International Journal of Research In Science & Engineering e-ISSN: 2394-8299
Volume: 3 Issue:2 March-April 2017 p-ISSN: 2394-8280

EMOTION RECOGNITION FROM SPEECH WITH GAUSSIAN MIXTURE MODELS & VIA BOOSTED GMM

Pavitra Patel[1], Anand Chaudhari[2], Ruchita Kale[3], M.A.Pund[4]

[1]*Assistant Professor, T & P Department, MIET Gondia, nimit_jmi585@yahoo.com*
[2]*Assistant Professor, Computer Science & Engineering, PRMIT&R- Badnera, aachaudhari@mitra.ac.in*
[3]*Assistant Professor, Computer Science & Engineering, PRMIT&R- Badnera, rakale@mitra.ac.in*
[4]*Professor, Computer Science & Engineering, PRMIT&R- Badnera, mapund@mitra.ac.in*

ABSTRACT

Speech has several characteristic features such as naturalness and efficient, which makes it as attractive interface medium. It is possible to express emotions and attitudes through speech. In human machine interface application emotion recognition from the speech signal has been current topic of research. 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. Then, classification is performed to minimize the classification error w.r.t. the estimated class-conditional distributions. We call this method the EM-GMM algorithm. 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. This is due to the fact that the boosting algorithm can lead to more accurate estimates of the class-conditional GMMs, namely the class-conditional distributions of acoustic features.

Keywords*: Emotion recognition, Gaussian mixture model, Bayesian optimal classifier, EM algorithm, boosting*

1. THE SELECTION OF SPEECH DATA FOR EMOTION ANALYSIS

In this paper, the selection of language sentence for experiment analysis mainly comes from two aspects followed. First, statements selected must not contain a particular aspect of emotional tendency; secondly, statements selected must contain high emotional freedom, for the same statement can exert all kinds of emotions. Moreover, to the length of the statement, composition of consonants and auxiliary components, all differences between male and female should be

considered. According to principles above, 60 sentences for sentiment analysis are selected [9]. In this paper, the emotion type is roughly divided into joy, anger, surprise and sadness, and all the common emotions are classified as much as possible into this category, which is considered as reasonable classification for computer sentiment analysis research. In order to obtain the original speech data, 60 statements from 10 male speakers with joy, anger, surprise and sadness is pronounced once again. At the same time, speakers are told to pronounce each sentence once again calmly as much as possible without emotion. Through the process above 3000 language sentences are collected for experiment. In the classification experiments, 2000 sentences are taken for training and 1000 sentences for recognition. To test the effectiveness of the speech data collected for emotion experiment, an audition experiment was carried out by the researchers. 5 speakers differing from the 10 above are required sitting in front of computer terminals and given collected statements with various emotions randomly. Then the speakers judge the emotion type of voices by subjective evaluation. After repeated listening and comparing, meaningful test in math (Mcnemar test) [10] is implemented. The unobvious emotion characteristics of sentence are deleted and redone.

1.1 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.

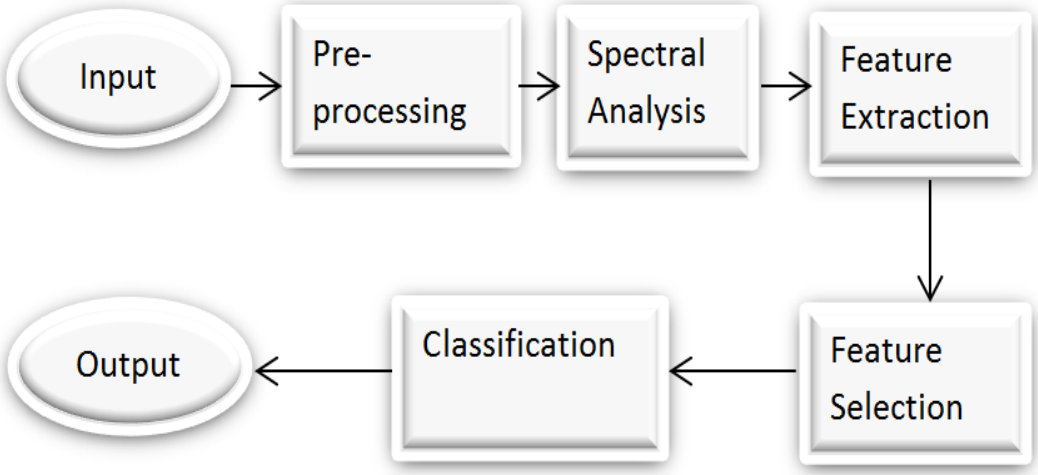


Fig-1: Speech Recognition.

