SPEECH EMOTION RECOGNITION SYSTEM USING RNN

Amal Joseph K A, Rohan H S, Surya V P, Yathin N P, Dr C M Patil

Amal Joseph K A, Department of Electronics and Communication Engineering, Vidyavardhaka College of Engineering Mysuru, India
Rohan H S, Department of Electronics and Communication Engineering, Vidyavardhaka College of Engineering Mysuru, India
Surya V P, Department of Electronics and Communication Engineering, Vidyavardhaka College of Engineering Mysuru, India
Yathin N P, Department of Electronics and Communication Engineering, Vidyavardhaka College of Engineering Mysuru, India
Dr C M Patil, Department of Electronics and Communication Engineering, Vidyavardhaka College of Engineering Mysuru, India

---***---

Abstract: *In the past, people used speech as a mode of communication, or the way a listener is conveyed through voice and emotion. However, for speech recognition in the context of machine interaction, the idea of machine learning and several approaches are necessary. Speech has become an important part of speech development, with the voice serving as a bio- metric via usage and importance. In this post, we attempt to explore a variety of speech and emotion recognition approaches. We discussed and separated speaking technologies in this study by focusing on needs, databases, classifications, feature extraction, augmentation, segmentation, and the method of Speech Emotion detection. Human intellect, on the other hand, has constructed a concealed mystery in the form of speech. It eliminates any uncertainty in speaking and comprehending other persons. Furthermore, comprehending what the voice decodes has been the most important element of the communication process. Wemay address one other and communicate deeper feelings with the use of words.*

Key Words: **Speech Emotion Recognition, Speech Processing, Biometric, Machine Learning, MLP**

1. INTRODUCTION

1.1 An overview

Speech cues can be used to discern emotions. Emotional differences are influenced by speech characteristics. This domain has a number of datasets available. Apart from feature extraction, emotion categorization in speech is becoming increasingly significant. Various classifiers are used to categorize emotions such as sadness, neutral, happiness, surprise, fury, and so on. Emotion identification systems can been hanced by using a deep neural network-based autonomous emotion detection system.

Multiple ML methods were applied to improve the accuracy of identifying speech emotions in English. Some of the researched speech characteristics are linear prediction cepstral coefficients (LPCC), fundamental frequencies, and Mel-frequency cepstral coefficients (MFCC).

The use of voice signals for human-machine interaction is the quickest and most efficient way. Emotional detection is a challenging assignment for machines, but it is intuitive for people. Speech emotion recognition may be used to distinguish between male and female speakers.

There is what is known as accosting variability, which alters the characteristics of speech due to the existence of varied speaking rates, styles, phrases, and speakers. The same speech can indicate several emotions, and different sections of the same utterance can reflect multiple emotions. This further complicates matters. There are alsocultural and environmental variances. There are two kinds of emotions: fleeting and persistent. Recognized emotions may be speaker independent or speaker dependent. For classification, we have some alternatives in the form of K- nearest neighbours (KNN), Support vector machine (SVM), Convolution neural network (CNN), Deep neural network (DNN) etc.

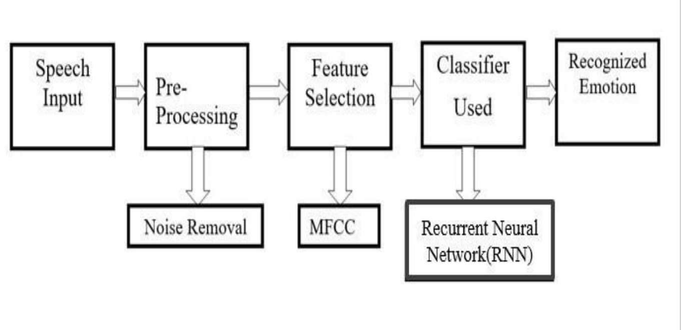

Fig 1. Block Diagram

The whole program would be built in the below given classification,

1.1.1 Speech Database

Various voice databases are used in this study to validate the offered methodologies in speech emotion recognition phrases. Anger, fear, neutrality, contempt, happiness, and sorrow are alldocumented emotions.

This helps us and the program to understand the difference between most the speech patterns of human voice and the underlying emotions in them. This is done byrecording the voice of the human and feeding it to the system, this in turn willbe further analysed be the program for further classification of its deferred notes and pitch format.

