Minimal feature set based classification of emotional speech

Agnes Jacob, P.Mythili

Abstract---This paper proposes the use of a minimum number of formant and bandwidth features for efficient classification of the neutral and six basic emotions in two languages. Such a minimal feature set facilitates fast and real time recognition of emotions which is the ultimate goal of any speech emotion recognition system. The investigations were done on emotional speech databases developed by the authors in English as well as Malayalam - a popular Indian language. For each language, the best features were identified by the KMeans, K-nearest neighbor and Naive Bayes classification of individual formants and bandwidths, followed by the artificial neural networks classification of the combination of the best formants and bandwidths. Whereas an overall emotion recognition accuracy of 85.28 % was obtained for Malayalam, based on the values of the first four formants and bandwidths, the recognition accuracy obtained for English was 86.15%, based on a feature set of the four formants and the first and fourth bandwidths, both of which are unprecedented. These results were obtained for elicited emotional speech of females and with statistically preprocessed formants and bandwidth values. Reduction in the number of emotion classes resulted in a striking increase in the recognition accuracy.

Index Terms: - formants, bandwidths, speech emotion recognition, recognition rate, artificial neural networks

1 INTRODUCTION

SPEECH is an evergreen area of research since the fundamental mode of human communication is still verbal. Several studies have been conducted in the past, in order to assess the emotional state of the speaker from the speech samples such as reported by Noel [1]. The significance of research in this area, apart from its well known relevance in modern human-computer interface, is illustrated by Ramakrishnan [2] in an overview of several interesting applications of emotional speech recognition. Most of such reported works in emotional speech recognition focus on reducing a very large set of features extracted from speech, to a much smaller and better manageable set of essential features by various techniques. This process often involves the application of complex algorithms that require greater computational effort than that applied for the emotional speech classification itself, and is therefore time consuming. Hence the authors were motivated to simplify the speech emotion recognition (SER) problem by identifying a small set of features extracted from emotional speech and to verify its efficiency in SER for English and Malayalam.

1.1Features- Formant Frequency and Bandwidth

Effective human and automatic processing of emotional speech requires recovery of not only the prosodic information, but also spectral content of the speech signal, of which the resonant frequencies of the vocal tract are the most important and are commonly referred to as formants. The vocal tract is a vital element for speech production. Quantitative information of the vocal tract that is provided by the formant frequencies has widely been used in SER along with several other features. These resonant frequencies change with the size and shape of the vocal tract which in turn is determined by the emotional state of the speaker at that moment. Therefore formants can be used as important features in SER. Each formant is characterized by its center frequency and its bandwidth. The formants (F) can be used to discriminate the improved articulated speech from the slackened one. Ververidis [3] has observed that the formant bandwidth during slackened articulated speech is gradual, whereas the formant bandwidth (B) during articulated speech is narrow with steep flanks.

This paper is organized as follows: Section 2 briefly outlines the methodology that has been adopted for the creation of the emotional speech database and the various approaches adopted by the authors for the selection of the best features efficient for classification of emotional speech. Section 3 presents the experimental results obtained for English and Malayalam and discusses the significance of these results. Section 4 concludes this paper, highlighting the contribution of our work.

2 METHODOLOGY

This section provides information on the design and development of the speech database along with the speaker

- *Agnes Jacob is currently pursuing doctoral research in Speech processing in Division of Electronics, School of Engineering, Cochin University, Kerala, India PH-+91-9446522793. E-mail: agneswills3@gmail.com*
- *Dr P.Mythili is currently the Head of Division of Electronics, School of Engineering, Cochin University, Kerala India. E-mail: mythili@cusat.ac.in*

profiles. It mentions the feature extraction method and briefly introduces the various classifiers used in our approach.

2.1 Database Development

The authors chose to investigate on female speech as it is emotionally more expressive than male speech. Since there are no public databases for these seven specific emotions in the selected Indian languages, first of all an exclusive database was developed for the two mentioned languages covering the seven emotions. Research on emotional speech relies on the richness and appropriateness of the databases which were therefore designed taking into account various gender, social and linguistic aspects as suggested by Giri [4] and Jones [5]. Since spontaneous emotions are very difficult to record and acted emotions have exaggerated expressions, a database of elicited emotions was developed. It consists of short, but often used utterances.

