Comparison between k-nn and svm method for speech emotion recognition

Muzaffar Khan, Tirupati Goskula, Mohmmed Nasiruddin ,Ruhina Quazi
Anjuman College of Engineering & Technology ,Sadar, Nagpur, India

Abstract Human - Computer intelligent interaction (HCII) is an emerging field of science aimed at providing natural ways for humans to use computer as aids. Machine intelligence needs to include emotional intelligence it is argued that for the computer to be able to interact with humans, it needs to have the communication skills of human. One of these skills is the ability to understand the emotional state of the person. Two recognition methods namely K-Nearest Neighbor (K-NN) and Support vector machine (SVM) classifier have been experimented and compared. The paper explores the simplicity and effectiveness of SVM classifier for designing the real-time emotion recognition system.

Keywords: HCII, Emotion states, SVM, K-NN classifier. Emotion classifier

1. INTRODUCTION

Emotions play an extremely important role in human mental life it is a medium of expression of one's perspective or his mental state to others. It is a channel of human psychological description of one's feelings. The basic phenomenon of emotion is something that every mind experiences and our paper make a specific hypothesis regarding the grounding of this phenomenon in the dynamics of intelligent systems. There are a few universal emotions-including Neutral, Anger, Surprise, Disgust, Fear, Happiness, and Sadness which any intelligent system with finite computational resources can be trained to identify or synthesize as required. In this paper, we present an approach to language-independent machine recognition of human emotion in speech [5]. The potential prosodic features are extracted from each utterance for the computational mapping between emotions and speech patterns. The selected features are then used for training and testing a modular neural network. Classification result of neural network and K-nearest Neighbors classifiers are investigated for the purpose of comparative studies.

2. SYSTEM DESCRIPTION

The functional components of the language and gender independent emotion recognition system are depicted in figure 1. It consists of seven modules speech input, preprocessing, spectral analysis, feature extraction, feature subset selection, neural network for classification, and the recognized emotion output. Emotional speech signal data is feed to the system as an input to the system [15]. As the database of the input sound contains noise signal/silent zone at the beginning and at the end of signal preprocessing of the signal is required to chop the silent zone after preprocessing of the signal the spectral analysis is done. The next stage of the system is to extract the speech features like Formant Frequencies, Entropy, Median, Mel-Frequency Cepstral coefficient, Variance, Minima etc. from the filtered emotional speech signal .Some of the speech extracted features may be redundant or even cause negative effects to the training of neural network for that feature selection method is applied, through which only that features which adds efficiency to the system is chosen so as to built an efficient system with greater accuracy. After selection of feature vector, a feature database is built up this is required as an input to classifier. On the basis of this database classifier which is vigorously train on the given input to recognize Human emotions with the accuracy.

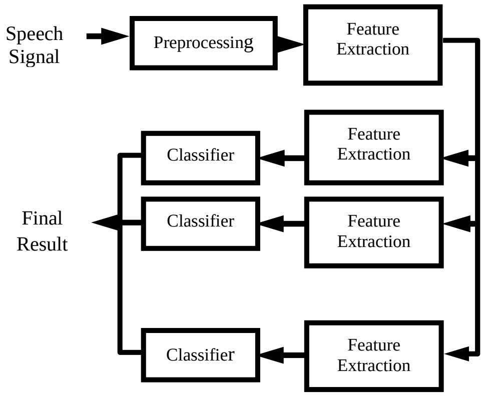

Figure 1: The structure of the speech recognition system

2.1 Preprocessing Of Audio Signal

Before giving Speech Data as an input to the system preprocessing of the signal is required. Preprocessing means filtering, cutting down the silent zone before the signals are normalized .All the data which is being feed to the system is processed in the same manner for which the complete silent zone prefixing the sentence and post fixing the sentence is chopped out.

2.2 Feature Extraction

Speech includes several kinds of factors about speaker, context, state of speech, such as emotion, stress, dialect and accent, are important problem.The rationale for feature selection is that new or reduced features might perform better than the base features because we can eliminate irrelevant features from the base feature set that small values decrease, large values increase. This can also reduce the dimensionality, which can otherwise hurt the performance of the pattern classifiers. In this work, we used the forward selection (FS) method. First, FS initializes to contain the single best feature with respect to a chosen criterion from the whole feature set. Here, classification accuracy criterion by nearest neighborhood rule is used, and the accuracy rate is estimated by leave-one-out method. The subsequent features are added from the remaining features which maximize the classification accuracy. In this work, we experimented with two sets of rank-ordered selected features from Formant Frequencies fo to Log Entropy as indicated in table 1, both male and female data have similar features in their best feature sets.

