GMM and LDA based Speech recognition using Sonogram

Dr. R. Thiruvengatanadhan
Assistant Professor
Department of Computer Science and Engineering
Annamlai University, Annamalainagar, Tamilnadu, India
Email Id: thiruvengatanadhan01@gmail.com

Abstract: *Automatic recognition of speech using computers is a challenging issue. This paper describes a techniques that uses Gaussian mixture model (GMM) and Linear Discriminate Analysis (LDA) to recognized speech based on features using Sonogram. Modeling techniques such as GMM and LDA were used to model each individual word which is trained to the system Each separated word Segment utilizing Voice Activity Detection (VAD) from the test sentence is coordinated against these models for finding the semantic portrayal of the test input discourse. Experimental results of GMM and LDA shows good performance in recognized rate.*

Keywords: **Feature Extraction, Voice Activity Detection (VAD), Sonogram, uses Gaussian mixture model (GMM) and Linear Discriminate Analysis (LDA)**

1. INTRODUCTION

An audio signal represents the sound as an electrical voltage. Signal stream is only a course taken by a sound sign for going towards the speaker from the source. Sound sign is described by transfer speed, force and voltage. Impedance of the sign way decides the connection among force and voltage [1]. Electrical sign is utilized by simple processors however computerized signals are numerically bargains by the advanced processors. Because of capacity requirements, research identified with discourse ordering and recovery has gotten a lot of consideration [2]. As capacity has become less expensive, huge assortment of spoken reports is accessible on the web, however there is an absence of satisfactory innovation to clarify them. Manual record of discourse is exorbitant and furthermore has security imperatives [3]. Subsequently, the need to investigate programmed ways to deal with look and recover spoken archives has expanded. Besides, a wide assortment of sight and sound information is accessible on the web and makes ready for advancement of new advances to file and look through such media [4]. Discourse acknowledgment is a primary center of communicated in language frameworks.

Proposed work expects to build up a framework which needs to change over verbally expressed word into text utilizing AANN displaying procedure utilizing acoustic element specifically Sonogram. In this work the transient encompass through RMS energy of the sign is determined for isolating individual words out of the consistent discourses utilizing voice action location strategy. Highlights for each separated word are removed and those models were prepared. SVM and GMM modeling techniques is used to model each individual utterance. Thus each isolated word segment from the test sentence is matched against these models for finding the semantic representation of the test input dialogue.

2. VOICE ACTIVITY DETECTION

Voice Activity Detection (VAD) is a procedure for finding voiced portions in discourse and assumes a significant job in discourse mining applications [5]. VAD disregards the extra sign data around the word viable. It tends to be likewise seen as a speaker autonomous word acknowledgment issue. The essential rule of a VAD calculation is that it removes acoustic highlights from the info sign and afterward contrasts these qualities and edges for the most part extricated from quietness. Voice movement is pronounced if the deliberate qualities surpass the edge. Something else, no discourse movement is available [6].

VAD discovers its utilization in an assortment of discourse correspondence frameworks like coding of discourse, perceiving discourse, hands free communication, sound conferencing, discourse improvement and retraction of sound [7]. It distinguishes where the discourse is voiced, unvoiced or maintained and gains smooth ground of the discourse cycle [8]. A framesize of 20 ms, with a cover of half, is considered for VAD. RMS is separated for each casing. Figure 1 shows the detached word partition.

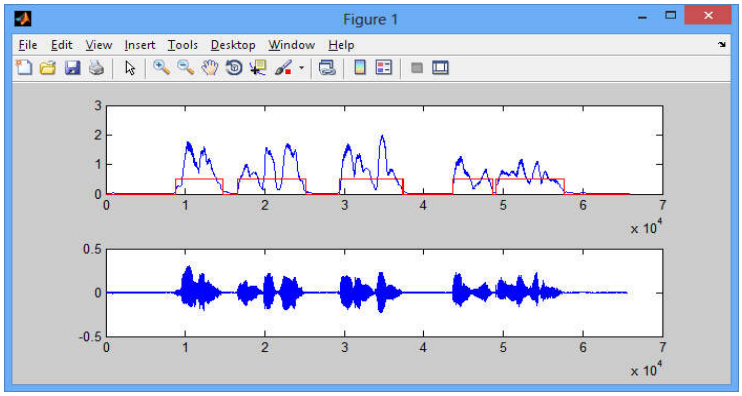

Figure 1. Isolated Word Separations.

