International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 04 Issue: 07 | July -2017 www.irjet.net p-ISSN: 2395-0072

EFFICIENT SPEECH EMOTION RECOGNITION USING SVM AND DECISION TREES

T. V. Vamsikrishna[1], P. Naga vyshnavi[2]

[1]Assitant Professor, Computer Science and Engineering Department, Vignan's lara, Andhra Pradesh , India,
[2]Student,Computer Science and Engineering Department, Vignan's lara , Andhra Pradesh, India

---

Abstract - *Speech emotion recognition(SER) is a hot research topic in the field of Human Computer Interaction (HCI). The interaction between human beings and computers will be more natural if computers are able to perceive and respond to human non-verbal communication such as emotions. In this paper- Speech Emotion Recognition Using Binary Support Vector Machines- seven emotional states are considered: anger, boredom, disgust, fear, happy, sad and neutral. The speech features extracted are variance, standard deviation, energy, pitch, timing, etc. Emotional Speech Corpus used in this work is Emo-DB and it contains 535 speech segments from 10 people, 5 male and 5 female. The speech segments are given as input to openSMILE for feature extraction. The extracted features may contain irrelevant features, so feature selection algorithms are used to eliminate those irrelevant features. The selected feature set is divided into training set and test set. Training set is used to train the classifier and the performance of the classifier is evaluated over test set. The overall evaluation results of the classifier over test set is 78.9% and the accuracy of the classifier over training set is 92.6%. Thus the average accuracy of the classifier is 85 percentage.*

***Key Words*: SER; Telugu Emo-DB; SVM**

1.INTRODUCTION

Speech is one of the most fundamental and natural communication means of human beings. With the exponential growth in available computing power and significant progress in speech technologies, spoken dialogue systems (SDS) have been successfully applied to several domains.

The goal of affective interaction via speech, several problems in speech technologies, including low accuracy in recognition of highly affective speech and lack of affect-related common sense and basic knowledge, still exist. So, in order to accomplish the goal of affective communication through speech, emotions are considered.

Emotions are fundamental for humans, impacting perception and everyday activities such as communication, learning, and decision making. They can be expressed through speech, facial expressions, gestures, and other nonverbal clues. For a better communication, emotions of the other person need to be recognized.

Many applications can benefit from an accurate emotion recognizer. For example, customer care interactions (with a human or an automated agent) can use emotion recognition systems to assess customer satisfaction and quality of service (e.g., lack of frustration).

For Speech Emotion Recognition, a better quality speech corpus is very important. From a lot of pre-recorded emotional speech corpuses, a German database, Emo-DB was selected to be used in this project. The speech samples from the corpus was given to a speech processing tool like Open SMILE to extract features like pitch, loudness, variance, and standard deviation, etc.

Selected features are used to train a classifier for seven different classes (anger, boredom, and disgust, fear, happy, sad, and neutral) of emotions. The classifier used in this project is Support Vector Machine (SVM). The classifier is then tested using a validation set to assess the recognition performance.The main stages in the proposed system are Input data collection, Preprocessing, Feature extraction, Feature selection, Classification and Recognition.

2.SPEECH EMOTION recognition(SER)

SER is the Speech signal is the fastest and most natural method of communication between humans. This fact has motivated researchers to think of speech as a fast and efficient method of interaction between human and machine.It involves the data collection, preprocessing, feature extraction, feature selection, classification and recognition. Data collection is a critical task in speech emotion recognition [2], and the collected data may contain lot of noises that must be cleaned in the preprocessing phase which involves filter,wrapper and embedded. The output obtained in the preprocessing phase is given as input to the next feature extraction phase.

For classification Support vector machine is used and the selected features are trained in LIBSVM, a tool for SVM classifier. Classification can be thought as of two separate problems like binary and multiclass classification. In binary classification only two classes are involved, whereas

multiclass classification often requires the combined use of multiple binary classifiers.

