World Journal of Engineering Research and Technology
WJERT
www.wjert.org
SJIF Impact Factor: 5.218

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

B. Meenapreethi, Deepika G. Krishnan* and Sivaranjani G.

Ramanathapuram Coimbatore Tamil Nadu India.

Article Received on 31/01/2018 **Article Revised on 21/02/2018** **Article Accepted on 14/03/2018**

***Corresponding Author**
Deepika G. Krishnan
Ramanathapuram
Coimbatore Tamil Nadu
India.

ABSTRACT

Speech has several endowment features such as naturalness and efficient, which makes it as winsome interface medium. It is possible to express emotions and attitudes via speech. In human machine interface application emotion recognition from the speech signal has been prevailing topic of research. Speech emotion recognition is an important issue which affects the human machine intercommunication. Automatic recognition of human emotion in speech angles at recognizing the primitive emotional state of a speaker from the speech signal. Gaussian mixture models (GMMs) and the scintilla error rate classifier (i.e. Bayesian optimal classifier) is embraced and effective tools for speech emotion recognition. Typically, GMMs are used to model the class-conditional distributions of acoustic visage and their parameters are outlined 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 judged class conditional distributions. This method is called the EM-GMM algorithm. In this paper, we discuss about boosting algorithm for reliably and accurately estimating the class-conditional GMMs. The resulting algorithm is named the Boosted-GMM algorithm. This speech recognition experiment shows better results than the prior algorithms available update. This is due to the fact that the boosting algorithm can lead to more scrupulous estimates of the class-conditional GMMs, namely the class-conditional dispersions of acoustic features.

KEYWORDS: Emotion recognition, Gaussian mixture model, Bayesian optimal classifier, EM algorithm, boosting.

1. INTRODUCTION

In this paper, the selection of language sentence for experiment exploration mainly comes from two countenances followed. At the start, statements selected must not contain a particular countenance of emotional tendency; to continue, statements selected must contain high emotional privilege. Moreover, to the span of the statement, composition of affricates and auxiliary components, all differences between male and female should also be premeditated. According to tenets above, 60 sentences for inclination analysis were selected.[9] In this paper, the emotion type is loosely divided into joy, anger, surprise and sadness, and all the common emotions are classified as much as probably into this category, which is deliberate as legitimate classification for computer sentiment analysis research. In order to obtain the aboriginal speech data, 60 statements from 10 male speakers with joy, anger, surprise and sadness is phonated once again. At the same time, speakers are told to pronounce each sentence once again motionlessly as much as possible without emotion. Through the process above 3000 language sentences were collected for experiment. In the classification experiments, 2000 sentences were taken for training and 1000 sentences for recognition. Then the speakers sneered at the emotion type of voices by subjective assessment. After repeated listening and comparing, meaningful test in math (Mcnemar test).[10] is implemented. The emotion which is not obvious characteristics of sentence are deleted and redone.

1.1 SPEECH RECOGNITION

In this section we first fleetingly review how the speech signal recognition is becoming. It is known that the speech signal is one of the most labyrinthine signals to recognize. Firstly the signal gets through some pre-processing for analyzing.

1.2 GMM AND MER CLASSIFIER

1. The GMM[14] connes the form of the PDF to be a rectilinear superposition of a nite number of Gaussian distributions where the mixture weight of the kth component Gaussian of the form is.
2. Prosodic feature extraction 1. Pitch Statistics related to pitch[13] totes considerable information about emotional status. For this project, pitch is elicited from the speech waveform using a amended version of the RAPT algorithm for pitch tracking instigated in the VOICEBOX toolbox. Using a frame span of 50ms, the pitch for each frame was calculated and placed in a vector to reciprocate to that frame. The various demographic

features are extricated from the pitch tracked from the samples. We use minimum value, maximum value, range and the moments- mean, variance, skewness and kurtosis. Hence 7 dimensional feature vector which is annexed to the end of the 39 dimensional super vector obtained from the GMM.

3. Loudness[14] is extracted from the tasters using DIN45631 implementation of loudness model in MATLAB. The function loudness resumes loudness for each frame length of 50ms and also one single unequivocal loudness value. Now the same minimum value, maximum value, range and the moments- mean, variance, skewness and kurtosis statistical features are used to consummate the loudness vector. Hence we get an 8 dimensional feature vector which is annexed 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.

4. Formant is the differentiating or meaningful frequency components of human speech and of singing. By definition, the information that a human necessitates to distinguish between vowels can be represented purely quantifiably by the commonness content of the vowel sounds. In speech, these are idiosyncratic partials that identify vowels to the listener. The formant with bottommost 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 legislate 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 advanced abundance for an open vowel (such as [a]) and a lower abundance for a close vowel (such as [i] or [u]); and the second formant f2 has a higher abundance 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 evident able formants; sometimes there are more than six. However, the first two formants are the most important in arbitrating vowel quality, and this is displayed in terms of a plot of the first formant against the second formant,[17] though this is not sufficing to capture some aspects of vowel quality, such as rounding.[18] Nasals usually have an appended 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) transmogrify 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 disentangling from the same 'pinch' as the velar is released; alveolar sounds (English /t/ and/d/) cause less systematic changes in neighboring vowel formants, reckoning on partially on exactly which vowel is present. The time-course of the changes in vowel formant abundances 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 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, 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

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

IJRISE JOURNAL| www.ijrise.org|editor@ijrise.org [47-53] 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).

CONCLUSION

Emotion recognition in speech is one of the trending research topics in field of human computer interaction. Emotion recognition in speech is a perplexing problem because the features available update is not still up to the mark for speech emotion recognition. In this paper we will extract the features by PCA thereby extending the explanation of the 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. In this paper we have implemented this algorithm to speech emotion recognition to show that the emotion recognition rates are effectively and also it significantly boosts the Boosted- GMM algorithm as compared to the EM-GMM algorithm. This is 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, 1-4.
4. T.L. Nwe, S.W. Foo and L.C. De Silva, “Speech emotion recognition using hidden markov models,” Speech Communication, 2003; 41: 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, 1968-1970.

7. Yalamanchili, B. S., et al. "Non Linear Classification for Emotion Detection on Telugu Corpus." International Journal of Computer Science & Information Technologies, 2014; 5(2).
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, 2012; 14(3): 597-607.
9. Nwe, Tin Lay, Say Wei Foo, and Liyanage C. De Silva. "Speech emotion recognition using hidden Markov models." Speech communication, 2003; 41(4): 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, 2009; 17(3): 446-458.