KNN based emotion recognition system for isolated Marathi speech

Rani Prakash Gadhe, Ratnadeep R. Deshmukh, Vishal B. Waghmare

Department of Computer Science and IT,
Dr. Babasaheb Ambedkar Marathwada University,
Aurangabad-431004 (MS) India

Abstract

This paper gives a comparison of two extracted features namely pitch and formants for emotion recognition from speech. The research shows that various features namely prosodic and spectral have been used for emotion recognition from speech. The database used for recognition purpose was developed on Marathi language using 100 speakers. We have extracted features pitch and formants. Angry, stress, admiration, teasing and shocking have been recognized on the basis of features energy and formants. The classification technique used here is K-Nearest Neighbor (KNN). The result for formants was about 100% which is comparatively better than that of energy which was 80% of accuracy.

Keywords-Database, Emotion recognition, Feature Extraction, Formants, KNN classification , Pitch, Speech signals.

I. Introduction

Speech is a complex signal which contains information about the message, speaker, language and emotions. An emotion makes speech more expressive and effective. Different ways like laughing, yelling, teasing, crying, etc, are used by humans to express their emotions [1].

With computers becoming an integral part of our daily life, there is a need for a more intelligent interaction between humans and machines. While speech recognition is already a fairly well established area of research, it lacks the “human touch” to respond appropriately to another person’s emotion. To overcome this set-back, it is desirable for computers to have the built-in capability to detect, interpret and respond to various emotional situations in the same way as a human does. This audio-emotion recognition (AER) is a fairly new field of research which is being given great importance in the area of human-machine interfaces (HMIs). Research studies have shown that emotion affects our decision making process [2]. It helps us to communicate with one another by expressing our feelings and providing feedback. This makes emotion recognition an increasingly important aspect in the design of a HMI model. AER is particularly useful in the area of human-robotic interaction and telephone-based customer service application. In a human-robotic interaction, service robots are developed to perform human tasks such as assisting in caring for the people in their daily life. Unlike the industrial robots which are typically found in manufacturing environments,

Service robots interact with many users in a variety of places such as hospitals, homes and offices. These robots should be able to recognize the emotional state of the users and respond accordingly to the users’ intentions [3]. For instance, robotic pets can be taught to interact with humans and recognize their emotions and be able to react appropriately to different situations [4].

In a smart call center, the AER system can help to detect irate callers and avert ill-feelings that may arise from an unsatisfactory interaction between the caller and the service provider. With AER, the problematic or agitated callers can be rerouted to the professional operators for assistance and the majority of the callers can then be attended to by less experienced or trainee operators. This can help to enhance customer satisfaction by creating a more efficient, friendly and pleasant customer services environment [5].

II. Database

There are general issues consider while recording the emotion speech corpora are as follows:

1. The number of the emotions and number of the subjects who are contributing to recording this should be decided properly.
2. The database which is recorded as natural or acted helps to decide the applications provided by database and quality of database.
3. Proper contextual information is essential; as expressions mainly depend on linguistic content and its context.
4. Labelling of emotions present in the speech databases is highly subjective.

   Size of the database matters more in speech emotion recognition for deciding properties such as scalability and reliability of the develop system.

The database we consists of five emotions each having 8 words. The database was recorded in noisy environment using PRAAT software. The frequency was set 16000HZ. Each word was recorded thrice. In total there were 100 speakers 20male and 50female.