3. SONOGRAM

Pre-accentuation is performed for the discourse signal followed by outline hindering and windowing. The discourse fragment is then changed utilizing FFT into spectrogram portrayal [9]. Bark scale is applied and recurrence groups are gathered into 24 basic groups. Ghostly covering impact is accomplished utilizing spreading capacity. The range energy esteems are changed into decibel scale [10]. Equivalent commotion shape is consolidated to ascertain the uproar level. The tumult sensation per basic band is registered. STFT is figured for each portion of pre-prepared discourse. An edge size of 20 ms is sent with half cover between the edges. The inspecting recurrence of 1 second term is 16 kHz. The square graph of sonogram extraction is appeared in Figure 2.

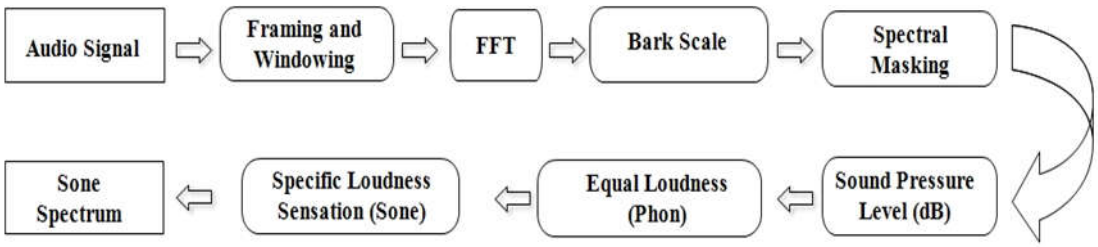

Figure 2. Sonogram Feature Extractions.

A perceptual scale known as bark scale is applied to the spectrogram and it bunches the frequencies dependent on the discerning pitch areas to basic groups. The impediment of one sound to another is demonstrated by applying a phantom concealing spread capacity to the sign [11]. The range energy esteems are then changed into decibel scale. Telephone scale calculation includes equivalent din bend which speaks to various view of clamor at various frequencies individually. The qualities are then changed into a sone-scale to mirror the commotion impression of the human hearable framework [12].

4. GAUSSIAN MIXTURE MODEL (GMM)

Parametric or non-parametric methods are used to model the distribution of feature vectors. Parametric models are based on the shape of probability density function [13]. In non-parametric displaying just insignificant or no suspicion with respect to the likelihood thickness capacity of highlight vector is made [14]. The Gaussian combination model (GMM) is utilized in arranging distinctive sound classes. The Gaussian classifier is an illustration of a parametric classifier. It is an instinctive methodology when the model comprises of a few Gaussian segments, which can be believed to demonstrate acoustic highlights. In arrangement, each class is spoken to by a GMM and alludes to its model. When the GMM is prepared, it tends to be utilized to foresee which class another example presumably has a place with [15].

The likelihood conveyance of highlight vectors is displayed by parametric or non-parametric techniques. Models which expect the state of likelihood thickness work are named parametric. In non-parametric displaying, insignificant or no suppositions are made with respect to the likelihood dissemination of highlight vectors. The capability of Gaussian combination models to speak to a hidden arrangement of acoustic classes by individual Gaussian parts, in which the unearthly state of the acoustic class is defined by the mean vector and the covariance grid, is critical.

Additionally, these models can frame a smooth estimate to the subjectively molded perception densities without other data [16]. With Gaussian blend models, each solid is demonstrated as a combination of a few Gaussian bunches in the element space. The reason for utilizing GMM is that the appropriation of highlight vectors extricated from a class can be displayed by a combination of Gaussian densities. The inspiration for utilizing Gaussian densities as the portrayal of sound highlights is the capability of GMMs to speak to a basic arrangement of acoustic classes by individual Gaussian segments in which the ghastly state of the acoustic class is defined by the mean vector and the covariance matrix[17]. Likewise, GMMs can frame a smooth estimation to the subjectively molded perception densities without other data. With GMMs, each solid is demonstrated as a combination of a few Gaussian groups in the component space [18].

5. LINEAR DISCRIMINATIVE ANALYSIS (LDA)

LDA classifies a dataset based on the relation between the dispersions within classes and between classes, in order to find the dimension that best classifies a dataset in a linear way [19]. The number of preparing tests in class, the quantity of unmistakable classes, the mean vector of tests having a place with class and speaks to the arrangement of tests having a place with class. The objective of the LDA is to acquire the grid that augments the connection between classes. The LDA gives the best dimension that describes a dataset dispersion given its features and therefore, is able to analyze the data by reducing its dimensionality [20]. LDA classifier was chosen to be utilized because of its elite in characterization, along with its power in long haul use and its low computational expense [21].

