MAT JOURNALS

Journal of Image Processing and Artificial Intelligence
Volume 3 Issue 1

Need of Boosted GMM in Speech Emotion Recognition System Implemented Using Gaussian Mixture Model

Prof. A. A. Chaudhari, Dr. M. A. Pund, Dr. G. R. Bamnote, Prof. S. V. Pattalwar
Department of CSE, PRMIT&R, Badenra, India

E-mail: aachaudhari@mitra.ac.in, mapund@mitra.ac.in, grbamnote@mitra.ac.in, svpattalwar@mitra.ac.in

Abstract

Speech feeling recognition is a vital issue that affects the human machine interaction. Automatic recognition of human feeling in speech aims at recognizing the underlying spirit of a speaker from the speech signal. Gaussian mixture models (GMMs) and therefore the minimum error rate classifier (i.e., theorem optimum classifier) is widespread and effective tools for speech feeling recognition. Typically, GMMs are wont to model the class-conditional distributions of acoustic options and their parameters are calculable by the expectation maximization (EM) algorithmic rule supported a coaching information set. During this paper, we have a tendency to introduce a boosting algorithmic rule for faithfully and accurately estimating the class-conditional GMMs. The ensuing algorithmic rule is known as the Boosted-GMM algorithmic rule. Our speech feeling recognition experiments show that the feeling recognition rates are effectively and considerably boosted by the Boosted-GMM algorithmic rule as compared to the EM-GMM algorithmic rule. During this interaction, human beings have some feelings that they want to convey to their communication partner with whom they are communicating, and then their communication partner may be the human or machine. This work dependent on the emotion recognition of the human beings from their speech signal. Emotion recognition from the speaker's speech is very difficult because of the following reasons: Because of the existence of the different sentences, speakers, speaking styles, speaking rates accosting variability was introduced. The same utterance may show different emotions. Therefore, it is very difficult to differentiate these portions of utterance. Another problem is that emotion expression is depending on the speaker and his or her culture and environment. As the culture and environment gets change the speaking style also gets change, which is another challenge in front of the speech emotion recognition system.

Human beings normally used their essential potentials to make communication better between themselves as well as between human and machine. During this interaction, human beings have some feelings that they want to convey to their communication partner with whom they are communicating, and then their communication partner may be the human or machine. This dissertation work dependent on the emotion recognition of the human beings from their speech signal. In this chapter introduction of the speech emotion recognition based on the problem overview and need of the system is provided. Emotional speech recognition aims at automatically identifying the emotional or physical state of a human being from his or her voice. Although feeling detection from speech could be a comparatively new field of analysis, it is several potential applications. In human-computer or human-human interaction systems, feeling recognition systems might give users with improved services by being adaptative to their emotions. The body of labor on sleuthing feeling in speech is sort of restricted. Currently, researchers area unit still debating what options influence the popularity of feeling in speech. There is conjointly appreciable uncertainty on the simplest algorithmic program for classifying feeling, and those emotions to category along.

Keywords: *Speech, human-computer, algorithm, communication, theorem*

INTRODUCTION

Problem Overview

There are many different ways through which emotion can be expressed. Emotion is expressed via facial movements; body and hand gestures and various biological signals such as heart rate and blood pressure or brain activity. Moreover, emotions can also be expressed in speech, e.g. by rise or fall in the voice, there may be change in the speech speed, speech tone or volume, and this is referred as a term speech emotion. Typical communication channels that indicate emotions are voice and facial expressions [1, 2]. Humans have the aptitude to acknowledge the emotions of their communication partner by victimization all their offered senses. They hear the sound, they scan lips, they interpret gestures and countenance and in fact they resolve the linguistics of the auditory communication. Through all the mentioned senses, folks truly sense the spirit of the speech communication partner and so area unit able to adapt to that. But feeling recognition from the speech signal is incredibly difficult task for machine, as a result of this needs that the machine ought to have the ample intelligence to

acknowledge human voices and feeling through it [1].

