Speech Emotion Detection A Review

Jyoti Sain[1] Dr. Sanjeev Dhull[2]
[1]PG Scholar [2]Professor
[1,2]Department of Electronics Communication & Engineering
[1,2]GJUST, Hisar, Haryana, India

Abstract— Emotion detection from voice has risen as a vital research range in the current past. In such manner, audit of existing work on passionate voice handling is helpful for completing further research. In this paper, the re-penny writing on voice emotion detection has been presented considering the issues identified with enthusiastic voice corpora, distinctive sorts of voice elements and models utilized for detection of emotions from voice. Thirty two representative voice databases are evaluated in this work from perspective of their dialect, number of speakers, number of emotions, and reason for gathering. The issues identified with enthusiastic voice databases utilized as a part of passionate voice recognition are additionally quickly talked about. Writing on various features utilized as a part of the undertaking of emotion detection from voice is exhibited. The significance of picking distinctive classification models has been talked about alongside the survey. The vital issues to be considered for advance emotion detection inquire about when all is said in done and in particular to the Indian setting have been featured any place important.

Key words: Speech Emotion Recognition, Classifiers, Recognition Process

I. INTRODUCTION

Voice is a perplexing sign containing data about message, speaker, dialect, emotion etcetera. Most existing voice frameworks process studio recorded, unbiased voice successfully, nonetheless, their execution is poor on account of enthusiastic voice. This is because of the trouble in demonstrating and portrayal of emotions introduce in voice. Nearness of emotions makes voice more normal. In a discussion, non-verbal correspondence conveys an imperative data like goal of the speaker. Notwithstanding the message passed on through content, the way in which the words are talked, passes on basic non-phonetic data. The same literary message would be passed on with various semantics (which means) by joining proper emotions. Talked content may have a few translations, contingent upon how it is said. For instance, "Approve" in English, is utilized to express profound respect, incredulity, assent, lack of engagement or an affirmation. Hence understanding the content alone is not adequate to translate the semantics of a talked expression. In any case, it is imperative that, voice frameworks ought to have the capacity to process the non-etymological data, for example, emotions, alongside the message. People comprehend the planned message by seeing the fundamental emotions notwithstanding phonetic data by utilizing multi-modular prompts. Non-phonetic data might be seen through (1) outward appearances on account of video, (2) articulation of emotions on account of voice, and (3) accentuation on account of composed content. The voice in this paper is limited itself to emotions or articulations identified with voice. Fundamental objectives of enthusiastic voice preparing are (an) understanding emotions exhibit in voice and (b) integrating wanted emotions in voice as per the planned message. From machine's point of view understanding voice emotions can be seen as characterization or segregation of emotions. Amalgamation of emotions can be seen as joining emotion particular information amid voice blend.

Voice is one of the common modalities of human mama chine collaboration. The present voice frameworks may achieve human proportional execution just when they can process underlying emotions successfully. Motivation behind advanced voice frameworks ought not be constrained to negligible message preparing, rather they ought to comprehend the fundamental goals of the speaker by recognizing expressions in voice. In the current past, handling voice motion for perceiving basic emotions is developed as one of the essential voice inquire about territories. Inserting the segment of emotion handling into existing voice frameworks makes them more regular and compelling.

The dynamic necessities of mechanized frameworks have driven the degree of acknowledgment framework to consider the exact method for summon rather to run just on order formats. The thought connects itself with the speaker distinguishing proof in the meantime perceiving the emotions of speaker. The acoustic preparing field can recognize „who" the speaker is as well as advise "how" it is addressed accomplish the most extreme regular collaboration. [1]

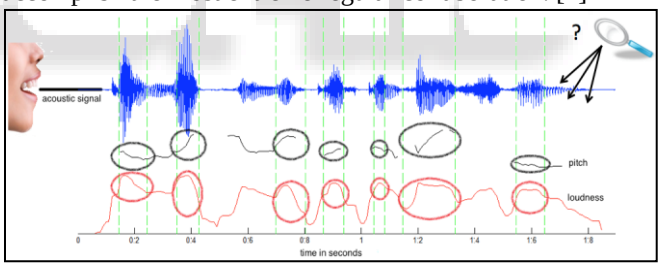

Fig. 1: Voice segmentation using pitch detection

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 nature recognizes emotions by observing both psycho-visual appearances and voice. Machines may not exactly follow this natural leaning as it is but still they are not behind to reproduce this human ability if speech processing is employed. Previous investigations on speech open the doors to exploit the sound properties that contract with the emotions. At the other side the signal processing tools like MATLAB and pattern identification researcher's community developed the number of algorithms (e.g. HMM, SVM) which completes needed resources to achieve the aim 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. We get properties

of the speech signals from the database 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 [12]. First of all we would fetch the properties of the voice samples. All those properties which are required would be calculated and then it would be stored into an array [13]. 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.

