ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 6, June 2014

Speech Emotion Recognition A Review and Related Terms

Ekta Garg[1], Madhu Bahl [2]

M Tech Scholar, CEC, LANDRAN, India

Assistant Professor, CEC, LANDRAN, India

ABSTRACT : Speech emotion recognition is a process where a speech file is recognized against the stored speech data set . It analyzes the data set according to the classifier and predicts results accordingly . In this scenario , a predicted output is one which matches the most with the data base. Several kinds of classifier have been used in this scenario . This paper represents different sections of the speech recognition process and classification methods are also discussed .

KEYWORDS*:* SPEECH RECOGNITION, CLASSIFIERS , RECOGNITION PROCESS

I. INTRODUCTION

The dynamic requirements of automated systems have pushed the extent of recognition system to consider the precise way of command rather to run only on command templates. The idea correlates itself with the speaker identification at the same time recognizing the emotions of speaker. The acoustic processing field not only can identify „who" the speaker is but also tell „how" it is spoken to achieve the maximum natural interaction. [1]

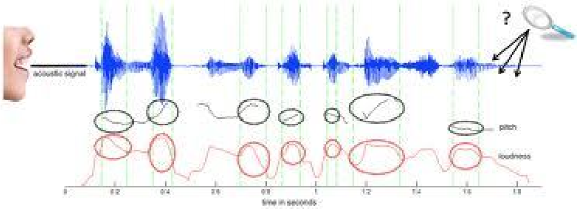
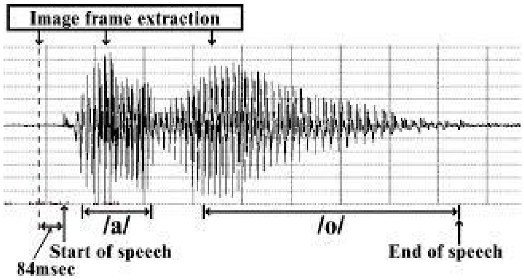


Fig.1

This can also be used in the spoken dialogue system e.g. at call centre applications where the support staff can handle the conversation in a more adjusting manner if the emotion of the caller is identified earlier. The human instinct recognizes emotions by observing both psycho-visual appearances and voice. Machines may not exactly emulate this natural tendency as it is but still they are not behind to replicate this human ability if speech processing is employed. Earlier investigations on speech open the doors to exploit the acoustic properties that deal with the emotions. At the other hand the signal processing tools like MATLAB and pattern recognition researcher's community developed the variety of algorithms (e.g. HMM, SVM) which completes needed resources to achieve the goal of recognizing emotions from speech.[2]

a) **DATABASE**: A data base is the collection of data .In our proposed work we have used speech samples for the database. In the database we find properties of the speech signals and then we store them into the database. The question comes that how we are going to store hundreds of files in the database. The procedure would be as follows. First of all we would fetch the properties of the voice samples. All those properties which are required would be computed and then it would be stored into an array. The array would move on as the files would move. We would fetch the features and would take the average by the end and then store them into the database for each category of the voice which we have taken i.e. **HAPPY, SAD, ANGRY AND FEAR.**



b) **VOICE FILES**: The voice files are the files which would be processed for the feature extraction.

c) **PROPERTIES**: When we would process the voice files their properties would be fetched .For the feature extraction there are several algorithms which can be used. In our approach we have used HMM algorithm for the training purpose.

1.1 SECTIONS OF THE RESEARCH WORK:

There are two sections in our research work. The sections are explained as follows.

A) TRAINING: The training section ensures that the database gets trained properly so that at the time of testing it produces extensive results. The features of the training are as follows.

a) Maximum Frequency: The maximum frequency of a file is the value which we get at the peak on a frequency map. When so ever we put a voice sample over the time and frequency pattern, the maximum peak is called the maximum frequency of the voice sample.

b**) Minimum Frequency**: The minimum frequency of a file is the value which we get at the peak on a frequency map. When so ever we put a voice sample over the time and frequency pattern, the minimum peak is called the minimum frequency of the voice sample.[3]

c) Average Frequency: The average frequency can be calculated using two techniques. The first technique is to add all the frequency samples and then divide the entire sum with the total number of frequency. The second method is a very ethical method in which we can add the minimum frequency and the maximum frequency and then we can divide them by two.