Emotion recognition from the speaker's speech is incredibly troublesome owing to the subsequent reasons: In differentiating between varied emotions that specific speech options area unit additional helpful is not clear. Owing to the existence of the various sentences, speakers, speaking designs, speaking rates accosting variability was introduced. Identical vocalization could show totally different emotions. Every feeling could correspond to the various parts of the spoken vocalization. Thus it is terribly troublesome to differentiate these parts of vocalization [1].

Another problem is that emotion expression is depending on the speaker and his or her culture and environment. As the culture and environment gets change the speaking style also gets change, which is another challenge in front of the speech emotion recognition system. Another important problem is that one may undergo a certain emotional state such as sadness for days, weeks, or even months. In such a case, other emotions will be transient and will not last for more than a few minutes. As a consequence, it is not clear which emotion the emotion recognizer should detect. Emotion does not have a commonly agreed theoretical definition. However, people know emotions when they feel them. For this reason, one has to study and define different aspects of emotions. In speech emotion recognition, the emotions of the male or female speakers are found out from their speech [1].

Scope of the Work

This dissertation work focuses on recognizing emotions from one of the previously mentioned methods that is speech. This work is focus on the speech emotion recognition system based on the statistical classifier that is Gaussian mixture model (GMM) ,.The need to find out a set of the significant emotions such as anger, happiness, sadness, fear, neutral state etc. to be classified by an automatic emotion recognizer is a main concern in speech emotion recognition system. For emotion recognition based on the acoustics signal, following broad lines have adopted:

- Consider an emotional model,
- Start with analyzing one or more of the available databases,
- Extract a set of suitable features from the speech signal,

- Train a classifier or use some reasoning techniques or probabilistic approaches in order to be able to make statements on test data.

Need of Emotion Recognition through Speech

The most necessary application of Speech feeling recognition (SER) is in intelligent human-machine interaction. In today's human-machine interaction systems, machines will acknowledge "what is aforesaid" and "who said it" exploitation speech recognition and recognition techniques. If equipped with feeling recognition techniques, machines may also apprehend "how it's said" to react a lot of suitably, and build the interaction a lot of natural. So, by exploitation speech feeling recognition (SER) system human machine interaction can get increased [16].

It is also useful for in-car board system where information of the mental state of the driver to initiate safety strategies, and provide aid or resolve errors in the communication according to the emotion of the driver [2]. It can be also employed as a diagnostic tool for therapists. It may be also useful in automatic translation systems in which the emotional state of the speaker plays an important role in communication between parties. In aircraft cockpits, it has been found that speech recognition systems trained to stressed-speech achieve better performance than those trained by normal speech [6].

SPEECH RECOGNITION

In this section we first briefly review how the speech signal recognition is becoming. It is known that the speech signal is one of the most complex signals to recognize. First of all the signal get through some pre-processing for analyzing.

