Emotion Recognition using Support Vector Machines

Armaan Ansari[1], Rahul Sharma[2], Niranjan Samudre[3]
[1, 2](BE Student, Department of Electronics Engineering, Atharva College of Engineering, Mumbai, India)
3(Assistant Professor, Department of Electronics Engineering, Atharva College of Engineering, Mumbai, India)

Abstract: *In the present day speech signal process has a very big selection of applications in several technical fields like human laptop interaction, biometrics, computing etc. In speech processing emotion recognition is major research area where different emotions of people are recognized in this paper the proposed system allows recognizing a person's emotional state from audio signals.*

The projected solution is aimed toward raising the interaction among humans and computers, therefore permitting effective human-computer intelligent interaction. The system is in a position to acknowledge six emotions (anger, boredom, disgust, fear, happiness and sadness). This set of emotional states is wide used for emotion recognition functions. It conjointly distinguishes one feeling versus all the opposite potential ones, as tried within the projected numerical results. The system consists of 2 subsystems specifically feeling recognition (ER) Gender recognition (GR).For this two support vector machines (SVMS) are used for the male and female speaker emotion recognition.

I. Introduction

Recently there has been growing interest to improve Human-computer interaction (HCI) means computers should interact to the humans in day to day life .In this context recognizing people emotional state and giving suitable feedback may play a crucial role. As a consequence, emotion recognition represents a hot analysis space in each industry and academic field. Usually emotion recognition based on facial or voice options.

This project proposes a solution, designed to be employed in a smart phone Environment able to capture emotional state of a person starting from registration of speech signals in the surrounding obtained by mobile devices like smart phones. People can use their voice to give command to car, cell phone, computer TV and many electrical devices. Hence build the device perceives human feeling and provides a stronger experience of interaction becomes a very interesting challenge.

Many researchers have been done for this purpose, for instance Fadi A Machtetal investigate an application of speech emotion recognition to avoid traffic accident. The work performed utilizes a recognition machine to classify the voice message in phone answering machine and gives priority.

Typically, the foremost common thing to recognize speech emotion is to first extract vital features that are associated with totally different emotion states from the voice signal (i.e. Energy is a important feature to distinguish happy and sad), then feed those features to the input end of a classifier and obtain different emotions at the output end. Although emotion detection from speech is a comparatively new field of research, it has many potential applications. In human-computer or human-human interaction systems, emotion recognition systems could provide users with improved services by being adaptive to their emotions. In virtual worlds, emotion recognition could help simulate more realistic avatar interaction.

The body of work on detecting emotion in speech is kind of restricted. Currently, researcher's area unit still debating what options influence the popularity of feeling in speech. There is conjointly substantial uncertainty on the most effective algorithmic program for classifying emotion, and which emotions to class together.

In this project, we tend to attempt to address these problems. We use K-Means and Support Vector Machines (SVMs) to classify opposing emotions.

II. Literature survey

A suitable alternative of speech information (corpora) plays a very vital role within the field of have an effect on detection. A context rich emotional speech database is preferred for a good emotion recognition system. Mainly three types of corpora are used for developing a speech system they are1) Elicited emotional speech database: This type of corpora is collected from speaker by creating artificial emotional situation.

Advantage of this kind of information is that it's very close to the natural information however there are some issues also. All emotions might not be accessible and if the speaker is alert to that they're being recorded, then the emotion expressed by him could also be artificial.2) Actor based speech database: This type of speech data set collected from professional and trained artists. Collecting of these form of data are very easy and a wide variety of emotion are accessible within the corpora .The main downside of this kind of information are

it is episodic in nature and it's considerably artificial in nature.3) Natural speech database: this type of database created from real world data. These forms of information are completely natural and extremely helpful for real world emotion recognition. The problem is that, all emotions may not be present and it consists of background noise. Implementation of emotional speech database depends on objective of the analysis.

For efficient affect detection system, it is important that the corpora must consist of real and natural emotional speech spoken by a large number of male and female persons.

Though there are different corpora exists, there is no standard, globally approved speech database available for emotion recognition. In Indian context, there are different speech corpora for speaker recognition however there's a scarcity of corpora for emotion recognition.