3.EMOTIONAL SPEECH CORPUSES

As mentioned in the previous chapters, there are many emotional speech corpuses. Selecting a better quality speech corpus is very important for an efficient Speech Emotion Recognition System as a low quality database may degrade the performance of our system.From the variety of existing emotional databases, a German database, Emo-DB was selected for this project. This particular database seems very prominent in its quality.Emo-DB contains 535 speech samples spoken by 10 speakers, 5 male and 5 female in seven basic emotions such as anger, boredom, disgust, fear, happy, sad and neutral.

As per literature, the first reported work on SER the database was recorded at the Faculty of Electrical Engineering and Computer science,University of Maribor,Slovenia. It contains emotional speech in six emotion categories, such as disgust, surprise, joy, fear, anger and sadness. Two neutral emotions were also included: fast loud and low soft. It is seen that the emotion categories are compliant with MPEG-4. Four languages (i.e. English, Slovenian, French and Spanish) were used in all speech recordings. The database contains 186 utterances per emotion category.

Emo-DB, this German database of emotional utterances was recorded in 1997 and 1999 spoken by actors. It contains 535 samples of speech spoken by 5 male and 5 female speakers each 10 sentences of different ages in seven different emotions anger, boredom, and disgust, fear, happy, sad and neutral.

W. F. Sendlmeier et al. at the Technical University of Berlin constructed another emotional speech database. The database consists of emotional speech in seven emotion categories. Each one of the ten professional actors expresses ten words and five sentences in all the emotional categories. The corpus was evaluated by 25 judges who classified each emotion with a score rate of 80%.

Nakatsu et al. at the ATR Laboratories constructed an emotional speech database in Japanese. The database contains speech in 8 emotion categories. The project employed 100 native speakers (50 male and 50 female) and one professional radio speaker. The professional speaker was told to read 100 neutral words in 8 emotional manners. The 100 ordinary speakers were asked to mimic the manner of the professional actor and say the same amount of words. The total amount of utterances is 80000 words.

F. Yu et al. at the Microsoft Research China recorded emotional speech database. It contains speech segments from Chinese teleplays in four emotion categories, such as anger, happiness, sadness, and neutral. Four persons tagged the2000 utterances. Each person tagged all the utterances. When two or more persons agreed in their tag, the utterance got their tag. Elsewhere the utterance was thrown away. After tagging several times, only 721 utterances remained.

Letter	Emotion
W	Anger
L	Boredom
E	Disgust
A	Fear
F	Happy
T	Sad
N	Neutral

Fig1.Coding of emotions in Emo- DB

4. PROPOSED WORK

An important issue in speech emotion recognition is the extraction of speech features that efficiently characterize the emotional content of speech and at the same time do not depend on the speaker or the lexical content. Although many speech features have been explored in speech emotion recognition, researchers have not identified the best speech features for this task. And since pattern recognition techniques are rarely independent of the problem domain, it is believed that a proper selection of features significantly affects the classification performance. Speech features can be grouped into four categories: continuous features, qualitative features, spectral features, and TEO (Teager energy operator)-based features.

4.1. pitch

The pitch signal, also known as the glottal waveform, has information about emotion. Two features related to the pitch signal are widely used, namely the pitch frequency and the glottal air velocity at the vocal fold opening time instant. The time elapsed between two successive vocal fold openings is called pitch period T, while the vibration rate of the vocal folds is the fundamental frequency of the phonation F0 or pitch frequency.

4.2. Teager energy operator

Another useful feature for emotion recognition is the number of harmonics due to the nonlinear air flow in the vocal tract that produces the speech signal. In the emotional state of anger or for stressed speech, the fast air flow causes vortices located near the false vocal folds providing

additional excitation signals other than the pitch [1]. The additional excitation signals are apparent in the spectrum as harmonics and cross-harmonics.

4.3. Vocal tract features

The shape of the vocal tract is modified by the emotional states. Many features have been used to describe the shape of the vocal tract during emotional speech production. Such features include

– The formants which are a representation of the vocal tract resonances,
– The cross-section areas when the vocal tract is modeled as a series of concatenated lossless tubes,
– The coefficients derived from frequency transformations.