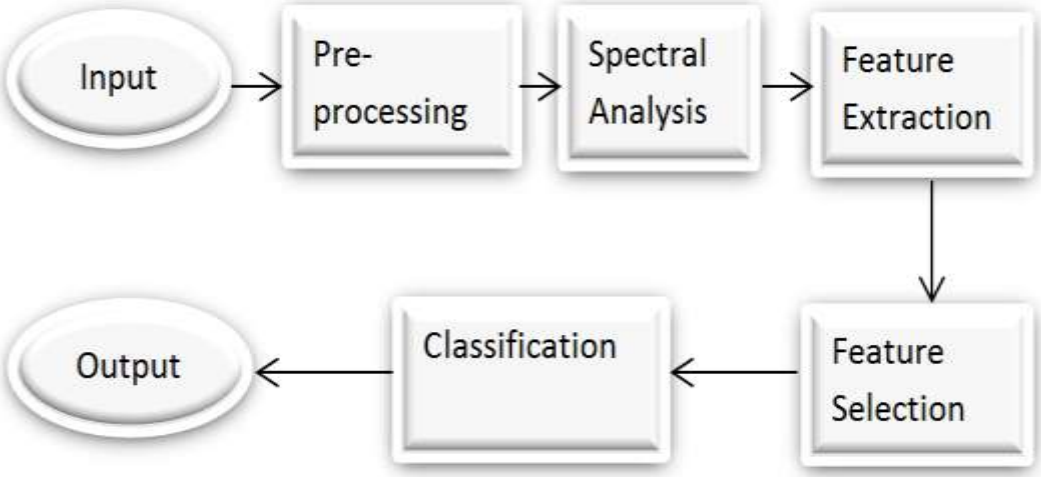


Fig. 1: *Speech Recognition.*

GMM AND MER CLASSIFIER

The GMM [14] cones the form of the PDF to be a linear superposition of a nite number of Gaussian distributions

$$p(\mathbf{x}) = \sum_{k=1}^{K} \alpha_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu_k}, \Sigma_k)$$

Where

α_k

is the mixture weight of the kth component Gaussian of the form

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu_k}, \Sigma_k) = \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma_k|^{\frac{1}{2}}}$$

2. Prosodic feature extraction

1. Pitch

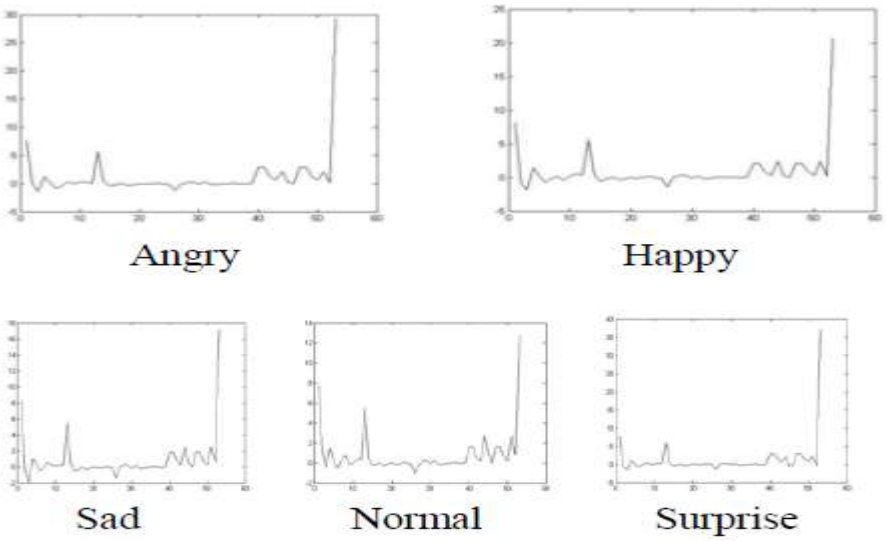
Statistics related to pitch [13] conveys considerable information about emotional status. For this project, pitch is extracted from the speech waveform using a modified version of the RAPT algorithm for pitch tracking implemented in the VOICEBOX toolbox. Using a frame length of 50ms, the pitch for each frame was calculated and placed in a vector to correspond to that frame. The various

statistical features are extracted from the pitch tracked from the samples. We use minimum value, maximum value, range and the moments-mean, variance, skewness and kurtosis. We hence get a 7 dimensional feature vector which is appended to the end of the 39 dimensional super vector obtained from the GMM.

Loudness

Loudness [14] is extracted from the samples mistreatment DIN45631 implementation of loudness model in MATLAB. The operate loudness () returns loudness for every frame length of 50ms and additionally one single specific loudness worth. currently a similar minimum worth, most worth, vary and therefore the moments- mean, variance, asymmetry and kurtosis applied math options are used to model the loudness vector. Hence we tend to get an eight dimensional feature vector that is appended to the already obtained 46 dimensional feature vector to get the ultimate 54 dimensional feature vector. This vector will currently lean as input to the SVM.

.



