International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Hybrid Classifiers for Gender Driven Emotion Recognition

Pravina Ladde

Savitribai Phule University, STE'S Smt. Kashibai Navale College of Engineering, Sinhagad Road, Pune 411041, India

Abstract: ***Recognizing human emotions by registering speech signals is always an interesting field in Artificial Intelligence. This paper describes a system which makes 1] gender recognition first, to get apriori knowledge about the speaker. 2] Then it uses hybrid of Hidden Markov Model [HMM] and Support Vector Machine [SVM] to make classification of emotions. This proposed system combines advantages of the classifiers to give more accurate results. Here HMM is used to model speech feature sequence i.e. this system is trained using HMM algorithm for considered emotion whereas SVM is used to make decision i.e. for classification. Literature survey and past results indicates that Gender Recognition apriori knowledge and use of hybrid HMM-SVM algorithms will considerably increase accuracy of emotion recognition.***

Keywords: Hybrid Classifiers, Multiple classifier systems, Hidden Markov model**,** Support vector machine

1. Introduction

Recognizing people emotional state and giving a suitable feedback may play crucial role in some context of Human Computer Intelligent Interaction [HCII] eg- giving a favorable response according to speaker's emotional state. Nowadays there is lot of work done to improve human computer interaction and to get the Human Copmuter Intelligent Interaction effectively Computers should be able to interact with users very naturally i.e. this interaction would be likeli as Human-Human interactions. HCII is becoming very necessary in all aspects of future peoples life, virtual reality, Smart phones, Smart Offices, Smart Homes. Emoton recognition is a growing research area in Academy and Industry fields. Some of the successfull products from this research area generally shows emotion recognition by facial or voice features. But this paper captures emotion state of a person by registering the speech signals in surrounding obtained from devices as Smart phones which can be implemented in Smart Environment easily. This paper describes Emotion recognition of six emotions [Happiness, Anger, boredom, Sadness, disgust, fear]. The main contribution of this paper is i] Gender Recognition Subsystem based on pitch of the input voice which provides additional knowledge of gender of the speaker. ii] SVM emotion classifier with the employed database named Reading Leads database to store speech samples. iii] Combined features of HMM & SVM Algorithm This system uses HMM for training considered emotions, while SVM is employed to make decisions i. e. for classification[3]

2. Related Work

Traditional speech based emotion recognition systems have worked according to following 4 main parts of a Emotion Recognition System as Feature Extraction, Feature Reduction and Selection, various Databases, different Classification schemes.

2.1 Feature Extraction

As structure of human vocal tract is same for every human being, speech samples obtained will be highly redundant. But there are some unique features associated with the specific speaker which are much useful to characterize different emotions such as pitch of the specific speaker, accent of the speaker, duration taken to make particular utterance, formant, energy used to make the utterance, his speaking rate, Mel frequency cepstrum coefficient [MFCC], Linear prediction cepstrum coefficient[LPCC] [4] etc. So extraction of such distinguishing features is very much important to efficiently recognize different emotions.

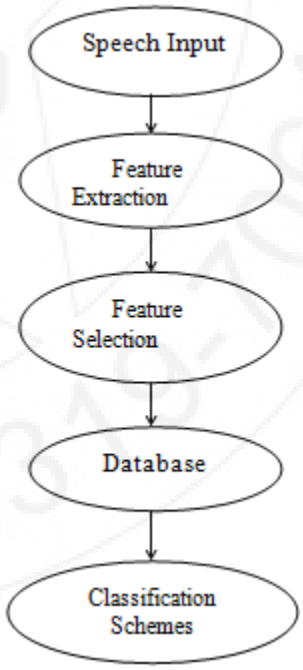

Figure 1: Proposed Structure of the Speech Emotion Recognition System

2.2 Feature Selection

Feature Extraction collects all the basic features of the speech. Among them all features may not be useful for

further processing and may cause additional load for the used classifier. So we need to remove such unwanted feature from extracted base features and also it is necessary to make a proper selection of some useful features. Feature selection reduces the number of base features used.

Forward Selection Method – In this method a single feature is selected as a best feature from the extracted base features. Remaining extracted features are further added to increase classification accuracy. The selection process stops, once the added features attains a preset number.

