Process of Speech Emotion Recognition and its Categories

Inderjeet Kaur[1], Dr. Rakesh Kumar[2], PalwinderKaur[3]
[1]*Research Scholar, Department of CSE,Sachdeva Engineering College for Girls, Gharuan*
[2]*Principal &Professor, Sachdeva Engineering College for Girls, Gharuan*
[3]*Assistant Professor, Sachdeva Engineering College for Girls, Gharuan*

Abstract: Speech emotion recognition is the automatic method of identifying feelings in the voice. There are world-wide applications in the field of phsyc-atrics and in robotics HCI (human computer interaction). The voice signal comprises the communication being vocal, the emotional state of the vocal and the information from the speaker, so that the voice signal could be cast off for the identification of the vocal and the emotional phase of the vocal. In this paper, we discuss about process of speech emotion and its types like isolated speech, Spontaneous speech and feature of speech emotion (Paralinguistic Features, Linguistic Features).

Keyword: *Process of speech emotion, isolated speech, Spontaneous speech, Paralinguistic Features, Linguistic Features.*

I. INTRODUCTION

Speech emotion identification objectives at involuntarily verification of the emotional/physical condition of a person via his/her tone. A vocal has different phases overall voice that are shown in figure 1, as a feeling facet of voice is combined with the so termed paralinguistic aspects [1]. The language satisfied could not adapt by express condition; in communication of entity this is an important factor, because response data is provided in numerous appliance. Voice is probably the normal capable path to agree with each other. These resources oblige boundary to collaborate with machineries. Some winning illustrations based-on it in the past years, they have awareness regarding electromagnetism; with the expansion of the telephone, megaphone. Yet in the previous times people were analyzing on speech combined on Kempelen developed machine, capable of 'speaking' arguments and expression.

In current time, speech emotion recognition has achieved not only to expand inspection and the run able voice recognition scheme, but also to have systems competent to real-time change of texture into speech. Unfortunately, on account of the wide quality progress made on that field, there are limitless applications that are the voice recognition process facing crack currently. These are some difficulty factors in speech identification are discussed as [2]:

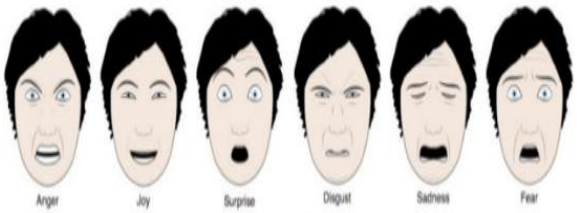

Figure 1: speech emotion [14].

Speaker Complete- matching expression is marked as a new way by diverse persons because masculinity, stage, fastness of voice, fluency of the utterer and language difference.

The various advantages and disadvantages of speech emotion recognition are:

Merits:

- Permits a consumer to function a computer by language in it;
- Open-up reasoning employed area;
- Permits transcript of texture, instructions;
- Removes handwriting, issues with spelling;
- Doesn't forever identify terms accurately.
- Voice is a natural path to interact and it's not required to be seated at a work through a remote manage.
- Training isn't necessary for consumers [10].

Demerits:

- Requires big quantity of recollection to supply speech records.
- Hard in usage in school room locations, owed to interference intrusion.
- Necessitates of every consumer to teach software to identifyspeech, though for deprived translators.
- The typing mistake could be frustrating.
- Contribution with individual phase of the script procedure, not an answer to the script issue.
- Even the best speech recognition systems sometimes make errors. If there is sound or some other sound, the amount of mistakes will increase.

- Recognition of the voice mechanism is best, if the microphone is near to the customer (an example, in a handset, or if the person is exhausting a microphone), more distant microphones (example, on a counter or partition) would incline to augment the amount of exceptions.

