Sound based human emotion recognition using MFCC & multiple SVM

Conference Paper · August 2017

DOI: 10.1109/ICOMICON.2017.8279046

3 authors, including:

Kishor Bhangale
Pimpri Chinchwad College Of Engineering and Research
23 PUBLICATIONS **107** CITATIONS

IEEE, International Conference on Information, Communication, Instrumentation and Control (ICICIC – 2017)
Paper Id:

Sound based Human Emotion Recognition using MFCC & Multiple SVM

Anagha Sonawane
Dept. of E&TC,
Siddhant College of Engineering,
Pune, India

M.U.Inamdar
Dept. of E&TC,
Siddhant College of Engineering,
Pune,India

Kishor B. Bhangale
Dept. of E&TC,
D. Y. Patil College of Engineering,
Pune, India

***Abstract*—Emotion recognition using human speech is one of the latest challenges in speech processing and Human Machine Interaction (HMI) for the purpose of addressing varied operational needs for the real world applications. Besides human facial expressions, speech has been proven to be one of the most valuable modalities for automatic recognition of human emotions. Speech is a spontaneous medium of perceiving emotions which provides in-depth. Here in this paper, we have used MFCC for extraction of features and Multiple Support Vector Machine (SVM) as a classifier. We have performed extensive experiment on happy, anger, sad, disgust, surprise and neutral emotion sound database. Performance analysis of multiple SVM revealed that non-linear kernel SVM achieved greater accuracy than linear SVM.**

Keywords*— *Automatic Speech Recognition, Mel Frequency Cepstrum Coefficients, Support Vector Machine, Speech Emotion Recognition.

I. INTRODUCTION

Human beings always do the communication with each other by expressive gestures of emotions and feelings which are recognised by some experiences and knowledge. These expressions are conveyed in speech form or through body language. Emotions are part and parcel of human life and among other things, highly influence decision making [5][2][7]. In this paper various kinds of features that might carry more information about the emotional meaning of each utterance are considered. The features that contribute to emotions may be different for different spoken languages. The approach is to calculate which features carry more information and to combine these features to get a better recognition rate. It also depends on which emotions we want a machine to recognize and its purpose. Active learning tries to select the most informative examples to build a training set for a predictive model. In this paper, we have used the Support Vector Machine inorder to model the phonetic units corresponding to sentences taken from the training base[3]. The results obtained are very encouraging given the size of the training set and the number of people taken to the registration. This algorithm is based on the flexibility of the Support Vector Machine for sentences by means of dynamic programming.

The speech Recognition (which is also known as Automatic Speech Recognition (ASR) or computer speech recognition) is the process of converting a speech signal to a sequence of words by means of an algorithm implemented as a computer program. Speech is a unique human characteristics used as a tool to communicate and express ideas. Research work in the field of automatic speech recognition (ASR) using machine has attracted a great deal of attention over the past few decades due to various reasons[1]. This has technological strong desire regarding the mechanisms for mechanical realization of human speech capabilities and the desire to automate simple tasks inherently requiring human machine interactions. Speech recognition technology has made it possible for computer to follow various human voice commands and try to understand human languages [12][6][9]. The main purpose of speech recognition field is developing techniques and systems for speech input to machine. Speech is the primary means of communication between humans[11][8].

II. FEATURE EXTRACTION

The generalized block diagram of voice recognition is shown in fig. 1.Main stages of automatic sound recognition are feature extraction and classification. For our implementation. We have used MFCC for the extraction of features and classification is done with Support Vector Machine (SVM). In this paper, we are building a English emotional speech corpus with various emotions like happy, anger, fear, neural and sad. The corpus has been evaluated using SVM based emotion recognition engine.

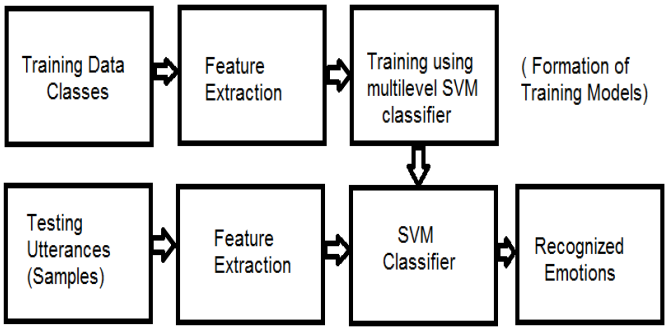


Fig.1 Generalized block diagram of automatic speech recognition

The ideal representation of parameter extraction of acoustic signals is an important task to produce a better performance in recognition . This phase's efficiency is important for the next phase since it affects its behavior. MFCC is dependent on

human hearing perceptions which cannot perceive frequencies over 1Khz. In other words, in MFCC depends on known variation of the human ear's critical bandwidth with frequency [9-11].