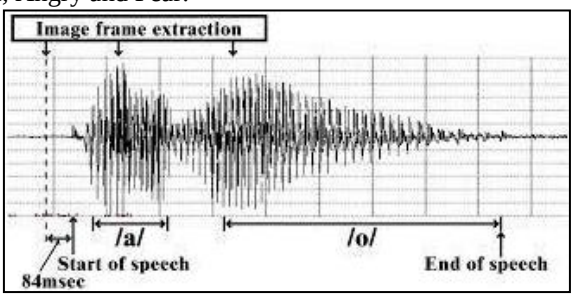

Fig. 2: The voice files are those files which would be processed for the feature extraction.

B. Properties

When we would process the voice files their properties would be fetched .For the feature extraction there are different algorithms which can be used. In our approach we have used HMM algorithm for the training use.

C. Sections of the research work

There are two sections in our research work. The sections are explained as follows [14].

1) Raining

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. The minimum peak is called the minimum frequency of the voice sample when voice sample is put over a time and frequency pattern [3].

c) Average Frequency

The average frequency can be calculated using two techniques. The first method is to include 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 frequency maximum and the minimum frequency and then we can divide them by two.

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. The noise level can be calculated by taking the difference of the threshold of the voice sample and each frequency sample when noise is uniform.

There are two categories of the noise level.

– Uniform Noise
– 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.

1) Pitch: It is the average value of the entire voice sample.
2) Spectral Frequency: The spectral frequency is the frequency of the voice pitch next to the highest voice sample.

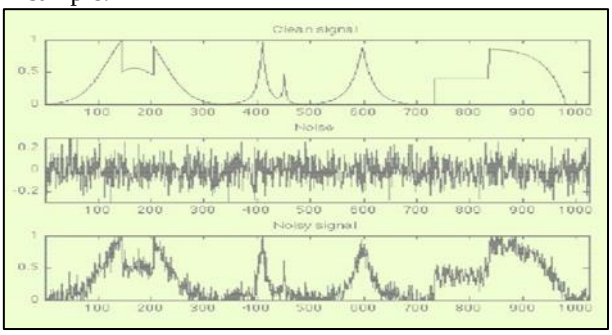

Fig. 3: Noisy Signal

D. Algorithm helpful in the feature extraction

1) HMM

HMM stands for HIDDEN'S MARKAV MODEL. It is a worldwide known algorithm for the training of the dataset. 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 rate 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]

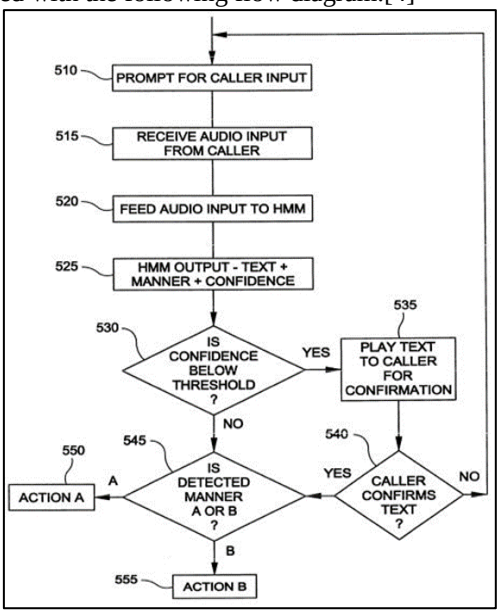

Fig. 4: Flow diagram of HMM

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 creating mapping between the basic speech units such as syllables, phones& tri-phones, an accurate 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 accent of the words in the particular 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 fair database is needed.

E. 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 Factors 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 involves 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 root deception can be used since the method of the SVM use the data train 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 fulfill these conditions are the polynomial kernel

B. GNB Classifier

GNB stands for Gaussian natïve based classifier. It is useful when the prediction has to be done on noisy speech.

C. Neural Network Classifiers

The neural system classifier is a standout amongst the most propels classifiers which takes two inputs .The first info is the preparation set and the second information is the objective set. The objective is drawn on the premise of which the preparation set has been updated [6].