II. PRIOR WORK

Ali, et al. (2013) [3] presented Involuntary recognition by speaking the words, was unique of the most stimulating tasks in the area of speech recognition. The exertion of this workowed to the auditory is same of several of the words and some programs. Enhancing the accuracy recognition necessitates of the system to achieve fine phonetic differences. The method for analysing spoken the words in Bangla is described. In this learning the initial derives feature from speaking thewords. It defines some method for analysing spoken words in Bangla. **Ververidis, et al. (2006) [4]** described an initial task to assemble information and update file where collection of emotional speech information is accessible. A File contains data regarding types of emotions of speakers, voice –type set. In the next step, area properties are represented that are used to take out features for sensitive speech recognition and to calculate how the emotions affect them. Normally, features that stay alive in the market are like ground, the vocal emotional states. Here dissimilar classification techniques are inspecting in which timing information was broken. The normal classification methods are HMM (Hidden Markov Models), Artificial Neural Network (ANN), k-nearest neighbor's method and support vector machine techniques. **Platt, et al. (1999)** described new **[5]** method for exercising SVM machine learning approach: Sequential Nominal Optimization, or SMO. Perquisite of the key for the problem of training a support vector machine is to be used on a very big quadratic programming (QP) optimization difficulty. Firstly large QP problem is divided by SMO into a sequence of small conceivable programming problems. Solution for these small QP issues is critically examined, in which a time-consuming numerical quadratic programming optimization as an inner circle was avoided. Necessity of storage for support vector model is linear in size for the training set, in which Sequential Minimal Optimization was allowed to handle large training sets. Because of the avoidance of matrix computation, for various test problems SMO scales somewhere lie between linear and quadratic in the training-set size. SVM evaluation represents SMO's approach is time consuming; hence as compared to linear SVMs, SMO are faster and uncovered data sets as compared to chunking algorithm, SMO can be thousand times faster. **Gevaert, et al. (2010) [6]** described an analysis that was completed on performance for classification of voice recognition. There are two standard Feed Forward neural networks arrangement that are used for performance estimate as classifiers and FFNN (Feed-forward Neural Network) is back propagation algorithm (BPNN) and Common methods Neural Networks. **Kaur, et al. (2014) [7]** presented that the emotion recognition from speech has established as a recent research area in Human–Computer Interaction. The objective of this paper is to use a BPNN classifier to classify four dissimilar emotions from 10 human user's speech Database. The main objective of this research work is to identify the users for which the database is trained and then identifying their emotions. Speech emotion detection refers to discovering the speech category based on the training & testing of the database provided.

III. PROCESS OF SPEECH EMOTION

Speech recognition process is basically done by the Speech Recognition System. In the speech recognition process, speech input signal is processed into recognition of speech as a text form. Speech Recognition Structure helps the technology to bring CPUs and humans more closely. There is basic terminology that one must know in order to implement or develop a Speech Recognition System.[1]

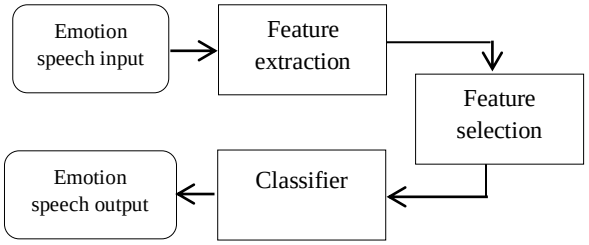

Figure 2: Processes of Speech Emotions [8]

- *Utterances*- Consumer input speech is called utterances, in simple words when consumer speaks something it is called utterances.
- *Pronunciations*- Single word has multiple meanings and multiple recognitions. It all depends on accent. A single word is uttered in dissimilar means in accordance to country, age etc.
- *Accuracy*- It is the performance measurement tool. It is measured by number of means , then in this case, if speaker utters "NO", then Speech Recognition Scheme must recognize it as a word "NO". If it is done precisely then correctness of the system is efficiently very good.

IV. SPEECH CATEGORIES

Speech recognition methods in dissimilar classes can be completed based-on the fact that type of utterance they have to identify.

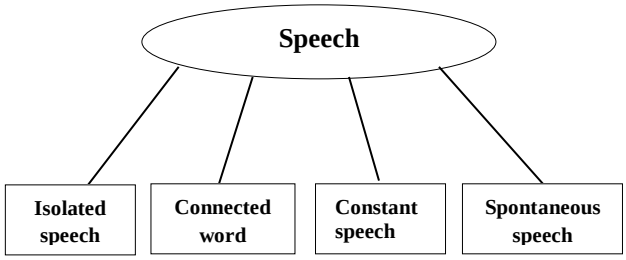

Figure 3: Types of speech

A. Insulated speech

Insulated phrase recognizer regularly set essential situations that every expression have small or no interference on both phases of model gap. It needs particular utterance at an interval of time. Frequently, this kind of voice has "Listen/Not-Listen states", where they require the speaker to have a break between words. Remote word may be a better term for this kind [9].

B. Connected word

It needs minimum pause among the utterances to make speech flow easily. They are almost alike in remote.

C. Constant speech

It is basically processor's dictation. It is usual human person speech, with no silent pauses between words. This kind of speech makes machine understanding more difficult.

D. Spontaneous speech

It can be attention of as speech that is natural sounding and no-tried out before hand. In this method with the spontaneous speech capability one should be able to manage a variety of usual speech properties such as words being run at the similar interval of time.

V. APPLICATION OF SPEECH RECOGNITION

Additional, to have a fine voice recognition technology, efficiency voice based applications, they mainly depends on many features, excluding:

- A user interface which creates the submission in normal use and robust to better models of dialogue the keep the discussion moving forward, even in similar the task to the technology.
- Some confusing situations that arise in human machine communication by speech tone.
- A time period of huge uncertainty on the part [13].
- Remind consumers what could be held at some point in the record.
- Reliability maintained across properties using vocabulary i.e., the almost available.
- Design Issues
- Give the capacity to barge in over pre-empts.
- The user implicit confirmation of speech input.