Algorithm 1 The Boosted-GMM algorithm

1: Input: $X = \{\mathbf{x}_i\}_{i=1}^{N}$, r, and T.
2: Initialize $W_1(\mathbf{x}_i) = 1/N, i = 1, ..., N, p_0 = 0$.
3: For $t = 1, ..., T$ or until $L(p_t) \leq L(p_{t-1})$

- Sample X_t from X according to W_t and estimate q_t from X_t using the F-J algorithm [24].
- Set $p_t = (1 - \alpha)p_{t-1} + \alpha p_t$ where $\alpha = \arg\max_{0 \leq \alpha \leq 1} L(p_t)$.
- Update $W_{t+1}(\mathbf{x}_i) = \frac{1}{p_t(\mathbf{x}_i)}, i = 1, ..., N$.

4: Output: Final density estimate p_T.

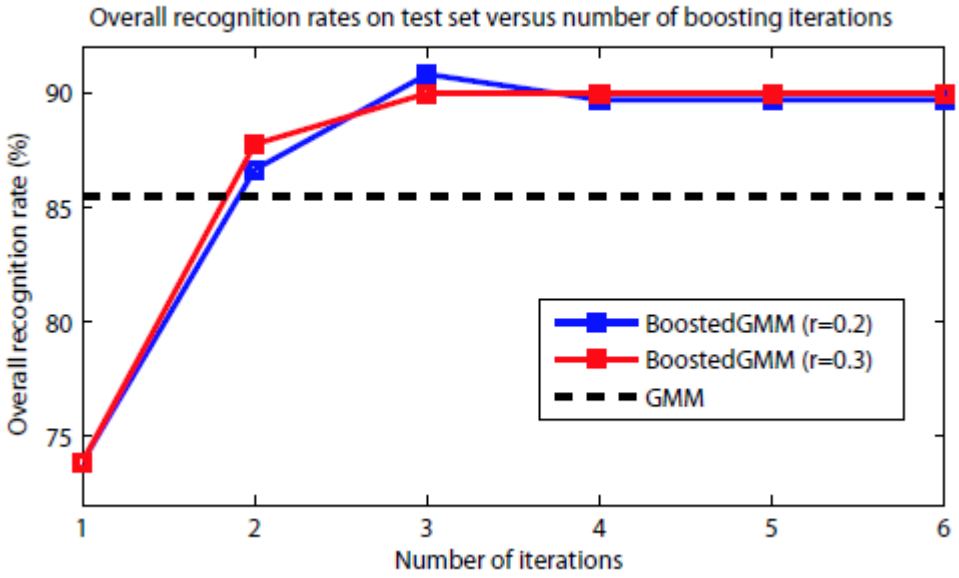


Fig. 2: *Comparison of Overall Emotion Recognition Rates of the Boosted-GMM Algorithm and the EM-GMM Algorithm.*

GAUSSIAN MIXTURE MODEL (GMM)

A Gaussian Mixture Model (GMM) 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. GMM parameters are estimated from training data using the iterative Expectation-Maximization (EM) algorithm or Maximum *A Posteriori* (MAP) estimation from a well-trained prior model.

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

$$P(x/\lambda)=\sum_{i=1}^{M} w_i\, g(\frac{x}{\mu_i}, \sum_i)$$

……
(5.1)

Where x is a D-dimensional continuous-valued data vector (i.e. measurement or features), wi, i = 1, . . . ,M, are the mixture weights, and $g(\frac{x}{\mu_i}, \sum_i), i = 1, \ldots, M$ are the component Gaussian densities. Each component density is a D-variate Gaussian function of the form,

$$g(\frac{x}{\mu_i}, \sum_i) = \frac{1}{(2\pi)^{D/2}\,|\sum_i|^{\frac{1}{2}}} \exp\{-\frac{1}{2}(x-\mu_i)'\sum_i^{-1}(x-\mu_i)\}$$

……(2)

With mean vector μi and covariance matrix $\sum_i$.the mixture weights satisfy the constraint that$\sum_{i=1}^{M}(w_i) = 1$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,

$$\lambda = \{w_i, \mu_i, \Sigma_i\} \quad i = 1, \ldots, M.$$

……(3)

There are several variants on the GMM shown in Equation (3). The covariance matrices can be full rank or constrained to be diagonal. Additionally, parameters can be shared, or tied, among the Gaussian components, such as having a common covariance matrix for all components, The choice of model configuration (number of components, full or diagonal covariance matrices, and parameter tying) is often determined by the amount of data available for estimating the GMM parameters and how the GMM is used in a particular biometric application [8][1][2].