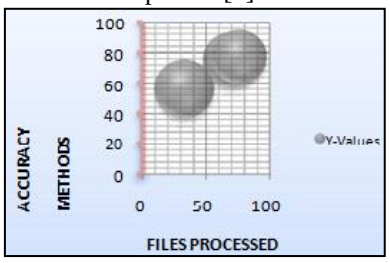

Fig. 5: Neural System Classifier

Neural nets are exceedingly interconnected systems of moderately straightforward preparing components, or hubs that work in parallel. They are intended to copy the capacity of neurobiological systems. Late work on neural systems raises the likelihood of new ways to deal with the discourse acknowledgment issue. Neural nets offer two potential favorable circumstances over existing methodologies. To begin with, their utilization of numerous processors working in parallel may give the computational power required to consistent speech recognition. Second, new neural net calculations, which could self-sort out and manufacture an inner discourse display that augments execution, would

perform far superior to existing calculations. These new calculations could imitate the sort of learning utilized by a kid who is acing new words and expressions.

III. CONCLUSION

With the above content, it can be reasoned that the discourse acknowledgment framework is a procedure which requires two periods of information. The primary stage is the preparation stage and the second stage is the trying stage. A testing stage can't be ideal if the preparation has not be given effectively. The testing should be possible utilizing distinctive kind of classifiers as of now specified in the setting composed previously. The preparation should be possible utilizing highlight extraction techniques.

REFERENCES

[1] Sathit Prasomphan, "Detect Human Emotion via words Recognition by Using Speech Spectrogram", 16th IEEE International journal on Human statement Recognition, January, Thailand, .2016.
[2] M.M.H.E. Ayadi, M.S., Kamel and F. Karray, "Survey on speech emotion identification: Features, arrangement schemes and databases", Pattern Recognition, pp. 572–587, 2011.
[3] E.H. Kim, K.H. Hyu, S.H. Kim and Y.K.Kwak, "Speech Emotion detection Using Eigen-FFT in fresh and Noisy environment", 16th IEEE International Conference on Robot Human Interactive Communication August, Jeju, Korea, 2007.
[4] Akash Shaw, Rohan Kumar Vardhan,"Emotion Recognition and Classification in Speech using ANN", International Journal of Computer Applications (0975 – 8887) Volume 145 – No.8, July 2016.
[5] Shambhavi S.S, Dr. V.N Nitnaware, "Emotion Speech Recognition using MFCC and SVM", International Journel of Engineering & Technology (ISSN:2278-0181) Vol.4 Issue 06,June-2015.
[6] L. He, M. Lech, M.C. Maddage and N.B. Allen," Time-frequency feature extraction from spectrogram and wavelet packets with application to automatic stress and emotion classification in speech Information", 7th International Conference on Communications and Signal Processing, ICICS 2009, Macau, pp. 1–5, 2009, .
[7] A. Harimi, A. Shahzadi, A.R. Ahmadyfard and K. Yaghmaie, "Classification of emotional speech through spectral pattern features", Journal of AI and Data Mining, vol.2, no.1, pp.53–61, 2014, .
[8] Japanese Emotion Database (Japanese emotional speech).
[9] M.M. Javidi and E.F. Roshan. ,"Speech Emotion Recognition by Using Combinations of C5.0, Neural Network (NN) and Support Vector Machines (SVM) Classification Methods". Journal of mathematics and computer Science. vol. 6, pp. 191-200, Apr. 2013.
[10] Mihir Narayan Mohanty, A Routray.,"Machine Learing Approach for Emotional Speech Classification",SEMCO-14, Book chapter, Springer Verlag Berlin Heidelberg, 2014.
[11] K.Z. Mao et.al.,"Probabilistic Neural-Network Structure Determination for Pattern Classification". IEEE Transactions on neural networks.,vol. 11, no. 4, pp. 1009-1016. Jul. 2000.
[12] Chuang, Z.J., Wu, C.H. "Emotion recognition using acoustic features and textual content", In Proc. ICME, Taipei, Taiwan. pp. 53- 56, 2004.
[13] H.Kaur and R.Talwar, "Performance and Convergence Analysis of LMS Algorithm," IEEE ICCIC, Dec.2012.
[14] J.Gorriz and J.Ramrez, "A Novel LMS Algorithm Applied to Adaptive Noise Cancellation," IEEE Signal Process Letters, vol. 16, no. 1, Jan.