VI. SPEECH FEATURES OF EMOTION DETECTION

Still, though the sound is a particular channel, there are two kinds of properties that can be extracted and studied:

1) Para-linguistic features and

2) Linguistic features.

The para-linguistic features can be classified in prosodic, spectral and voice quality features.

1. Para-linguistic properties

a) Prosodic features

It analyzed the rhythms of speech and objectives on bigger segments of speech, like words, phonemes.There is a good suggestion among the prosodic features. [11] The pitch signal is produced by the program of the vocal cord. It has carried data about emotions of the strain on the vocal cords. There is a term pitch-period; it is the time between the beginnings of the verbal words. Sound power is also produced due to the pitch measure.

b) Spectral features

Some features have been generated of the airflow from the vocal words, as in the case of anger the airflow is very fast and in case of calm the airflow is very deliberate. So, it depends on the flow of air. The power option is used to consider the flow of air.

c) Voice quality features

It is a very solid network between voice-quality features and speech emotions. Speech feature based on the activity. Various types of speeches like high voice or pitch, slow speech and calm, all are based on the mood [12].

2. Linguistic Features

Sometimes deliver of phrases give the speech emotion quality. Most reliable used are diagrams and uni-grams.

VII. CLASSIFICATION IN EMOTIONAL STATE

The speech, emotional state classification has an important role in emotion recognition system using speech. The classification accuracy is based on dissimilar features extracted from the voice samples of dissimilar emotional state. The classifier is given by proper feature points to classify the speech emotions. In overview section we discuss the type of classifiers, out of which K Nearest Neighbor (KNN) and Gaussian mixture model (GMM) &

Support Vector machine (SVM) classifiers were used for emotion recognition.

1. K Nearest Neighbor (KNN) classifier

KNN is simplest and an influential method of classification of an emotional state, similar observations belong to similar classes is the key idea behind KNN classification. The NN (Nearest Neighbor) is the most custom methods in diverse supervised statistical pattern recognition methods. If error of cost is equal for individual classes, the valued class of not defined sample is selected to be the class that is most normally defined in the collection of its KNNs. The nearest neighbor technique based on the classification of only single nearest neighbor, considering the classification of an unknown sample on the "votes" of K nearest neighbor. In this, k is a constant value, and an unlabeled vector is used for classification transmission the label which is most common among the k-training samples nearest to that query value, in which the input consists of the k closest training examples in the future space. It includes Euclidean distance as the continuously variable as distance [15]. In the training dataset the effects of noisy points reduced by Larger K values where cross validation performs the choice of K. The classification of the samples of speech signal with the nearest training distance is calculated. It involves a training set of all cases. The classifier finds the KNN classifier to the unlabeled data from the training section based-on the chosen distance measure. Here we have considered six emotional states, namely anger, happiness, sadness, fear, disgust and Neutral [16]

2. Gaussian mixture model classifier

GMM is extensively used classifier for the task of speech emotion recognition and speaker identification. It is a model for probability density assessment using a convex arrangement of multiple normal densities. It is the parametric probability density function characterize as a weighted sum of Gaussian component densities. GMM is parameterized by the mean vectors, covariance matrices and mixture weights from all component densities. Gaussian Mixture Models are broadly used as probability distribution features, such as fundamental prosodic features and vocal-tract related spectral features in an emotion recognition system as well as in speaker recognition systems. It is estimated from the training data phase using the epoch's expectation maximization (EM) method and using a convex combination of multiple densities. In this model, probability density function of consider the data values by a multiple GMM density. After, the set of defined inputs is provided to the model, by using EMA calculates the weights of individual distributions. The computational of probabilities evaluated for defining test input designs only when a GMM model is faced. In these 6-different emotional states like anger, happy, sad, disgust and intensifier are considered [17][18][19].

3. Machine Learning Classifier (SVM)

It is a computer algorithm that studied by illustrations to assign the makes two objects. The Machine Learning is well-known in the design recognition communal and highly general due to their overview capabilities reached by structural risk minimization concerned with training. In various phases, it's evaluated is considerably better than that of competing methods. Nonlinear problems are explicated by a change of the input feature vectors into a usually higher-dimensional feature spaces by a mapping function, the Kernel function or hyper-plane. The highest discrimination is found by an optimal assignment of the maximum departure between the borders of two groups [20]. It can manage two-class issues, but a variety of phases, exist for multiple class discrimination. To construct an optimal hyper-plane, SVM employs an exchange the training algorithm, used to reduce the means error function at the training phase. A large number of kernels can be used in machine learning models, including linear, polynomial, radial basis function (RBF) and Activation function. We determined on an RBF and Polynomial kernel function, because both give auspicious consequences. Non-linear machine learning can be applied in an effective way through the kernel trick function that replaces the inside product calculated in linear vector machine by a kernel function. The basic idea behind the vector machine is to transforming the innovative input set to a high-dimensional feature space by using kernel method.