Implementation Using Gaussian Mixture Model

The probability density functions of distorted features caused by different emotions are different. As a result, we can use a set of GMMs to estimate the probability that the observed utterance from a particular emotion.

Maximum Likelihood Estimation: In construction of a Bayesian classifier the class-conditional probability density functions need to be determined. The initial model selection can be done for example by visualizing the training data, but the adjustment of the model parameters requires some measure of goodness, i.e., how well the distribution fits the observed data. Data likelihood is such goodness value [2].

Assume that there is a set of independent samples X = $\{x_1, x_2, \ldots\ldots x_N\}$ drawn from a single distribution described by a probability density function p(x; θ) where θ is the PDF parameter list.

The likelihood function

$$\mathcal{L}(\mathcal{X};\theta)=\prod_{x=1}^{N} P(x_N:\theta) \quad \ldots\ldots (4)$$

tells the likelihood of the data X given the distribution or, more specifically, given the distribution parameters θ. The goal is to find $\hat{\theta}$ that maximizes the likelihood:

$$\hat{\theta}=\arg max_{\theta}= \mathcal{L}(X;\theta) \quad \ldots\ldots (5)$$

Usually this function is not maximized directly but the logarithm

$$\mathcal{L}(X;\theta)=\ln \mathcal{L}(X;\theta)=\sum_{n=1}^{N} lnp(x_N:\theta) \quad \ldots\ldots (6)$$

Called the log-likelihood function which is analytically easier to handle. Because of the monotonicity of the logarithm function the solution to Eq. 8 is the same using $\mathcal{L}(X;\theta)$.

Steps for GMM classification

1] Initialize parameters Expectation step: Compute the posterior probability for i=1, n, k=1 ...K.

$$P_{i,k}= \frac{a_k^{(r)} \emptyset\,(x_i \mu_k^{(r)}, \sum_k^{(r)})}{\sum_{k=1}^{(r)} a_k^{(r)} \emptyset\,(x_i \mu_k^{(r)}, \sum_k^{(r)})} \quad \ldots\ldots (7)$$

2] Maximization step

$$a_k^{(r+1)} = \frac{\sum_{i=1}^{n} P_{i,k}}{n} \quad \ldots\ldots . (8)$$

$$\mu_k^{(r+1)} = \frac{\sum_{i=1}^{n} P_{i,k}\, X}{\sum_{i=1}^{n} P_{i,k}} \quad \ldots\ldots .. (9)$$

$$\mu_k^{(r+1)} = \frac{\sum_k^{(r+1)} P_{i,k} (x_i \mu_k^{(r+1)})(x_i \mu_k^{(r+1)})^{t}}{\sum_{i=1}^{n} P_{i,k}} \ldots \quad (10)$$

3] Repeat steps 2) and 3) until convergence.

RESULTS

Recognition Accuracy

This measure signifies the recognition accuracy in percentage for each known test speech input to the total trained emotional speech data.

$$\text{Accuracy} = \frac{\text{Correctly detected Emotions inputs}}{\text{Total trained emotions inputs}} \times 100\%$$

The accuracy for each classifier for the six emotions is calculated on the basis of above relation. It is calculated for both the Berlin emotional database (BES) and a recorded non-standard database

Table 1: *Recognition Accuracy for Berlin Emotion Database (BES).*

Emotion / Classifier	Angry	Happy	Sad	Neutral	Fear
GMM	100%	67%	89%	73%	50%

Confusion Matrix

The confusion matrix of the speech feeling recognition system signifies the slack of classifiers to settle on the simplest correct feeling within the testing part. It depicts the confusion in choice among the trained patterns of feeling options having similarities in feature pattern.

