An Improved Hindi Speech Emotion Recognition System

Agnes Jacob, P. Mythili

Abstract— This paper presents the results of investigations in speech emotion recognition in Hindi, using only the first four formants and their bandwidths. This research work was done on female speech data base of nearly 1600 utterances comprising neutral, happiness, surprise, anger, sadness, fear and disgust as the elicited emotions. The best of the statistically preprocessed formant and bandwidth features were first identified by the KMeans, K nearest Neighbour and Naive Bayes classification of individual features. This was followed by artificial neural network classification based on the combination of the best formants and bandwidths. The highest overall emotion recognition accuracy obtained by the ANN method was 97.14%, based on the first four values of formants and bandwidths. A striking increase in the recognition accuracy was observed when the number of emotion classes was reduced from seven. The obtained results presented in this paper, have not been reported so far for Hindi, using the proposed spectral features as well as with the adopted preprocessing and classification methods.

Index Terms— Formant, Emotion, KMeans, K nearest Neighbour, Naive Bayes, Artificial neural network.

I. INTRODUCTION

Speech and emotions have attracted the attention of psychologists, researchers in language and speech processing for several decades. Automatic detection of emotions in speech is crucial in man machine interaction. Speech emotion recognition (SER) is all the more challenging due to its extensive application potential as well as the involvement of diverse factors such as psychology, sociolinguistics and various speech signal processing techniques. Besides, different languages may have varied styles for expressing emotions due to various reasons such as the origin of the language considered as well as regional and cultural differences. This investigation was focused on Hindi, which is the fifth most spoken language in the world with about 188 million native speakers. Hindi is the common spoken language in India and is written in the Devanāgarī script [1]. This paper is organized as follows: The subsection A of the Introduction states the motivation for this research. Sub section B defines the problem investigated. Subsection C provides necessary background to the reader, with emphasis on the most recent relevant reported work. Subsection D introduces the formant and bandwidth features used in this work. The next section explains the methodology of this work. The obtained results are presented and analyzed in Section on Results and Discussion, so as to bring out their significance. The last section concludes this paper pinpointing the merits of this method and gives directions for future work.

Manuscript Received November, 2013.

Agnes Jacob , Research Scholar, Division of Electronics, School of Engineering, Cochin University of Science and Technology, Kochi, Kerala, India.

Dr P. Mythili, Head of Division of Electronics, School of Engineering, Cochin University of Science and Technology, Kochi, Kerala India.

A. Motivation for this research

In spite of the very many regional languages across various parts of India, Hindi is very popular, being the national language. The present educational and job scenario require people to move away from their native places or interact with non natives in their own locality. Both these situations call for a common language for oral communication which is often found to be Hindi.

Emotions are unavoidable in vocal interactions and it is very important to be aware of emotions and understand these correctly. Hence the researchers were motivated to focus on emotion recognition of Hindi speech.

B. Problem statement

The objectives of the present investigations on the Hindi speech database were threefold. These were (i) identifying the best spectral features among the first four formants and bandwidths, with the KMeans, K Nearest Neighbour(KNN) and Naive Bayes (NB) classifiers. (ii) Verifying the performance of the best formants and bandwidth features for SER, by means of the artificial neural network classifier. (iii) Analyzing the effect of the increasing complexity of the SER problem on the performance of the final ANN classifier based on formants and bandwidths. It was proposed to remove those features with very poor overall classification rates from the final feature set given to the ANN classifier. The next section briefs the reader on recent research in this area.

C. Existing work

Research in Hindi SER is relatively young, the main reason being the need to develop the speech databases for research. The development of an emotional speech database for happy, surprise, neutral, anger and sad emotions in Hindi have been reported earlier [2]. Koolagudi et al developed the simulated Hindi emotion speech corpus, IITKGP-SEHSC consisting of sarcastic, neutral and six basic emotions which was used for the detection of emotions from prosodic and spectral features[3]. The average emotion recognition performance was investigated based on the Mel frequency cepstral coefficients (MFCCs), using GMM method.The results of text dependent and text independent emotion recognition in Hindi using MFCCs along with their velocity and acceleration coefficients have been presented in [4]. The results were based on the IITKGP-SEHSC speech corpus and obtained with Gaussian mixture models. Around 72% and 82% of emotion recognition rates were reported for text independent and dependent cases respectively.

D. Spectral features – formants and bandwidths

Formants are resonant frequencies of the vocal tract. Each formant is characterized by its center frequency and its bandwidth.