6. EXPERIMENTAL RESULTS

6.1 Dataset Collection

Experiments for ordering discourse sound utilizing Television broadcast discourse information gathered from Tamil news stations utilizing a tuner card. A complete dataset

of 100 distinctive discourse exchange cuts, going from 5 to 10 seconds term, tested at 16 kHz and encoded by 16-bit is recorded. Voice movement location is performed to confine the words in every discourse record utilizing RMS energy envelope. For every discourse record, an information base of the secluded words is acquired utilizing VAD.

6.2 Feature Extraction

VAD the disengaged words are removed from the sentences. In this manner traces which are unvoiced excitations are killed by thresholding the segment size. Feature DWT are isolated from each edge of size 320 window with a front of 120 models. During getting ready measure each segregated word is secluded into 20ms covering windows for isolating 6 DWT features.

6.3 Classification

Using VAD isolated words in a speech is separated. GMM and LDA are made for each segregated word. For preparing, disengaged words from were thought of. The preparation cycle dissects discourse preparing information to locate an ideal method to group discourse outlines into their individual classes. For testing 22 dimensional Sonogram highlight vectors were given as information. . Figure 3 shows the exhibition of GMM for discourse and music arrangement dependent on the quantity of combinations.

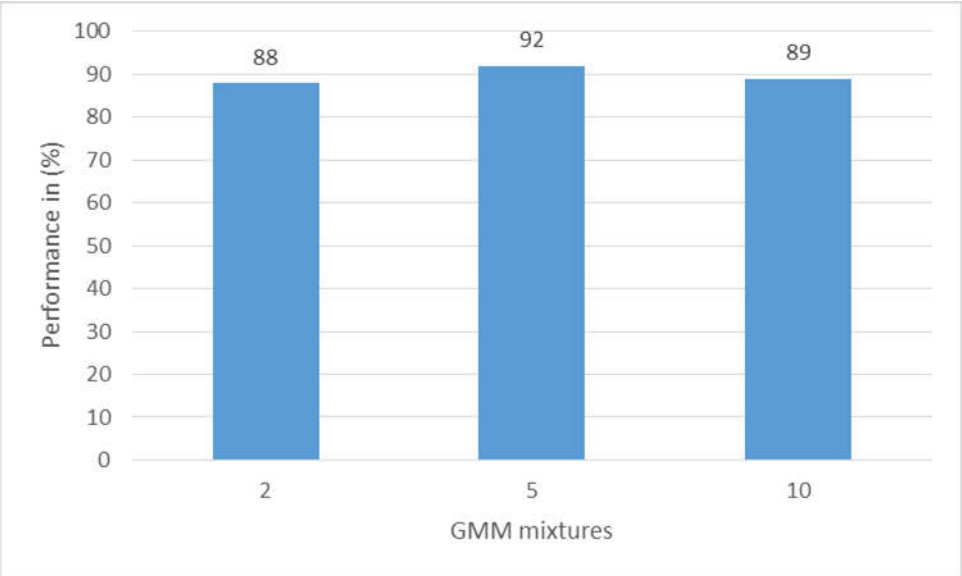

Figure 3 The performance of GMM for speech and music classification based on the number of mixtures.

Table 1 shows the performance of Speech recognition using GMM and LDA for different duration respectively.

Table 1 The performance of Speech recognition using GMM and LDA for different duration respectively.

	Speech Recognition Rate
GMM	92%
LDA	88%

7. CONCLUSIONS

In this paper, Voice Activity Detection (VAD) is utilized for isolating individual words out of the persistent addresses. Highlights for each disengaged word are removed and those models were prepared effectively. Sonogram is determined as highlights to portray sound substance. GMM and LDA are utilized to models every Individual expression. Sonogram is determined as highlights to describe sound substance. GMM and LDA learning calculations has been utilized for the perceived discourse by gaining from preparing information. Experimental results show that the proposed audio GMM learning method has good performance in 92% speech recognized rate compared with LDA.