Avg. Frequency = (Minimum frequency + Maximum Frequency)/2

d) Spectral Roll off: The spectral roll off in terms of development can be said as the difference between the maximum frequency differences with the adjacent frequency. The position of the frequency (max) can be stored into an array and similar of the adjacent node and then the difference can be calculated.

e) Noise Level: Ethically the noise level is the extra number of bits which has been added into the voice sample. If the noise is uniform then the noise level can be calculated by taking the difference of each frequency sample and the threshold of the voice sample.

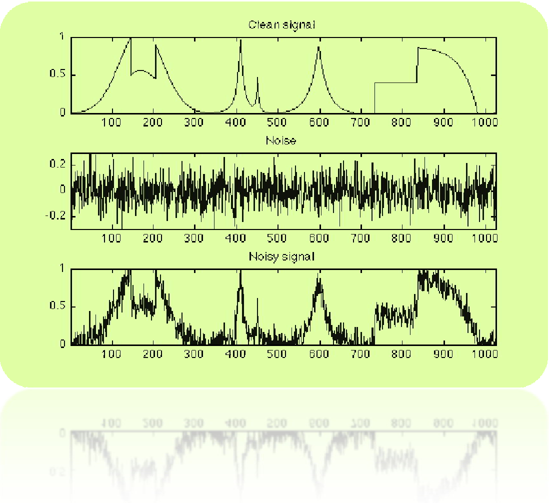
There are two categories of the noise level.

1) UNIFORM NOISE
2) NON UNIFORM NOISE

UNIFORM NOISE: Uniform noise is the noise which is simultaneously same all over the voice sample.
NON UNIFORM NOISE: The non uniform noise does not remain constant all over the sample.

f) Pitch: It is the average value of the entire voice sample.

g) Spectral Frequency: The spectral frequency is the frequency of the voice pitch next to the highest voice sample.



The above image shows normal and the noisy signal [10]

1.2 ALGORITHM HELPFUL IN THE FEATURE EXTRACTION:

1) **HMM**: HMM stands for HARCOV'S META MODEL. It is a worldwide known algorithm for the training of the data set. It extracts the features of the voice sample and saves them to the database for the future use. The maximum frequency of a file is the value which we get at the peak on a frequency map. When so ever we put a voice sample over the time and frequency pattern, the maximum peak is called the maximum frequency of the voice sample. It is viewed as the counter part of the training and it is used to sample size the data for the further processing. In this approach we take each sample of data set as a unique item which has to be processed. The extraction of the feature and saving it to the data base can be classified with the following flow diagram.[4]

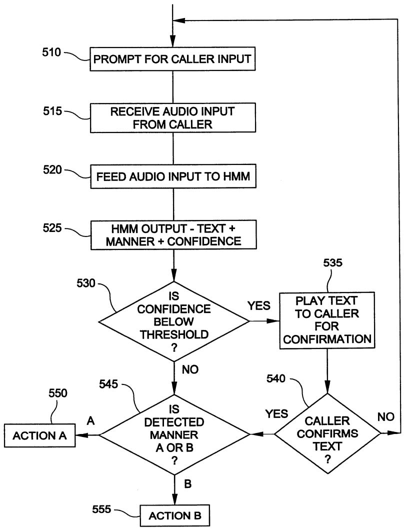

510
PROMPT FOR CALLER INPUT
515
RECEIVE AUDIO INPUT FROM CALLER
520
FEED AUDIO INPUT TO HMM
525
HMM OUTPUT - TEXT + MANNER + CONFIDENCE
530
IS CONFIDENCE BELOW THRESHOLD ?
YES
535
PLAY TEXT TO CALLER FOR CONFIRMATION
NO
540
545
550
ACTION A
A
IS DETECTED MANNER A OR B ?
YES
CALLER CONFIRMS TEXT ?
NO
B
555
ACTION B

Fig. Flow diagram of HMM

2. Acoustic Modelling: Acoustic models are developed to link the observed features of the speech signals with the expected phonetics of the hypothesis word/sentence. For generating mapping between the basic speech units such as phones, tri-phones & syllables, a rigorous training is carried. During training, a pattern representative for the features of a class using one or more patterns corresponding to speech sounds of the same class.

