SVM and GMM based Speech recognition using DWT

Dr. R. Thiruvengatanadhan
Assistant Professor
Department of Computer Science and Engineering
Annamlai University, Annamalainagar, Tamilnadu, India

Abstract- Automatic recognition of speech using computers is a challenging issue. This paper describes a techniques that uses Gaussian mixture model (GMM) and Support Vector Machine (SVM) to recognized speech based on features using Discrete Wavelet Transform (DWT). Demonstrating strategies, for example, GMM and SVM were utilized to show every individual word which is prepared to the framework. 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 SVM shows good performance in recognized rate.

Keywords – Feature Extraction; Voice Activity Detection (VAD); Discrete Wavelet Transform (DWT), Gaussian mixture model (GMM) and support vector machines (SVM)

I. 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.

II. 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.

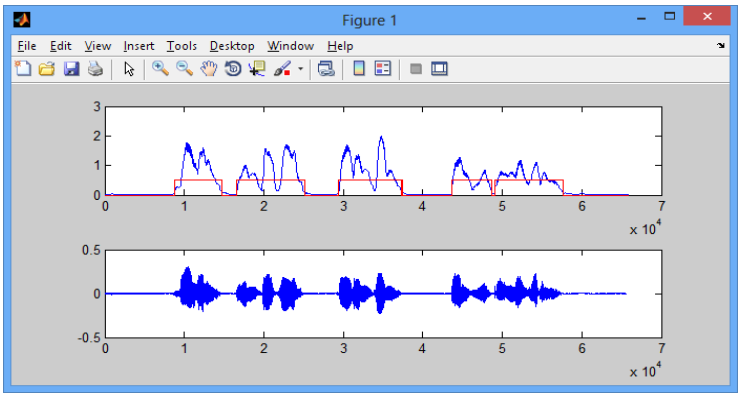

Figure 1. Isolated Word Separations.

III. Discrete Wavelet Transform (DWT)

Discrete Wavelet Transform (DWT) is a time scale representation of the audio signal which is computed using digital filtering techniques. DWT depends on subband coding and it diminishes the computational time needed to yield the wavelet change. In the mid 1970's new methods were acquainted with deteriorate discrete time signals. Following this, gigantic work was done in 1980's in the zone of discourse signal coding utilizing subband coding and pyramidal coding. After numerous extemporizations the coding plans prompted multi-goal examination. DWT include extraction utilizes channels with shifting cut-off frequencies at various scales. A breaking down wavelet called mother wavelet executes a model capacity in its wavelet investigation measure [9].

The first sign can be created utilizing a straight blend of wavelet work and fitting information activities with the wavelet coefficients. The decision of the quantity of wavelet coefficients relies on the application and the excess coefficient under a specific edge should be shortened. The wavelet scale is registered utilizing subsampling activities specifically here and there testing [10]. The detailed information in the signal is known as the resolution of the signal and is computed using filtering operations. Figure 2 shows two level wavelet decomposition techniques.

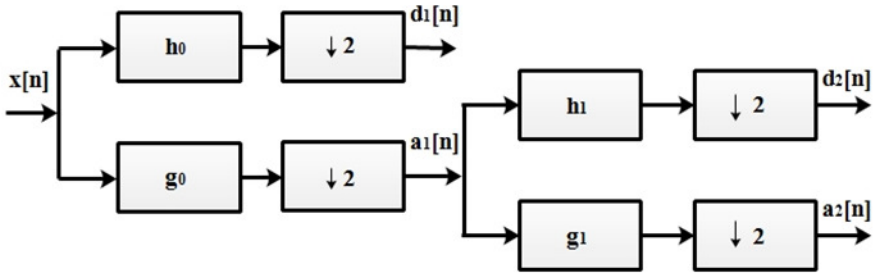

Figure 2. Two Level Wavelet Decomposition Techniques.

IV. Support Vector Machine

A machine learning strategy which depends on the standard of structure hazard minimization is uphold vector machines. It has various applications in the territory of example acknowledgment [11]. SVM constructs linear model based upon support vectors in order to estimate decision function. If the training data are linearly separable, then

SVM finds the optimal hyper plane that separates the data without error [12]. Figure 3 shows an illustration of a non-straight planning of SVM to develop an ideal hyper plane of detachment. SVM maps the information designs through a non-straight planning into higher measurement include space. For straightly detachable information, a direct SVM is utilized to arrange the informational indexes [13]. The examples lying on the edges which are boosted are the help vectors.