The ten female subjects selected were educated, non-professional, urban, Indian speakers of English and Malayalam; in the age group of 32–42 years and well aware of the purpose of the recordings. Since emotions had to be elicited, several trials were required in order to get sufficient good samples of each utterance. The recordings were done in different sets on to a computer hard disk and the corresponding wave files were then segmented, labeled and stored. The database of nearly 1600 wav files in each language was subjected to perceptual listening tests to verify the correctness of emotional content in the recordings under each of the seven classes of emotion for the selected languages.

Feature extraction was done using the Pratt software. The formant and bandwidths values were tabulated and statistical preprocessing of these values was done by the repeated measures analysis of variance. Only sample values that differed between various emotion classes, with a high significance level were used for classification.

2.2 Classifiers used in this work

Classifications of formants and bandwidths were done using Matlab. For each language, the selection of features to the final minimal feature set was made by assessment of the recognition rate (RR) obtained using three different classifiers with feature values in their raw or modified format. The RR is adopted as a measure of accuracy and defined as

$$\text{Recognition rate} = \frac{\text{Number of successfully detected emotional samples}}{\text{Total number of utterances in a class}} \quad (1)$$

The Kmeans classifier is a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster. G.S.Nayak [6] observed that clustering is done such that patterns in the same cluster are alike and patterns belonging to different clusters are different. *The Naive Bayes classifier* is a probabilistic one, working on the principle of the Bayes theorem and assumes feature independence across classes**. T**hough a very simple classifier, it has the advantage of being very fast without performing much worse than more sophisticated classifiers such as support vector machines, as reported by Lugger and Yang [7]. *The K-Nearest Neighbor (KNN)* classification algorithm is based on the assumption that samples residing closer in the sample space belong to the same class. When a new sample data x arrives, KNN finds the k neighbors nearest to the unlabeled data from the training space based on some distance measure such as the Euclidean distance. This supervised technique yields accurate results in most of the cases and has been widely used in SER [8], [9]

A*rtificial Neural network classifications* are popular in SER systems [10], [11], [12]. A two-layer feed-forward, back propagation, network with sigmoid hidden and output neurons and sufficient neurons in its hidden layer was used for classification. First, the classification problem is defined through the set of inputs and corresponding targets. In this supervised learning technique the network is trained to classify the inputs according to the targets. The Mean Squared Error (MSE) is the average squared difference between the outputs and targets. Percent Error (%E) indicates the fraction of samples which are misclassified.The total number of samples was randomly divided into three classes for training, validation and testing. For both languages, classification was done for various combinations of formants and bandwidths separately. After identifying the best combinations in the separate groups of formants and bandwidths that yield the least classification error, ANN classification was done for the best combination of both formants as well as bandwidths. The classification was repeated for different network sizes in order to arrive at the optimum size of the network.

3. EXPERIMENTAL RESULTS AND DISCUSSION

This section provides the reader with the necessary information regarding the performance of the proposed spectral features in terms of SER rates obtained using the aforesaid classifiers.

3.1 English

Selection of features to the minimal set: The features to be included as input to the ANN classifier were selected on the basis of their respective individual RR obtained using the first three classifiers mentioned above. The K Means classifier gave 100% recognition of disgust on the basis of F1 and gave

nonzero, but very small RR for all other emotions based on each feature.

The Naïve Bayes classifier gave RR as high as 92.3% in the case of anger on the basis of the B2 but failed to recognize even a single instance of neutral , sad or disgust on the basis of B2 itself. Thus the Naive Bayes classifier performed poorly in this class seven emotion recognition problem in English, as it failed to recognize certain emotions for each of the eight features considered.

The KNN classifier gave the highest overall RR based on the various features. Besides, the KNN classifier was able to identify all the seven emotions on the basis of the formant and bandwidth features considered individually. With this classifier, the highest RR of 90.9% was obtained for disgust on the basis of B1. Based on the above, the KNN classifier is the best among the three classifiers used in this investigation for English. All these three classifiers gave higher RR in a mixed manner, for the raw and modified values of the various features and deciding on the optimum feature format has not come under the purview of this work.

Rejection of B2 and B3 from the feature set: Across the three classifiers and eight features, the overall recognition rate was the least (21.9% only) for B2 and it gave very poor recognition rates for disgust and fear. The emotions best recognized based on B3 are happy, anger and surprise, all of which have been better recognized by B4, F1, F2, F3 and F4. Disgust was poorly recognized on the basis of B3. Further details of the classification as given in Table 1 below show the average recognition rates for each emotion (under each of the 8 features, across the 3 different classifiers) as belonging to one of the four classes - very good, good, fair and poor.