III. Proposed methodology

In this paper we intend to propose an effective method for speaker emotion recognition. The proposed technique has three main phases namely, feature extraction, feature selection and emotion recognition. Initially we select the input signal from the speech signal database. From the fetched input signal the features are extracted in the first phase. In second phase, the selected features are optimally selected by means of optimization algorithm. After feature selection, the selected features are fed to the emotion recognition technique for recognizing various speaker emotions such as Anger, Surprise, Fear, Happiness, Sadness, Disgust and Neutral state. The overall flow diagram is shown in fig.1

The overall procedure of the proposed work is classified into three important steps such as,

1. Feature Extraction
2. Feature Selection
3. Emotion Recognition

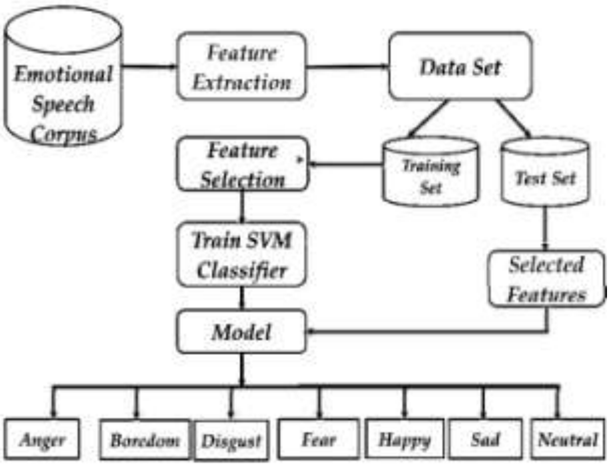

Fig.3. Block Diagram of Speech Emotion Recognition

Low Level Features for Emotion Recognition Using HMM:

In our work we've got decided to use those raw features that would result in statistical measures a lot of the same as those projected in literature for feeling recognition. We first considered the following possible features: short time pitch and energy, the contours of pitch and energy, the spectral shape, and duration and silence related measures. Of these measures we have only tried four: the instant values and contours of pitch and energy. These features are easy to estimate in real time frameworks, while being known to carry a large amount of information about the emotional state. Their means, standard deviations, etc. are measures proposed in almost each work related to emotion recognition. Besides, if careful chosen, pitch and energy features can be made quite robust to channel distortion, speaker, sex and even language.

Spectral measures were discarded in our first approach to emotion recognition because they need complex frameworks to be characterized. This is thus as a result of the spectrum depends heavily on the phonetic content of the sentence. Pitch and energy also do, but we can expect them to depend only on broad classes of sounds, rather than on phonemes. Another reason for discarding spectral measures is that their phonetic dependency would be a main drawback for building language independent feeling recognizers. Although many often named, we also discarded direct use of temporal and silence related measures because they need a previous recognition step in order to get a phone/silence segmentation/ recognition, increasing the complexity of the overall system.

Yet, the HMM structure beside sensible an honest {decent} alternative of the pitch and energy features can give a quite good illustration of this type of measures. Articulator rate, and frequency and duration of silences, for instance, will have direct implications of pitch and energy, and their derivatives. Absolute values and long term evolution of some parameters square measure avoided because of their dependency on factors that don't have anything to try and do with the speaker’s emotional state.

For instance, the absolute value of energy reflects not only the intentional level, but the sex and age of the speaker and the gain of the recording chain as well. On the other hand, whether a sentence is affirmative or interrogative, or its length, will probably play a determinant role on the whole sentence contour of pitch. For both energy and pitch, we have a tendency to think about two types of temporal scope: instantaneous values and syllabic contour

Speech recognition with support vector machines in a hybrid system:

While the temporal dynamics of speech are often represented very efficiently by Hidden markov Models (HMMs), the classification of speech into single speech units (phonemes) is sometimes done with Gaussian mixture models which do not discriminate well. Here, we use Support Vector Machines (SVMs) for classification by integrating this method in a HMM-based speech recognition system. In this hybrid SVM/HMM system we tend to translate the outputs of the SVM classifiers into conditional probabilities and use them as emission probabilities in an exceedingly HMM-based decoder. SVMs are terribly appealing because of their association with statistical learning theory. They have already shown excellent classification results indifferent fields of pattern recognition .We train and test our hybrid system on the darpa Resource Management (RM1) corpus. Our results show better performance than HMM-based decoder using Gaussian mixtures.

Energy features:

In order to model the instantaneous values of energy without relying on the absolute value of energy we have a tendency to use the primary and second derivatives of the power of the mean energy in the frame. The acoustical meaning of those measures is said to the sharpness of the energy state, reflecting each the articulation speed and also the dynamic range. Besides, effects such a stream or small frequent variations in voice intensity are also easily characterized by the instantaneous energy levels.

Pitch features:

Pitch features present a similar behavior as energy ones. In this case we are neither interested in the global pitch which is heavily influenced by the speakers nature, nor its global evolution along the utterance which will depend on the sentence structure. Besides, we can expect the syllabic contour of pitch and its instantaneous levels to provide profitable information about the emotion. In order to characterize instantaneous pitch, a simple autocorrelation analysis is performed at every frame. The maximum of the long term auto-correlation is determined and used to form five different parameters: the value of the maximum of the long term auto-correlation, along with its first and second derivatives; and the first and second derivatives of the logarithm of the pitch log.