A. MFCC Feature Extraction

There are two types of filter of MFCC which are spaced linearly at low frequency below 1000 Hz and logarithmic spacing above 1000Hz. A subjective pitch is present on Mel Frequency Scale to capture important characteristic of phonetic in speech [10]. Pre-emphasis is a process of passing the input signal through a filter which emphasizes higher frequencies. This process will increase the energy of signal at higher frequency. Framing is the step next to preemphasis in which the speech samples obtained from analog to digital conversion (ADC) are segmented into a small frame with the length within the range of 20 to 40 msec [6]. Hamming window is used as window shape by considering the next block in feature extraction processing chain and integrates all the closest frequency lines.The process of FFT converts each frame of N samples from time domain into frequency domain. The frequencies range in FFT spectrum is very wide and voice signal does not follow the linear scale. Mel frequency reduces this scale.

$$F(Mel) = [2595 * \log 10 [1+ f] * 700] \quad (1)$$

DCT convert the log Mel spectrum into time domain. The voice signal and the frames changes, such as the slope of a formant at its transitions. Hence , there is a need to include features related to the change in cepstral features over time . 13 delta or velocity features (12 cepstral features plus energy), and 39 features a double delta or acceleration feature are added [10][9].

B. Classification using Support Vector Machine:

One of the strong tools for pattern recognition that uses a discriminative approach is a SVM.SVMs use linear and nonlinear separating hyper-planes for data classification[3] [12] [1] [10]. Since SVMs can only classify fixed length data vectors, this method cannot be readily applied to task involving variable length data classification. Variable length data is needed to be transformed to fixed length vectors before using SVMs .It is a generalized linear classifier with maximum-margin fitting functions. This fitting function provides regularization which helps the classifier to be generalized better. The classifier ignores many of the features. Conventional neural network and statistical methods control model complexity by using a small number of features (the problem dimensionality or the number of hidden units) [12].

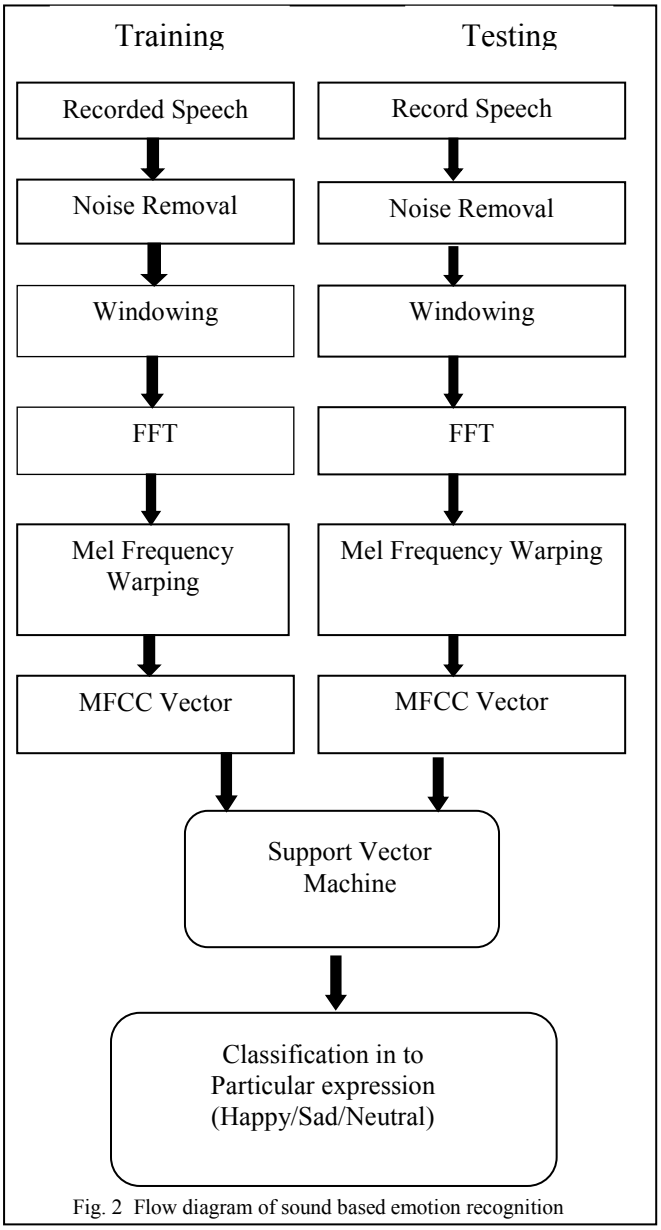


Fig. 2 Flow diagram of sound based emotion recognition

This method is not dependent on dimensionality and can utilize spaces of very large dimensions spaces, which permits a construction of very large number of non-linear features and then performing adaptive feature selection during training.

