IRA-International Journal of Technology & Engineering ISSN 2455-4480
Proceedings of the
International Conference on Science & Engineering for Sustainable Development(2017)
Pg. no.56-64
Published by: Institute of Research Advances
https://research-advances.org/index.php/IRAJTE

Speech Emotion Recognition System Using Gaussian Mixture Model and Improvement proposed via Boosted GMM

Ms. Pavitra Patel*[1], A. A. Chaudhari[2], M. A. Pund[3], Ms. D. H. Deshmukh[4]

*[1,2,3,4]PRM Institute of Technology and Research, Badnera-Amravati (MS) India 444701

Type of Review: Originality Check & Peer Review under the responsibility of the Scientific Committee of the Conference and The Institution of Engineers (India).
DOI: http://dx.doi.org/10.21013/jte.ICSESD201706

How to cite this paper:
Patel, P., Chaudhari, A., Pund, M., Deshmukh, D. (2017). Speech Emotion Recognition System Using Gaussian Mixture Model and Improvement proposed via Boosted GMM. *Proceedings of the International Conference on Science & Engineering for Sustainable Development (2017)*, 56-64. doi: http://dx.doi.org/10.21013/jte.ICSESD201706

ABSTRACT

Speech emotion recognition is an important issue which affects the human machine interaction. Automatic recognition of human emotion in speech aims at recognizing the underlying emotional state of a speaker from the speech signal. Gaussian mixture models (GMMs) and the minimum error rate classifier (i.e. Bayesian optimal classifier) are popular and effective tools for speech emotion recognition. Typically, GMMs are used to model the class-conditional distributions of acoustic features and their parameters are estimated by the expectation maximization (EM) algorithm based on a training data set. In this paper, we introduce a boosting algorithm for reliably and accurately estimating the class-conditional GMMs. The resulting algorithm is named the Boosted-GMM algorithm. Our speech emotion recognition experiments show that the emotion recognition rates are effectively and significantly boosted by the Boosted-GMM algorithm as compared to the EM-GMM algorithm.

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.

Keywords—Speech, Emotion, Gaussian, Boosted

INTRODUCTION

Emotional speech recognition aims at automatically identifying the emotional or physical state of a human being from his or her voice. Although emotion detection from speech is a relatively 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. The body of work on detecting emotion in speech is quite limited. Currently, researchers are still debating what features influence the recognition of emotion in speech. There is also considerable uncertainty as to the best algorithm for classifying emotion, and which emotions to class together [1].

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 emotion detection from speech is a relatively 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. The body of work on detecting emotion in speech is quite limited. Currently, researchers are still debating what features influence the recognition of emotion in speech. There is also considerable uncertainty as to the best algorithm for classifying emotion, and which emotions to class together [1].

PROBLEM STATEMENT

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 [2]. Humans have the natural ability to recognize the emotions of their communication partner by using all their available senses. They hear the sound, they read lips, they interpret gestures and facial expression and of course they find out the semantics of the utterance. Through all the mentioned senses, people actually sense the emotional state of the conversation partner and therefore are able to adapt to it. However emotion recognition from the speech signal is very challenging task for machine, because this requires that the machine should have the sufficient intelligence to recognize human voices and emotion through it [1].

Emotion recognition from the speaker's speech is very difficult because of the following reasons: In differentiating between various emotions which particular speech features are more useful is not clear. Because of the existence of the different sentences, speakers, speaking styles, speaking rates accosting variability was introduced. The same utterance may show different emotions. Each emotion may correspond to the different portions of the spoken utterance. Therefore it is very difficult to differentiate these portions of utterance [1].

NEED OF EMOTION RECOGNITION THROUGH SPEECH

The most important application of Speech emotion recognition (SER) is in intelligent human-machine interaction. In today's human-machine interaction systems, machines can recognize "what is said" and "who said it" using speech recognition and speaker identification techniques. If equipped with emotion recognition techniques, machines can also know "how it is said" to react more appropriately, and make the interaction more natural. So by using speech emotion recognition (SER) system human machine interaction will get enhanced [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].

The GMM [14] cones the form of the PDF to be a linear superposition of a finite 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

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

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 using DIN45631 implementation of loudness model in MATLAB. The function loudness() returns loudness for each frame length of 50ms and also one single specific loudness value. Now the same minimum value, maximum value, range and the moments- mean, variance, skewness and kurtosis statistical features are used to model the loudness vector. Hence we get an 8 dimensional feature vector which is appended to the already obtained 46 dimensional feature vector to obtain the final 54 dimensional feature vector. This vector can now be given as input to the SVM.

