Emotion Recognition from Speech using GMM and VQ

Miss. Surabhi Agrawal
Computer Department
G. H. R. C. E. M.
Pune, India

Mrs. Shabda Dongaonkar
Computer Department
G. H. R. C. E. M.
Pune, India

Abstract**—** In this paper, there is a tendency to study the effectiveness of anchor models applied to the multiclass drawback of Emotion recognition from speech. Within the anchor models system, Associate in nursing emotion category is characterized by its line of similarity relative to different emotion categories. Generative models like Gaussian Mixture Models (GMMs) are typically used as front-end systems to get feature vectors wont to train complicated back-end systems like support vector machines (SVMs) or a multilayer perceptron (MLP) to enhance the classification performance. There is a tendency to show that within the context of extremely unbalanced knowledge categories, these back-end systems will improve the performance achieved by GMMs as long as Associate in nursing acceptable sampling or importance coefficient technique is applied. The experiments conducted on audio sample of speech; show that anchor models improve considerably the performance of GMMs by half dozen.2 % relative. There is a tendency to be employing a hybrid approach for recognizing emotion from speech that may be a combination of Vector quantization (VQ) and mathematician Mixture Models (GMM). A quick review of labor applied within the space of recognition victimization VQ-GMM hybrid approach is mentioned here.

keywords—*Anchor models, emotional speech, emotion recognition, GMM model, VQ.*

*****

Introduction

Automatic Emotion Recognition (AER) from speech has garnered increasing interest in recent years given the broad field of applications that may enjoy this technology. as an example, a speaker emotion recognition system is commonly used to develop a further natural associate degreed effective human-machine interaction system that comes with an interface exhibiting larger sensitivity toward user behavior [6]. used during a distance learning context, a tutoring system may discover bored users and allow for a amendment of fashion and level of the equipped material, or provide Associate in nursing emotional encouragement [3].

During this method, emotions square measure expressed in terms of the presence or absence of a group of half emotions like anger, happiness, neutrality, and unhappiness. The EPs square measure created victimization SVM with Radial Basis performs (RBF). Emotion-specific SVMs square measure trained for each class as self-versus various classifiers. each EP contains n-components, one for the output of each emotion-specific SVM. The profiles square measure created by weight every of the n-outputs by the area between the individual purpose and conjointly the hyper plane boundary [5]. the ultimate emotion is chosen by classifying the generated profile in associate degree passing speaker-dependent fashion practice Naıve man of science. during a precursor technique supported the similarity conception, named WOC-NN, and has been planned. throughout this new framework, every emotion is depicted by a vicinity pattern composed of a group of emotion classes class-conscious to keep with their closeness or distance to each various. Classification is dispensed by computing distances between the check information neighborhood pattern and conjointly the particular patterns of each emotion class issued from coaching job [2].

For emotion recognition from speech, anchor model was experimented as a mixture methodology of various classifiers so as to boost the system performance. The experimental framework utilized in was adopted from language recognition and was composed of 2 parts: front-end and back-end systems. The anchor model was used as back-end to fuse 2 sub-systems, namely, a delivery GMM-SVM and delivery statistics-SVM systems. Finally, associate degree SVM classifier was accustomed train the back-end emotions within the anchor model area [1]. The reported results show that the anchor-model fusion methodology considerably improves recognition performance compared to the total rule fusion once tested on 2 out of 3 corpora.

Literature Survey

In this section we discussed about literature survey on emotion recognition from speech.

In [1], W. Li, Y. Zhang, and Y. Fu et al. aiming at emotion deficiency in gift E-learning system, plenty of negative effects were analysed and corresponding countermeasures were projected during this paper they combined emotive computing with the standard E-learning system. The model of E-learning system supported emotive computing was made by victimisation speech emotion that took speech feature as computer file. Their simulation experiment results show that neural networks is effective in emotion recognition, and that they achieved a recognition rate of roughly five hundredth once testing eight emotions. Different key techniques of realizing the system like following the amendment of emotion state and adjusting teaching methods were additionally introduced additionally.

