Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 70 (2015) 29 – 35

Procedia Computer Science

4th International Conference on Eco-friendly Computing and Communication Systems

Emotion Detection using MFCC and Cepstrum Features

S Lalitha[a], D Geyasruti[a], R Narayanan[a], Shravani M[a]

Dept. of ECE, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Bangalore, Karnataka, 560035, India.

Abstract

A tremendous research is being done on Speech Emotion Recognition (SER) in the recent years with its main motto to improve human machine interaction. In this work, the effect of cepstral coefficients in the detection of emotions is performed. Also, a comparative analysis of cepstum, Mel-frequency Cepstral Coefficients (MFCC) and synthetically enlarged MFCC coefficients on emotion classification is done. Using a compact feature vector, our algorithm depicted better recognition rates of identifying seven emotions from Berlin speech corpus compared to the earlier work by Firoz Shah where only four emotions were recognized with good accuracy. The proposed method has facilitated a considerable reduction in the misclassification efficiency which outperforms the algorithm by InmaMohino, where the feature vector included only synthetically enlarged MFCC coefficients.

Peer-review under responsibility of the Organizing Committee of ICECCS 2015

*Keywords:*Enlargement; MFCC; Cepstral coefficients; Neural Networks; Berlin Database.

1. Introduction

It is very difficult to predict human emotions quantitatively. Though facial expressions and gestures are the best ways to figure out one's emotions, it becomes difficult to identify them as the age of a person increases, because people learn to control their expressions with age and experience. Moreover, the expressions and gestures reveal only the external emotions like anger, happy and sad but fail to do in case of emotions like disgust, boredom etc. To overcome this, different methods are discovered for emotion recognition. SER is one among them. Speech signal can be used to articulate various kinds of emotions. SER system identifies emotions on paralinguistic basis. It also plays an important role in finding out the psychological state of a person. The urge to improvise the efficiency and naturalness of Human Machine Interaction (HMI) derives major motivation for the work. MFCC coefficients

Peer-review under responsibility of the Organizing Committee of ICECCS 2015
doi:10.1016/j.procs.2015.10.020

derived from human speech samples play a vital role in the field of speech signal processing. They are used in applications including speaker verification, speaker recognition, emotion detection etc.

Earlier researchers have incorporated MFCC coefficients in the feature vector for identifying the paralinguistic content but could recognize only three emotions [1] and four emotions with poor recognition accuracy [2]. The feature vector consisted of first 19 MFCC coefficients [2] and a total of 63 MFCC features [3]. The available raw signal is too small to use it both for training and testing. Hence, the speech signals are synthetically enlarged so that the signal will be sufficient both for training and testing. Using Synthetic enlargement of MFCC's, the emotion misclassification efficiency reduces [4]. Compared to the MFCC method, the use of Subband based Cepstral parameters increases the classification efficiency by 19% [3].

The task of emotion classification involves two stages. The first stage is feature extraction followed by classification. Here MFCC, Cepstrum and MFCC enlarged coefficients are the speech features considered. The effect of these features and their possible combinations on SER is analyzed. Neural Networks does emotion identification and recognition work.

2. Emotion Corpus

The training and testing is done using Berlin Emotional database. It consists of 5 male and 5 female speakers (total 10 speakers) who were asked to speak 10 different sentences in German. It is an acted database consisting of 535 speech signals which are divided into 7 categories anger, boredom, disgust, fear, happy, neutral and sad.

Table1 gives the number of samples in the database corresponding to each emotion while Fig1 summarizes the proposed algorithm in this work.

Table 1.The Berlin Database

EMOTIONS	# OF SAMPLES
Anger	127
Boredom	81
Disgust	46
Fear	69
Happy	71
Neutral	79
Sad	62

3. Feature Vector

3.1 *Mel Frequency Cepstral Coefficients*

Mel Frequency Cepstrum (MFC) is a representation of linear cosine transform of a short-term log power spectrum of speech signal on a non-linear Mel scale of frequency. Mel-frequency cepstral coefficients (MFCCs) are together make up an MFC. MFCC extraction is of the type where all the characteristics of the speech signal are concentrated in the first few coefficients [3].

3.2 Cepstrum

Cepstrum is obtained by taking the inverse transform of the logarithm of Fourier transform of the signal [5].