3. Language & Lexical Modelling: Word ambiguity is an aspect which has to be handled carefully and acoustic model alone can't handle it. For continuous speech, word boundaries are major issue. Language model is used to resolve both these issues. Generally ASR systems use the stochastic language models. These probabilities are to be trained from a corpus. Language accepts the various competitive hypotheses of words from the acoustic models and thereby generates a probability for each sequence of words. Lexical model provides the pronunciation of the words in the specified language and contains the mapping between words and phones. Generally a canonical pronunciation available in ordinary dictionaries is used. To handle the issue of variability, multiple pronunciation variants for each word are covered in the lexicon but with care. A G2P system- Grapheme to Phoneme system is applied to better the performance the ASR system b predicting the pronunciation of words which are not found in the training data. [5]

4. Model Adaptation: The purpose of performing adaptation is to minimize the system's performance dependence on speaker's voice, microphones, transmission channel and acoustic environment so that the generalization capability of the system can be enhanced. Language model adaptation is focused at how to select the model for specific domain. Adaptation process identifies the nature of domain and, thereby, selects the specified model.

5. Recognition: Recognition is a process where an unknown test pattern is compared with each sound class reference pattern and, thereby, a measure of similarity is computed. Two approaches are being used to match the patterns: First one is the Dynamic Time Warping based on the distance between the acoustic units and that of recognition. Second one is HMM based on the maximization of the occurrence probability between training and recognition units. To train the HMM and thereby to achieve good performance, a large, phonetically rich and balanced database is needed.

C. Performance Parameters

Accuracy and Speed are the criterion for measuring the performance of an automatic speech recognition system which are described below:

1. Accuracy Parameters
Word Error Rate (WER): The WER is calculated by comparing the test set to the computer-generated document and then counting the number of substitutions (S), deletions (D), and insertions (I) and dividing by the total number of words in the test set[5]

2. Speed Parameter
Real Time Factoris parameter to evaluate speed of automatic speech recognition. Formula: P RTF = ----------------- I where P: Time taken to process an input Duration of input I e. g. RTF= 3 when it takes 6 hours of computation time to process a recording of duration 2 hours. RTF ≤ 1 implies real time processing.

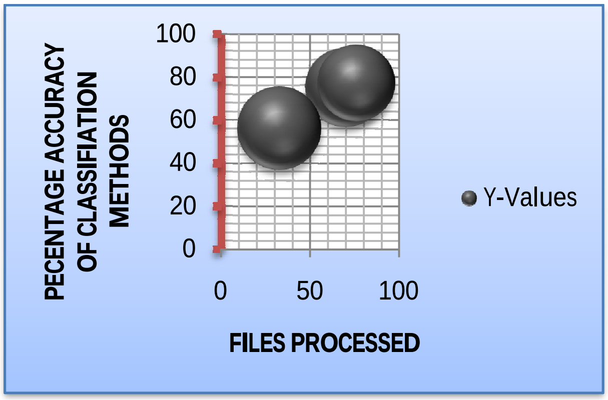
II. TESTING METHOD

The testing module of the speech processing includes the testing of the speech file on the basis of the trained data set . To perform a testing operation over the speech files different types of classifiers are used to analyze the services of the speech samples . Some of the classifiers are explained as follows .

A) SVM : SVM stands for support vector machine . It takes the entire data set as the binary input and classifiers for the same . The SVM classifier generates the FAR and FRR ratio successfully to determine the matching percentage . SVMs are linear classifiers (i.e. the classes are separated by hyper planes) but they can be used for non-linear classification by the so-called *kernel trick*. Instead of applying the SVM directly to the input space Run they are applied to a higher dimensional *feature space* F, which is nonlinearly related to the input space: _ : Run ! F. The kernel trick can be used since the algorithms of the SVM use the training vectors only in the form of Euclidean dot-products (x _ y). It is then only necessary to calculate the dot-product in feature space (_(x) __(y)), which is equal to the so-called *kernel function* k(x; y) if k(x; y) fulfils the Mercer's condition. Important kernel functions which fulfil these conditions are the polynomial kernel

B) GNB CLASSIFIER : GNB stands for Gaussian naïve based classifier . It is useful when the prediction has to be done on noisy speech.