In [2], E. Mower, M.J. Mataric, and S.S. Narayanan et al. Automatic recognition of emotion is changing into Associate in nursing progressively vital part within the style method for affect-sensitive human-machine interaction (HMI) systems. Well-designed emotion recognition systems have the potential to enhance HMI systems by providing extra user state details and by informing the planning of showing emotion relevant and showing emotion targeted artificial behaviour. This paper describes Associate in nursing emotion classification paradigm, supported emotion profiles (EPs). This paradigm is Associate in nursing approach to interpret the emotional content of realistic human expression by providing multiple probabilistic category labels, instead of one arduous label. EPs give Associate in nursing assessment of the emotion content of Associate in nursing vocalization in terms of a group of straightforward categorical emotions: anger; happiness; neutrality; and disappointment. This methodology will accurately capture the final emotional label (attaining Associate in nursing accuracy of sixty eight.2% in our experiment on the IEMOCAP data) additionally to characteristic underlying emotional properties of extremely showing emotion ambiguous utterances. This capability is helpful once managing realistic human emotional expressions, that square measure typically not well represented by one linguistics label.

In [3], Y. Mami and D. Charlet, et al. Speaker illustration by location may be a new technique of speaker recognition and adaptation. It consists in representing a replacement speaker, not in Associate in nursing absolute manner, however comparatively to a group of well-trained speaker models. every new speaker is depicted by its location in Associate in Nursing best illustration area. This paper addresses the situation task. It describes an illustration area engineered either by cluster speakers or by choosing Associate in Nursing best set of them. During this illustration area, speaker location is then performed by the anchor models technique to search out vector of coordinates. Associate in nursing orthogonalization method is then applied to the vector of coordinates, thus on reason the gap properly. This orthogonalization method (PCA or LDA) proves through an experiment to boost considerably the popularity.

In [4], P. Dumouchel, N. Dehak, Y. Attabi, R. Dehak, and N. Boufaden, et al. describe systems that were developed for the Open Performance Sub-Challenge of the INTERSPEECH 2009 emotion Challenge. They participate to each two-class and five-class emotion detection. For the two-class downside the simplest performance is obtained by logistical regression fusion of 3 systems. These systems use short- and long speech options. This fusion achieved Associate in nursing absolute improvement of two, 6 June 1944 on the un-weighted recall worth. For the five-class downside, we tend to submitted 2 individual systems: cepstral GMM vs. long GMM-UBM. The simplest result comes from a cepstral GMM Associate in Nursing created an absolute improvement of three, 5%.

PROBLEM STATEMENT

There are many approaches were investigated to boost emotion recognition performance significantly the discriminative and generative ones. Doubtless promising ways that are nevertheless to be deeply explored are those supported the similarity approach. The simplest technique, selected to beat the matter of skew category distribution, is classifier and options dependent. Thus, by virtue of its recursive simplicity that doesn't need any parameter calibration, it's low time execution quality, and at last its unfitness toward unbalanced information, the anchor models system supported distance metrics represent a sexy resolution to boost on the performance of generative models like VQ-GMM.

PRPOSED SYSTEM

After researching the bottom paper, here come back to conclude that, paper is predicated on basic GMM model for emotion recognition. The a lot of work is finished here during this paper is largely on anchor models, a similarity-based methodology, to resolve the multiclass emotion recognition downside. however if additional we tend to conjointly work on to improved GMM model, then this recognition performance ought to be increase as mentioned finally conjointly. so projected work we tend to area unit attending to use GMM + VQ [Vector Quantization] model for options extraction.

Aims and Objective:

during this project we've main aim is to gift the extended climbable technique to acknowledge emotion from speech. except this below are the most objectives of this project:

- To gift literature review over emotion recognition.
- To gift the sensible simulation of projected algorithms and valuate its performances.
- To gift the comparative analysis of existing and projected algorithms so as to assert the potency of emotion recognition system.