2.3 Database

Database plays a very important role in speech emotion recognition. It is a collection of emotions represented in sentences format. Strength of database also helps to increase the accuracy of recognized emotions. The discussion about database arises four questions. First, what should be the scope of speech and emotion database will be, Second which type of speeches can be used – natural or acted, Third what type of context additional information need to be provided [such as verbal, facial etc.] with the vocal signs/speeches, fourth which descriptors we should attach to the speech and to the emotional content of the database [5]. Some of the databases which answer few of above questions are as - Reading-Leeds Database, Belfast Database, CREST-ESP [Expressive Speech Database], Berlin Emotional Speech [BES][5].

a) Reading Leeds Database

This database is aimed to meet the apparent need for a large, well-annotated body of natural or near natural speech stored in an orderly way on computer. It focuses on 3 broad areas : First, it identified types of natural material where phonetic marking of emotion was (and was not) evident. Second, it established some broad characteristics of that kind of material. Third, it developed principled techniques for annotating both the emotional content of the material and the features of speech that might carry emotional information. The essential aim of the database is to collect speech that was genuinely emotional rather than acted or simulated[5].

b) The Belfast database

The aim of the database is to develop a system capable of recognising emotion from facial and vocal signs. The system was to be based on hybrid computing, i.e. a combination of neural net techniques and traditional symbolic computing. The core function of the data was to train the neural net component. It was assumed that the system was unlikely to achieve real-world applications unless the training material was naturalistic. Hence, collection was guided by four principles. (i) The material should be spoken by people who at least appeared to be experiencing genuine emotion. (ii) The material should be derived from interactions rather than from reading authored texts, even in a genuinely emotional state. (iii) The primary concern was to represent emotional states of the type that occur in everyday interactions rather than archetypal examples of emotion (such as full-blown fear or anger). (iv) The material collected was audio–visual as opposed to audio alone. The decision was partly driven by the specific needs of the PHYSTA project, but they converge with general ecological principles in this respect.The ideal goal was that the system should form the same emotional judgements as people would. Hence objective knowledge about a speaker's true emotional state was not considered critical[5].

c) CREST-ESP [Expressive Speech Database]-

Goal for this database is to collect a database of spontaneous, expressive speeches.

Which includes i) collecting a database of spontaneous, expressive speech that meets the requirements of speech technology (particularly concatenative synthesis) ii) statistical modeling and parameterization of paralinguistic speech data iii) developing mappings between the acoustic characteristics of speaking style and speaker-intention or speaker-state; and iv) the implementation of prototypes and testing of the software algorithms developed in (ii) and (iii) in real-world applications[5].

d) Berlin Emotional Speech [BES]-

BES is a public database of acted speeches. The sentences are recorded by 10 German actors (5 male and 5 female) that produced 10 utterances each (5 short and 5 long phrases). The sentences were evaluated by 20 listeners to check the emotional state and only those that had recognition rate of 80% or above were retained, getting about 500 speeches. Additionally, two more perception tests were carried out: one to rate the strength of the displayed emotion for each speech, the other to judge the syllable stress of every speech[5].

2.4 Classification Schemes

Following are the few main classifiers found mostly used in speech Recognition system

1] Hidden markov model

HMM is so long used in much of the speech applications. HMM is the Markov chain of hidden states which contains internal behavior of the model. Hidden states of the model capture temporal data and gives statistical model that describes sequences of events. In HMM temporal data of the speech features is trapped in the state transition matrix [4].The main strong two reasons for more accuracy and efficiency of HMM are, first the models are very rich in mathematical structure so can form theoretical basis for its use in wide range of applications. Second the models, when applied properly, work very well in practice for several important applications

2] Gaussian mixture models

Gaussian mixture model is a probabilistic model for density estimation using a convex combination of multi-variant normal densities. It can be considered as a special continuous HMM which contains only one state. GMMs are very efficient in modeling multi-modal distributions and their training and testing requirements are much less than the requirements of a general continuous HMM. Therefore, GMMs are more appropriate for speech emotion recognition

when only global features are to be extracted from the training utterances[2].