TABLE 1
CONSOLIDATED RR FOR FORMANTS AND BW- ENG

Emotions	F1	F2	F3	F4	B1	B2	B3	B4
Hap	G	P	F	P	F	F	G	G
Surp	F	G	G	G	F	G	F	F
Neut	P	F	F	F	G	F	F	F
Ang	√	F	P	G	F	F	√	F
Sad	F	F	F	G	G	F	F	F
Fear	G	√	G	F	G	P	F	F
Disg	F	F	G	F	G	P	P	F

Key:very good (√)-RR > 50; good (G) - 50 > RR >35; Fair (F)-RR <35 Poor (P) RR<15

Table 1 indicates that across the various features and emotions, there were only three cases of very good RR. The authors noted the feasibility of reduction of features for SER in English by omitting B2 and B3 since the few emotions that were well recognized using these features were better recognized by means of the other 6 features as evident from Table 1. The ANN RRs were found out for each of the individual formants as well as for various combinations of formants. The best result in terms of the least percentage error was obtained when using all the four formant values and is as given in Table 2 below. Likewise separate classifications based on each of the four bandwidths and their various combinations was carried out, resulting in significantly less classification rates than those obtained with formants ,thereby indicating bandwidth to be less efficient than formants in the classification of emotions. Subsequently, classification was carried out for various combinations of formants and bandwidths. A slight improvement in performance was noted for a combination of formants with B1 and B4 over that based on formants alone. The salient results are presented in Table 2. The performance of the classifier was noted by changing the network size that is, for different number of neurons in the hidden layer. The optimum number was found to be 60 neurons as it yielded the least MSE as well as the least percentage error on the test data.

TABLE 2
FEATURES WITH BEST RR IN ENGLISH

Features	Identity	MSE	%E
BW	B4	1.16 e^{-1}	60.0
Formants and BW	F1 toF4 and B1	4.1e^{-2}	14.72

Whereas emotion RR of 85.3% was obtained for formant based ANN classification, it was only 80.8% with hundreds of other features of the bench marked emotional speech database eNTERFACE comprising elicited basic emotions and by applying the deep neural network based Generalized Discriminant analysis technique propose by Andre et al [13] .

The output of the classifier for the minimal feature set is presented in the confusion matrix given in Table 3 below.

TABLE 3
CONFUSION MATRIX OF RR (PERCENTAGE) FOR FORMANTS, B1, B4 IN ENGLISH

Emo	Hap	Surp	Neut	Ang	Sad	Fear	Disg
Hap	90.3	0	6.5	0	3.3	0	0
Surp	5.2	*84.5*	10.3	0	0	0	0
Neut	9.1	0	*81.8*	6.1	0	3	0
Ang	0	0	0	*79.5*	0	2.5	17.9
Sad	0	0	0	0	*90.3*	0	9.6
Fear	0	0	0	0	7.9	*81.6*	10.5
Disg	0	0	0	0	5	0	*95*

Surprise has no false hits whereas disgust which has the highest recognition rate has the maximum number of false hits

too. Anger has the least recognition rate and it has been confused mostly with disgust. Table 4 shows the improvement in performance obtained by reducing the number of emotion classes

TABLE 4
FFBPNN PERFORMANCE OF FORMANT BASED SER OF VARIOUS CLASSES IN ENGLISH

Problem class and description	No. of neuron	MSE	%E
-positive and negative	45	$4.97e^{-3}$	0
3-positive, neutral and negative	60	$1.98e^{-2}$	2.86
4- 2 surp, neut, ang, sad,	60	$2.64e^{-2}$	5
5- surp,neut, ang,sad, fear	60	$4.18e^{-2}$	8
6-hap, surp, neut, ang, sad, fear	60	$3.57e^{-2}$	10
7- neutral and six basic emotions	60	$4.45\ e^{-2}$	16.88

The overall RR of the above neural network classifier for English (with all formants, B1 and B4) is 86.14% , with a maximum recognition accuracy of 95% for disgust and 90.3% for sad and happy.