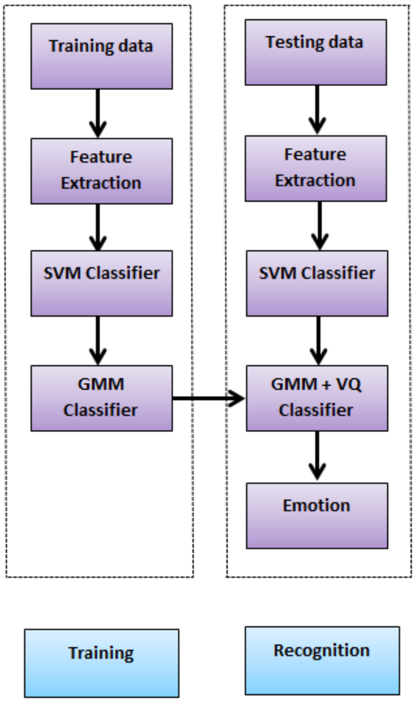


Figure1: Speech Emotion Recognition System

I. ARCHITECTURE DIAGRAM

A. *General Feature Extractions*

First extract all general options like audio mechanical device, sample rate, audio Channel Type: Mono or Stereo, no of channels, Content-Type: audio/mpeg version.

B. *Preprocess the Audio*

Now preprocess the audio file means that take away the silence, noise.

C. *MFCC Feature Extraction*

Mel-Frequency Cepstral Coefficients (MFCC) is used as choices to model the various nature of speech with connation the kind of feeling. The MFCC vector is formed by the first twelve coefficients along with C0 (energy component) measured at a rate of 10-ms and by applying 25-ms acting window

.D. *Anchor Models*

In anchor model system, feeling category is characterized by its relative live of similarity to different feeling classes. three steps characterize come with the looks of anchor model system: Construction of anchor area, representing acoustic options of AN anchor area and empowerment check feeling speech.

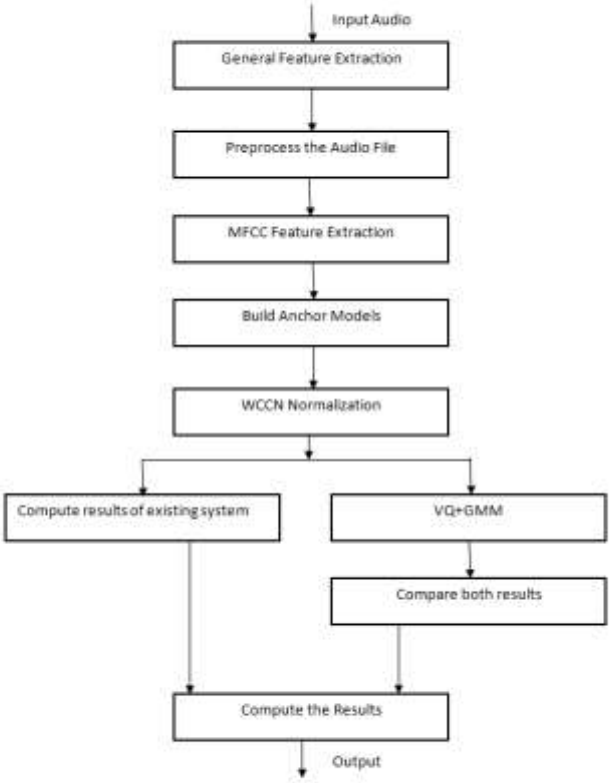


Figure 2: System Architecture Diagram

1. Construction of anchor space:

In this pattern recognition cons with limitless vary of subcategories like within speaker verification task, we'd prefer to hunt down audio of speeches of feeling. generally it needs a classified vary of classes, as for feeling recognition task, we've got amendment to model full set of feeling classes. thus we'll illustrate 2 main changes between the anchor models in identification and in feeling identification. first off identification anchor space contains a high dimensions, created of several speaker models. For feeling recognition space, anchor space dimension is relatively very little. second in speaker recognition, the speaker to qualify inside the anchor models, work or check stage doesn't belong to line of anchor models. On the opposite side in feeling recognition, the feeling pertains to line of anchor models, outstanding that every one feeling class models square measure used as anchor models.