TABLE 1: LIST OF 14 FEATURE VECTORS

Sr. No.	**Feature**	**Sr. No.**	**Feature**
1	Formant0	8	Threshold Entropy
2	Formant1	9	Sure Entropy
3	Formant2	10	Norm Entropy
4	Formant3	11	Median
5	Formant4	12	Mel-Frequency Cepstral coefficient
6	Pitch	13	Variance
7	Shanon Entropy	14	Log Entropy

3. ENGLISH SPEECH DATABASE

We have developed our own database in English for this work. The recording is done using five speech texts spoken in seven emotions by male and female actors. We have recorded audio speech signals with well equipped audio recording equipments. The sentences were designed to use for recording the seven emotions (Neutral, Anger, Surprise, Disgust, Fear, Happiness, and Sadness) by each speaker. The author has prepared own simulated speech database. This database contains speech 350 samples. The length of speech samples is up to 5 Seconds.

4.1SUPPORT VECTOR MACHINE (SVM)

SVM is a binary classifier An approach to solve this problem was to build Seven different SVMs one for each emotion and choose the class (emotion) which gives the highest output score. If the highest output score was negative, a testing sample could not be classified. Based on this approach, different experiments with different kernel functions were performed during this research. Kernel employed polynomial is given by

$$K_p(X,Y) = (X.Y+1)^p \ldots\ldots\ldots\ldots\ldots\ldots 1$$

Where p is the order of the polynomial, Classifier employs $K_p(.)$ have polynomial decision function, polynomial functions whose orders ranged from 2 to 3 respectively. Radial basis functions whose gamma values ranged from 2 to 6 respectively [17].

4.2 K-Nearest Neighbor Technique as an Emotion Recognizer

A more general version of the nearest neighbor technique bases the classification of an unknown sample on the "votes" of K of its nearest neighbor rather than on only it's on single nearest neighbor. The K-Nearest Neighbor classification procedure is denoted is denoted by K-NN. If the costs of error are equal for each class, the estimated class of an unknown sample is chosen to be the class that is most commonly represented in the collection of its K nearest neighbors. Among the various methods of supervised statistical pattern recognition, the Nearest Neighbor is the most traditional one, it does not consider a priori assumptions about the distributions from which the training examples are drawn. It involves a training set of all cases. A new sample is classified by calculating the distance to the nearest training case, the sign of that point then determines the classification of the sample. The K-NN classifier extends this idea by taking the K nearest points and assigning the sign of the majority. It is common to select K small and odd to break ties (typically 1, 3 or 5). Larger K values help reduce the effects of noisy points within the training data set, and the choice of K is often performed through cross-validation. In this way, given a input test sample vector of features *x* of dimension *n*, we estimate its Euclidean distance *d* equation 3 with all the training samples (*y*) and classify to the class of the minimal distance.

$$q(x,y) = \sqrt{\sum_{j=1}^{n}(x_j - y_j)^2} \qquad \ldots\ldots\ldots \quad 2$$

The training examples are vectors in a multidimensional feature space, each with a class label. The training phase of the algorithm consists only of storing the feature vectors and class labels of the training samples. In the classification phase, K is a user-defined constant, and an unlabelled vector (a query or test point) is classified by assigning the label which is most frequent among the K training samples nearest to that query point. Usually Euclidean distance is used as the distance metric, however this is only applicable to continuous variables.

4.3 K-NN Algorithm:

The *k*-NN algorithm can also be adapted for use in estimating continuous variables. One such implementation uses an inverse distance weighted average of the *k*-nearest multivariate neighbors. This algorithm functions as follows: Compute Euclidean or Mahalanobis distance from target plot to those that were sampled.

1. Order samples taking for account calculated distances.
2. Choose heuristically optimal K nearest neighbor based on root mean square error q(x, y) done by cross validation technique.
3. Calculate an inverse distance weighted average with the *k*-nearest multivariate neighbors.

5. Results:

Table 2: Classification result of SVM and K-NN

States	Neutral	Anger	Fear	Disgust	Sadness	Happine ss	Surprise
Speech Samples	50	50	50	50	50	50	50
SVM	47	34	48	35	34	35	35
KNN	40	48	45	44	47	48	49
Performance SVM (%)	94	68	96	70	68	70	70
Performance KNN (%)	80	96	90	88	94	96	98
Overall performance SVM =76.57% Overall performance K-NN=91.71%							

TABLE 3: CONFUSION MATRIX FOR K-NN

States	Neutral	Anger	Fear	Disgust	Sadness	Happiness	Surprise
Neutral	40	4	0	0	4	2	0
Anger	0	48	0	2	0	0	0
Fear	0	0	45	3	2	0	0
Disgust	0	3	2	44	0	0	1
Sadness	1	0	0	2	47	0	0
Happiness	0	0	0	0	0	48	2
Surprise	0	0	0	0	0	0	49

TABLE 4: CONFUSION MATRIX FOR SVM

States	Neutral	Anger	Fear	Disgust	Sadness	Happine ss	Surprise
Neutral	47	0	0	0	3	0	0
Anger	0	34	0	6	2	2	6
Fear	0	0	48	0	0	0	2
Disgust	0	8	2	35	0	0	5
Sadness	0	0	8	8	34	0	0
Happiness	8	0	0	0	0	35	7
Surprise	0	6	0	2	0	7	35

6. APPLICATIONS

The emotion recognition using speech signals have wide applications. The proposed work can be implemented in the in the following fields.

