency Cepstral Coefficients (MFCCs) are a feature widely used in automatic speech and speaker recognition[5]. It is a representation of short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a non-linear mel scale of frequency. They were introduced by Davis and Mermelstein in the 1980’s, and have been state-of-the-art ever since. Prior to the introduction of MFCCs, Linear Prediction Coefficients (LPCs) and Linear Prediction Cepstral Coefficients (LPCCs) were the main feature type for automatic speech recognition (ASR).

The process of solving MFCC is divided into following steps:

1. Pre-emphasis, Frame blocking, hamming window, Fast Fourier Transform (FFT), mel Filter banks, Discrete cosine transform (DCT) and Log energy[1].

These steps are shown in the figure 1.

At the Frame Blocking Step, we split the continuous speech signal into N frames. Each sound frame multiplied by the Hamming window in order to increase the continuity of the left and right ends of the sound box. We assume that the sound box of the signal is $S(n)$, $n=0,..,N-1$. Multiplied by Hamming window, the formula is $S(n) = S(n) * h(n)$. The form of $h(n)$ is as follows:

$$h(n) = 0.54-0.46 \cos(\frac{2\pi n}{N-1}) \hspace{1cm} 0 \leq n \leq N-1$$  \hspace{1cm} (4)

After being multiplied by the Hamming window, each frame is also necessary to be processed by FFT which can obtain the frequency spectrum energy distribution.

$$s_i(k) = \sum_{n=1}^{N} s_i(n) e^{-j2\pi kn/N} \hspace{1cm} 1 \leq k \leq N$$  \hspace{1cm} (5)

Where $s_i(n)$ represents the ith frame, and $k$ is the length of FFT.

The signal change in time domain is often difficult to indicate the characteristics of the signal, so we usually convert the signals into energy distribution in the frequency domain. Different energy distribution can represent the characteristics of different speech. Typically the mel filter banks is used. The energy spectrums multiplied by a group of 24 filters which are evenly distributed in the Mel frequency and obtain the log energy from each filter. And then Take the DCT of the log filter bank energies. Mel frequency and the average frequency have the following relationship:

$$\text{mel}(f) = 2595 \log_{10}(1+f/700)$$  \hspace{1cm} (6)

B. Linear Predictive Coding (LPC)

LPC method is most widely used in speech coding, speech synthesis. It provides extremely accurate estimate of speech parameters[19]. Basic idea of Linear Prediction is that the current speech sample can be closely approximated as a linear combination of past samples, i.e.,

$$S(n) = \sum_{k=1}^{P} a_k s(n-k)$$  \hspace{1cm} (7)

for some value of $p, a_k$’s.

Assume that present sample of the speech is predicted by the past $M$ samples of the speech such that

$$\tilde{x}(n) = a_1x(n-1) + a_2x(n-2) + \ldots \ldots + a_Mx(n-M) = a_1x(n-1)$$  \hspace{1cm} (8)

Where $\tilde{x}(n)$ is the predictive of $x(n)$, $x(n-1)$. $a$ is the linear prediction coefficients. To solve the resultant $M$ equations an
effective of all called Levinson and Durbin algorithm is used to estimate linear predictive coefficients.

C. Discrete Wavelet Transform

The DWT[19] of a signal $x(n)$ is calculated by passing it through a series of filters. First the samples are passed through a low pass filter with impulse response $g(n)$ which gives the convolution of two signals as

$$y(n) = (x * g)(n) = \sum_{k=0}^{\infty} x(k) g(n - k)$$  \hspace{1cm} (9)

The signal is also decomposed simultaneously using a high-pass filter $h(n)$. The outputs of the high pass filter are detail coefficients and the outputs of the low pass filter are approximation coefficients. It is important that the two filters are related to each other and they are known as a quadrature mirror filter. Since half the frequencies of the signal have now been removed, half the samples can be discarded according to Nyquist’s rule. The filter outputs are then sub sampled by 2.

$$y_{\text{low}}(n) = \sum_{k=0}^{\infty} x(k) g(2n - k)$$  \hspace{1cm} (10)

$$y_{\text{high}}(n) = \sum_{k=0}^{\infty} x(k) g(2n + 1 - k)$$  \hspace{1cm} (11)

This decomposition has halved the time resolution since only half of each filter output characterises the signal. However, each output has half the frequency band of the input so the frequency resolution has been doubled. This decomposition is repeated to further increase the frequency resolution and the approximation coefficients decomposed with high and low pass filters and then down sampled. This is represented as a binary tree with nodes representing a sub-space with a different time-frequency localisation. The tree is known as a filter bank.