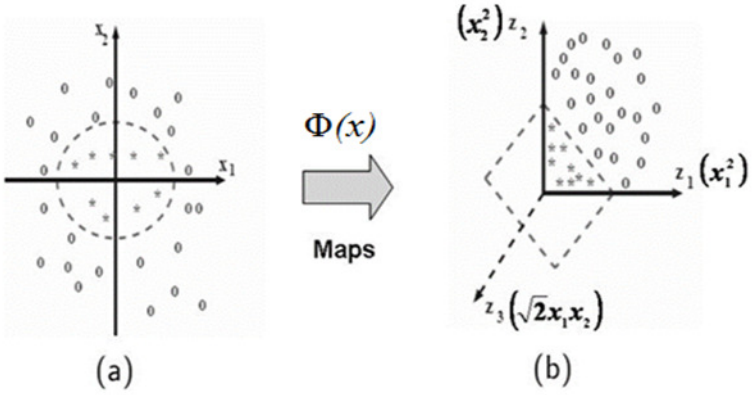

Figure 3. Example for SVM Kernel Function Φ(x) Maps 2-Dimensional Input Space to Higher 3-Dimensional Feature Space. (a) Nonlinear Problem. (b) Linear Problem.

The help vectors are the preparing designs and are similarly near hyperplane of partition. The help vectors are the preparation tests that characterize the ideal hyperplane and are the most troublesome examples to group [14]. Casually, they are the examples generally instructive of the characterization task. The portion work creates the internal items to develop machines with various kinds of non-straight choice surfaces in the information space [15].

V. Gaussian Mixture Model

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 [16]. In non-parametric displaying just insignificant or no suspicion with respect to the likelihood thickness capacity of highlight vector is made [17]. 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 [18].

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 [19]. 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[20]. 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 [21].

VI. 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 disconnected words are extricated from the sentences. In this way outlines which are unvoiced excitations are eliminated by thresholding the section size. Highlight DWT are separated from each edge of size 320 window with a cover of 120 examples. During preparing measure each detached word is isolated into 20ms covering windows for separating 6 DWT highlights.

6.3 Classification

Utilizing VAD segregated words in a discourse is isolated. SVM and GMM are made for each disengaged word. For preparing, confined words from were thought of. The preparation cycle examines discourse preparing information to locate an ideal method to order discourse outlines into their individual classes. For testing 6 dimensional DWT include vectors were given as information.

N-SVMs are made for each segregated word. For preparing, secluded words from were thought of. The preparation cycle examines discourse preparing information to locate an ideal method to group discourse outlines into their individual classes. The inferred uphold vectors are utilized to arrange discourse information. For testing 6 dimensional DWT highlight vectors were given as contribution to SVM model and the distance between every one of the element vectors and the SVM hyperplane is gotten. The normal distance is determined for each model. The content relating to the question discourse is chosen dependent on the greatest distance. A similar cycle is rehashed for various question discourse, and the exhibition is contemplated. The exhibitions of discourse acknowledgment for various SVM portions are looked at for DWT acoustic highlights are appeared in Table 1.

Table - 1 Performance of speech recognition rate in different SVM kernel function.

SVM Kernels	Speech Recognition Rate
Polynomial	**84%**
Gaussian	**86%**
Sigmoidal	**82%**

Gaussian blends for the speec are demonstrated for the highlights extricated. For characterization the element vectors are separated and every one of the element vectors is given as contribution to the GMM model. The dispersion of the acoustic highlights is caught utilizing GMM. We have picked a combination of 2, 5, 10 blend models. The class to which the sound example has a place is chosen dependent on the most elevated yield.

The conveyance of the acoustic highlights is caught utilizing GMM. The class to which the discourse test has a place is chosen dependent on the most elevated yield. Figure 4 shows the exhibition of GMM for discourse and music arrangement dependent on the quantity of combinations.