Mel-frequency Cepstral Coefficients (MFCC) and formant frequency used in speech recognition and speech processing applications have also been studied for the purpose of emotion recognition [5]. Experimental analysis has shown that the first and second formants are affected by the emotional states of speech more than the other formants [6], [7].

II. METHODOLOGY

This section outlines the various steps adopted to achieve the objectives of this research, as mentioned in the problem statement. The first step was the design and development of the speech database, the emotional quality of which was evaluated by means of perceptual listening tests. This was followed by statistical preprocessing to ensure differences amongst the feature values subsequently given to the various classifiers. The schematic representation of the sequence of tasks proposed in order achieves the research objectives is as given below.

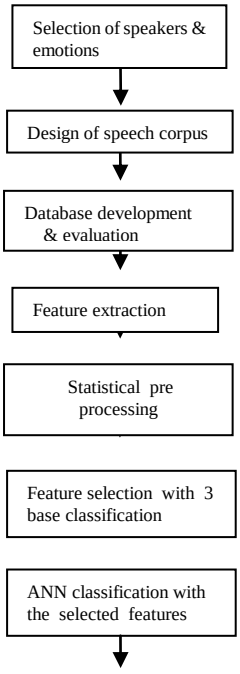

Fig. 1 Schematic of the proposed system for speech emotion recognition in Hindi

A. Database development and evaluation

The Hindi emotional speech database was developed from the recordings of ten female speakers in the age group of 24 to 42 years as mentioned in [8] for anger, sad, fear and disgust. The speakers were volunteers who were briefed on the purpose of the recordings. Due the difficulty in obtaining spontaneous speech and because of the exaggerated affects in acted speech, the investigators chose to record elicited emotional speech, for which the speakers were asked to imagine certain situations matching the seven emotions. These investigations were based on scripted dialogues in order to ensure appropriate and adequate emotional content semantically. All seven emotions were recorded using the Realtek audio system, with 16 bits resolution and 16 KHz sampling frequency.

The next step was the subjective evaluation of the database. The recordings were played back to ten listeners in order to verify their perceptual emotional quality. All the listeners were in the age group 15-50 and had no hearing pathologies. Also they had no training in perceptual phonetics. Listeners listened over headphones at a comfortable listening level to the utterances. They were required to indicate the perceived emotion from a list of the seven emotions. 85.4% recognition rate was obtained as a result of the listening tests. The sound files that were rated poorly were not included for further analysis. This was followed by the extraction of formant and bandwidths, by the linear prediction method.

The tabulated formant and bandwidth were then analyzed for statistical differences with respect to different emotions. The investigators applied statistical preprocessing to the input features, in order to enhance their quality before these were given to the various classifiers. Hence, only those values that were found statistically different by the analysis of variance (anova) were used for further classification.

B. Base Classifiers for identification of best features

The formants and bandwidths to be given to the ANN classifier for the final identification of emotions were selected on the basis of the performance of the individual features, using three base classifiers, namely the KMeans, the Naïve Bayes and the K nearest neighbor classifiers

C. Artificial Neural Networks

Artificial neural networks are popular in speech emotion recognition [9]. A two-layer feed-forward network, with sigmoid hidden and output neurons and sufficient neurons in its hidden layer was used for classification. We have used a feed forward back propagation neural network which is trained to classify the inputs according to the targets. Training automatically stops when generalization stops improving, as indicated by an increase in the mean square error of the validation samples. 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. Training multiple times generated different results due to different initial conditions and sampling. Since feed-forward neural network based on the gradient descent algorithm have a poor convergence rate, we decided to use a Scaled Conjugate Gradient (SCG) algorithm with super linear convergence rate as in [10].The total number of samples was randomly divided into three classes for training, validation and testing purposes in the proportion 70% for training and 15% each for validation and testing. The network was tuned according to the recognition error on the training data. Validation samples measure network generalization, and halt training when generalization stops improving. Test samples provide an independent measure of network performance during and after training.

III. III. RESULTS AND DISCUSSION

This section sequentially presents the results of ANOVA followed by the classification results of the three base classifiers used for selecting optimum features for the subsequent ANN classification.

A. Results of statistical analysis

Sample results of the repeated measures ANOVA are presented below in Table I. for the first formant values.

