International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Importance of Artificial Intelligence in Neural Network: Speech Signal Segmentation Using K-Means Clustering with Kernelized Deep Belief Networks

V.Kakulapati*[1] , Gagandeep Singh Gill[2] , Chandramma R[3] , Sambhrant Srivastava[4] , Dr. Meenakshi Sharma[5] , Vijay Kumar [6]

Submitted: 03/11/2022 **Accepted:** 03/02/2023

Abstract: There has been a tonne of study on use of ML for speech processing applications, particularly voice recognition, over the past few decades. This study suggested an innovative method in speech signal processing and segmentation relied upon deep learning configurations. As input, this speech signal has been accumulated from the crime scene and this signal has been pre-processed using using K-means clustering (K-means C)for cluster the fragments of the input speech signal and process them for noise removal and signal artifacts removal. Here the segmentation is carried out for processed signal using Kernel based deep belief networks (KDBN). Experimental results demonstrate that proposed method outperforms the input speech signal based on both weighted accuracy (WA) and unweighted accuracy (UA).

Keywords: *speech processing, segmentation, deep learning, K-means C,KDBN*

1. Introduction

Speech can be segmented into phonemes manually by phonetic specialists, but this is known to be tedious, expensive and subjective. The use of accurate and reliable automatic segmentation algorithms is a desirable alternative. Since the location of the phoneme boundary is not taken into account when estimating parameters, the commonly used HMM-based forced alignment is not optimal for speech synthesis [3] and frequently necessitates human correction after the forced alignment. When only the phoneme sequences and not their boundary locations are given, accurate phonetic segmentation becomes a challenge. Syllable boundaries can be determined for syllable-timed languages using signal processing cues that are independent of the speaker [4]. When the settings are chosen so that the boundaries are overstated, signal processing cues cause false alarms but rarely cause deletions. Speech data has been segmented in TTS systems for syllable-timed Indian languages using signal processing cues and HMM-based alignment [5]. Typically, a GMM determines the posterior probability of how well an HMM state fits a frame. Deep neural networks and convolutional deep neural networks, which have found success in recent times in automatic speech recognition (ASR), perform better in acoustic modelling than generalised linear models (GMMs) because they can handle highly non-linear relationships between the input and output. Despite being widely employed in speech recognition, neural networks are not used for voice segmentation for TTS [6].

The contribution of this paper is given below:

- To collect the speech signal from crime scene for forensic identification
- To pre-process the speech signal using K-means clustering
- To segment processed signal using Kernel based deep belief networks

[1] *Sreenidhi institute of science and Technology,Yamnampet, Ghatkesar, Hyderabad-501301. Telangana, India.*
vldms@yahoo.com,
orchid id :0000 0002 1753 3298
[2] *Department of Instrumentation, Kurukshetra University. Kurukshetra.*
gsgill@kuk.ac.in
Orchid id : 0000-0002-7109-0379
[3]*Assistant Professor, Department of Computer Science and Engineering, Jain (Deemed-to-be University), Bangalore, India*
r.chandramma@jainuniversity.ac.in
Orchid id : 0000-0002-2427-0246
[4]*Assistant Professor, Mechanical Engineering department, Rajkiya Engineering College Azamgarh, Uttar Pradesh, India.*
sambhrantsrivastava4@gmail.com
Orchid id : 0000-0002-5868-1561
[5]*Department of Education, Sanskriti University, Mathura, Uttar Pradesh, India.*
osd@sanskriti.edu.in
Orchid id : 0000-0002-7596-1784
[6]*Assistant Professor, Mechanical Engineering Department, Rajkiya Engineering College, Azamgarh, Uttar Pradesh, India.*
vijayakgarg@gmail.com
Orchid id : 0000-0002-1107-6946

2. Related Works:

The feature maps of voice sounds may be obtained using deep learning methods containing convolutional layers. The frequency component's temporal variation is dynamically shown by time-frequency representations of speech sounds. These representations are produced utilising the short-time Fourier transform plus single-channel input tensors in various speech processing applications. By synthesizing continuous wavelet transform, Melspectrograms, with Gammatone spectrograms, the author of this study [1] proposes a way to merge three distinct time-frequency representations of the signals. This might be utilised for phoneme class classification to extract phone-attribute information and automated detection of speech deficiencies in cochlear implant users. Convolutional neural networks as well as recurrent neural networks with convolutional layers are two separate deep learning-based models that are taken into account for this.