![Figure 2: DWT 3 level Decomposition](image)

At each level in the above diagram the signal is decomposed into low and high frequencies. Due to the decomposition process the input signal must be a multiple of 2n where n is the number of levels. The wavelet decomposition (WADE) was performed by using a Haar wavelet with a decomposition level of 3.

D. Database

In this work Savee database is explored for analyzing the emotion. Surrey Audio-Visual Expressed Emotion (Savee) database has been recorded as a pre-requisite for the development of an automatic emotion recognition system. The database consists of recordings from 4 male actors in 7 different emotions, 480 British English utterances in total. The data were recorded in a visual media lab with high quality audio-visual equipment, processed and labeled. To check the quality of performance, the recordings were evaluated by 10 subjects under audio, visual and audio-visual conditions. Classification systems were built using standard features and classifiers for each of the audio, visual and audio-visual modalities, and speaker-independent recognition rates of 61%, 65% and 84% achieved respectively. These are trained to neural networks[19]. Neural nets offer an approach to computation that mimics biological nervous systems. Algorithms based on neural nets have been proposed to address speech recognition tasks which humans perform with little apparent effort.

III. RESULTS AND DISCUSSION

The two experiments were done to evaluate the recognition accuracies of the two different emotions from speech viz happy, and anger. The databases were divided in to two sets for training and testing of the classifier respectively. We used a proportion of 80% for training and remaining 20% for testing of the classifier in all the experiments. In the first method, we have analysed using MFCC, and Linear predictive co-efficient with the database consisting of 40 utterances. During testing of the classifier to recognize the speech from the two different emotional classes and obtained a recognition accuracy of approximately 72%. For recognizing the emotion happy the machine can attain only a recognition accuracy of approximately 60% and obtained a confusion of 40% with the emotion angry. In recognizing the emotion angry the machine attained a recognition accuracy of approximately 85% and obtained a confusion of 15% with the emotion happy. The confusion matrix for emotional speech database indicating the recognition accuracies of two emotions using MFCC and LPC is given in Table I.

<table>
<thead>
<tr>
<th>Emotion Class</th>
<th>Emotion recognition performance(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angry</td>
<td>85</td>
</tr>
<tr>
<td>Happy</td>
<td>15</td>
</tr>
</tbody>
</table>

In the second method, we have analysed using MFCC, DWT and Linear predictive co-efficient with the database consisting of 40 utterances. During testing of the classifier to recognize the speech from the two different emotional classes and obtained a recognition accuracy of approximately 85%. For recognizing the emotion happy the machine can attain only a recognition accuracy of approximately 78% and obtained a confusion of 22% with the emotion angry. In recognizing the emotion angry the machine attained a recognition accuracy of approximately 95% and obtained a confusion of 5% with the emotion happy. The confusion matrix for emotional speech
database indicating the recognition accuracies of two emotions using MFCC, DWT and LPC is given in Table II.

<table>
<thead>
<tr>
<th>Emotion Class</th>
<th>Emotion recognition performance(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Angry</td>
</tr>
<tr>
<td>Angry</td>
<td>95</td>
</tr>
<tr>
<td>Happy</td>
<td>22</td>
</tr>
</tbody>
</table>

This Recognition performance is calculated using FAR method (false acceptance rate), which is a measure of incorrectly identifying sample, person, or user etc. Here emotion angry is represented by 1, and happiness by -1. Suppose for a given angry speech i.e. 1, if the machine identifies it as happiness i.e. -1, then the false acceptance rate is 1. Similarly for a given angry speech, if the machine identifies it as angry itself, then the FAR rate is 0. This is shown in table III. This FAR is calculated for all the test speech voices, based on which emotion recognition performance is calculated. This is done for both the methods then the performance (%) is compared.

Table III: Emotion Recognition Performance Calculation Based On FAR

<table>
<thead>
<tr>
<th>Emotion class</th>
<th>FAR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Angry</td>
</tr>
<tr>
<td>Angry</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Happy</td>
<td>-1</td>
</tr>
<tr>
<td></td>
<td>-1</td>
</tr>
<tr>
<td></td>
<td>-1</td>
</tr>
<tr>
<td></td>
<td>-1</td>
</tr>
<tr>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

IV. CONCLUSION

The recognition performance of different emotions from speech by using different features are carried out in this experiment. Artificial Neural Network was used for the machine learning purpose. Percentage recognition obtained for each emotions in the each case are compared. The results obtained from the experiments shows that emotion recognition from speech strictly depends on the database used. Experiment results demonstrate that this method can provide a recognition rate of 85% with high computational time, but it still has room to improve.

REFERENCES