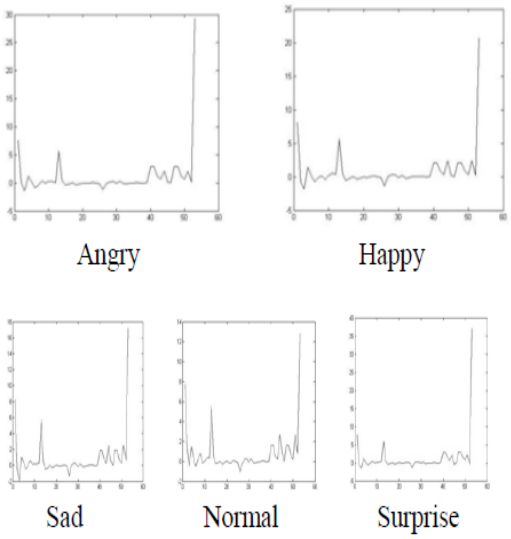

Fig 1: Different Emotion States

Algorithm 1 The Boosted-GMM algorithm

1: Input: $X = \{x_i\}_{i=1}^{N}$, r, and T.
2: Initialize $W_1(x_i) = 1/N, i = 1, ..., N, p_0 = 0$.
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 $W_{t+1}(x_i) = \frac{1}{p_t(x_i)}, i = 1, ..., N$.

4: Output: Final density estimate p_T.

Speech Input

As described in the last chapter the emotional speech database for the speech emotion recognition system is an important factor. This is due to the fact that the performance of the speech emotion recognition system is based on the naturalness of the speech from which emotion has to be extracted. So the collection of the emotional speech database for the speech emotion recognition experiment should be very carefully performed. A typical set of emotions contains 300 emotional states. Therefore to classify such a great number of emotions is very complicated. According to "Palette theory" any emotion can be decomposed into primary emotions similar to the way that any color is a combination of some basic colors. Primary emotions are anger, disgust, fear, joy, sadness and surprise.

An overview of existing databases (mostly in German or English) indicates that speech emotions are dominantly described using the categorical framework, and simulated emotions are more prevalent relative to spontaneous or elicited emotions. In addition, recognition is shown to be the major purpose for database creation. Moreover, relative to automatic speech recognition tasks, no standardized speech corpora and test conditions exist for speech emotion recognition, making performance comparison under the same settings difficult [1].

There are three kinds of emotional databases with regard to the authenticity of emotion. Databases with acted speech include portrayals of emotions by professional or amateur actors. In general actors are asked to speak some given utterances while expressing a certain emotion and the recording is labeled as containing the specified desired emotion. Another kind of databases contains elicited emotions. This kind of emotion is neither real, nor simulated. The last types are databases of spontaneous speech which contain real emotions [7]. Therefore during the collection of the emotional speech database for the speech emotion recognition system care has to be taken that there should be naturalness in the speech recording. For the speech emotion recognition experiment emotional speech data should be collect from the all type of speaker [10].

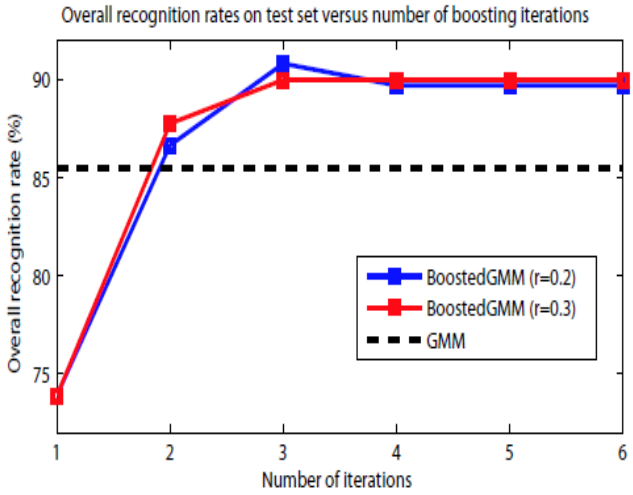

Fig 2: Comparison of overall emotion recognition rates of the Boosted-GMM algorithm and the EM-GMM algorithm.

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,

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

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,

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

(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,

$$\lambda = \{w_i, \boldsymbol{\mu}_i, \Sigma_i\} \qquad i = 1, \ldots, M. \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

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

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

$\hat{\theta}$=argmax_{θ}=$\mathcal{L}(X;\theta)$..... (5)

Usually this function is not maximized directly but the logarithm

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

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

2] Maximization step

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

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

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

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.

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

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

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:

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"