In order to diagnose and treat children with sigmatism with the use of computers, a method as a means of analysing audio data is presented in this work [2]. The focus of the study is on the articulation problems in particular that are present in sibilant sounds. 923 examples of the /s/ and / consonants from 98 five- along with six-year-old children are included in the collected speech corpus. Here, the author suggested two methods for the enlargement of data by means of the available dataset and prevent overfitting, as well as three methods for combining the multichannel data into the acoustic volume. Capacity of the mechanism to identify the examined pronunciation abnormalities with respectable accuracy has been demonstrated by classification trials using various data subsets. The architecture that organises speech data geographically into five channels offers the most effective categorization.

A common job is speech music discrimination, which entails splitting audio streams into speech- and music-only segments. What Convolutional Neural Networks are capable of in relation to the speech-music discrimination challenge are examined in this research [5]. Deep structures are being utilized to learn visual feature relationships when they exist in the spectrogram domain, as opposed to utilising to depict the audio material manually created audio features as is done with previous approaches. In light of these findings, CNNs are capable of greatly outperforming the cutting-age in regards to effectiveness, particularly when transfer learning is used.

A component of artificial intelligence called machine learning pits machines against one another to create analytical models that let them respond autonomously to novel situations. The term "deep learning" refers to the process of learning using numerous (deep) layers of neural networks. In this work [6], the technique of deep learning as well as machine learning in large data is essentially examined.

Using more complex acoustic models with multiple layers of characteristics is one area of investigation. In the work in [7], a hierarchical structure is proposed, with each layer intended to capture a specific group of recognisable feature landmarks. A unique auditory representation that makes each feature simple to recognise is built for each feature. In [8], a probabilistic generative model is presented in which long-span contextual effect across phonetic units is characterised by the dynamic structure in the concealed vocal tract resonance space. Numerous ASR systems have utilised feedforward neural networks [9]. The TRAP architecture [11], which was developed in response to findings from [10], methods a whole second of speech utilizing a feature vector. For each important band, feedforward neural networks are used to estimate the sub-word posterior probability. These probabilities are then combined to create final evaluate of posterior probabilities utilizing another FFNN. The split temporal context method, which modifies TRAP method by integrating splits over time and over frequency bands in middle layer of method before the final merging neural network, is introduced in [12].

3. System Model:

This section discuss the proposed speech signal processing and segmentation employing deep learning techniques. Firstly, the input speech signal was gathered from the crime scene and this signal has been pre-processed using using Convolutional K-means clustering (K-means C)for cluster the fragments of the input speech signal and process them for noise removal and signal artifacts removal. Here the segmentation is carried out for processed signal using Kernel based deep belief networks (KDBN). Figure-1 depicts the entire intended architecture of suggested strategy..