VIII. CONCLUSION

In conclusion, most current work is done in the area of voice emotion recognition is conversed. Mostly used approaches of unique features extraction and several classifier performance parameters are reviewed. Achievement of voice emotion recognition is reliant on appropriate feature extraction as well as proper classifier collection from the model emotional speech. It can be seen that addition of various features can give the improved recognition rate. Classifier performance parameter is needed to be increased for recognition of presenter independent systems. The presentation area of emotion appreciation from voice is increasing as it opens the new means of communication between human and machine learning. It is needed to model operative method of speech feature extraction so that it can even offer emotion recognition of real-time voice.

IX. REFERENCES

[1]. I. Patel, Dr. Y. S. Rao, (2010) "Speech Recognition Using Hmm With Mfcc-An Analysis Using Frequency Spectral Decomposion Technique", Signal & Image Processing: An International Journal (SIPIJ) Vol.1, No.2.

[2]. W. HAN, C. CHAN, C. CHOY, K. PUN, (2006), "An Efficient MFCC Extraction Method in Speech Recognition", IEEE.

[3]. Ali, MdAkkas, M. Hossain, M. N. Bhuiyan, (2013)"Automatic speech recognition technique for Bangla words." International Journal of Advanced Science and Technology.
[4]. Ververidis, Dimitrios, C. Kotropoulos, (2006) "Emotional speech recognition: Resources, features, and methods." Speech communication 48.9: 1162-1181.
[5]. Platt, C. John, (1999) "12 fast training of support vector machines using sequential minimal optimization." Advances in kernel methods: 185-208.
[6]. W. Gevaert, G. Tsenov, V. Mladenov, (2010) "Neural Networks used for Speech Recognition", JOURNAL OF AUTOMATIC CONTROL, UNIVERSITY OF BELGRADE 20: 1-7.
[7]. J. Kaur, A. Sharma, (2013) "SPEECH EMOTION-SPEAKER RECOGNITION USING MFCCAND NEURAL NETWORK.".
[8]. B. Schuller, G. Rigoll, M. Lang, (2004) "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine - belief network architecture," Proc. Internat. Conf. Acoust. Speech Signal Process. (ICASSP '04), vol. 1, pp. 577–580.
[9]. B. Schuller, G.Rigoll, M. Lang, (2004) "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture." Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on.Vol.1.IEEE.
[10]. HL J. Hansen, M. A. Clements, (1991), "Constrained iterative speech enhancement with application to speech recognition." IEEE Transactions on Signal Processing 39.4: 795-805.
[11]. Ververidis, Dimitrios, C. Kotropoulos, (2006) "Emotional speech recognition: Resources, features, and methods." Speech communication48.9: 1162-1181.
[12]. D. Krom, Guus, (1994) "Consistency and reliability of voice quality ratings for different types of speech fragments." Journal of Speech, Language, and Hearing Research 37.5: 985-1000.
[13]. S. K. Shevade, S. S. Keerthi, C. Bhattacharyya, K. R. K. Murthy, (2000) "Improvements to the SMO Algorithm for SVM Regression", IEEE Transactions on Neural Networks, Vol. 11, No. 5, September.
[14]. http://www.slideshare.net/LakshmiSarvani1/net-upload
[15]. N. Sugunai, Dr. K. Thanukodia (2010) "An improved k-nearest neighbor classification using genetic algorithm", IJCSI International Journal of Computer Science Issues, vol. 7, pp. 18-21, July.
[16]. R. O. Duda, P. E. Hart, D. G. Stork, (2001) Pattern Classification, 2nd ed, Wiley Interscience.
[17]. V. K. Govindan, A. P. Shivaprasad, (1990) "Character recognition –a review", Pattern Recognition, vol. 23, pp. 671-683.
[18]. S. Mori, C. Y. Suen, K. Yamamoto, (1992) "Historical review of OCR research and development", Proceedings of the IEEE, vol. 80, pp. 1029- 1058, July.
[19]. C. J. C. Burges, (1998) A Tutorial on Support Vector Machines for Pattern Recognition, Knowledge Discovery and Data-Mining, pp. 1-43.
[20]. M. Sanghamitra, H. N. D. Bebartta, (2011) "Performance comparison of svm and k-nn for oriya character recognition." International Journal of Advanced Computer Science and Applications, Special Issue on Image Processing and Analysis: 112-116.