Table I. Summary statistics of first formant values in Hindi, with emotion discrimination levels

Emotions	Distinct from emotions	P<0.001	P<0.01	Emotions not discriminated
Happy	disgust, surprise fear, neutral, anger	sad		-
Surprise	sad, anger, happy	disgust	Neutral	fear
Neutral	anger, sad, disgust, happy,	-	Surp, fear	-
Anger	neutral, surprise sad, disgust, fear, happy	-	-	-
Sad	anger, surprise neutral fear,	disgust, happy	-	-
Fear	sad, surprise, happy		disgust neutral	Surprise
Disgust	neutral, happy, anger	sad, surp	fear	-

Statistical analysis revealed anger as the best discriminated from the rest of the emotions with a high significance level, followed by happiness and sadness. Anger had high first formant values. Statistically only surprise and fear were not discriminated based on F1. ANOVA of F2 showed no significant differences in statistical parameters for surprise –happy, sad-disgust and disgust –neutral emotion pairs. Anger and fear were the best discriminated statistically. Sadness was the best discriminated emotion based solely on the third formant statistics. Repeated measures ANOVA indicated highly significant difference in statistics ($p<0.0001$) within all pairs of emotions except between surprise-fear, neutral -disgust, and happy-angry. Statistical analysis of the fourth formants showed sad and happy emotions to be the best discriminated from the rest of the emotions. Anger could not be discriminated from neutral or fear. Fear could not be differentiated from neutral. So also, disgust could not be statistically differentiated from surprise. Summarizing, the individual, statistical analysis of the first four formants showed that happiness, anger and sadness were well discriminated.

The repeated measures ANOVA of the log of the first bandwidth of Hindi utterances, revealed surprise to be differentiated from all other emotions. Happiness was discriminated from anger, neutral and disgust and fear was discriminated from anger. Repeated measures anova gave better discrimination results with log of B2. Sad could not be distinguished from anger, happy and neutral emotions. Fear -surprise and neutral anger emotion pairs could also not be distinguished using B2 or its logB2 values. With raw B2, disgust could not be distinguished from emotions other than happiness. Neutral and sad could also not be distinguished. The statistical analysis of the log of the bandwidth, rather than the raw B3 values of the third formants showed better discrimination among emotions. Disgust was the worst discriminated as there was no significant difference between surprise, neutral and sad and anger. Fear could not be differentiated from happiness and neutral could not be discriminated from surprise. Statistical analysis of the log of B4 values gave better results than the raw B4 values. Anger was the best discriminated and happiness was the worst discriminated from the rest of emotions. Happiness could not be discriminated from disgust, surprise and fear. Surprise could not be statistically distinguished from fear and sad could not be differentiated from neutral. In short, only the two positive valence emotions and anger were well discriminated statistically based on the bandwidths.

B. Results of feature selection by the three base classifiers

This section presents the results of the selection of features by the base classifiers based on the performance of each individual feature. This is followed by the results of the final ANN classification done on the basis of the selected formants and bandwidths. The summary of the classification results obtained with the Kmeans, Naive Bayes and KNN Classifiers are as follows:

The KMeans classifier gave non zero, though very poor recognition rates across the various formant bandwidth features for emotions like sad, surprise, disgust. The highest RR obtained at the output of this classifier is 53.13% for sad and fear, based on log B1.

The NB classifier could not recognize all emotions based on individual formants/ bandwidths. The highest RR of 81.8% was obtained for anger, based on F1. But anger was not at all recognized by the NB classifier for certain other features such as B1, B2 and B3. Hence the NB classifier too is not a good solution for this class seven emotion recognition problem in Hindi.

The KNN Classifier was able to recognize all emotions based on each feature and gave higher overall RR than the other two classifiers. The highest RR obtained with this classifier is 84.6% for happy emotion based on B1. Table. V provides information on the effectiveness of each feature in the recognition of emotions, using the aforesaid three classifiers. It is seen that very good overall recognition rates were obtained with F1, F2, B1, B2, and B3. F3 as well as F4 contribute to the recognition of sad emotion. With B4, moderate recognition rates were obtained for each emotion. This suggests the complete formant bandwidth feature set to be useful for speech emotion recognition in Hindi.

Table II. Consolidated table of contribution of features-Hindi

Emotions	F1	F2	F3	F4	B1	B2	B3	B4
Happy	G	G	F	F	G	√	F	G
Surprise	F	F	P	F	F	F	F	F
Neutral	P	G	F	F	F	F	F	F
Anger	√	√	G	F	F	F	F	G
Sad	F	P	G	G	F	P	P	F
Fear	G	F	F	F	√	G	G	F
Disgust	F	G	F	P	G	G	√	F

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

Table III below provides details of the final classification by a two layer feed forward back propagation artificial neural network.

The features were given to the network in separate groups of formants and bandwidths as well as in a third group of the combined formant bandwidth values. The obtained recognition accuracy of 97.14% for the combined group proves all the formants as well as bandwidths in Hindi, to be very good markers of emotion.