3] Neural networks

Another common classifier, used for many pattern recognition applications is the artificial neural network (ANN). ANNs have some advantages over GMM and HMM. They are known to be more effective in modeling nonlinear mappings. Also, their classification performance is usually better than HMM and GMM when the number of training examples is relatively low. Almost all ANNs can be categorized into three main basic types: MLP, recurrent neural networks (RNN), and radial basis functions (RBF) networks. The latter is rarely used in speech emotion recognition. However, ANN classifiers in general have many design parameters, e.g. the form of the neuron activation function, the number of the hidden layers and the number of neuron in each layer, which are usually set in an ad hoc manner. In fact, the performance of ANN heavily depends on these parameters [2].

4] Support vector machine

SVM is used to classify sentences. The main idea of SVM is to transform the original input set into high dimensional feature space by using a kernel function and then to achieve optimum classification in this new feature space. SVM is a supervised learning process of two steps-

1] Learning [Training] : Learn the model using training data.

2] Testing: Test the model using unseen test data to assess the model accuracy. The standard SVM takes a set of input data and predicts, for each given input, which of two possible classes the input is a member of, which makes the SVM a non-probabilistic binary linear classifier [2].

5] Multiple classifier systems

There are three approaches for combining classifiers : hierarchical, serial, and parallel. In the hierarchical approach, classifiers are arranged in a tree structure where the set of candidate classes becomes smaller as we go in depth in the tree. At the leave-node classifiers, only one class remains after decision. In the serial approach, classifiers are placed in a queue where each classifier reduces the number of candidate classes for the next classifier. In the parallel approach, all classifiers work independently and a decision fusion algorithm is applied to their outputs.

3. Proposed Approach

The general architecture of our Speech Emotion Recognition system has following steps:

1) Our speech processing system extracts some appropriate features from signal.
2) Database is prepared for different emotions in excel spreadsheet.
3) Using HMM algorithm our system is trained in a supervised manner with example data how to associate the features to the different emotions.
4) SVM classifier is used to recognize different emotions by matching the features of uploaded audio file with the features of trained system. It means that System is trained using HMM and tested using SVM classifier i.e. output of HMM is taken as input of SVM for classification. Figure 2 shows the system architecture of proposed system which indicates the exact processing of gender driven speech emotion recognition.

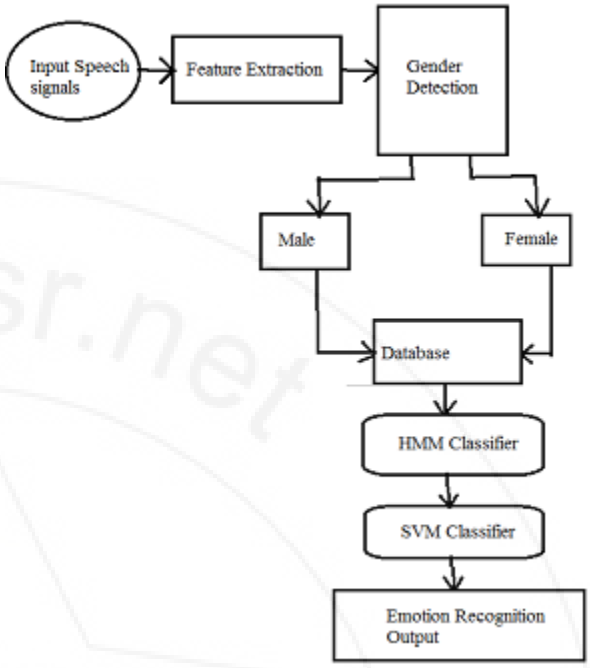

Figure 2: System Architecture of Proposed Approach

4. Mathematical Equation

1] Gender Recognition

Let us consider

S(n) : Speech signal input and

γ_{thr} : Threshold value computed by using the training set.

In particular, given a real-value discrete-time signal s(n); where n € [1. . . N] we have:

$$R(\tau) = \sum_{n=0}^{N-1} S(n)S(n+\tau) \quad \ldots\ldots\ldots (1)$$

Where τ € [0, 1 . . . N-1]. R(τ) in (1) is the autocorrelation of lag τ .