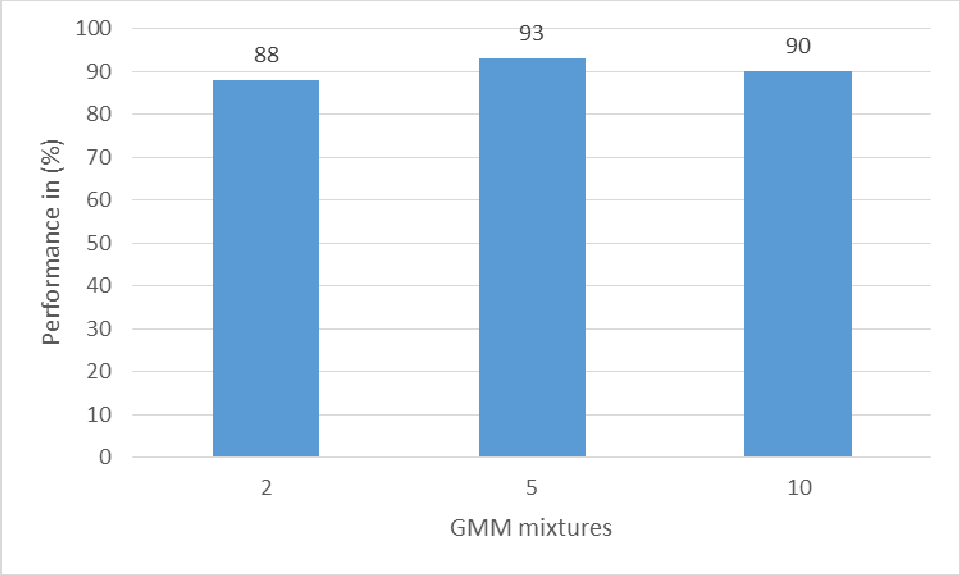

Figure 4. The performance of GMM for speech and music classification based on the number of mixtures.

VII. CONCLUSIONS

In this paper, Voice Activity Detection (VAD) is utilized for isolating individual words out of the persistent talks. Highlights for each separated word are removed and those models were prepared effectively. DWT is determined as highlights to describe sound substance. SVM and GMM is used to model each Individual utterance. DWT is calculated as features to characterize audio content. Experimental results show that the proposed audio GMM learning method has good performance in 93% speech recognized rate compared with SVM.

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] Rekik, S., D. Guerchi, H. Hamam and S.A. Selouani, "Audio Steganography Coding using the Discrete Wavelet Transforms," International Journal of Computer Science Security, vol. 6, pp. 79-83, 2012.

[10] Patil, V. D. and S. D. Ruikar, "Wavelet-Based Image Enhancement using Nonlinear Anisotropic Diffusion," International Journal of Advance Research Computer Science Software Engineering, vol. 2, pp. 158-162, 2012.

[11] Chungsoo Lim Mokpo, Yeon-Woo Lee, and Joon-Hyuk Chang, "New Techniques for Improving the practicality of a SVM-Based Speech/Music Classifier," IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 1657-1660, 2012.

[12] Hongchen Jiang, JunmeiBai, Shuwu Zhang, and Bo Xu, "SVM-Based Audio Scene Classification," IEEE International Conference Natural Language Processing and Knowledge Engineering, Wuhan, China, pp. 131-136, October 2005.

[13] Lim and Chang, "Enhancing Support Vector Machine-Based Speech/Music Classification using Conditional Maximum a Posteriori Criterion," Signal Processing, IET, vol. 6, no. 4, pp. 335-340, 2012.

[14] Md. Al Mehedi Hasan and Shamim Ahmad. predSucc-Site: Lysine Succinylation Sites Prediction in Proteins by using Support Vector Machine and Resolving Data Imbalance Issue. International Journal of Computer Applications 182(15):8-13, September 2018.

[15] Hend Ab. ELLaban, A A Ewees and Elsaeed E AbdElrazek. A Real-Time System for Facial Expression Recognition using Support Vector Machines and k-Nearest Neighbor Classifier. International Journal of Computer Applications 159(8):23-29, February 2017.

[16] 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.

[17] 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.

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

[19] 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.

[20] 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.

[21] 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