TABLE I. SOME OF THE SAMPLES OF DATABASE

Angry Emotion	English
समजताय काय स्वत:ला	What do you think of yourself
चल निघ	Get out
येड लागल का	Are you mad

a. Some of the samples of Angry emotion

Stress Emotion	English
आता काय	Now what
नशिबच् फुटक्	Bad luck
मिच् का?	Why me?

b. Some of the samples of Stress emotion

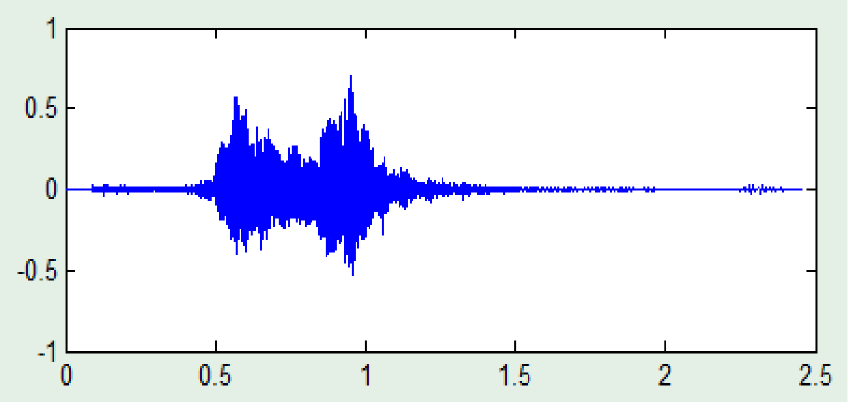
Admiration Emotion	English
वा वा	wow
किती सुंदर	How beautiful
जब्बरदस्त	Fantastic

c. Some of the samples of Admiration emotion

III. FEATURE EXTRACTION

A. Pitch

The fundamental frequency (F0), often referred to as the pitch, is one of the most important features for determining emotion in speech. One of the most popular methods of pitch extraction of a periodic signal in the time domain is to calculate the distance between the zero crossing points of the signal. For pitch analysis, the speech signal is segmented into speech frames using a Hamming window of 512 points with 50% overlap as stated in. A fast Fourier transform is then applied to obtain the spectral information. Figure 1 shows the analysis plots obtained for the pitch features. It shows that emotion recognition will work if thresholding methodology is applied on the pitch mean feature. The term pitch refers to the ear’s perception of tone height. For most purposes, this is just the fundamental frequency f0, though the two terms are not identical since f0 can be measured as a property of an acoustic wave, while pitch is grounded by human perception. Pitch is a very obvious property of speech, also for non-experts, and it is often erroneously considered to be most important for emotion perception. Pitch does definitely have some importance for emotions, but it is probably not as huge as typically assumed. Generally, a rise in pitch is an indicator for higher arousal, but also the course of the pitch contour reveals information on affect. In this paper, pitch was extracted from the speech waveform [6].



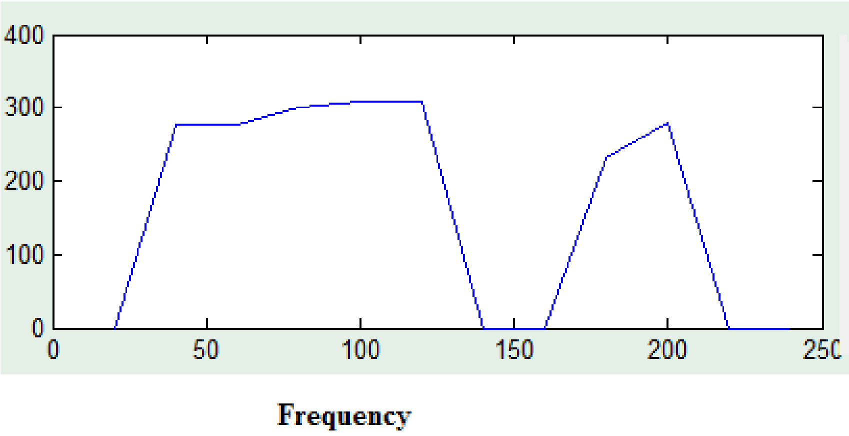


Fig. 1 original sound signal and pitch

B. *Formant*

Formant is a very important parameter to reflect sound track features. First of all, linear prediction is applied to calculate the 14 order prediction coefficients, and then, those coefficients are used to estimate the sound track frequency response curve, and last, the peak picking method is adopted to calculate the frequency of every formant [6]. The average formant frequency and formant frequency changing rate of the first formant; the average peak values and the average slope of regression curve of formant peak of the first 4 formants. Choose the difference between the first average formant frequency of every frame, the average values and slopes of the peak value regression curves of the first 4 formants, and the parameters of the corresponding peaceful Sentences; also choose the rate of first formant frequency changing rate over corresponding sentence [7].

