Nehru E-Journal
A Journal of Nehru Arts and Science College(Autonomous)
Journal Home Page: http://nehrucolleges.net/jo.html
ISSN: 2349-9052

Research Article
Vol 6 issue2(2018)
(Aug - Dec 2018)

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

I.Jeevitha[#1],Sherley.M.P [*2],S.Meghala[#3]

[#1]ASSISSTANT PROFESSOR, BCA & M.Sc. SS Department, Sri Krishna Arts and Science College, Coimbatore;
[*2] II M.Sc. SS, Department of BCA & M.Sc. SS, Sri Krishna Arts and Science College, Coimbatore
[#3] II M.Sc. SS, Department of BCA & M.Sc. SS, Sri Krishna Arts and Science College, Coimbatore
[1]jeevithai@skasc.ac.in
[2]sherleymp15mss037@skasc.ac.in
[3]meghalas15mss024@skasc.ac.in

***Keywords*:**

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

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 concluded 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 classconditional 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.

Introduction

THE SELECTION OF SPEECH DATA FOR EMOTION ANALYSIS

In this paper, the selection of language sentence for research analysis mainly comes from two aspects followed. First, statements selected must not contain a particular aspect of emotional leaning; secondly, statements selected must contain high expressive freedom, for the same statement can exert all kinds of feelings. Moreover, to the length of the statement, composition of consonants and support components, all differences between male and female should be considered. According to principles above, 60 sentences for sentimentality 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 type, which is considered as sensible 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 evident once again. At the same time, speakers are told to pronounce each sentence once again calmly for instance much as possible without emotion. Through the process above 3000 language sentences are collected for research. In the classification experiments, 2000 sentences are taken for training and 1000 sentences for appreciation. To test the efficiency of the speech data collected for emotion experiment, an audition experiment was carried out by the investigators. 5 speakers differing from the 10 above are required sitting in front of computer stations and given collected statements with various emotions

randomly. Then the speakers judge the emotion type of voices by particular calculation. After repeated attending and comparing, meaningful test in math (Mcnemar test) [10] is implemented. The unclear emotion characteristics of sentence are removed and redone.

- **Speech Recognition**

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

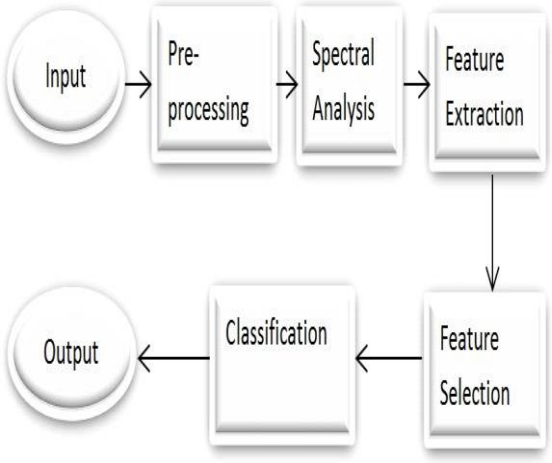


Fig-1: Speech Recognition.

- **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

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

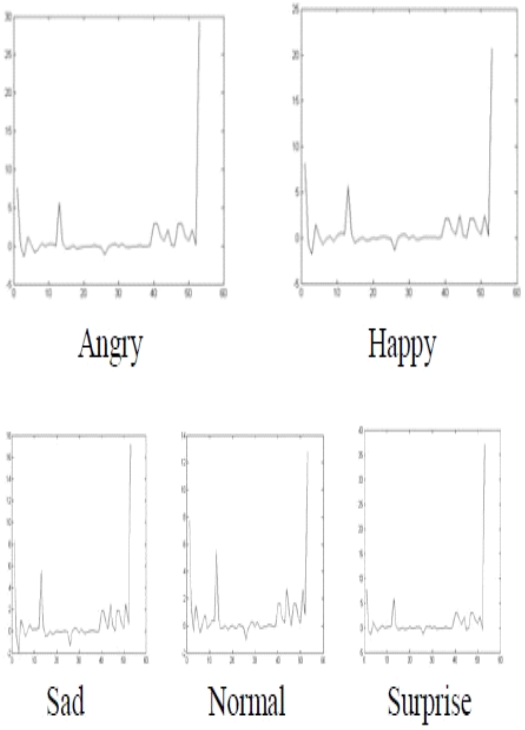
Prosodic feature extraction

Pitch

Statistics related to pitch [13] carriessignificant information about expressive status. For this project, pitch is removed from the speech waveform using aaltered version of the RAPT procedure for pitch followingapplied in the VOICEBOX toolbox. Using a border length of
50ms, the pitch for each frame was calculated and placed in a vector to match to that frame. The various numerical features are removed from the pitch tracked from the examples. We use least value, extreme value, range and the moments- mean, change, skewers and kurtosis. We hence get a 7 dimensional feature vector which is attached to the end of the 39 dimensional super vector obtained from the GMM.