Table 2: *Confusion Matrix for GMM Classifier.*

Responded / Presented	Angry	Happy	Sad	Neutral	Fear
Angry	100%	-	-	-	-
Happy	-	67%	-	-	-
Sad	-	-	89%	11%	-
Neutral	-	19%	-	73%	8%
Fear	-	-	-	-	50%

CONCLUSION

Experimental results of the speech emotion recognition system based on the GMM, has explained in the previous chapter. The results were obtained by performing classification of the emotions of the different speakers. Here the final conclusion of the dissertation work is drawn, which is based on the theoretical and practical implementation of the system. The scope for the future work for the further improvement in the speech emotion recognition system is also discussed here. Three speech emotion recognition systems based on the Gaussian mixture model were studied in this dissertation. Features based on the fundamental frequency, energy, formants, and Mel frequency cepstrum coefficient would be extracted as the input to the GMM classifier. In these systems obtained relatively high accuracy in classifying the five emotional states. However, appropriate features that can efficiently carry the characteristics of signals are of great importance in the problems of emotion recognition classification. Thus, classification algorithms should be

followed by an efficient feature set extraction and feature selection processes. Again the emotional speech database used in the emotions recognition from the speech is an important factor, because wrong conclusion can be derived from the incorrect speech database.

REFERENCES

1. Ayadi M. E., Kamel M. S. and Karray F., 'Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases', Pattern Recognition, 44 (16), 572-587, 2011.

2. Pai C.Y. and Pao T. L., *'Analysis and Detection of Emotion Change in Continuous Speech'*, Master of Science Thesis, Department Of Computer Science And Engineering, Tatung University, 2008.

3. Emerich S., Lupu E. and Apatean A., 'Emotions Recognitions by Speech and Facial Expressions Analysis', *Proceedings ofConference on European Signal Processing*, 1617-1621, 2009.

4. Zhou y., Sun Y., Zhang J, Yan Y., 'Speech Emotion Recognition using Both Spectral and Prosodic Features', IEEE, 23(5), 545-549, 2009.

5. Chiriacescu I., *'Automatic Emotion Analysis Based On Speech'*, M.Sc. Thesis, Department of Electrical Engineering, Delft University of Technology, 2009.

6. Vogt T., Andre E. and Wagner J., 'Automatic Recognition of Emotions from Speech: A Review of the Literature and Recommendations for Practical Realization', *Proceedings ofLNCS* 4868, 75-91, 2008.

7. Ververidis D. and Kotropoulos C., 'Emotional Speech Recognition: Resources, Features and Methods', Speech Communication, 48 (9), 1162-1181, 2006.

8. Ciota Z., "Feature Extraction of Spoken Dialogs for Emotion Detection", *Proceedings ofInternational Conference on Signal processing*, 727-731, 2006.

9. Lee C. M. and Narayanan S. S., 'Towards Detecting Emotions in Spoken Dialogs', IEEE, 13(2), 293-303, 2005.

10. Rabiner L. R. andJuang, B., '*Fundamentals of Speech Recognition'*, Pearson Education Press, Singapore, 2nd edition, 2005.

11. Albornoz E. M., Crolla M. B. and Milone D. H. "Recognition of Emotions in Speech". *Proceedings of17th European Signal Processing Conference,*2009.

12. VibhaTiwari , "MFCC and its applications in speaker recognition" , International Journal on Emerging Technologies 10 Feb., 2010

13. Reynolds D. A., "Speaker identification and verification using Gaussian mixture speaker models", *Speech Commun.* 17 (1995), 91–108.

14. Dimitrios Ververidis and Constantine Kotropoulo, " A Review of Emotional Speech Databases"