The formants are one of the quantitative characteristics of the vocal tract. In the frequency domain, the location of vocal tract resonances depends upon the shape and the physical dimensions of the vocal tract. Since the resonances tend to "form" the overall spectrum, speech scientists refer to them as formants. Each formant is characterized by its center frequency and its bandwidth. The formant bandwidth during slackened articulated speech is gradual, whereas the formant bandwidth during improved articulated speech is narrow with step flanks.

4.4. Speech energy

The short-term speech energy can be exploited for emotion recognition, because it is related to the arousal level of emotions [1].

5 PREPROCESSING TECHNIQUES

The speech processing tool involves four basic operations: spectral shaping, feature extraction, parametric transformation, and statistical modeling. Feature extraction is process of obtaining different features such as power, pitch, and vocal tract configuration from the speech signal. Parameter transformation is the process of converting these features into signal parameters through process of differentiation and concatenation. Statistical modelling involves conversion of parameters in signal observation vectors.

In this tool is capable of loading any sound file saved in WAV file format. After the loading of the file in the computer memory, the operations applied to the speech signal are pre-emphasis, frame blocking, windowing, lpc analysis. The main advantage of the WinSPT tool is that it supports the not only processing of a single file but also of a group of files in a batch mode operation. After the loading of the selected WAV file, the program automatically performs all the speech processing steps.

5.1. OpenSMILE

The Munich Open Speech and Music Interpretation by Large Space Extraction (openSMILE)Toolkit is a modular and it is feature extractor for signal processing and machine learning applications. The primary focus is clearly put on audio-signal features. openSMILE features platform independent live audio input and live audio playback, which enabled the extraction of audio features in real-time.

To facilitate interoperability, openSMILE supports reading and writing of various data formats commonly used in the field of data mining and machine learning. These formats include PCM WAVE for audio files, CSV (Comma Separated Value, spreadsheet format) and ARFF (Weka Data Mining) for text-based data files, HTK (Hidden-Markov Toolkit) parameter files, and a simple binary float matrix format for binary feature data. Open SMILE is used by researchers and companies all around the world, which are working in the field of speech recognition (feature extraction front-end, keyword spotting, etc.),the area of affective computing (emotion recognition, affect sensitive virtual agents, etc.), and Music Information Retrieval (chord labeling, beat tracking, onset detection etc).

5.2. SPICE Toolkit

Speech Processing - Interactive Creation and Evaluation, a web based toolkit for rapid language adaptation to new languages and RLAT (Rapid Language Adaptation Toolkit), an extension to SPICE for web harvesting and language model evaluation. The methods and tools implemented in SPICE and RLAT will enable the attendees to develop speech processing components, to collect appropriate data for building these models,. By archiving the data gathered on-the-fly from many cooperative users to significantly increase the repository of languages and resources and make the data and components for under-supported languages available at large to the community.

Among all these speech processing tools we worked on openSMILE because it is open source software which lets users to extract needed features such as frequency, loudness, energy, pitch etc. from the speech signal.

6. Feature extraction

In feature extraction, many tools are there like Open SMILE, SPICE toolkit, and so on. OpenSMILE is used in this project. The OpenSMILE feature extraction tool enables you to extract large audio feature spaces in real-time. It combines features from Music Information Retrieval and Speech Processing. It is written in C++ and is available as both a standalone command line executable as well as a dynamic

library. The main features of open SMILE are its capability of on-line incremental processing and its modularity. Feature extractor components can be freely interconnected to create new and custom features, all via a simple configuration file. New components can be added to openS MILE via an easy binary plugin interface and a comprehensive API.

6.1. Curse of Dimensionality

The **curse of dimensionality** refers to various phenomena that arise when analyzing and organizing data in high-dimensional spaces (often with hundreds or thousands of dimensions) that do not occur in low-dimensional settings such as the three-dimensional physical space of everyday experience [20].The "curse of dimensionality" is not a problem of high-dimensional data, but a joint problem of the data and the algorithm being applied. In machine learning problems that involve learning a "state-of-nature" (maybe an infinite distribution) from a finite number of data samples in a high-dimensional feature space with each feature having a number of possible values. When facing the curse of dimensionality, a good solution can often be found by changing the algorithm, or by pre-processing the data into a lower-dimensional form.