1.1.2 Feature Extraction (MFCC)

In recent years, a variety of parameters have been proposed to enhance human emotional screening findings, including Mel- frequency cepstral coefficients (MFCC), linear predictive cepstral coefficients (LPCC), and perceptual linear predictive coefficients (PLP). Recently, there has been considerable interest in separating speech using the filter bank and the auditory filter in order to extract variables from the long-term spectro-temporal- inspired signal. Although these features gave acceptable results, certain limits necessitated researchers developing their own set of features because most conventional features are focused on short-term analysis of non-stationary audio signals.

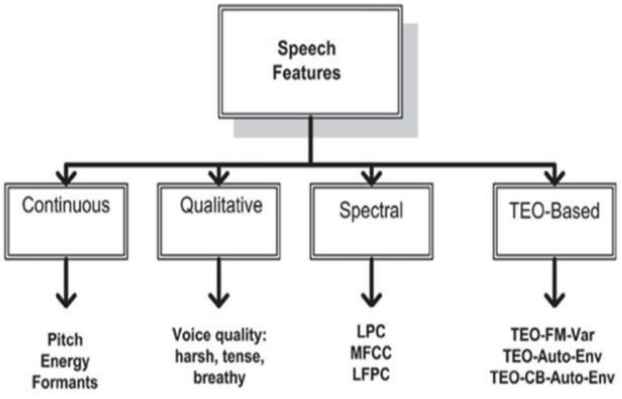

Fig 2. Feature categories

Based on this classification, the input voice will be completely desynchronized to a limit where the increase in any pitch or note would develop an understanding of the conclusion. And finally, this evident conclusion would be grouped together to gather a single pre-dominant conclusion. Which also changes and updates based on the further experimentations and trails.

1.1.3 Classification Approach (RNN)

Some of the classification algorithms used to build successful classifiers for collecting emotional states include support vector machine (SVM), Hidden Markov models (HMM), Deep Neural Network, K-nearest neighbour, and Gaussian mixed model(GMM). A basic level classifier, on the other hand, may strugglewith highly emotional situations. In comparison to a combination of SVM and radial basis function, a ranking SVM technique can result in significant improvements in emotion identification (RBF). Several hybrid/fusion-based systems outperform solo strategies in terms of identification rate.

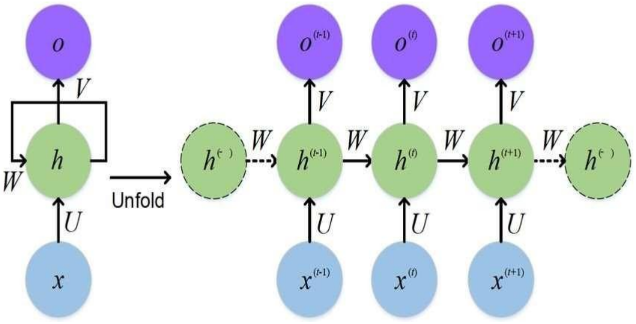

Fig 3. RNN as a classifier

2. Literature Review

In answer to the binary split issue, Cao et al suggested a typical SVM strategy for integrating information linked to emotional perception. This strategy teaches SVM algorithms for certain emotions, takes each input as a distinct query, then combines all results from partners to employ multiple class predictions. SVM measurement has two advantages. It starts with gathering speaker-specific data for the training and evaluation phases. Second, it acknowledges that multiple emotions might be produced by each speaker in order to decide the dominating emotion. The measuring techniques outperform the conventional SVM in the two public data sets of heart-to-heartspeech, Berlin and LDC. When both performance and default data are incorporated, standard-based SVM algorithms recognize emotional expression substantially better than standard SVM algorithms. The average accuracy, also known as the weighted average (UA), was 44.4 percent.

Chen et al. used a three-level emotion identification approach to improve speaker-independent speech emotion detection. After sorting emotional reactions from coarse to fine, the Fisher rate is utilized to choose the appropriate feature. The Fisher rate output is used as an input parameter in a multi-level SVM- based classifier. Four comparable trials are also identified using principal component analysis (PCA) and artificial neural networks (ANN). Fisher Plus SVM, PCA Plus SVM, Fisher + ANN, and PCA + ANN are the four comparative experiments. As a result, Fisher outperforms PCA in dimension reduction, although SVM outperforms ANN in emotion classification for speaker independence. In BHUDES, the recognition rate for three levels is 86.5 percent, 68.5 percent, and 50.2 percent.