Processing of the derived features:

The features are freed of their mean value and normalized to their standard deviation. They are classified by single state HMMs (GMM), which are able to approximate the probability distribution, function of each derived feature by means of a mixture of Gaussian distributions. Up to four mixtures have been used. No further gain could be observed exploitation more than these. Each emotion is sculptured by one GMM in our approach. The maximum probability model are considered because the recognized emotion at a time throughout the recognition process.

Classification of System

In the speech emotion recognition system after calculation of the features, the most effective features are provided to the classifier. A classifier recognizes the emotion within the speaker’s speech utterance. Various kinds of classifier have been proposed for the task of speech emotion recognition. Gaussian Mixtures Model (GMM), K-nearest neighbours (KNN), Hidden Markov Model (HMM) and Support Vector Machine (SVM), Artificial Neural Network (ANN), etc. are the classifiers used in the speech emotion recognition system. Each classifier has some advantages and limitations over the others. In speech recognition system like isolated word recognition and speech emotion recognition, hidden markov model is generally used; the main reason is its physical relation with the speech signals production mechanism.

In speech emotion recognition system, HMM has achieved great success for modelling temporal information in the speech spectrum. The HMM is doubly stochastic process consist of first order markov chain whose states are buried from the observer. For speech emotion recognition usually one HMM is trained for every emotion associated an unknown sample is classed in line with the model that illustrates the derived feature sequence best.

HMM has the important advantage that the temporal dynamics of speech features can be caught second accessibility of the well-established procedure for optimizing the recognition frame work. The main downside in building the HMM based mostly recognition model is that the features choice method. Because it's not enough that features carries info concerning the emotional states, however it should match the HMM structure further.

Usually, in the literature of the field, a Support Vector Machine (SVM) is used to classify sentences. SVM is a relatively new machine learning algorithm introduced by Vapnik [11] and derived from statistical learning theory in the 90s.The main idea is to transform the original input set into a high dimensional feature space by using a kernel function and then, to achieve optimum classification in this new feature space, where a clear separation among features obtained byte optimal placement of a separation hyper plane under the precondition of linear reparability. Differently from the previously proposed approaches two different classifiers, both kernel-based Support Vector Machines (SVMs), have been employed in this paper. Neural networks are chosen for the solution because a basic formula cannot be devised for the problem. The neural networks are also quick to respond which is a requirement as the emotion should be determined almost instantly. The training takes a long time but is irrelevant as the training is mostly done off-line. Deep learning has been applied to SER in prior work, as discussed. However, with different data subsets and under various experiment conditions involved in prior studies, it is difficult to directly compare various deep learning models. To the best of our knowledge, our work provides the first empirical exploration of various deep learning formulations and architectures applied to SER. As a result, we report state-of-the-art results on the popular Interactive Emotional Dyadic Motion Capture (IEMOCAP) database (Busso et al., 2008) for speaker independent SER

IV. Results

In this paper, an HMM based approach to emotion recognition has been presented. Results a good accuracy confirm both the usefulness of the approach and the convenience of the low level features used. Given the reduced scope of the scenario considered it may be arguable if the results can be generalized to other speakers and/or languages, yet we believe that the results achieved are encouraging: at least, they show the usefulness of the approach for multi-speaker emotion recognition besides, we believe that this is a good baseline for more ambitious tasks. Since the introduced approaches tend to strongly depend on the speaker, no cross-speaker evaluation results are presented.

Classification	Happiness	Neutral	Boredom	Sadness	Anger
Happiness	99.7	0	0	0	1.7
Neutral	2.9	90.5	1.5	0	0
Boredom	0	9.5	91.0	11.4	0
Sadness	0	0	6.0	88.6	0
Anger	4.5	0	0	0	90.1

IV. Conclusion

The proposed system, able to recognize the emotional state of a person starting from audio signals registrations, is com-posed of two functional blocks: Gender Recognition (GR) and Emotion Recognition (ER). The former has been implemented by a Pitch Frequency Estimation method, the latter by two Support Vector Machine (SVM) classifiers (fed by properly selected audio features), which exploit the GR subsystem output. Various deep learning architectures were explored on a Speech Emotion Recognition (SER) task. Experiments conducted illuminate how feed-forward and recurrent neural network architectures and their variants could be employed for paralinguistic speech recognition, particularly emotion recognition. Convolution Neural Networks (Convents) demonstrated better discriminative performance compared to other architectures.

Reference

[1]. Kamran Soltani and Raja Noor Ainon. Speech emotion detection based on neural networks. In 9th International Symposium on Signal Processing and its Applications, 1 4244-0779-6/07, IEEE, 2007.

[2]. Jouni Pohjalainen and Paavo Alku. Multi-scale modulation filtering in automatic detection of emotions in telephone speech. International Conference on Acoustic, Speech and Signal Processing, 980-984, 2014.

[3]. https://www.irjet.net/archives/V3/i4/IRJET-V3I460.pdf

[4]. http://www.ijesit.com/Volume

[5]. https://www.sciencedirect.com/science/article/pii/S089360801730059X