6.2. Dimensionality Reduction

In machinelearning and statistics, dimensionalityreduction o r dimension reduction is the process of reducing the number of random variables under consideration, and can be divided into feature selection and feature extraction. For high-dimensional datasets (i.e. with number of dimensions more than 10), dimension reduction is usually performed prior to applying a K-nearest neighbours algorithm(k-NN) in order to avoid the effects of the curse of dimensionality [21].

Feature extraction and dimension reduction can be combined in one step using principal component analysis (PCA), lineardiscriminant analysis (LDA).In machine learning this process is also called low-dimensional embedding [21].For very-high-dimensional datasets (e.g. when performing similarity search on live video streams, DNA data or high-dimensional Time series) running a fast approximate K-NNsearch using locality sensitive hashing, “random projections", "sketches" or other high-dimensional similarity search techniques from the VLDB toolbox might be the only feasible option [21].

6.3. Feature Selection: CFS

CFS is a simple filter algorithm that ranks feature subsets according to a correlation based heuristic evaluation function. The bias of the evaluation function is toward subsets that contain features that are highly correlated with the class and uncorrelated with each other. Irrelevant features should be ignored because they will have low correlation with the class. Redundant features should be screened out as they will be highly correlated with one or more of the remaining features.

$$Merit_{S_k} = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}}.$$

The implementation of CFS used in the experiment in this project allows the user to choose from three heuristic search strategies: forward selection, backward elimination, and best first. Forward selection begins with no features and greedily adds one feature at a time until no possible single feature addition results in a higher evaluation. Backward elimination begins with the full feature set and greedily removes one feature at a time as long as the evaluation does not degrade.

6.4. Learning and Generalization of SVMs

Early machine learning algorithms aimed to learn representations of simple functions and the goal of learning was to output a hypothesis that performed the correct classification of the training data and early learning algorithms were designed to find such an accurate fit to the data. SVM performs better in term of not over generalization when the neural networks might end up over generalizing easily. Another thing to observe is to find where to make the best trade-off in trading complexity with the number of epochs; the illustration brings to light more information about this. The below illustration is made from the class notes.

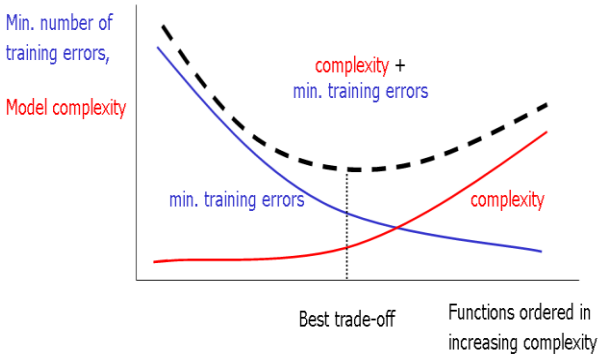

Fig.2 Number of Epochs Vs Complexity

7.Feature selection

In machine learning and statistics, feature selection, also known as variable Selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features for use in model construction. The central assumption when using a feature selection technique is that the data contains many redundant or irrelevant features. Feature selection techniques are to be distinguished from feature extraction. Feature selection techniques provide three main benefits when constructing predictive models:

- improved model interpretability,
- shorter training times,
- Enhanced generalisation by reducing over fitting.

Feature selection is also useful as part of the data analysis process, as it shows which features are important for prediction, and how these features are related. A feature selection algorithm can be seen as the combination of a search technique for proposing new feature subsets, along with an evaluation measure which scores the different feature subsets. The simplest algorithm is to test each possible subset of features finding the one which minimises the error rate.