Nwe and her colleagues suggested a novel method for detecting emotions in audio data. The system employs accurate HMM and short-term log energy coefficients to define speech and separation signals (LFPC). Before training and assessing the system with secret information,divide emotions into six stages. The suggested approach is compared to Mel-frequency Cepstral coefficients (MFCC) and direct prediction of Cepstral coefficients (LPCC) to assess its efficacy (LPCC). Statistically, the best category's accuracy was 78 and 96 percent, respectively. Furthermore, the findings demonstrate that LFPC is a far superior alternative to traditional emotional isolation.

Wu et al. developed a unique set of modulation spectral indicators for detecting emotions in human speech (MSFs). Tocomplete the speech, a suitable feature is discovered in the long-term spectro-temporal-inspired database by merging a modified filter bank and a hearing filter bank. These methods record acoustic frequency and momentarily changed frequency components in order to offer sensitive data with typical short- term properties that cannot be transferred. In the separation approach, SVM with radial base function (RBF) is applied. In Berlin and Vera am Mittag, MSFs are being probed (VAM). According to the findings, MSFs outperform the MFCC and perceptual linear prediction coefficients (PLPC). Recognition performance increases dramatically when MSFs are employedto add prosodic aspects. Furthermore, the class acquisitionaccuracy is 91.6 percent.

Rong et al. proposed using an ensemble random forest to trees (ERFTrees) technique with a high number of emotional identification variables without using any language or grammar, however the grammar issue remains unresolved. Thesuggested approach was evaluated on the Chinese emotional expression database and shown to enhance the degree of emotional recognition. ERFTrees goes beyond classic size reduction algorithms like PCA and multi-dimensional scaling (MDS), as well as the recently published ISO Map. The predictability of the 16-item women's data set was 82.54 percent, whereas the 84- factor performance was just 16 percent.

Wu et al. created a mix-based technique for detecting speech emotions by combining several class separators, acoustic- prosodic (AP) characteristics, and semantic labels (SLs). Following the extraction of AP features, three types of basic level separators are used: GMMs, SVMs, MLPs, and Meta decision trees (MDT). A high-end entropy model (MaxEnt) was applied in a semantic labelling approach. MaxEnt modelled the interaction between the rules of emotional organization (EARs) and emotional states in the identification of emotions. In the latter situation, combined data from SL and AS are utilized to create emotional awareness outcomes. According to secret data test findings, replies based on MDT archives are 80%, SL-based diagnostic archives are 80%, and a mix of AP and SL archives is 83.55 percent.

Narayanan has recommended using the contact center app's voice support suggestions to identify domain-specific emotions. The primary purpose of this research was to differentiate between negative and positive emotions. Emotional perception is employed in many situations, including auditory, lexical, and spoken data. Emotional theoryelements are also supplied in order to collect sensory data at the language level. Both k-NN and direct separator separators are used to work with distinct sorts of features. According toresearch findings, integrating acoustic and oral data yields the greatest outcomes. According to the study, using three sources instead of one boost the accuracy of the sections by 40.7 percent for men and 36.4 percent for women. Male accuracy varied from 1.4 percent to 6.75 percent in a prior study, while femaleaccuracy ranged from 0.75 percent to 3.96 percent.

Yang and Lugger presented a new set of harmonic notes to represent emotion in speech. These characteristics are based on music theory's psychoacoustic awareness. To begin, combine the spherical autocorrelation of the pitch histogram with the specified voice input pipeline. It determines whether two tone durations produce harmonic or inharmonic perception. The Bayesian class with the conditional condition of the Gaussian class is significant in the separation process. The recognition function has improved, according to test findings employing consensus characteristics from Berlin neuroscience. Theaverage visibility level increased by 2%.

Albornoz and colleagues investigate a distinct region of the spectrum that may be utilized to identify emotions and separategroups. This function employs sensory features as well as a distinct program area to categorize emotions. A few classdividers, including as HMM, GMM, and MLP, have experimented with different configurations and input data to develop a hierarchical approach for a unique emotional separation. The recommended approach differs in two ways: first, it selects the most efficient features; second, it exploits the best performance of the category editor in all areas, including the separator. According to the results, the text editing approach outperforms the standard separator in the Berlin database utilizing decuple validation. The normal HMM technique, for example, earned 68.57 percent, but whereas in the position model received 71.75 percent.

For visualizing emotions, Lee et al. propose a hierarchical calculating approach. This approach translates input audio signals into associated emotion categories by using many layers of binary categories. The primary idea behind the tree's numerous levels is to perform the sorting operation

as quickly as feasible while limiting mistake spread. AIBO and USC IEMOCAP data sets are used to evaluate the partition algorithm. The real result improves accuracy by 72.44 percent- 89.58 percent when compared to the SVM basis. The results demonstrate that the suggested textual sequencing technique is successful in categorizing emotional expression into various databases.