Table III. Formant /bandwidth features giving the best classification rates in Hindi

Feature identity	MSE	Percentage error (%E)
F1 to F4	$3.73e^{-2}$	5.71
B1 to B4	$3.10\ e^{-2}$	17
F1 to F4 & B1 to B4	$2.073e^{-2}$	2.86

Table IV. Presents the confusion matrix of formant and bandwidth based emotion classification accuracies for Hindi, as percentage belonging to each class.

Table IV. Confusion matrix of formant and bandwidth based emotion Classification accuracies as percentage belonging to each class

Emotions	Happy	Surprise	Neutral	Anger	Sad	Fear	Disgust
Happy	100	0	0	0	0	0	0
Surprise	0	100	0	0	0	0	0
Neutral	0	0	100	0	0	0	0
Anger	0	0	0	90	0	0	10
Sad	0	0	0	0	100	0	0
Fear	0	0	0	0	0	100	0
Disgust	0	0	0	0	0	0	100

It is seen that all emotions except anger were recognized correctly. There was confusion between disgust and anger only. Figure 2. shows the variation in the percentage error on the training, validation and test data for five different sizes of ANN .

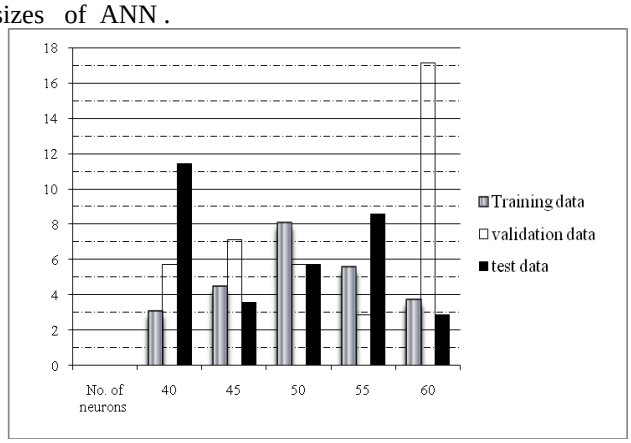

Fig. 2 Percentage error for various datasets and network sizes

C. Comparison with certain other SER systems

Very few studies have been reported for Indian languages. An overall accuracy of 87.9% has been reported for Hindi SER using MFCC [2]. An average emotion recognition rate around 81% for female speech utterances has been published in [11]. The highest overall accuracy obtained in this work with the selected feature set of all four formants and bandwidths was 97.14%.

The formants feature group used in this work gave a recognition rate of 94.29% with the ANN classifier. A maximum recognition rate of 63.73% and 41.79% only, have been reported on the formant features of the emotional speech database EMO-DB and the Danish emotional speech database DES respectively, with SVM classifiers[12].
A recent review cited the average recognition rate for happy, anger, neutral and sad as 85.7% obtained with acoustic prosodic and semantic label information [13].

The best results of the ANN classification for recognition of different number of emotional classes are given in Table V. The classification was repeated for different network sizes in order to arrive at the optimum size of the network.

Table V. FFBPNN performance measures for formant based Hindi Emotion recognition of various problem classes to each class

Problem class and description	No. of neurons	MSE	Error percentage
2 -positive and negative	20	$2.36e^{-2}$	0
3-positive, neutral and negative	60	$1.86e^{-2}$	0
4- hap,surp, sad, fear	60	$1.64e^{-2}$	0
5- hap,surp,neut,sad, fear	30	$2.81e^{-2}$	0
6-hap, surp, neut, sad, fear ang	30	$1.96e^{-2}$	0
7- neutral and six basic emotions	60	2.073e	2.86

By this approach we have been able to identify the combinations of emotions in groups of size ranging from 2 to six that are fully recognised. Not all combinations of emotions were recognised fully. This implies that other features are needed to give complete recognition in those cases. Improvement in performance of the ANN was obtained by reducing the number of emotion classes from seven.

IV. CONCLUSION AND FUTURE WORK

In this paper we have presented the results and discussed the performance of a simple, yet efficient system for class seven speech emotion recognition in Hindi, using the proposed minimal feature set of formants and bandwidths. The recognition rate obtained with this approach was 97.14%, which is the highest reported so far, for Hindi, supporting the use of the proposed minimal feature set for Hindi speech emotion recognition. The classification results of the three base classifiers were validated by the performance of ANN classifier. The high quality emotional speech database as indicated by the score of perceptual listening tests, contributed to the obtained recognition rate. Statistical preprocessing of the minimal feature set values have also helped to obtain this improved recognition rate. The speech emotion recognition rate increased with decrease in the number of emotion classes, from seven. This method is easy to implement and requires lesser time due to the use of only minimal features. As such, it has potential applications in research in speech as well as in sociolinguistics. Future work in this direction is to implement real time emotion recognition by this approach.