2. Emotion Speech Classification:

The distance between ECV to categoryify of the check knowledge and people of every class representative is computed mistreatment either of following:

1. Euclidean Distance Metrics.
2. Cosine Distance Metrics.

A. *WCCN Normalization*

WCCN may be a technique introduced to coach a generalized linear kernel of AN SVM based mostly system to attenuate the exclusion of false positive and false negative errors.

II. MATHEMATICAL MODEL

Input: Audio Files
Output: Emotion Detection
System:

$$L(x) = \begin{bmatrix} \frac{1}{T}\log P(X \mid \lambda 1) \\ . \\ . \\ \frac{1}{T}\log P(X \mid \lambda c) \end{bmatrix}$$

Where, log P (X|) is the log likelihood of the feature vectors X given a GMM that belongs to the set of class models {A, E, N, P, R} and L(X) represents the ECV of X. log P (X|) is computed according to

$$\log P(X \mid \lambda i) = \sum_{n=1}^{T} \log P(X \mid \lambda i)$$

$$L^i = \frac{1}{n_i} \sum_q^{n_i} L(X_q^i)$$

Where X_q^i represents the q^th utterance of class i and n_i the number of training utterances of class i.

- Euclidean metric:

$$d(L_1, L_2) = \sqrt{| L_1 - L_2 |^2}$$

- Cosine metric:

$$d(L_1, L_2) = 1 - \frac{(L_1, L_2)}{||L_1||L_2||}$$

Where, <L_1, L_2>is the dot product of the vectors and L_1 and L_2

$$\text{emotion} = \text{argmin}_{i=1,\ldots,C}\ (d(L_T, L_i))$$

where, d represents the metric used to compute the distance between L_T, the ECV of the test data, and, L_i, the representative ECV of the emotion class *i*.

$$\text{emotion} = \text{argmin}_{i=1,\ldots,C}\ \log(P|\lambda_i)$$

$$\min = V^T V + c \sum_j €_j$$

Where, c is the slope of the hinge function, V is a vector normal to the decision boundary and is a slack variable.

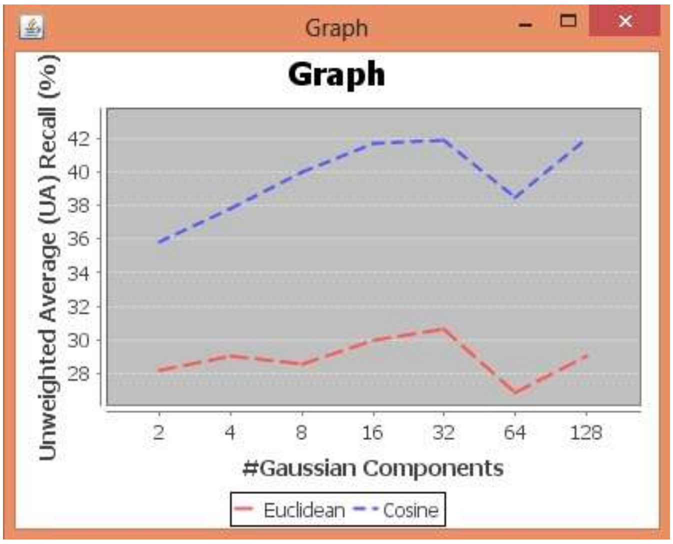


Figure 3: Result achieve using emotion training with respect to GMM

III. ALGORITHAM

```
1.  1.components = number of gaussian components
                p_ij = buffer array
                points = array of gaussian points
                upper_boundary = upper boundary of gaussian
                        components
              dimension = dimension

2.  2. for Gaussian Component component : components
              sum = sum of Gaussian Components
              if component=null
                   throw exception
              end if
              sum += component
        end for

3.  p = 0
        for j=0;j<points.size();j++
                p+= Math.log(getProbability((Matrix)
                        points.get(j)));
        end for
        return p

4.  for  GaussianComponent component : components
           p += component.getWeightedSampleProbability(x);
        end for
        return p;

5.  print the components
        for int i = 0; i < components.length; i++
              components[i].print()
              return components
        end for
```

```
6.  get mean of gaussian components
          return mean
7.  reading gaussian mixture model
            read gmm
            return gmm.
```

IV. Expected Result & Discussion

A. *Emotion speech classification:*

Fig three shows that the result discovered for every system GMM judge in coaching knowledge, during this we tend to sees that geometer distance got poor result as compared to cos based mostly system.