REFERENCES

[1] *Reda Elbarougy. Speech Emotion Recognition based on Voiced Emotion Unit. International Journal of Computer Applications 178(47):22-28, September 2019*

[2] *Ayushi Y Vadwala, Krina A Suthar, Yesha A Karmakar and Nirali Pandya. Survey paper on Different Speech Recognition Algorithm: Challenges and Techniques. International Journal of Computer Applications 175(1):31-36, October 2017.*

[3] *Iswarya, P. and Radha, V, "Speech and Text Query Based Tamil - English Cross Language Information Retrieval system," International Conference on Computer Communication and Informatics, pp. 1-4, Coimbatore, 2014.*

[4] *Chien-Lin Huang, Chiori Hori and Hideki Kashioka, "Semantic Inference Based on Neural Probabilistic Language Modeling for Speech Indexing," IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 8480-8484, 2013.*

[5] *Ivan Markovi, Sre´ckoJuri´ Kavelj and Ivan Petrovi, "Partial Mutual Information Based Input Variable Selection for Supervised Learning Approaches to Voice Activity Detection," Applied Soft Computing Elsevier, vol. 13, pp. 4383-4391, 2013.*

[6] *Khoubrouy, S. A. and Panahi, I.M.S., "Voice Activation Detection using Teager-Kaiser Energy Measure," International Symposium on Image and Signal Processing and Analysis, pp. 388-392, 2013.*

[7] *Saleh Khawatreh, Belal Ayyoub, Ashraf Abu-Ein and Ziad Alqadi. A Novel Methodology to Extract Voice Signal Features. International Journal of Computer Applications 179(9):40-43, January 2018.*

[8] *Tayseer M F Taha and Amir Hussain. A Survey on Techniques for Enhancing Speech. International Journal of Computer Applications 179(17):1-14, February 2018.*

[9] *Xiaowen Cheng, Jarod V. Hart, and James S. Walker, "Time-frequency Analysis of Musical Rhythm," Notices of AMS, vol. 56, no. 3, 2008.*

[10] *Ausgef¨uhrt, Evaluation of New Audio Features and Their Utilization in Novel Music Retrieval Applications, Master's thesis, Vienna University of Technology, December 2006.*

[11] *Eberhard Zwicker and Hugo Fastl, "Psychoacoustics-Facts and Models," Springer Series of Information Sciences, Berlin, 1999.*

[12] *M. R. Schroder, B. S. Atal, and J. L. Hall, "Optimizing Digital Speech Coders by Exploiting Masking Properties of the Human Ear," Journal of the Acoustical Society of America, vol. 66, pp. 1647-1652, 1979.*

[13] *Tang, H., Chu, S. M., Hasegawa-Johnson, M. and Huang, T. S., "Partially Supervised Speaker Clustering", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 5, pp. 959-971, 2012.*

[14] *Chunhui Wang, Qianqian Zhu, Zhenyu Shan, Yingjie Xia and Yuncai Liu, "Fusing Heterogeneous Traffic Data by Kalman Filters and Gaussian Mixture Models," IEEE International Conference on Intelligent Transportation Systems, pp. 276-281, 2014.*

[15] *Sourabh Ravindran, Kristopher Schlemmer, and David V. Anderson, "A physiologically inspired method for audio classification," Journal on Applied Signal Processing, vol. 9, pp. 1374–1381, 2005.*

[16] *Menaka Rajapakse and Lonce Wyse, "Generic audio classification using a hybrid model based on GMMs and HMMs," in IEEE Int'l Conf. Multimedia Modeling, February 2005, pp. 1550–1555.*

[17] *Poonam Sharma and Anjali Garg. Feature Extraction and Recognition of Hindi Spoken Words using Neural Networks. International Journal of Computer Applications 142(7):12-17, May 2016.*

[18] *Sujay G Kakodkar and Samarth Borkar. Speech Emotion Recognition of Sanskrit Language using Machine Learning. International Journal of Computer Applications 179(51):23-28, June 2018*

[19] *Iswarya, P. and Radha, V, "Speech and Text Query Based Tamil -English Cross Language Information Retrieval system," International Conference on Computer Communication and Informatics, pp. 1-4, Coimbatore, 2014.*

[20] *K R. M. Aarts and R. T. Dekkers, "A real-timespeech-music discriminator," J. Audio Engi-neering Society, vol. 47, no. 9, pp. 720–725,September 1999.*

[21] *K. Englehart, B. Hudgin, and P. A. Parker, "A wavelet-based continuous classification scheme for multifunction myoelectric control," IEEE Transactions on Biomedical Engineering, vol. 48, no. 3, pp. 302–311, 2001.*