C) NEURAL NETWORK CLASSIFIERS : The neural network classifier is one of the most advance classifiers which takes two inputs.the first input is the training set and the second input is the target set . The target is drawn on the basis of which the training set has been updated .[6]



Neural nets are highly interconnected networks of relatively simple processing elements, or nodes, that operate in parallel. They are designed to mimic the function of neurobiological networks. Recent work on neural networks raises the possibility of new approaches to the speech recognition problem. Neural nets offer two potential advantages over existing approaches. First, their use of many processors operating in parallel may provide the computational power required for continuous-speechrecognition. Second, new neural net algorithms, which could self-organize and build an internal speech model that maximizes performance, would perform even better than existing algorithms. These new algorithms could mimic the type of learning used by a child who is mastering new words and phrases.

III. CONCLUSION

With the above text , it can be concluded that the speech recognition system is a process which requires two phases of data . The first phase is the training phase and the second phase is the testing phase . A testing phase cannot be optimal if the training has not be provided efficiently . The testing can be done using different sort of classifiers as already mentioned in the context written above . The training can be done using feature extraction methods .

REFERENCES

[1]http://www.cs.uiuc.edu/~hanj/pdf/ency99.pdf
[2] 1 K.A.Senthildevi, 2 Dr.E.Chandra, "Speech Data Mining & Document Retrieval", publication of the IEEE signal processing.
[3]www.asru2007.org/submission/EDICS.txt.
[4] IEEE BjornSchuler, Gerhard Ripoll, and Manfred Lang "Hidden Markov Model based speech emotion recognition". Institute for Human-Computer Communication Technische UniversidadMunched (Schuler | ragdoll | Lang)@ei.tum.de, 0-7803-7663-3/03/$17.00 ©2003
[5] Tin Lay New a,*, Say Wei Food b, Liyanage C. De Silva a "Speech Emotion Recognition using hidden Markov models", in Elsevier Speech Communications Journal Vol. 41, Issue 4, pp. 603-623, 2003
[6] Fixing Pan, PeepedSheen and LopingSheen "Speech Emotion Recognition Using Support Vector Machine", Department of Computer Technology Shanghai Jiao Tong University, Shanghai, China. International Journal of Smart Home Vol. 6, No. 2, April, 2012
[7] Mustafa Khan et al. / International Journal on Computer Science and Engineering (IJCSE) "Comparison between k-nun and sum method for speech emotion recognition", Inhuman College of Engineering & Technology ,Sadder, Nagpur, India
[8] Anural Kumar1, Paul Agarwal1, Panay Dighe1, Subsalt Subhechha1, Bhiksha Raj2, Inshore Prahallad3, "Speech Emotion Recognition by Gadabouts Algorithm and Feature Selection for Support Vector Machine", Indian Institute of Technology, Kanpur, 2LTI, School of Computer Science, Carnegie Mellon University, Pittsburgh, 3International Institute of Information Technology, Hyderabad
[9] 1Ashish B. Ingle, 2Dr.D.S.Chaudhari, International Journal of Advanced Engineering Research and Studies E-ISSN2249–8974 "Speech Emotion Recognition using Hidden Markov Model and Support Vector Machine".
[10]https://www.google.co.in/search?q=noisy+signal&safe=active&source=lnms&tbm=isch&sa=X&ei=rrVXU4PQMcyBrQfY_oDoDQ&ved=0CAYQ_AUoAQ&biw=1366&bih=624#facrc=_&imgdii=_&imgrc=Wy_xlQboyshnmM%253A%3BWHlOQeNaIwPXYM%3Bhttps%253A%252F%252Fwww.ceremade.dauphine.fr%252F~peyre%252Fnumericaltour%252Ftours%252Fdenoisingsimp_1_noise_models%252Findex_01.png%3Bhttps%253A%252F%252Fwww.ceremade.dauphine.fr%252F~peyre%252Fnumerical-tour%252Ftours%252Fdenoisingsimp_1_noise_models%252F%3B560%3B420