7.1. Subset selection

Subset selection evaluates a subset of features as a group for suitability. Subset selection algorithms can be broken up into Wrappers, Filters and Embedded. Wrappers use a search algorithm to search through the space of possible features and evaluate each subset by running a model on the subset. Wrappers can be computationally expensive and have a risk of over fitting to the model. Filters are similar to Wrappers in the search approach, but instead of evaluating against a model, a simpler filter is evaluated. Embedded techniques are embedded in and specific to a model .Alternative search-based techniques are based on targeted projection pursuit which finds low-dimensional projections of the data that score highly: the features that have the largest projections in the lower-dimensional space are then selected.

7.2. Entropy

In information theory, entropy is the average amount of information contained in each message received. Here, *message* stands for an event, sample or character drawn from a distribution or data stream. Entropy thus characterizes our uncertainty about our source of information. Shannon defined the entropy H (Greek letter Eta) of a discrete random variable *X* with possible values {*x*1... *xn*} and probability mass function P(*X*) as:

$$\mathrm{H}(X) = \mathrm{E}[\mathrm{I}(X)] = \mathrm{E}[-\ln(\mathrm{P}(X))].$$

Here E is the expected value operator, and I is the information content of *X*. (*X*) is itself a random variable.When taken from a finite sample, the entropy can explicitly be written as

$$\mathrm{H}(X) = \sum_i \mathrm{P}(x_i)\mathrm{I}(x_i) = -\sum_i \mathrm{P}(x_i)\log_b \mathrm{P}(x_i),$$

k

Where *b* is the base of the logarithm used. Common values of *b* are 2, Euler's number e, and 10, and the unit of entropy is Shannon for *b* = 2, nat for *b* = *e*, and Hartley for *b* = 10.When *b* = 2, the units of entropy are also commonly referred to as bits.In the case of *p* (*xi*) = 0 for some *i*, the value of the corresponding summand 0 log*b* (0) is taken to be 0, which is consistent with the limit:

$$\lim_{p\to 0^+} p\log(p) = 0.$$

One may also define the conditional entropy of two events *X* and *Y* taking values *xi* and *yj* respectively, as

$$\mathrm{H}(X|Y) = \sum_{i,j} p(x_i, y_j)\log\frac{p(y_j)}{p(x_i, y_j)}$$

Where *p* (*xi*, *yj*) is the probability that *X* = *xi* and *Y* = *yj*. This quantity should be understood as the amount of randomness in the random variable *X* given that you know the value of Y.

7.3. Information Gain

Information gain (IG) measures the amount of information in bits about the class prediction, if the only information available is the presence of a feature and the corresponding class distribution. Concretely, it measures the expected reduction in entropy (uncertainty associated with a random feature). Given SX the set of training examples, xi the vector of ith variables in this set, |Sxi=v|/|SX| the fraction of examples of the ith variable having value v:

IG (SX, xi) = H (SX) − |Sxi=v| |S X| v=values (xi) H (Sxi=v)

with entropy:

H(S) = −p+(S) log2 p+(S) − p−(S) log2 p−(S) p±(S) is the probability of a training example in the set S to be of the positive/negative class.

7.4. Correlation Feature Selection

The Correlation Feature Selection (CFS) measure evaluates subsets of features on the basis of the following hypothesis: "Good feature subsets contain features highly correlated with the classification, yet uncorrelated to each other". The following equation gives the merit of a feature subset *S* consisting of *k* features:

Here, $\overline{r_{cf}}$ is the average value of all feature-classification correlations, and $\overline{r_{ff}}$ is the average value of all feature-feature correlations. The CFS criterion is defined as follows:

$$\mathrm{CFS} = \max_{S_k}\left[\frac{r_{cf_1} + r_{cf_2} + \cdots + r_{cf_k}}{\sqrt{k + 2(r_{f_1f_2} + \cdots + r_{f_if_j} + \cdots + r_{f_kf_1})}}\right].$$

The r_{cf_i} and $r_{f_if_j}$ variables are referred to as correlations, but are not necessarily Pearson's correlation

coefficient or Spearman's ρ. Dr. Mark Hall's dissertation uses neither of these, but uses three different measures of relatedness, minimum description length (MDL), symmetrical uncertainty, and relief.Let *xi* be the set membership indicator function for feature *fi;* then the above can be rewritten as an optimization problem:

$$\mathrm{CFS} = \max_{x\in\{0,1\}^n}\left[\frac{\left(\sum_{i=1}^{n} a_i x_i\right)^2}{\sum_{i=1}^{n} x_i + \sum_{i\neq j} 2b_{ij}x_i x_j}\right].$$

The combinatorial problems above are, in fact, mixed 0–1 linear programming problems that can be solved by using branch-and-bound algorithms.