B. *WCCN Normalization:*

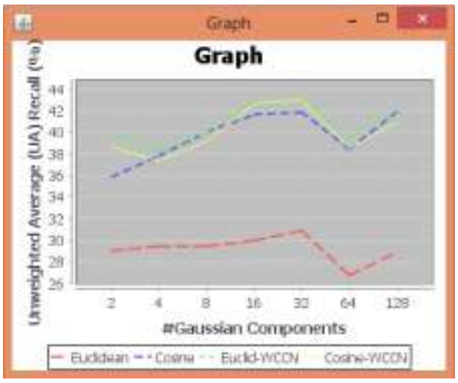


Figure 4: results of WCCN standardization on dataset performance of anchor model with relevance GMM.

Fig four shows that result classified of anchor model is employed on coaching knowledge cos and geometer distance matrix, when and before WCCN standardization. thus in this when applying WCCN on geometer and cos matrix performances square measure up to forty and 3.3.

C. *Class representative vector:*

In fig 5 shows that performance beastly reach with 2 clusters then the performance is drop forcefully for higher no of cluster.

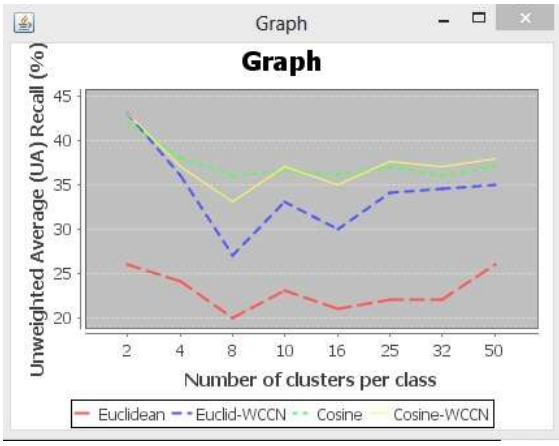


Figure 5: UA mean result of anchore model with respect to number of cluster per class By using clustring based class

Fig 6 shows that performance is became less delicate to campossed cluster so we get a best result with a less time required.

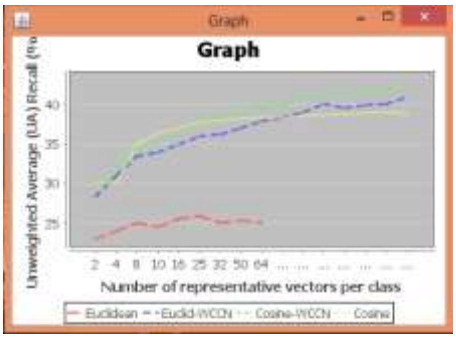
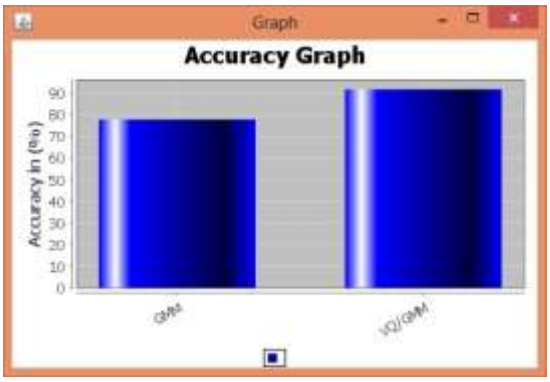


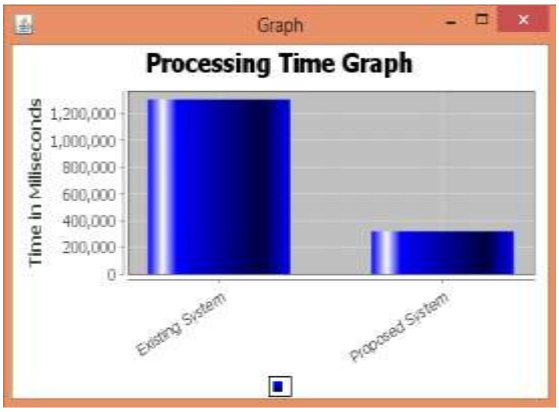
Figure 6: UA mean result of anchor model with respect to number of representative vectore per class by clsturing and weighing based method .

