SPEECH EMOTION RECOGNITION USING ITERATIVE CLUSTERING TECHNIQUE

[1]R.ANILA, [2]DR. A. REVATHY

Saranathan College of Engineering Trichy, Tamilnadu-620 012
E-mail: [1]anilarajagopal09@gmail.com, [2]revathidhanabal@rediffmail.com

Abstract- This paper proposes a method to recognize the emotion present in the speech signal using Iterative clustering technique. We propose Mel Frequency Perceptual Linear Predictive Cepstrum (MFPLPC) as a feature for recognizing the emotions. This feature is extracted from the speech and the clustering models are generated for each emotion. For the Speaker Independent classification technique, pre- processing is done on test speeches and features are extracted. In K-means clustering algorithm, the classification is based on the minimum distance between each test vector and centroid of clusters. Mean of the minimum distances for each speech is found out. The test speech belongs to the model which has minimum of averages corresponds to particular emotion. We obtained better recognition rate by using MFPLPC feature for a cluster size 1024. The results are obtained for SAVEE database using data of seven emotional states such as anger, disgust, fear, happy, sad, surprise and neutral.

KeyWords- Emotion recognition (ER), Mel Frequency Perceptual Linear Predictive Cepstrum (MFPLPC), Vector Quantization (VQ).

I. INTRODUCTION

Despite the great progress made in artificial intelligence, we are still far from being able to naturally interact with machines, partly because machines do not understand our emotion states. Recently, speech emotion recognition, which aims to recognize emotional states from the speech signals, has been drawing increasing attention. Speech emotion recognition is a very challenging task of which extracting effective emotional features is an open task [1, 2].

Usually human beings can easily detect various kinds of emotions. This can be achieved by human mind through years of practice and observation. The human mind captures all kinds of emotions since childhood and is taught to differentiate between the emotions based on its observations. A human machine interface that can process speech having emotional content makes use of similar concept – training and then testing [3]. In training phase, the interface is fed with samples of each emotion. The classifier used in the interface extracts the features from all the samples and forms a model for each emotion. The classifier extracts the features from the input and compares it to all the models. The input is classified into that emotion to which it is closest.

Recent years have been marked by the increasing development of personal robots, either used as new educational technologies (Druin and Hendler, 2000) or for pure entertainment (Kusahara, 2000). These robots take the shape of young children such as humanoids. The interactions with these machines are radically different from the way we interact with the traditional computers. Among the capabilities these personal robots need, the most basic is the ability to grasp human emotions and in particular, they should be able to recognize human emotions as well as to express their own emotions. This system can also be applied to human machine interface in call centers where the computers have to respond to the users based on their emotional state.

This paper is organized as follows. Section II describes about the database which we have used for our work. Section III describes about the several previous works related to this study are presented. In Section IV feature extraction is presented. Section V & VI describes about the vector quantization, k-means clustering and emotion recognition based on our proposed system respectively. The results obtained for our proposed system is discussed in Section VII. Conclusion is summarized in Section VIII.

II. DATABASE DESCRIPTION

The data used for our proposed system is SAVEE database. Basically in this dataset there are 4 different speakers. Each one of them is asked to speak 15 different texts (in English).

Codes for emotions:
- a-Anger.
- d- Disgust.
- f- Fear.
- h- Happiness.
- sa- Sadness.
- Su- Surprise
- n- Neutral version.

Codes for text:

- 01- She had your dark suit in greasy wash water all year.
- 02- Don't ask me to carry an oily rag like that.
- 03- Will u tell me why?
- 04- Who authorized the unlimited expense account?

05- Destroy every file related to my audits.
06- Cory and Trish played tag with beach balls for hours.
07- He will allow a rare lie.
08- Withdraw all phony accusations at once.
09- Right now may not be the best time for the business mergers.
10- Kindergarten children decorate their class rooms for all holidays.
11- A few years later the dome fell in.
12- But in one section we welcomed auditors.
13- Lot people will roam the streets in costumes and masks, and having a ball.
14- In many of his poems, death comes by train; a strongly evocative visual image.
15- Then he would realize they were really things that only he himself could think.

Moreover, each one of the speakers speaks these sentences with different emotions. The different emotions shown during speaking these words are: anger, disgust, fear, happiness, surprise, sadness and neutral.