3.3 Malayalam

The K Means classifier gave the highest RR of 66.7% with B4 and B3, for happy and fear. This classifier recognized all emotions based on each of the eight individual features, but at a lesser rate, based on raw feature values. The *NB classifier* failed to recognize certain emotions at the expense of high RR of certain other emotions. Among formants, the NB classifier gave the best results on the basis of F4 and surprise was the most recognized. Anger was the worst recognized. For each feature, the KNN classifier recognized all the seven emotions and gave the highest overall classification accuracy, exhibiting superior performance over the other classifiers used in this investigation on single formant- bandwidth feature based Malayalam emotional speech recognition. The highest recognition rate obtained was 80% for neutral and based on F2 values. In the bandwidth based classification, the KNN classifier gave 100% recognition rates for sad and fear based on B1 leading to the inclusion of the latter in the feature group for the final ANN based classification of emotional speech. The consolidated RR is given in Table 5. The first three bandwidths are seen to give very good RR for several of the six emotions. F1 and B2 are seen to give very good recognition rates. For Malayalam, the fourth formant was only moderate performer in speech emotion recognition. Since there were no instances of either very good RR or poor recognition for F4, the authors decided to include it with other seven features in the final ANN classification.

TABLE 5
CONSOLIDATED RR IN MALAYALAM

Emotions	F1	F2	F3	F4	B1	B2	B3	B4
Hap	√	G	G	F	G	F	F	√
Surp	√	F	F	F	F	√	F	G
Neut	G	√	F	G	F	G	F	F
Ang	F	F	F	F	√	√	√	G
Sad	G	G	G	F	G	√	√	F
Fear	F	F	√	G	G	F	F	P
Disg	√	G	P	F	√	G	G	G

Key:very good (√)-RR > 50; good (G) - 50 > RR >35; Fair (F)- RR <35; Fair (XP)- 15<RR <35; Poor (P) RR< 15

Very good RRs were obtained for each emotion on the basis of the various individual features other than F4. Since there were no instances of either very good RR or poor recognition for F4, the authors decided to include it with other seven features in the final ANN classification. Table 6 gives results of ANN classification based on combinations of formants alone, bandwidths alone and both formants as well as bandwidths.

TABLE 6
FEATURES GIVING THE BEST RR IN MALAYALAM

Features	Feature identity	MSE	Percentage error (%E)
Formants	F1 to F4	$6.196e^{-2}$	21.65
BW	B2	$1.247e^{-1}$	60
	B1	$1.200e^{-1}$	65.71
Formants and BW	F1 to F4 & B1 to B4	$4.04e^{-2}$	*13.85*
	F1 to F4 & B1, B2	$3.79\ e^{-2}$	15.15

The first four bandwidths considered in this study proved to be poor stand- alone features for ANN based emotion classification since better classification were obtained with their combinations. The feature group of the first four formants alone resulted in 78.35% overall classification accuracy in recognizing seven emotions. The classification accuracy obtained with all formants and two select bandwidths were reasonably good at 85.85%. In Malayalam, the least classification error for test data was obtained for classification based on a feature set comprising all four formants as well as bandwidths and the overall RR with the first four formants and bandwidths was 86.15% which has not been reported so far.

Table 7 below gives the confusion matrix of the classification accuracies (as percentage in each class) based on the complete set of eight features in Malayalam. 100% recognition of fear in this class 7 emotion recognition

problem is significant. Surprise and sad were also well recognized.

TABLE 7.
CONFUSION MATRIX OF RR BASED ON FORMANTS AND BANDWIDTHS IN MALAYALAM

Emotion	Hap	Surp	Neut	Ang	Sad	Fear	Disg
Hap	*78.6*	0	0	0	0	0	21.4
Surp	0	*94.1*	0	0	2.94	2.94	0
Neut	16.7	0	*73.8*	0	2.4	2.4	4.8
Ang	2.6	2.6	5.2	*84.6*	2.6	2.6	0
Sad	10	0	0	0	*90*	0	0
Fear	0	0	0	0	0	*100*	0
Disg	0	0	0	0	10.3	3.4	*86.2*

Surprise and sad were also well recognized. Neutral was the least recognized in the group. Table 8 gives the ANN performance measures for various number of emotion classes.

TABLE 8
FFBPNN PERFORMANCE FOR MALAYALAM SER

	Problem class and description	No. of neurons	MSE	%E
2	positive and negative	45	$1.49e^{-2}$	**0**
	positive, neutral and negative	60	$8.63e^{-2}$	8.57
3	Surp, neut, fear	45	$6.13e^{-2}$	**0**
4	Surp,neut ,ang,sad	55	$4.21e^{-2}$	**0**
5	surp,neut, sad, ang, fear	55	$2.75e^{-2}$	**0**
6	hap, surp, neut, ang, sad, fear	60	$3.05e^{-2}$	3.33
7	neutral and six basic emotions	60	$4.04e^{-2}$	13.85

Thus fear, surprise and sad have been very well recognized in Malayalam. In all cases considered here the optimum network had sixty neurons in the hidden layer.