REFERENCES

1. Anik Dey, Ying Li, Pascale Fung, "Using English Acoustic Models for Hindi Automatic Speech Recognition". Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing (SANLP), pp. 123–134, COLING 2012, Mumbai, December 2012. Published in Proceedings of the International Conference Speech Database and Assessments (Oriental COCOSDA), 2011.
2. Anurag Jain, Nupur Prakash, S.S. Agrawal, "Evaluation of MFCC for Emotion Identification in Hindi Speech" IEEE .978-1-61284-486-2/111 ©2011 IEEE.
3. Shashidhar G. Koolagudi, Ramu Reddy, Jainath Yadav, K. Sreenivasa Rao, "IITKGP-SEHSC: Hindi speech corpus for emotion analysis." 978-1-4244-9190-2/11/ ©2011 IEEE.
4. Rahul Chauhan, Jainath Yadav, S. G. Koolagudi, K. Sreenivasa Rao, "Text Independent Emotion Recognition Using Spectral Features" Communications in Computer and Information Science –Springer, Volume 168, 2011, pp 359-370 Contemporary Computing - 4th International Conference, IC3 2011, Noida, India, August 8-10, 2011. Proceedings.
5. Yongjin Wang, and Ling Guan, Recognizing Human Emotional State from Audiovisual Signals. IEEE Transactions on Multimedia, Vol 10, No.5, August 2008. pp.936-946**.**
6. Tolkmitt, F. J., Scherer, K. R., Effect of experimentally induced stress on vocal parameters. Journal of Experimental Psychology: Human Perception and Performance 12 (3), 302–313. 1986.
7. France, D. J., Shiavi, R. G., Silverman, S., Silverman, M., Wilkes, Acoustical properties of speech as indicators of depression and suicidal risk. IEEE Trans.Biomedical Engineering 7, 829–837, 2000.
8. Agnes Jacob, Dr. P. Mythili, "Development and Evaluation of Emotional Speech Databases in the Indian Context." Proceedings of the International conference on Intelligent Systems and Control, 4 pages, B3 -09, ISCO 2010, February 2010, Coimbatore.
9. Yegnanarayana and S. P. Inshore, "An ANN an alternative to GMM for pattern recognition," Neural Networks, vol. 15, pp. 459– 469, April 2002.
10. Jana Tuckova, Martin Sramka, Emotional Speech Analysis using Artificial Neural Networks Proceedings of the International Multiconference on Computer Science and Information Technology pp. 141–147 ISBN 978-83-60810-27-9 2010 IEEE .
11. Shashidhar G. Koolagudi, K. Sreenivasa Rao, "Emotion recognition from speech: a review". International Journal of Speech Technology, vol 15, pp. 99–117, 2012. DOI 10.1007/s10772-011-9125-1.
12. Björn Schuller, Dejan Arsi, and Frank Wallhoff, Manfred Lang, and Gerhard Rigoll, Bioanalog acoustic emotion recognition by genetic feature generation based on low-level-descriptors. EUROCON 2005,Serbia and Montenegro, Belgrad, November 22-24,2005.1-4244-0049- X/05 IEEE pp. 1292-1295.
13. Rahul. B. Lanjewar, D. S. Chaudhari, Speech Emotion Recognition: A Review International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-2, Issue-4, March 201368.

AUTHORS PROFILE

Agnes Jacob, received her M.Tech degree from the National Insti tute of Technology, Kozhikode, Kerala and MBA in Human Resources Management from IGNOU. She is pursuing Ph. D. at the Cochin University of Science and Technology, Kochi, Kerala. She has published several research papers in various International conferences and journals. Her research interests include speech processing, emotional intelligence and human resources management

Dr. Mythili, received her M. E. and Ph. D. degrees from the College of Engineering, Guindy, and the Anna University, Chennai. She is currently Head of the Division of Electronics, School of Engineering. She was awarded the prestigious BOYSCAST fellowship, by the Department of Science and Technology. She is the Principal Investigator and Co-Investigator of several funded projects of UGC, KSCSTE and AICTE. She has more than 60 publications in International Journals and conferences. Her areas of interest are Signal Processing, Genetic Algorithms, Microwaves and Speech Processing.