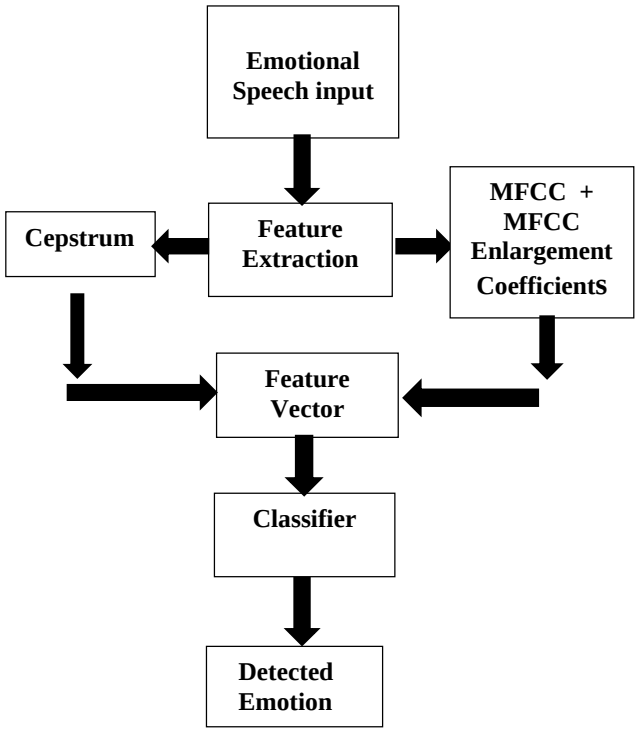

Fig1. Block diagram of the proposed algorithm

4. Classifier

Artificial neural networks (ANNs), inspired by biological neural networks (the central nervous systems of animals, in particular the brain), are a family of statistical learning models. They are used to calculate or approximate functions that may depend on a large number of unknown inputs. Artificial neural networks are represented as system of “neurons” which are interconnected and send messages to each other. The connections have numeric weights that can be assigned or tuned on experience, thus making neural nets adaptive to the inputs and capable of learning.

In this work, neural network pattern recognition tool (NPR tool) has been used. NPR tool performs supervised training and testing [2][6].

5. Results and Analysis

The emotion samples from the database are divided into training and testing samples. The training data consists of 280 samples (40 samples of each emotion) and the testing data consists of 42 samples (6 samples of each emotion). In each of the training and testing data, signals are arranged in sets, each set having one signal of each emotion in the order anger, boredom, disgust, fear, happy, neutral and sad. Corresponding target matrices are constructed for training and testing data. The emotion labels and target values corresponding to each emotion are represented in Table2 and Table 3.

Table 2.Emotion labels in the NPR tool

EMOTIONS	LABEL
Anger	1
Boredom	2
Disgust	3
Fear	4
Happy	5
Neutral	6
Sad	7

Table 3.Target Values for each Emotion

Emotion	1	2	3	4	5	6	7
1	1	0	0	0	0	0	0
2	0	1	0	0	0	0	0
3	0	0	1	0	0	0	0
4	0	0	0	1	0	0	0
5	0	0	0	0	1	0	0
6	0	0	0	0	0	1	0
7	0	0	0	0	0	0	1

Neural networks classifier performs supervised training and testing. The amount and extent of training depends upon the number of iterations which in turn depend upon the number of hidden neurons and number of layers in the network. The training of the classifier consists of three stages: training, validation and testing.

In Table 4, the frequency scaled MFCC's are the MFCC coefficients after enlargement.

Table 4.Emotion Recognition Accuracy of different Combinations of Feature Vector

FEATURE SET	ACCURACY
MFCC	57.1%
Frequency scaled MFCC	42.9%
Cepstrum	71.4%
Frequency scaled MFCC + Cepstrum	28.6%
MFCC + Cepstrum	57.1%
MFCC + Frequency scaled MFCC	71.4%
MFCC + Cepstrum+ Frequency scaled MFCC	**85.7%**

Table 5.Change in Recognition Accuracy

Case	Train %	Validate %	Test %	# of neurons	ER %
1	70	15	15	20	20.3
2	70	15	15	50	45.6
3	80	10	10	55	85.7

*ER = Efficiency

It can be observed from Table 4 that individually, the effects of Cepstrum coefficients and MFCC with frequency scaled coefficients have the same effect on the recognition accuracy of 71.4%. Best results were obtained when the feature vector consisted of the combined features (2 MFCC coefficients, 2 MFCC coefficients after enlargement

and 3 cepstral coefficients i.e total of 7 coefficients) resulting an accuracy of 85.7% with an improvement in the classification accuracy of 14%.

From Table 5, the best recognition rate of 85.7% was obtained in case 3 where a two layered neural network was formed consisting of 55 hidden neurons thus increasing the number of iterations and efficiency of the network. Increase in the training samples and neurons have given the best results. The confusion matrix for this efficiency is depicted in Table 6.

Table 6.Final Confusion Matrix

Emotion	1	2	3	4	5	6	7
1	6	0	0	0	0	0	0
2	0	6	0	0	0	0	0
3	0	0	6	0	0	0	0
4	0	0	0	6	0	0	0
5	0	0	0	0	6	0	0
6	0	0	0	0	0	6	0
7	0	0	0	0	6	0	0

It can be observed that Sad is classified as Happy. The misclassification is happening as the feature vectors of happy and sad signals are almost in the same range.