8.Classification

SVMs (Support Vector Machines) are a useful technique for data classification. A classification task usually involves separating data into training and testing sets. Each instance in the training set contains one target value (i.e. the class labels) and several attributes (i.e. the features or observed variables). The goal of SVM is to produce a model (based on the training data) which predicts the target values of the test data given only the test data attributes .

9.RESULTS

The performance evaluation of the SVM was done with the multiclass classifier and Binary classifier over test set. The lower accuracy indicates that the class is more subjective quality and is more difficult to classify.
The evaluation results of SVM multi-class classifier are as follows

- Test set: 78.9%
- Training set: 92.6%
- Average: 85%

Binary Classifier	TRAIN	TEST	AVERAGE
Anger-Not Anger	97.85	98.46	98.12
Boredom- Not Boredom	100	97.22	98.61
Disgust-Not Disgust	100	95.65	97.82
Fear-Not Fear	100	91.42	95.71
Happy-Not Happy	100	97.29	98.64
Sad-Not Sad	100	96.87	98.43
Neutral-Not Neutral	100	97.43	98.71

Fig3. Results of Binary classifier

The Binary classifier implemented was based on plain majority voting. We are further trying to improve the accuracy by implementing based on weighted majority.

10. CONCLUSION AND FUTUREWORK

The current work has been aimed to present a high-accuracy training and classification framework for emotion recognition from speech. The process of speech emotion recognition requires the creation of a reliable database to extract features from the speech corpus. The accuracy of the system, is highly depends on emotional speech database used in the system therefore it is necessary to record correct emotional speech database. In our research, the database used was a pre-recorded one. For a better result one can use their own emotional speech corpuses. Therefore, the next generation of human-computer interfaces might be able to perceive humans feedback, and respond appropriately and opportunely to changes of user's affective states for improving the performance of results.

Three important issues have been studied: the features used to characterize different emotions, the classification techniques, and the important design criteria of emotional speech databases. There are several conclusions that can be drawn from this study. Most of the current body of research focuses on studying many speech features and their relations to the emotional content of the speech utterance. New features have also been developed such as the TEO-based features. There are different studies are not consistent. The main reason may be attributed to the fact that only one emotional speech database is investigated in each study.

FUTUREWORK

The speech emotional corpus used in this project was a pre-recorded one. It can built for a better result. In this project, only SVM classifier was used. For a better result one can use other possible classifiers like HMM, ANN, Decision Trees, and so on.CFS algorithm can be implemented instead of using it from Weka tool for feature selection. The pairwise classification is done with the plain majority voting of all the pairwise machines. Another possibility is using a weighted majority for more accurate results.

REFERENCES

[1] Moataz El Ayadi, Mohamed S. Kamel,FakhriKarray, "Survey on speech emotion recognition: Features, classification schemes, and databases", Pattern Recognition 44 (2011) 572–587.
[2] Tomas Pfister and Peter Robinson, "Real-Time Recognition of Affective States from Nonverbal Features of Speech and Its Application for Public Speaking Skill Analysis", IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, VOL. 2, NO. 2, APRIL-JUNE 2011.

[3] Dimitrios Ververidis and Constantine Kotropoulos, "Emotional speech recognition: Resources, features, and methods", Artificial Intelligence and Information Analysis Laboratory, Department of Informatics, Aristotle University of Thessaloniki, Box 451, Thessaloniki 541 24, Greece..

[4] Chi-Chun Lee, Emily Mower, Carlos Busso, SungbokLee,Shrikanth Narayanan, "Emotion recognition using a hierarchical binary decision tree approach", Speech Communication 53 (2011) 1162–1171.