- Human-computer intelligent interaction (HCII) for make machines more user friendly
- Project can be implemented as a Lie Detector.
- Designing intelligent Robotics.
- Develop learning environments and consumer relations.
- Entertainment etc.

7. CONCLUSION

Human emotions can be recognized from speech signals when facial expressions or biological signals are not available. In this work Emotions are recognized from speech signals using real time database. In this work we presented an approach to emotion recognition from speech signal. Our results indicate that the K-NN classifier average accuracy 91.71% forward feature selection while SVM classifier has accuracy of 76.57%.Table 3 and 4 show SVM classification for neutral and fear emotion are much better than K-NN.The future work will be to conduct comparative study of various classifier using different parameter selection method to improve performance accuracy .

REFERENCES

[1] Lawrence S. Chen & Thomas S. Huang, “Emotional Expressions in Audiovisual Human Computer Interaction”, 0-7803-6536-4/00 2000 IEEE.

[2] Yi-Lin Lin, Gang Wei, “Speech Emotion Recognition Based on HMM and SVM ”, Proceedings of the 4th International Conference on Machine Learning and Cybernetics, Guangzhou, pp.4898-4901. 18-21 August 2005 IEEE.

[3] Zhongzhe Xiao, Emmanuel Dellandrea, Weibei Dou and Liming Chen, “Features Extraction and Selection for Emotional Speech Classification”, 0-7803-9385-6/05/2005 IEEE

[4] Frank Dellaert, Thomas Polzin and Alex Waibel , “Recognizing Emotion In Speech”, Fourth Internation Conference on spoken language ICSPL 1996 pp 1970-1973 ISBN 0-7803-3555-4

[5] Fatema N Julia, Khan M Iftekharuddin, “Detection of Emotional Expressions in Speech” 0-4244-0169-0/06, pp.307-312. 2006 IEEE.

[6] Chul Min Lee, and Shrikanth S. Narayanan, “Toward Detecting Emotions in Spoken Dialogs”, IEEE Transactions on Speech and Audio Processing, Vol. 13, No. 2, pp.1970-1974 March 2005.

[7] Tsang-Long Pao, Yu-Te Chen, Jun-Heng Yeh, “Mandarin Emotional Speech Recognition Based on SVM and, and NN”, Proceedings of the 18th International Conference on Pattern Recognition 2006

[8] S.Ramamohan and S. Dandapat, Member, IEEE, “Sinusoidal Model-Based Analysis and Classification of Stressed Speech”, IEEE Transactions on Audio, Speech, And Language Processing, Vol. 14, No. 3, pp.737-746. May 2006 IEEE.

[9] M.M.H.; Kamel, M.S.; Karray, F.; EI Ayadi “Speech Emotion Recognition using Gaussian Mixture Vector Autoregressive Models”, Acoustics, Speech and Signal Processing”, 2007. ICASSP 2007. IEEE International Conference on volume 4, 15-20 April 2007 Page(s): IV-957-IV-960.

[10] Lili Cai, Chunhui Jiang, Zhiping Wang, Li Zhao, Cairong Zou, “A Method Combining The Global And Time Series Structure Features For Emotion Recognition In Speech”, IEEE Int. Conf. Neural Networks & Signal Processing Nanjing, China, December 14-17, 2003 0-7803-7702-8/03/ 2003 IEEE.

[11] Schuller, B. Seppi, D. Batliner, A. Maier, A.; Steidl, S.; “Toward More Reality in the Recognition of Emotional Speech”, Acoustics, Speech and Signal Processing”, 2007. ICASSP 2007. IEEE International Conference on volume 4, 15-20 April 2007 PP: IV-941-IV-944**.**

[12] Banse, R. & Scherer, K. R., “Acoustic Profiles in Vocal Emotion Expression”, Journal of Personality and Social Psychology”, Vol. 70, No. 3, pp. 614-636, 1996.

[13] Michael Lyons and Shigeru Akamatsu, Miyuki Kamachi and Jiro Gyoba, Coding “facial Expression with gabor wavelets. Proceedings, third IEEE International conference on automatic face and Gesture Recognition”, April 14-16 1998, Nava Japan, IEEE computer Society, pp 200-206.

[14] Kharat and Dudul, “ Design of Neural Network Based Human Emotion state Recognition System From Facial Expressions, International Journal of emerging technology and applications In Engineering Technology And Science” (IJ-ETA-ETS) pp 55-60, January 2009 - June 2009 (ISSN: 0974-3367).

[15] Talieh Seyed Tabatabaei, Sridhar Krishnan “Emotion Recognition Using novel Speech Signal ”circuits and syste” 2007 . pp. 345 - 348 25 june 2007 (ISSN: 1-4244-0920-9).

[16] Yongjin Wang ,Ling Guan “An investigation of speech based human emotion recognition” 2004 IEEE 6th Workshop on signal processing.

[17] Iris Bas Thao Nguyen “Investigation of Combining SVM and Decision Tree for Emotion lassification” Proceedings of the Seventh IEEE International Symposium on Multimedia 2005