Pitch of a speech signal, due to physiological reasons, is contained in a limited range [P1; P2] (typically P1= 50 [Hz] and P2 = 500 [Hz]) and limits the range between τ1 and τ2, defined in (2).

$$\tau 1 = \left\lfloor \frac{F_s}{P_2} \right\rfloor \text{ and } \tau 2 = \left\lfloor \frac{F_s}{P_1} \right\rfloor \quad \ldots\ldots (2)$$

Fs is the sampling frequency applied to the original analog signal to obtain the Discrete- time signal s(n).In practice, the applied autocorrelation is defined in (3):

$$\hat{R}(\tau) = \sum_{n=0}^{N-1-\tau} S(n)S(n+\tau) \quad \ldots . (3)$$

Where, τ € [$\tau_1, \tau_1 + 1, \tau_1 + 2 \ldots \tau_2$].

The pitch period can be defined as

$\tau_{pitch} = \arg\max \hat{R}(\tau)$

The frequency of pitch is computed as.

$\rho_{pitch} = \frac{F_s}{\tau_{pitch}}$.

ρ_{pitch} is the useful parameter to decide gender of the speaker.

2] HMM Classifier

In HMM a sequence of observable data vectors is x1, . . . ,xT , so we are providing ρ_{pitch} as a observable data vector to the HMM-

$\mathbf{P}(\rho_{pitch\ 1}, S_1 \ldots \rho_{pitch\ T}, S_T) =$

$= \pi_{S_1} B_{S_1}(\rho_{pitch\ 1})\ a_{S_1 S_2} B_{S_2}(\rho_{pitch\ 2}) \ldots\ a_{S_{T-1} S_T} B_{S_T}(\rho_{pitch\ T})$

$= \pi_{S_1} B_{S_1}(\rho_{pitch\ 1}) \prod_{t=2}^{T} a_{S_{T-1} S_T} B_{S_T}(\rho_{pitch\ T})$

where $B_i(\rho_{pitch\ t}) = \mathbf{P}(\rho_{pitch} \mid S_t = i)$.

3] SVM Classifier

SVM classifier is used to recognize different emotions by matching the features of uploaded audio file with the features of trained system and recognizes the correct emotion.

$y(t) = \vec{x}(t).\vec{w} + b$ *with* $\vec{w} = \sum_{i=1}^{N_s} \alpha_i \hat{y}_i \vec{s_i}$

$\vec{x}(t)$ the sequence provided by HMM classifier.

$\vec{s_i}$ and $\hat{y}_i$ are the class labels where as N_s is number of samples. It means that System is trained using HMM and tested using SVM classifier.

5. Conclusion

Emotion recognition systems are based on facial or voice features. This paper proposes a solution, designed to be employed in a Smart Environment, able to capture the emotional state of a person starting from a registration of the speech signals in the surrounding obtained by mobile devices such as smart phones. This paper tells about the system which is able to recognize the 6 emotional states [anger, boredom, disgust, fear, happiness, and sadness] of a person by hybrid of Hidden Markov Models (HMMs) and Support Vector Machines (SVM). Combining advantage on capability of HMM and pattern recognition of SVM. HMMs, which export likelihood probabilities and optimal state sequences, have been used to model speech feature sequences i.e. our proposed system is trained using HMM algorithm for emotions considered, while SVM has been employed to make a decision i.e. for classification.

References

[1] Igor Bisio, Alessandro Delfino, Fabio Lavagetto, Mario Marchese, AND Andrea Sciarrone, “Gender-Driven Emotion Recognition Through Speech Signals for Ambient Intelligence Applications” Date of publication 21 January 2014.

[2] M. El Ayadi, M. S. Kamel, and F. Karray, ``Survey on speech emotion recognition : Features, classification schemes, and databases," Pattern Recognition, vol. 44, no. 3, pp. 572 587, 2011.

[3] Aastha Joshi,“Speech Emotion Recognition Using Combined Features of HMM & SVM Algorithm” ISSN: 2277 128X, Volume 3, Issue 8, August 2013

[4] Speech Emotion Recognition, Ashish B. Ingale, D. S. Chaudhari, International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-1, March 2012.

[5] Ellen Douglas-Cowie,Nick Campbell, Roddy Cowie, Peter Roach ``Emotional speech: Towards a new generation of databases" 2003.

Author Profile

Pravina P. Ladde receiving the Master of Engineering degree in Computer Engineering from STE’S Smt. Kashibai Navale College of Engineering, Savitribai Phule University, Pune during 2013-2015.