III. SPEECH EMOITON RECOGNITON

Most of the researches on SER have concentrated on feature-based and classification-based approaches [11]-[13]. Feature- based approaches aim at analyzing speech signals and effectively estimating feature parameters representing human emotional states. Many kinds of acoustic features have been introduced and verified. According to [11] and [12], short-term acoustic features such as pitch, energy, duration and MFCC (Mel- Frequency Cepstral Co-efficients) are frequently used in SER systems.

The researchers have investigated the relationship between languages and emotional characteristics. They emphasize the necessity for multi-lingual emotional data collection, since each language has a different way of expressing emotional states. Most current research on SER is performed based on distinctive language databases [11] [14].

Some methods for speech emotion recognition focus on finding the relevant features of emotions in speech. Recently, a wide range of physiological features form various analysis domains such as time/frequency, entropy, geometric analysis, sub band spectra, multi-scale entropy have been proposed to find the emotion- relevant features in music listening [15].

IV. FEATURE EXTRACTION

The important part of any emotion recognition system is to extract the features from the emotional speech and it should not change for the particular emotion and it should not change with time. Fig 3.1 shows the PLPC feature extraction model.

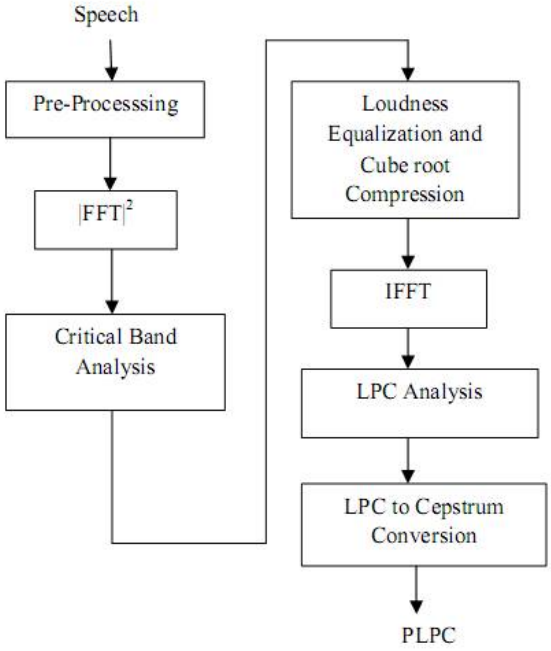


Fig. 3.1 PLPC extraction model

a. PLPC and MFPLPC feature extraction:

PLPC speech analysis is based on short-term spectrum. In quasi- steady state voiced regions of speech in which all pole model of LPC provides good approximation. The detailed procedure for PLPC and MFPLPC is given below.

1. Compute the power spectrum of windowed speech.
2. Perform grouping to 21 critical bands in bark scale and 47 critical bands in mel scale for sampling frequency of 16 KHz.
3. Perform loudness equalization and cube root compression to simulate the power law of hearing.
4. Perform IFFT.
5. Perform LPC analysis using Levison-Durbin procedure.
6. Convert LPC coefficients to cepstral coefficients.

The relationship between frequency in bark and frequency in hertz is given by

$$f_{(bark)} = 6 * \arcsin h(f_{(HZ)} / 100) \qquad (1)$$

The relationship between frequency in mel and frequency in hertz is given by

$$f_{(mel)} = 2595 \log_{10}\left(1 + f_{(HZ)} / 100\right) \qquad (2)$$

V. VECTOR QUANTIZATION AND K- MEANS CLUSTERING

a.Vector Quantization:

The concept of building a codebook of distinctive analysis vectors with significantly more than the basic set of phoneme is the attractive idea and basis behind a set of techniques called Vector Quantization methods.

b. Elements of Vector Quantization.

1. A large set of spectral vectors $\{V_1,V_2 \ldots V_n\}$ which forms the training set. The training set is used to create the optimal set of codebook vectors for representing the spectral variability observed in training set. The size of VQ codebook should be $M=2^B$ vectors. L >> M to find the best match of M codebook vectors in the robust manner.

2. A measure of similarity distance between a pair of spectral analysis vector. Spectral distance $d(V_i,V_j)$ between two vectors V_i and V_j is given by

$$d(v_i, v_j) = d_{ij} \{= 0 \; if \; v_i = v_j \quad (3)$$

3. A centroid computation procedure as described in Fig.1using K- means clustering algorithm.

4. A classification procedure. This is essentially a quantizer that accepts, as input, a speech spectral vector and provides, as output, the codebook index of the codebook vector that matches the input.