With reference to the percentage error for classification of different numbers of emotions in Malayalam, 100% correct classification was observed for the broad classes of positive and negative emotions. With the addition of neutral samples to the group, however the maximum percentage of overall recognition accuracy of the class three problem decreased to 91.43%, due to the mix up of the formant- bandwidth values of neutral with those of certain emotions from both groups. However without avoiding neutral, for the specific combination of surprise neutral and fear, 100% was obtained as the highest accuracy for a class three emotion recognition problem. The highest recognition accuracy for a class four problem was 100% and had been reported by other researchers with a maximum of accuracy of 72.1% only, with a different feature set. For the class 5 problem the highest accuracy of 100% obtained only by avoiding happiness and disgust has to be viewed from the perspective of the confusion matrix for the class seven problem which clearly indicates high misses and false hits for happiness and disgust. For class six problems, the overall accuracy is 96.67%.

4 CONCLUSION

In this paper we have proposed minimal feature sets of formants and bandwidths for simple, yet efficient class - seven- SER in English and Malayalam and implemented the same. In both cases, the classification results of the first three classifiers with this approach are quantitatively superior to those reported so far, especially in malayalam. The artificial neural network (ANN) used for classification of such a minimal set of features outperformed the highest reported in these languages with a larger feature set. Statistical preprocessing of the minimal feature set values have also contributed to the improved RR since statistical difference was ensured for the sample values of various emotion classes fed to the classifiers For both languages, the SER rate increased with decrease in the number of emotion classes, from seven. Further investigations may be done to implement a complete real time SER system with this approach.

REFERENCES

[1] Noel Chateau, Valerie Maffiala, Thibaut Ehrette, Christophe d'Alessandro, "Modeling the Emotional Quality Of Speech In A Telecommunication Context." *Proceedings of the 2002 International Conference on Auditory Display*, Kyoto, Japan, pp.1-6, July 2002.

[2] Ramakrishnan S., & Ibrahim M.M,El Emary. "Speech Emotion Recognition Approaches in Human Computer Interaction". *Telecommunication Systems*. pp.1-5, doi:10.1007/S.11235-011-9624z

[3] Dimitrios Ververidis and Constantine Kotropoulos, "Emotional Speech Recognition : Resources Features And Methods". *Elsevier. Speech Communication* vol 48 pp. 1162–1181 April 2006.

[4] Vijai N.Giri. *Gender Role In Communication Style*. New Delhi: Concept Publishing House. 2004 pp.19-45

[5] Daniel Jones. *Cambridge English pronouncing dictionary*. Cambridge: Cambridge University Press. 2004.pp.3-55

[6] G. Subramanya Nayak, Ottolina Davide, Puttamadappa C, "Classification of Bio Optical signals using K- Means Clustering for Detection of Skin Pathology" . *International Journal of Computer Applications* (0975 – 8887),Volume 1 – No. 2, pp. 91-99. 2010

[7] Lugger, M., & Yang, B , "The relevance of voice quality features in speaker independent emotion recognition". In *ICASSP*,Honolulu, Hawaii, 2007

[8] Naïve Bayes B. Schuller, S. Reiter, R. Muller, "Speaker Independent Speech Emotion Recognition by Ensemble Classification," *Proc. ICME,* Amsterdam, Netherlands, pp. 864-867 2005,.

[9] Dellaert, F., Polzin, T., & Waibel,A, "Recognising emotions in speech." In *ICSLP 96*, pp. 1970-1973, Oct. 1996.

[10] Schuller, B., Rigoll, G., & Lang, M, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture". In *Proceedings of the IEEE international Conference on Acoustics, Speech and Signal processing* pp. 577–580. New York: IEEE Press 2004.

[11] Nicholson J., Takahashi K., & Nakatsu, R. "Emotion recognition in speech using neural networks. In *6th International Conference on Neural Information Processing (ICONIP-99)*, Perth, WA, Australia, pp. 495–501, Aug. 1999.

[12] Nakatsu, R., Nicholson, J.,& Tosa, N. Emotion recognition and its application to computer agents with spontaneous interactive capabilities. *Knowledge-Based Systems*, *13*,pp. 497–504 , 2004.

[13] Andre Stuhlsatz, Christine Meyer, Florian Eyben, Thomas Zielke, Gunter Meier, Bjorn Schuller , "Deep Neural networks for acoustic emotion recognition: aising the benchmarks" *ICASSP 2011 IEEE* – 978-1-4577-0539-7/11,pp.5688-5692, 2011