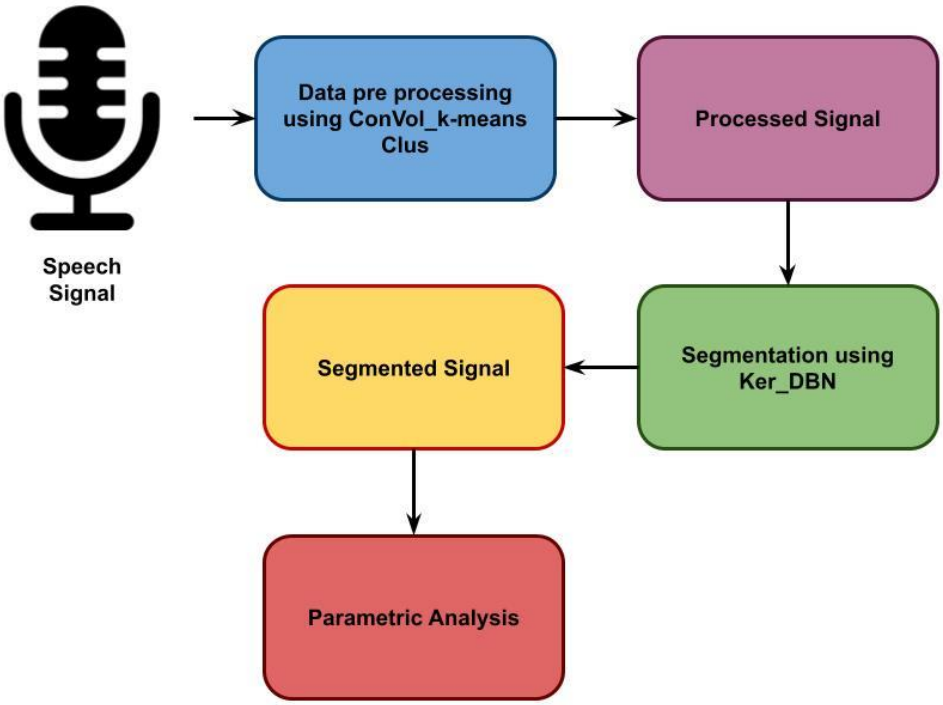

Fig.-1 System Architecture

Signal Pre-processing using Convolutional K-means clustering (ConVol_K-means Clus):

Each audio frame has a feature vector and a corresponding spectrogram, where is the number of frames and is the number of spectrograms that match each frame. There is a spectrogram that, as we previously discussed, corresponds to each of these important fragments. Therefore, it makes sense intuitively that the K-means clustering method might be used as a pre-processing defence to boost the resilience of classifiers. Following this intuition, we recreate input samples in accordance with the clustering algorithm's designation. Algorithm 1 presents an illustration of the K-means clustering procedure.

K-means clustering algorithm

Input: input example p that is a nxm matrix, and the number k of clusters

Output: reconstructed example q that is a nxm matrix

1. Obtain cluster assignment C(p) that is a nx1 vector, and corresponding centroids U that is a kxm matrix
2. Start q as a nxm matrix
3. For i=0 to n-1 do
4. Q[i]=U[C(p)[i]]
5. End for
6. Return q

Signal segmentation using Kernel based deep belief networks:

KDBN is made up of numerous RBMs that are stacked together, giving it a significant capacity for learning high level representations useful for speech emotion recognition. By using the greedy layer-wise method, it may be trained effectively. Hidden layer, with linkages allowed between aspects that are both concealed and obvious, but not both components in same layer. Figure 2 depicts the KDBN flowchart.