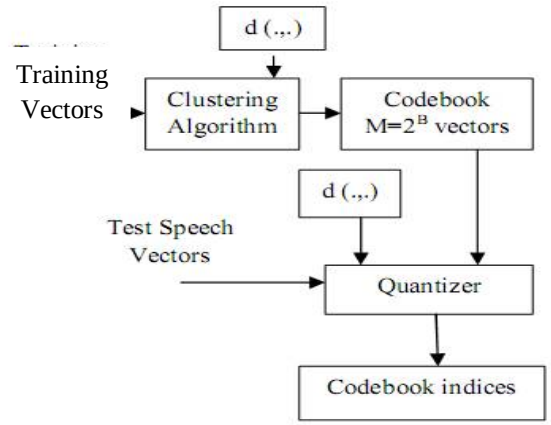


Fig.5.1. Block diagram of VQ training and classification structure.

If codebook vectors of M-vector codebook are taken as y_m, $1 \leq m \leq M$ and the new spectral vector to be classified is denoted as v, then the index m of the best codebook entry is given in (4)

$$m^* = \arg(\min(d(v, y_m))) \; for 1 \leq m \leq M \quad (4)$$

c. K- means Clustering Algorithm:

K- means clustering algorithm also known as Lloyd algorithm is described as follows.

1. Initialization: Arbitrarily choose M vectors as the initial set of code words in the codebook.
2. Nearest- Neighbour search: For each training vector, find the code word in the current codebook that is closest, and assign that vector to the corresponding cell.
3. Centroid Update: Update the code vector in each cell using the centroid of the training vectors assigned to that cell.
4. Iteration: Repeat steps 2 & 3 until the average distance fall below a preset threshold.

K- means is an well-known technique in unsupervised learning and vector quantization. The K- means clustering is formulated by minimizing a formal objective function, mean squared- error distribution.

$$MSE(P) = \sum_{i=1}^{N} \left\| x_i - c_{p(i)} \right\|^2 \quad (5)$$

Where

N is number of data samples;
K is number of clusters;
d is the dimension of data vector;
$X=\{x_1,x_2,\ldots..x_n\}$ is a set of N data samples.

$P= \{p(i)|i=1,2\ldots..N\}$ is a class of label N.
$C= \{c_j| \; j=1,2\ldots.K\}$ are k cluster centroids.

Due to its simplicity for implementation, the conventional K- means can be applied to a given clustering algorithm as a post processing stage to improve final solution [4].

VI. EMOTION RECOGNITION BASED ON INTERATIVE CLUSTERING

The recognition system involves the extraction of feature vectors from the training and the testing data, models are generated with the training data. Testing the emotional speech against these models is done to recognize the particular emotion.

For creating the training model, the speech signal is pre-emphasized to spectrally flatten the input speech using difference operator. Hamming window is applied on speech frames of 16 msecs duration with 8 msecs overlapping. Then the feature MFPLPC is extracted. For each emotion, VQ (Vector Quantization) codebook model is generated based on K- means clustering procedure. In K- means clustering algorithm there is a mapping of L training vectors into M clusters. Finally one model is created for each emotion.

For testing, the MFPLPC features are extracted for the test speech. The extracted features are fed to the training models. The minimum distance is found between test vector and centroid of clusters of speech models. Average of minimum distances for each speech model is determined. The test speech best matches with the emotional speech model which has minimum average value.

VII. RESULTS AND DISCUSSION

The performance of the emotion recognition system based on perceptual features is analyzed by applying test speech vectors to the training models of each emotion. Feature vectors of the test speech are applied to the clustering models and the emotion is identified based on the minimum average distances. The testing is done by computing average of

minimum distances for all the models and the model which produces minimum of averages corresponds to the particular emotion. Fig. 7.1 shows the comparison of recognition rates obtained for different cluster size.

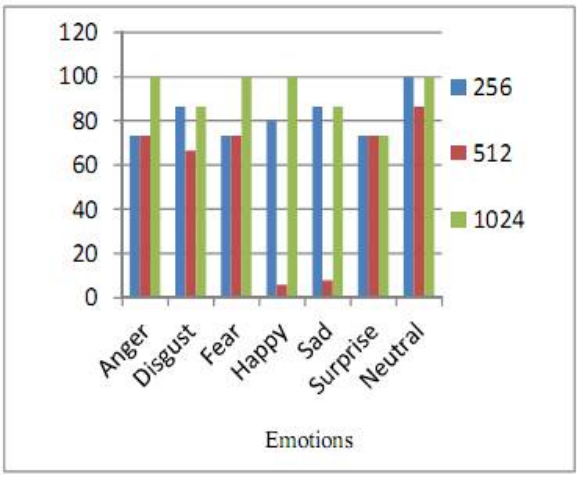


Fig. 7.1 comparison of recognition rates obtained for different cluster size.