1.2 GMM AND MER CLASSIFIER

The GMM [14] connes 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

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

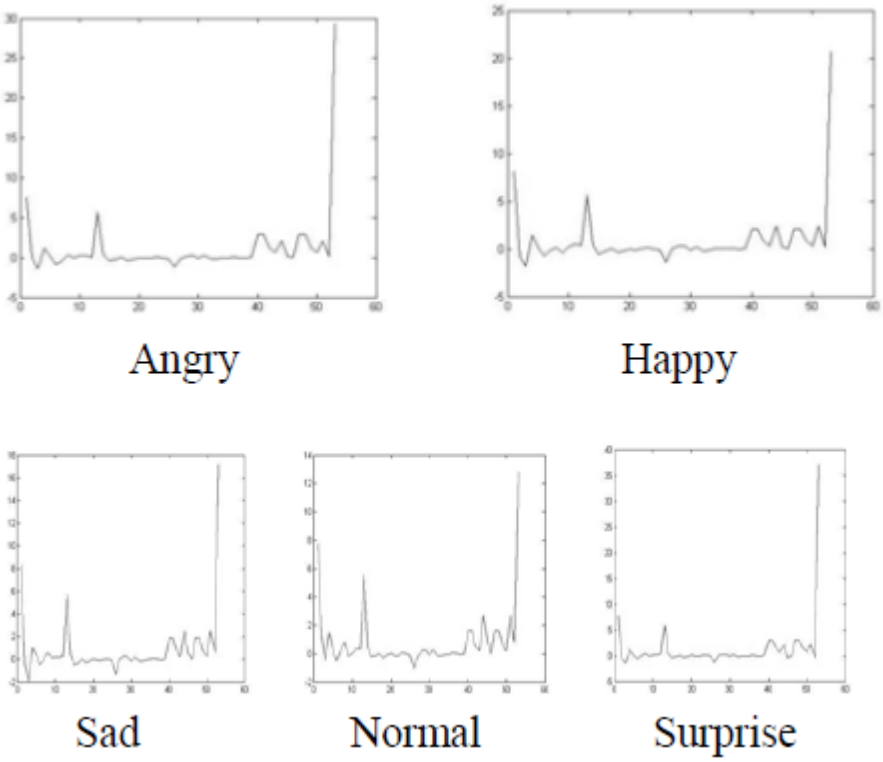
2. Prosodic feature extraction

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.

2. 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.



3. Formant

Formants are the distinguishing or meaningful frequency components of human speech and of singing. By definition, the information that a human requires to distinguish between vowels can be represented purely quantitatively by the frequency content of the vowel sounds. In speech, these are characteristic partials that identify vowels to the listener. The formant with lowest frequency is called f1, the second lowest called f2, and the third f3. Most often the first two formants, f1 and f2, are enough to disambiguate a vowel. These two formants determine quality of vowels in terms of the open/close and front/back dimensions (which have traditionally, though not accurately, been associated with position of the tongue). Thus first formant f1 has a higher frequency for an open vowel (such as [a]) and a lower frequency for a close vowel (such as [i] or [u]); and the second formant f2 has a higher frequency for a front vowel (such as [i]) and a lower frequency for a back vowel (such as [u]).[15][16] Vowels will almost always have four or more distinguishable formants; sometimes there are more than six. However, the first two formants are the most important in determining vowel quality, and this is displayed in terms of a plot of the first formant against the second formant,[17] though this is not sufficient to capture some aspects of vowel quality, such as rounding.[18] Nasals usually have an additional formant around 2500 Hz. The liquid [l] usually has an extra formant at 1500 Hz, while the English "r" sound ([ɹ]) is distinguished by virtue of a very low third formant (well below 2000 Hz).

Plosives (and, to some degree, fricatives) modify the placement of formants in the surrounding vowels. Bilabial sounds (such as /b/ and /p/ in "ball" or "sap") cause a lowering of the formants; velar sounds (/k/ and /g/ in English) almost always show f2 and f3 coming together in a 'velar

pinch' before the velar and separating from the same 'pinch' as the velar is released; alveolar sounds (English /t/ and /d/) cause less systematic changes in neighbouring vowel formants, depending partially on exactly which vowel is present. The time-course of the changes in vowel formant frequencies are referred to as 'formant transitions'.