The Linear predictive coding technique (LPC) has been used for estimation of the formant frequencies. The analog signal is converted in .wav digital format. The signal is transformed to frequency domain using FFT and the power spectrum is further calculated. Then the signal is passed through a Linear Predictive Filter (LPC) with 11 coefficients and the absolute values are considered. The roots of the polynomial are obtained which contain both real and imaginary parts. The phase spectrum is further displayed which clearly shows the formant frequencies. The first five formant frequencies are displayed in the graph. Figure 2 shows the formant frequency plot along with the original speech signal. The five formant frequencies obtained are 230 Hz, 800 Hz, 1684 Hz, 2552 Hz, 3159 Hz respectively [8].

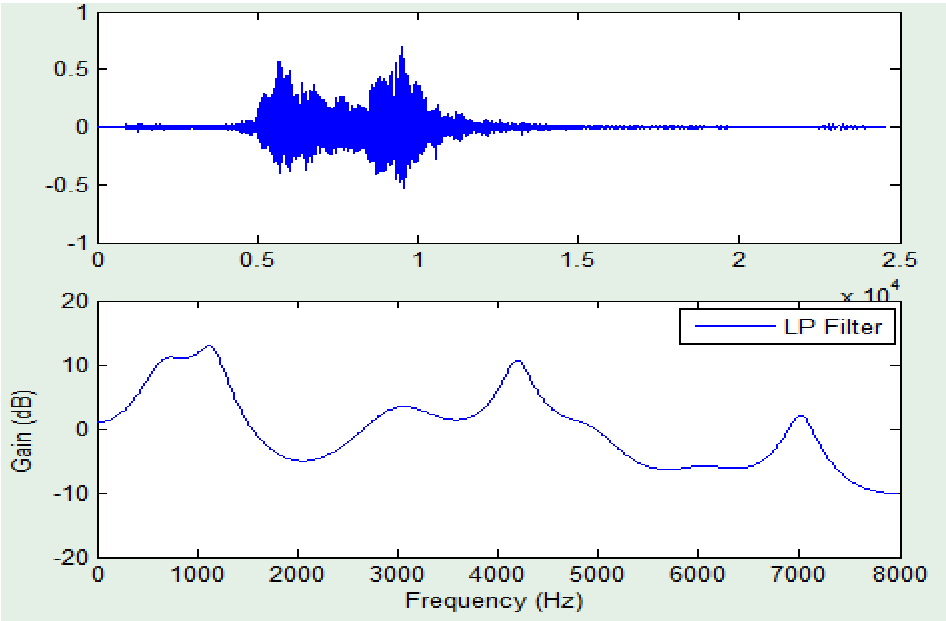


Fig. 2 Original sound signal and formant.

IV. CLASSIFICATION

A. KNN

In pattern recognition, KNN is the simplest algorithm only based on memory. Being simple, elegant and straightforward, many researchers often adopt KNN as a classifier for their applications today. 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 [9].

In recent years in speech emotion recognition, researchers proposed many classification algorithms, such as Neural Networks (NN), Gaussian Mixture Model (GMM),Hidden Markov Model (HMM), Maximum Likelihood Bayesian classifier (MLC), Kernel Regression and K-nearest Neighbors (KNN) [1] and Support vector machines (SVM). In pattern recognition, the k-Nearest Neighbors algorithm (or k-NN for short) is a non-parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the future space. The output depends on whether k-NN is used for classification or regression [10].