III. EXPERIMENTAL RESULTS

We have implemented the system on a system having 6GB RAM with Intel (R) Core(TM) i5-4200U CPU @ 1.60GHz 2.30Ghz processor.

We have created the Neutral emotion sound database from the voice samples of BBC news in male and female voices in British accent. For the Happy and Sad samples, we have collected the samples of male, female, children and infants with the variety of emotion sounds. For the training purpose

we have taken 500 samples of Happy, Sad and Neutral emotion sound samples and for testing purpose for every group 200 samples are selected. The database sound is recorded in clean and noise free environment to ensure good results.

Performance of the multiple SVM is measured on the basis of total cross validation accuracy.

Accuracy =

$$\frac{\text{True Positive(TP)}}{\text{True Positive(TP)+True Negative(TN)}} \quad (2)$$

Performance analysis reveals that non-linear multiple SVM has better ability of class separation which results in greater accuracy.

TABLE I. PERFORMANCE ANALYSIS OF MULTIPLE SVM (RBF KERNEL)

Database	Training Samples	Testing Samples	TP	TN	Accuracy
Happy	500	200	180	20	90.00 %
Sad	500	200	193	07	96.66 %
Neutral	500	200	183	17	91.66 %
Anger	500	200	176	24	88.33 %
Surprise	500	200	173	27	86.66 %
Disgust	500	200	183	17	91.66 %

TABLE II. PERFORMANCE ANALYSIS OF MULTIPLE SVM (POLYNOMIAL)

Database	Training Samples	Testing Samples	TP	TN	Accuracy
Happy	500	200	185	15	92.50%
Sad	500	200	186	14	93.33%
Neutral	500	200	180	20	90.00%
Anger	500	200	178	22	89.16%
Surprise	500	200	166	34	83.33%
Disgust	500	200	180	20	90.00%

TABLE III. PERFORMANCE ANALYSIS OF MULTIPLE SVM (QUADRATIC)

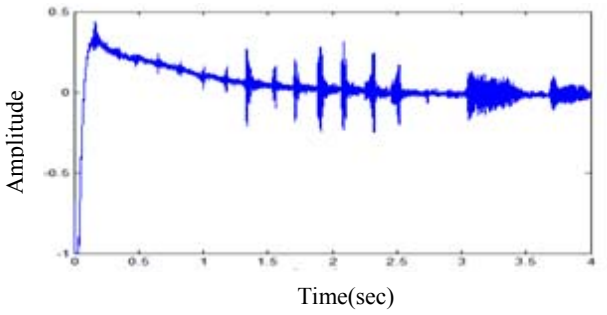
Database	Training Samples	Testing Samples	TP	TN	Accuracy
Happy	500	200	180	20	90.00%
Sad	500	200	181	19	90.83%
Neutral	500	200	186	14	93.33%
Anger	500	200	185	15	92.50%
Surprise	500	200	171	29	85.83%
Disgust	500	200	171	29	85.50%

TABLE IV. PERFORMANCE ANALYSIS OF MULTIPLE SVM (LINEAR)

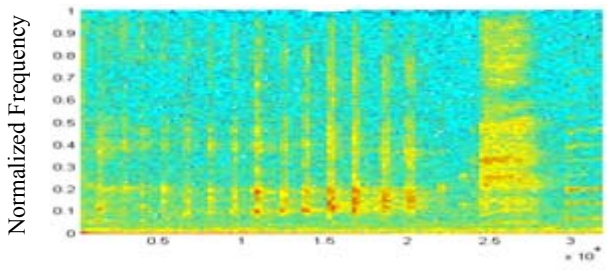
Database	Training Samples	Testing Samples	TP	TN	Accuracy
Happy	500	200	176	24	88.33%
Sad	500	200	163	37	81.66%
Neutral	500	200	171	29	85.83%
Anger	500	200	161	39	80.83%
Surprise	500	200	173	27	86.66%
Disgust	500	200	150	50	75.00%

IV. CONCLUSION

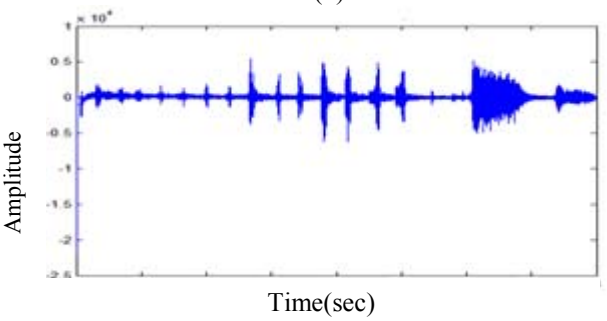
In this paper, emotion sound features are extracted using MFCC and classification is done using linear and nonlinear multiple SVM. Performance analysis of nonlinear SVM overcomes over linear SVM. Still this method suffers problem due to multiple languages and speaking tones.