PROPOSED FUTURE WORK AND SCOPE

There is a lot of work on emotional intelligence, and there are also separate work on extracting other information like age, gender etc. But it has been proved that the voice features keep on changing by age. Similarly for different genders the emotion matching parameters should be different. It can be felt easily that when we hear a sound, first thing comes in our mind whether the speaker is boy or a girl, then we estimate the age of person, then we guess the meaning and emotion flowing through the voice. There are different physiological aspects related to the both gender and similar is the case with the age of person. So the machine needs to be trained to differentiate between the gender as well as the age groups. If a lady shouts, it shows anger of fear, but this the same perception cannot be applied to the shouting baby. There is a lot of scope of using all the works combined to increase the accurateness of the emotion detection by the machine.

The goal of GMM model estimation (or model estimation in a very general sense) is to seek a set of model parameters that maximizes the data log likelihood. Given a training data set X = {xi}N i=1 and a probability density function p(x) to be estimated, the data log likelihood is given by

$$L(p) = \sum_{i=1}^{N} \log p(\mathbf{x}_i)$$

Here, in this paper, p(x) is the probability density function of a GMM given by Equation. Instead of directly optimizing Equation as in the EM algorithm, we start with an initial estimate p0 (a GMM) and iteratively add to this estimate a small component qt at round t. That is,

$$p_t = (1-\alpha)p_{t-1} + \alpha q_t$$

$$q_t = \arg\max_{q_t \in Q} \sum_{i=1}^{N} \frac{q_t(\mathbf{x}_i)}{p_{t-1}(\mathbf{x}_i)}$$

we can seek the that yields maximum increase in L(pt). From Equation 10, it is obvious that qt can be obtained through performing maximum likelihood estimation on the training examples weighted by Wt = 1/pt-1. This meets our intuition of boosting that more focus is put on the examples with low probabilities under the previous estimate, and Wt can be deemed as the distribution over the training set at round t in a boosting algorithm [26]. The Boosted- GMM

algorithm is summarized in Algorithm 1. The sampling procedure in Algorithm can be done as follows. At each round, we sort the training examples by their weights in the descending order and keep only a fraction r of them (e.g. r = 0:3).

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.

3. CONCLUSION

In the field of human computer interaction automatic speech emotion recognition is a current research topic. Emotion recognition in speech is a challenging problem because it is unclear that which features are effective for speech emotion recognition. In this paper we will extract the features by PCA therefore dimension of the processing is less than before existing approaches and we will compare the results of GMM classifier with other classifiers. We introduce the Boosted-GMM algorithm, which embeds the EM algorithm in a boosting framework and which can be used to reliably and accurately estimate the class-conditional probabilistic distributions in any pattern recognition problems based on a training data set. We apply the Boosted-GMM algorithm to speech emotion recognition and our experiments show that the emotion recognition rates are effectively and significantly boosted" by the Boosted- GMM algorithm as compared to the EM-GMM algorithm due to the fact that boosting can lead to more accurate estimates of the class-conditional GMMs, namely the class-conditional distributions of acoustic features.

REFERENCES

[1] Petrushin, V., " Emotion recognition in speech signal: experimental study, development, and application," Proc. ICSLP'00.

[2] Oudeyer, P., " Novel Useful Features and Algorithms for the Recognition of Emotions in Human Speech," Proc. ICSP'02.

[3] Schuller R., Rigoll G., Lang M., " Hidden Markov modelbased speech emotion recognition," Proc. ICASSP'03, pp. 1-4.

[4] T.L. Nwe, S.W. Foo and L.C. De Silva, " Speech emotion recognition using hidden markov models," Speech Communication 41, (2003), pp. 603-23.

[5] Lee, C.M., Yildirim, S., Bulut, M., Kazemzadeh, A., Busso,C., Deng, Z., Lee, S., Narayanan, S., " Emotion recognition based on phoneme classes," Proc. ICSLP'04.

[6] Dan-Ning Jiang, Lian-Hong Cai, " Speech emotion classification with the combination of statistic features and temporal features," Proc. ICME'04, pp. 1968-1970.

[7] Yalamanchili, B. S., et al. "Non Linear Classification for Emotion Detection on Telugu Corpus." International Journal of Computer Science & Information Technologies 5.2 (2014).

[8] Wang, Yongjin, Ling Guan, and Anastasios N. Venetsanopoulos. "Kernel cross-modal factor analysis for information fusion with application to bimodal emotion recognition." Multimedia, IEEE Transactions on 14.3 (2012): 597-607.

[9] Nwe, Tin Lay, Say Wei Foo, and Liyanage C. De Silva. "Speech emotion recognition using hidden Markov models." Speech communication 41.4 (2003): 603-623.

[10] Barker J. and X. Shao, "Energetic and informational masking effects in an audiovisual speech recognition system", Audio, Speech, and Language Processing, IEEE Transactions on, vol. 17, no. 3, (2009), pp. 446-458.