In the classification phase, k is a user-defined constant, and an unlabeled 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. A commonly used distance metric for continuous variables is Euclidean distance. In our research, we use the system to extract the speech's feature. After the feature extraction, we give each speech sample with the corresponding emotion class label. After that we input them to the K-NN classifier and gain a result class which classifies emotion.

B. 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.

Accuracy = (Correctly classified samples/Total number of samples) X 100.

The following table 5.1 displays the approximately achieved accuracy for Angry, stress, admiration, teasing and shocking using pitch and formants as discriminate factor.

TABLE II. ACCURACY RESULR

Emotion	Pitch	Formant
Angry	40%	100%
Stress	60%	90%
Admiration	50%	90%
Teasing	50%	100%
Shocking	50%	100%

V. CONCLUSION

Using pitch and formant stress can be 60% and 90% recognized where as angry gives the lowest accuracy that is of 40% accuracy with pitch. Admiration, teasing and shocking gives 50% 50% and 50 % of accuracy with pitch. Whereas using formant we get 90%, 100% and 100% for the same. Thus formant plays a vital role in recognizing emotions and we can conclude that using formant we can accurately recognize emotions compare to that of pitch.

REFERENCE

[1] Rani P. Gadhe, R. R. Deshmukh, V. B. Waghmare, P. P. Shrishrimal, “Emotion Recognition From Speech:A Survey,” ISSN 2229-5518, International Journal of Scientific & Engineering Research, Volume 6, Issue 4, April-2015.

[2] Busso, S. Lee And S. Narayanan, “Analysis Of Emotionally Salient Aspects Of Fundamental Frequency For Emotion Detection”, IEEE Transactions On Audio, Speech, And Language Processing, Vol. 17,No.4, Pp. 582–596, May 2009.

[3] J. S. Park, Ji-H. Kim And Yung-H. Oh, “Feature Vector Classification Based Speech Emotion Recognition For Service Robots” IEEE Transactions On Consumer Electronics, Vol. 55, No. 3, Pp. 1590-1596, Aug. 2009.

[4] Y. Sidorova, “Optimization Techniques for Speech Emotion Recognition”, Phd Thesis, Departament De Traducci´O I Ci`Encies Del Llenguatge, Http.Www.Tesisenxarxa.Net/ Tesis_Upf/Available/Tdx...//Tys.Pdf. Dec. 2009.

[5] A. Wahab, Q. Chai And S. De, “Speech Emotion Recognition Using Auditory Cortex”, IEEE Congress On Evolutionary Computation 2007, Pp. 2658-2664, Sep. 2007.

[6] J. Clark, C. Yallop, And J. Fletcher, An Introduction To Phonetics And Phonology, 3rd Ed. Malden, Ma, Usa: Blackwell Publishers, January 2007.

[7] Xin Min Cheng ,Pei Ying Cheng, Li Zhao,” A Study On Emotional Feature Analysis And Recognition In Speech Signal,” International Conference On Measuring Technology And Mechatronics Automation, 978-0-7695-3583-8/09 © 2009 IEEE Doi 10.1109/Icmtma.2009.89.

[8] Bageshree V. Sathe-Pathak, Ashish R. Panat” Extraction Of Pitch And Formants And Its Analysis To Identify 3 Different Emotional States Of A Person,” Issn (Online): 1694-0814, Ijcsi International Journal Of Computer Science Issues, Vol. 9, Issue 4, No 1, July 2012.

[9] Muzaffar Khan, Tirupati Goskula, Mohmmed Nasiruddin ,Ruhina Quazi,” Comparison Between K-nn And svm Method For Speech Emotion Recognition,” Muzaffar Khan Et Al. / International Journal On Computer Science And Engineering (Ijcse), X`Issn : 0975-3397 Vol. 3 No. 2 Feb 2011.

[10] Anuja Bombatkar, Gayatri Bhoyar, Khushbu Morjani, Shalaka Gautam,Vikas Gupta,” Emotion recognition using Speech Processing Using k-nearest neighbor algorithm,” International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622, April 2014.