3. CONCLUSION

Speech Emotion Recognition (SER) is the endeavor to detect human emotions and affective responses from speech. This takes use of the fact that tone and pitch of the voice frequently reflect underlying emotion. This acts as tool for the machine tounderstand human interaction and respond to user based on the emotion towards next generation machine learning model for user interaction with machines.

The results obtained would be useful in understanding theunderlying idea about the human speech and computed digital signals. Due to which the program can be upgraded to higher versions and other development criteria.

Based on the classification to be used, the input voice will be completely desynchronized to a limit where the increase in any pitch or note would develop an understanding of the conclusion. And finally this evident conclusion would be grouped together to gather a single pre-dominant conclusion. Which also changes and updates based on the further experimentations and trails.

The use of MFCC as the feature extraction, has given us certain ideas of how a SER works in its under-coverings. Therefore, we use an MFCC as an feature extraction system. This is due the higher extent of usage in major industries and the possibility of higher accuracies in Machine Learning and data science.

As the classifiers, the RNN system brings upon a coupled reaction towards the parallel use of the MFCC. Therefore, we ought to use an RNN classifier. The RNN has the capacity to link nodes in a directed or undirected network in response to atemporal sequence of data inputs.

ACKNOWLEDGEMENT

We would like to convey our wholehearted thanks of gratitudeto our guide and HOD, Dr. C M Patil, and our esteemed principal Dr. B Sadashive Gowda, under whose inspiration and the opportunity they had provided us to conduct this project 'SPEECH EMOTION RECOGNITION SYSTEM USING RNN' .Through this opportunity to perform this project, we are thankful and embed our gratitude in Prof. Audrey Arlene A, who had also helped us to complete this project.

REFERENCES

[1] H. Cao, R. Verma, and A. Nenkova, "Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech," Comput. Speech Lang., vol. 28, no. 1, pp. 186–202, Jan. 2015.

[2] L. Chen, X. Mao, Y. Xue, and L. L. Cheng, "Speech emotion recognition: Features and classification models," Digit. Signal Process., vol. 22, no. 6, pp. 1154–1160, Dec. 2012.

[3] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden Markov models," Speech Commun., vol. 41, no. 4, pp. 603–623, Nov. 2003.

[4] S. Wu, T. H. Falk, and W.-Y. Chan, "Automatic speech emotion recognition using modulation spectral features," Speech Commun., vol. 53, no. 5, pp. 768–785, May 2011.

[5] J. Rong, G. Li, and Y.-P. P. Chen, "Acoustic feature selection for automatic emotion recognition from speech," Inf. Process. Manag., vol. 45, no. 3, pp. 315–328, May 2009.

[6] J. Rong, G. Li, and Y.-P. P. Chen, "Acoustic feature selection for automatic emotion recognition from speech," Inf. Process. Manag., vol. 45, no. 3, pp. 315–328, May 2009.

[7] S. S. Narayanan, "Toward detecting emotions in spoken dialogs," IEEE Trans. Speech Audio Process.,vol. 13, no. 2, pp. 293–303, Mar. 2005.

[8] B. Yang and M. Lugger, "Emotion recognition from speech signals using new harmony features," Signal Processing, vol. 90, no. 5, pp. 1415–1423, May 2010

[9] E. M. Albornoz, D. H. Milone, and H. L. Rufiner, "Spoken emotion recognition using hierarchical classifiers," Comput. Speech Lang., vol. 25, no. 3, pp. 556–570, Jul. 2011.

[10] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," Speech Commun., vol. 53, no. 9–10, pp. 1162–1171, Nov. 2011.

BIOGRAPHIES

Amal Joseph K A
Department of Electronics and Communication Engineering
Vidyavardhaka College of Engineering, Mysuru, India

Rohan H S
Department of Electronics and Communication Engineering
Vidyavardhaka College of Engineering, Mysuru, India

Surya V P
Department of Electronics and Communication Engineering
Vidyavardhaka College of Engineering, Mysuru, India

Yathin N P
Department of Electronics and Communication Engineering
Vidyavardhaka College of Engineering, Mysuru, India

Dr C M Patil
Department of Electronics and Communication Engineering
Vidyavardhaka College of Engineering, Mysuru, India