- **Loudness**

Loudness [14] is removed from the examples using DIN45631 execution of loudness classic in MATLAB. The function loudness() returns loudness for each border length of 50ms and also one single detailed loudness value. Now the same least value, extreme 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 path can now be known as input to the SVM.



Formant

Formants are the unique or meaningful rate components of human communication and of humming. By definition, the information that a human requires to distinguish between vowels can be characterized purely quantitatively by the frequency content of the vowel sounds. In speech, these are characteristic partials that identify vowels to the auditor. The formant with lowermostrate is called f1, the second lowermost called f2, and the third f3. Most often the first two formants, f1 and f2, are enough to disambiguate a vowel. These two formants regulatevalue of vowels in terms of the open/close and front/back proportions (which have traditionally, though not exactly, been connected with position of the tongue). Thus first formant f1 has a developedrate 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 rate for a front vowel (such as [i]) and a lower rate for a back vowel (such as [u]).[15][16] Vowels will almost always have four or more different formants; sometimes there are more than six. However, the first two formants are the most important in defining vowel value, and this is displayed in terms of a plot of the first formant against the second formant,[17] though this is not enough to capture some aspects of vowel value, 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 eminent by virtue of a very low third formant (well below 2000 Hz).

Plosives (and, to some degree, fricatives) alter the placement of formants in the nearby vowels. Bilabial sounds (such as /b/ and /p/ in "ball" or "sap") cause a depressing 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 splitting from the same 'pinch' as the velar is free; alveolar sounds (English /t/ and /d/) cause less organized changes in neighbouring vowel formants, dependentpartly on exactly which vowel is present. The time-course of the changes in vowel formant rates are referred to as 'formant transitions'.

PROPOSED FUTURE WORK AND SCOPE

There is a lot of work on expressiveintellect, and there are also separate work on removing other material like age, gender etc. But it has been showed that the voice structures keep on altering by age. Similarly for dissimilar genders the emotion matching restrictions 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 appraisal the age of person, then we guess the significance and feelingsmooth through the voice. There are different biological aspects related to the both gender and similar is the case with the age of person. So the machine needs to be skilled to distinguish between the gender as well as the age groups. If a lady shouts, it shows anger of fear, but this the same awareness cannot be applied to the shouting baby. There is a lot of choice of using all the works combined to increase the correctness of the emotion recognition by the machine.

The goal of GMM model estimate (or model estimation in a very general sense) is to pursue a set of model limitations that exploits the data log possibility. Given a exercise data set X = {xi}Ni=1 and a prospectthickness function p(x) to be projected, 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 improving Equation as in the EM algorithm, we start with an initial estimate p0 (a GMM) and iteratively add to this evaluation a small moduleqt at round t. That is,

we can seek the that producesextreme increase in L(pt). From Equation 10, it is obvious that qt can be found through executionextreme likelihood estimate on the training samplessubjective by Wt = 1/pt-1. This meets our perception of improving that more focus is put on the samples with low likelihoods under the precedingestimation, and Wt can be believed as the circulation over the training set at round t in a boosting algorithm [26]. The Boosted- GMM algorithm is shortened in Algorithm 1. The sampling procedure in Algorithm can be prepared as follows. At each round, we class the training examples by their burdens in the descendant order and keep only a division r of them (e.g. r = 0:3).

Algorithm 1 The Boosted-GMM algorithm

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

4: Output: Final density estimate p_T.

CONCLUSION

In the field of human computer communicationprogrammed speech emotion gratitude is a current research topic. Emotion recognition in speech is ainteresting problem because it is unclear that which features are active for speech emotion recognition. In this paper we will extract the features by PCA therefore width of the processing is less than before presentmethods and we will link the results of GMM classifier with other classifiers. We introduce the Boosted-GMM algorithm, which inserts the EM algorithm in a boosting framework and which can be used to dependably and exactly estimate the class-conditional probabilistic circulations in any outlinecredit problems based on aexercise data set. We apply the Boosted-GMM algorithm to speech emotion recognition and our experiments show that the emotion recognition rates are successfully and meaningfully boosted" by the Boosted- GMM algorithm as compared to the EM-GMM algorithm due to the fact that boosting can lead to more exact estimates of the class-conditional GMMs, namely the class-conditional circulations of auralstructures.

REFERENCES

- Petrushin, V., " Emotion recognition in speech signal: experimental study, development, and application," Proc. ICSLP'00.
- Oudeyer, P., " Novel Useful Features and Algorithms for the Recognition of Emotions in Human Speech," Proc. ICSP'02.
- Schuller R., Rigoll G., Lang M., " Hidden Markov modelbased speech emotion recognition," Proc. ICASSP'03, pp. 1-4.
- 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.
- 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.
- Dan-Ning Jiang, Lian-Hong Cai, " Speech emotion classification with the combination of statistic features and temporal features," Proc. ICME'04, pp. 1968-1970.
- Yalamanchili, B. S., et al. "Non Linear Classification for Emotion Detection on Telugu Corpus." International Journal of Computer Science & Information Technologies 5.2 (2014).
- 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.
- 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.
- 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.