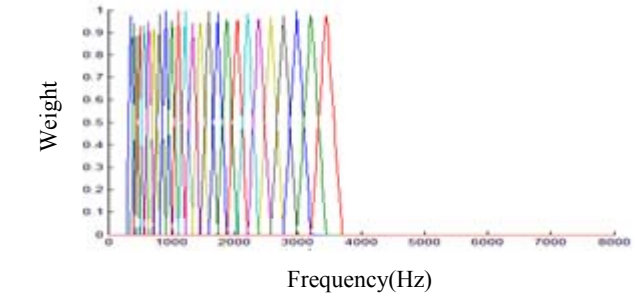
(a)



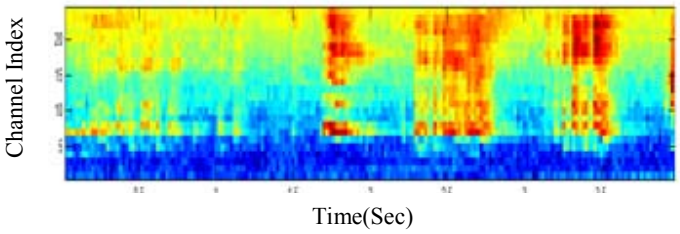
(b)



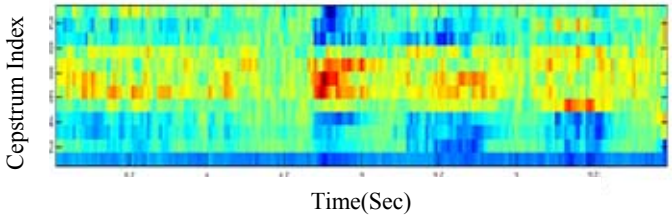
(c)



(d)



(e)



(f)

Fig. 3 a) Original Happy Sound sample b) Spectrogram of sound signal c) Pre-emphasis filter output d)Triangular filter bank output e) Log (Mel) filterbank energies f) Mel frequency cepstrum

References

[1] D. Ververidis, C. Kotropoulos, and I. Pitas, "Automatic emotional speech classification", in Proc. 2004 IEEE Int. Conf. Acoustics, Speech and Signal Processing, vol. 1, pp. 593-596, Montreal, May 2004.

[2] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, B. Weiss, "A database of German emotional speech", Proc. Interspeech, 2005.

[3] F. Yu, E. Chang, Y.Q. Xu, and H.Y. Shum, "Emotion detection from speech to enrich multimedia content", in Proc. 2nd IEEE Pacific-Rim Conference on Multimedia 2001, pp.550-557, Beijing, China, October 2001.

[4] Hermansky, H., "Perceptual linear predictive (PLP) analysis of speech", The Journal of the Acoustical Society of America, Vol. 87, No. 4, pp. 1738-1752, 1990.

[5] I. S. Engberg, and A. V. Hansen, "Documentation of the Danish Emotional Speech Database (DES)", Internal AAU report, Center for Person Kommunikation, Department of Communication Technology,Institute of Electronic Systems, Aalborg University, Denmark,September 1996.

[6] J. Kreiman and B. R. Gerrat, "Perception of aperiodicity in pathological voice", Acoustical Society of America, vol.117, pp. 2201-2211, 2005.

[7] J. M. Montero, J. Gutierrez-Arriola, J. Colas, E. Enriquez, and J. Pardo, "Analysis and modelling of emotional speech in Spanish", in Proc. ICPhS'99, pp. 957-960, San Francisco 1999.

[8] Johnson, M.T., Clemins, P.J., Trawicki, M.B., "Generalized Perceptual Features for Vocalization Analysis Across Multiple Species", ICASSP.2006 Proceedings.,2006.

[9] L. R. Rabiner and M. R. Sambur, "An algorithm for determining the endpoints of isolated utterances," Bell System Technical Journal, vol. 54, no. 2, pp. 297-315, February 1975.

[10] Lindasalwa Muda, Mumtaj Begam and I. Elamvazuthi, "Voice Recognition Algorithms using Mel Frequency Cepstral coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques," Journal of Computing, Volume 2, ISSUE 3, MARCH 2010

[11] Specht, D. F., "Probabilistic neural networks for classification, mapping or associative memory", Proceedings of IEEE International Conference on Neural Network, Vol. 1, pp.525-532, Jun. 1988.

[12] Tsuyoshi Moriyama, Shinya Mori, and Shinji Ozawa. "A synthesis method of emotional speech using subspace constraints in prosody". Journal of Information Processing Society of Japan, 50(3):1181–1191, 2009. (in Japanese).