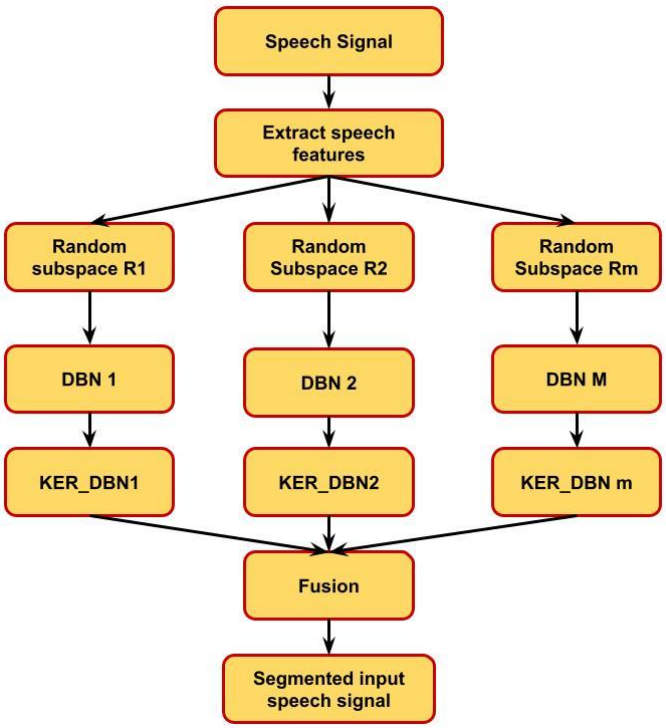

Fig. 2 flowchart of KDBN

$$P(v,h) = \frac{e^{-E(v,h)}}{\sum_{v,h} e^{-E(v,h)}}$$

where energy function is defined as

$$E(v,h) = -\left(\sum_{i=1}^{n}\sum_{j=1}^{m}\left(h_i * v_j * w_{ij}\right) + \sum_{j=1}^{m}\left(b_j * v_j\right) + \sum_{i=1}^{n}\left(c_i * h_i\right)\right).$$

Using technique covered in the next subsection, Ker DBN first extracts characteristics from input speech signals, which are then used to generate a large number of random subspaces Ri. The Ker DBN algorithm, often known as Algorithm 2, consists of two stages: training and testingThe same technique is used to collect features from the testing speech stream, which is then supplied to all base classifiers during testing.

Algorithm of Ker_DBN:

Input. Training speech signals $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, ensemble size M, and input speech signal x Output. emotion label y

Training Stage

(1) Extract features for every speech signal in $DD_v = \{(v_1, y_1), \ldots, (v_n, y_n)\}$ where v_i is feature vector of x_i

(2) Produce random subspaces $R_i (1 \le i \le M)$ from D_v

(3) ProduceDBN_i from R_i

$R_i \rightarrow \mathrm{DBN}_i$

(4) Produce base classifiers SVM_i for ensemble $\mathrm{DBN}_i \rightarrow \mathrm{SVM}_i$

Testing Stage

(5) Extract features for speech signal $x: x \rightarrow v$

(6) Produce$M$ random subspaces from $v: v \rightarrow R_i (1 \le i \le M)$

(7) Input each random subspace R_i into $\mathrm{DBN}_i R_i \rightarrow \mathrm{DBN}_i$

(9) Designate emotion label y for x by majority voting, where f is Boolean function $y = \mathrm{argmin}_{c_j} \sum_{i=1}^{M} f(\mathrm{SVM}_i == c_j)$

Change SVM to KDBN

4. Performance Analysis:

The experimental setup is described as follows: In our research, we used the open-source Python and R distribution Anaconda, with Spyder (version 3.3.4) serving as the working IDE.For the computational work and to work with files like comma-separated value (CSV) files, Numpy and Pandas libraries were utilised.The MatPlot package is utilised for the results and plotted graphs' display. We have utilised the Time module to record start and end times as well as to compute the algorithms' execution times.

Dataset description:

The tests were carried out using three benchmark datasets: RML, SAVEE, and eNTERFACE'05.Audio-visual modals are supported by all three datasets. Several factors were taken into account when selecting the datasets.To demonstrate the adaptability of our concept, we used datasets that ranged in size. First of all, since the same emotional states are covered across all three datasets, their comparability is greatly increased.

Table- 1 A comparison of the suggested method with the current one

Specifications	**ASR**	**TRAP**	**AI_NN_SSS_KDBN**
Signal errorrate	55	56	59
Signal to noise ratio	45	49	51
Accuracy	88	92	96
Minimum variance distortion	39	42	45

The table 1 shows comparative analysis in speech signal processing and segmentation for RML, SAVEE, and eNTERFACE'05 datasets. In this instance, the analysis has been shown with regards of accuracy, signal error rate, Signal to noise ratio, minimum variance distortion. From the above comparison, proposed technique obtained optimal result in processing and segmenting the speech signal.