For all the emotions except Disgust, Surprise and Sad, clustering model with cluster size 1024 gives 100% as accuracy. Table 1 indicates the confusion matrix of speaker independent emotion recognition for the cluster size 1024.

	1024 Clusters							
	A	D	F	H	SA	SU	N	RR
A	**100**	0	0	0	0	0	0	100
D	0	**86.6**	0	0	13.4	0	0	86.6
F	0	0	**100**	0	0	0	0	100
H	0	0	0	**100**	0	0	0	100
SA	0	13.4	0	0	**86.6**	0	0	86.6
SU	13.4	0	0	13.3	0	**73.3**	0	73.3
N	0	0	0	0	0	0	**100**	100

Tab. 1 Confusion matrix for 1024 clusters.

CONCLUSIONS

This paper proposes MFPLPC feature for recognizing the emotion present in the speech. The recognition rates obtained for each emotion is based on SAVEE database. Training models are developed using clustering technique and test speeches are applied to the individual models and performance is evaluated. The cluster size of training models is varied and the recognition rate is obtained for each cluster size. The cluster size 1024 gives the better recognition rate when compared to the 256 & 512 cluster size. For all the emotions except Disgust, Surprise and Sad, clustering model with cluster size 1024 gives 100% as accuracy.

REFERENCES

[1] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," Pattern Recognition , vol. 44, no. 3, pp. 572–587, 2011.

[2] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," Speech Communication, vol. 53, no. 9, pp. 1062–1087, 2011.

[3] L. R. Rabiner. & B. H. Juang ," Fundamentals of Speech Recognition", Prentice Hall, NJ1993.

[4] P. Friinti, J. Kivijirvi and 0. Nevalainen: ''Tabu search algorithm for codebook generation in VQ," Pattern Recognition31 (8),pp. -1148,August 1998.

[5] B.H.Juang; L.R.Rabiner,Hidden Markov Models for speech Recognition, Technometrics,Vol.33,No.3.(Aug,1991),pp.251-272.

[6] A.Revathi, Y. Venkataramani,"Speech Independent Continuous speech and Isolated Digit Recognition using VQ and HMM" International Journal on next generation networks, AIRCC, 2009

[7] R. Cowie, E. Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W Fellenz, and J. Taylor, "Eomotion recognition in human-computer interaction", IEEE Signal Processing Magazine, vol.18, pp. 32-80,2001.

[8] M. Shami and W. Verhelst,"An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech." Speech Communication, vol. 49, no.3, pp. 201-212, Mar.2007.

[9] T. New, S. Foo, and L. Silva, " Speech emotion recognition using hidden Markov models," Speech Communication, vol.41, nno.4 pp. 603-623, Nov 2003.

[10] P. Oudeyer, "The production and recognition of emotions in speech: Features and algorithms, " Int. Journal of Human-Computer Studies," vol. 59, pp. 157-183, 2003.

[11] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," Speech Communication, vol. 48,no. 9, pp. 1162-1181, Sep, 2006.

[12] O.Kwon, K. Chan, J. Hao, and T. Lee, "Emotion recognition by speech signals," Proc. Eurospeech, pp 125-128, 2003.

[13] R. Tato, R. Santos, R. Kompe and J. Pardo, "Emotion space improves emotion recognition,"Proc. ICSLP, pp. 2029-2032, 2002.

[14] E. Cowie, N. Campbell, R. Cowie and P. Roach, "Emotional speech: Towards a new generation of databases," Speech Communication, vol. 40 no. 1, pp. 33-60, Apr. 2003.

[15] J. Kim and E. Andre, "Emotion recognition based on physiological changes in music lestening," IEEE Trans. Pattern Analysis, Machine Intelligence, vol. 30, no. 12, pp. 2067-2083, December 2008.

[16] L.R.Rabiner, A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition - Proceedings of the IEEE, vol. 77, no. 8, Feb 1989.

[17] M. Richins, "Measuring emotions in the consumption experience", Journal of Consumer Research,Vol. 24, pp.127-146,1997.

★★★