Table 7.Recognition Accuracy of each Emotion

EMOTIONS	ACCURACY
Anger	100%
Boredom	100%
Disgust	100%
Fear	100%
Happy	100%
Neutral	100%
Sad	0%
Average	**85.7%**

The overall recognition efficiency is 85.7% which can be observed from Table 7. This work yielded better results compared to reference paper [4] as the error has changed from 33% to 14.3% yielding around 20% reduction in misclassification efficiency. Next, a comparative analysis is done with the four emotions considered in reference paper [2] and is tabulated in Table 8.

Table 8.Comparison of Recognition Accuracy

EMOTIONS	ACCURACY (PROPOSED ALGORITHM	ACCURACY (REFERENCE PAPER[2])
Anger	100%	95%
Happy	100%	80%
Neutral	100%	10%
Sad	0%	35%
Average	**75%**	**55%**

From Table 8, among the four emotions, Sad was not recognized while other three emotions were best detected. However, in the earlier work [2], recognition rates for Neutral and Sad depicted low. Thus, our feature set has resulted in an overall improvement of 20% accuracy.

The results in Table 7 were obtained for seven emotions inferring the fact that our algorithm is better in terms of number of emotions as well as recognition efficiency compared with the work in reference paper [2]. Here, the emotion Sad is not recognized but the proposed algorithm should be better since it is poorly recognized in the earlier work [2]. Also, the overall recognition accuracy is calculated including all the seven emotions in which Sad is

totally unrecognized.

6. Cross Validation

Cross validation is a method or technique of predicting the working of a model. It helps in estimating the working of a model for an unknown dataset. Here a part of the training data is used for cross validation, thus estimating an approximate output or efficiency that can be obtained when an unknown data is tested.

As mentioned in Case3 of Table IV, 10% of the training data is used for cross validation.

Table 9.Cross Validation Matrix

E	1	2	3	4	5	6	7	O
1	0	0	0	0	0	0	0	NaN
2	0	4	0	0	0	0	0	100
3	0	0	3	0	0	0	0	100
4	0	0	0	4	0	0	0	100
5	0	0	0	0	5	0	0	100
6	0	0	0	0	0	4	0	100
7	3	0	0	0	0	0	5	64.5
T	0	100	100	100	100	100	100	89.3

*E = Emotion, O = output class efficiency,
T = Target class efficiency

The validation efficiency is 89.3%. It can be concluded that the overall output efficiency will be around 89.3% ±5%. This is true as the output efficiency is 85.7% which lies in the validation range.

7. Conclusion and Future work

One method of SER has been presented in this paper and an accuracy of 85.7% is obtained in detecting 7 emotions. These results are achieved using cepstral based features compact feature vector, and a simple Neural Network Classifier. Since a supervised testing is done, it is better compared to six. In this work, seven emotions are considered but the emotion Sad could not be recognized. Sad and Happy are two extreme emotions having a very narrow feature set and this is leading to a misclassification.

Future scope may be to further improve the efficiency of this work. The misclassification could be overcome by making further changes in training and testing ratio of speech samples. Here, only cepstral features have been considered for emotion recognition. The work can be extended to combine both time domain and frequency domain features along with the proposed features. Also the algorithms may be tested with different databases.

References

1. Bedoya-Jaramillo, E.Belalcazar-Bolanos, T.Villa-Canas, J.R.Orozco-Arroyave, J.D.AriasLondono, J.F.Varagas-Bonnilla " Automatic Emotion detection in Speech using Mel frequency Cepstral Coefficients", XV Simposio De Tratamiento De Senales, Images, Vision, Artificial-STSIVA 2012.
2. FirozShah.A, Vimal Krishnan V.R., RajiSukumar.A, AthulyaJayakumar, BabuAnto.P "Speaker Independent Automatic Emotion Recognition from Speech:-A Comparison of MFCCs and Discrete Wavelet Transforms" 2009 International Conference on Advances in Recent Technologies in Communication and Computing, pp:528-531,2009
3. K.V.Krishna Kishore, P. Krishna Satish "Emotion Recognition in Speech using MFCC and Wavelet Features",3rd IEEE International Advance Computing Conference.
4. Inma Mohino-Herranz1, Roberto Gil-Pita1, Sagrario Alonso-Diaz2 and Manuel Rosa-Zurera1 "MFCC Based Enlargement of the Training set for Emotion Recognition in Speech" International Journal (SIPIJ) Vol.5, No.1, February 2014.
5. J. SirishaDevi ,Dr. Srinivas Yarramalle,Siva Prasad Nandyala "Speaker Emotion Recognition Based on Speech Features and Classification Techniques" I.J. Image, Graphics and Signal Processing, 2014, No:7, pp: 61-77 ,June 2014.

6. Mehmet S. Unluturk, Kaya Oguz, CoskunAtay “Emotion Recognition Using Neural Networks” , World Academy of Science, Engineering and Technology, Vol:7, No:3, 2013