D. *Accuracy & time required:*

In this fig seven shows that once we use GMM i.e. existing system seventy eight accuracy that is a smaller amount than GMM+VQ i.e. planned system ninety two accuracy. planned system needed less time to run than existing system that is shown in fig eight.



X-axis: Existing system and Proposed System
Y-axis: Accuracy
Figure 7: Accuracy of the existing and proposed system



X-axis: Existing system and Proposed System
Y-axis: Time (MS)
Figure 8: Efficiency between existing & proposed systems

CONCLUSION

In this paper, we've got bestowed anchor models, a similarity-based technique, to unravel the multiclass emotion recognition drawback. We've got shown that once WCCN normalization, Euclidian or Cosine distances will be indifferently used as call metric to considerably improve performance of the front-end system, specifically the GMM model. After going through the base paper, we come to conclusion that, paper is based on basic GMM model for emotion recognition. The more work is done here in your paper is basically on anchor models, a similarity-based method, to solve the multiclass emotion recognition problem. But if further we also work on Improved GMM model, then this recognition performance should be increase as mentioned in conclusion also. Therefore proposed work we are going to use GMM + VQ [Vector Quantization] model for features extraction.

REFERENCES:

[1] Speech Emotion Recognition in ELearning system Based on Affective Computing, Proc. 3rd International Conf. Natural Computation (ICNC '07), pp. 809-813, 2007 by W. Li, Y. Zhang, and Y. Fu.

[2] A Framework for Automatic Human Emotion Classification using Emotional Profiles, IEEE Trans. Audio, Speech, and Language Processing by E. Mower, M.J Matric and S.S Narayanan.

[3] Speaker Identification by Anchor Model with PCA/LDA Post-Processing, Proc. IEEE Int'1 Conf. Acoustics, Speech, and Signal Processing (ICASSP'09), 2009 by Y. Mami and D. Charlet.

[4] Cepstral and Longterm Features for Emotion Recognition, Proc. 10th Ann. Conf. Int'l Speech Comm. Assoc. 2009 by P. Dumouchel, N. Dehak, Y. Attabi.

[5] Emotion Recognition using a Hierarchical Binary Decision Tree Approach, Speech Comm. Sensing Emotion and Affect Facing Realism in Speech Processing, vol. 53, Nov./Dec. 2011 by C. Lee, E. Mower, C. Busso and S. Narayanan.

[6] Weighted Ordered Classes Nearest Neighbors: A New Framework for Automatic Emotion Recognition from Speech, Proc. Conf. Int'1 Speech Comm. Assoc., 2011 by Y. Attabi and P. Dumouchel.

[7] Emotion Recognition from Speech: WOC-NN and Class Interaction, Proc. 11th Int'1 Conf. Information Science, Signal Processing and their Application (ISSPA'12), 2012 by Y. Attabi and P. Dumouchel.

[8] Emotion Recognition from Children's Speech using Anchor Models, Proc. Workshop Child, Computer and Interaction(WOCCI'12), 2012 by Y. Attabi and P. Dumouchel.

[9] Gang Liu, John H.L. Hansen, "Supra- Segemotional Feature Based Speaker Trait Detection", Odyssey 2014: The Speaker And Language Recognition Workshop 16-19 June 2014, Joensuu, Finland

[10] Firoj Alam, Giuseppe Riccardi, "Comparative Study Of Speaker Personality Traits Recognition In Conversationaland Broadcast News Speech", Conference: Interspeech