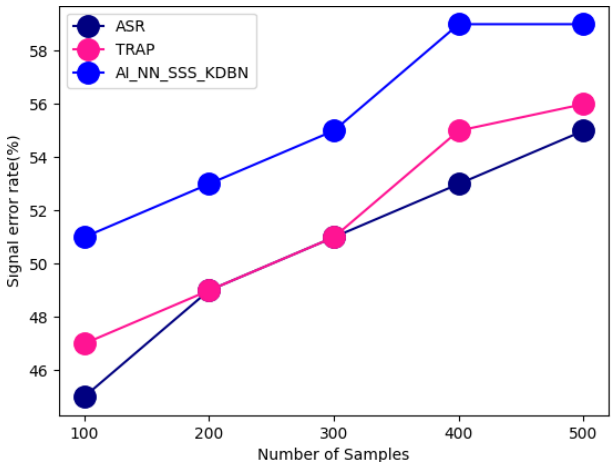

Fig.3 Comparative analysis of SRR

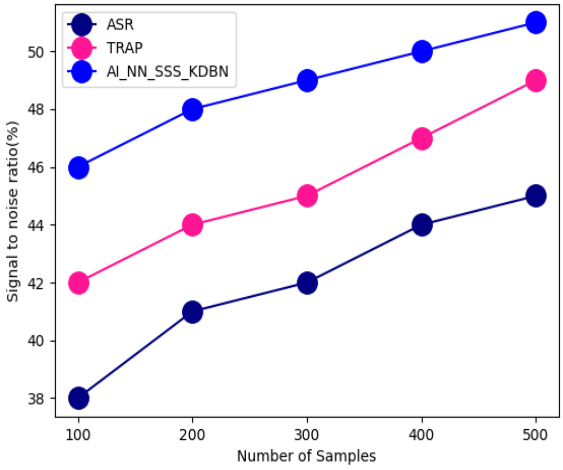

Fig. 4 Comparative analysis of SNR

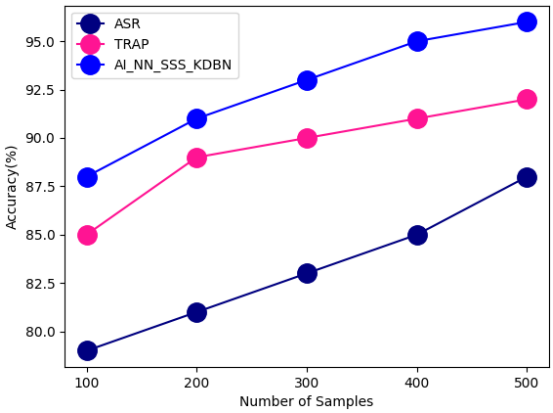

Fig. 5 Comparative analysis of accuracy

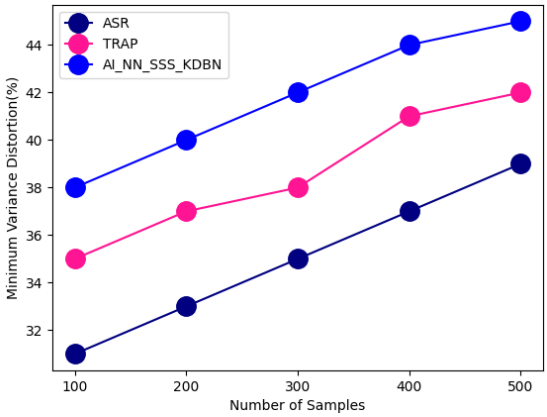

Fig. 6 Comparative analysis of MVD

The figure 3-6 illustrates the assessment of new and current techniques in comparison with regards of accuracy, signal error rate, Signal to noise ratio, minimum variance distortion. Here the proposed technique obtained optimal results in processing the input speech signal and segmenting the signal.

5. Conclusion:

This study suggested an innovative method in processing and segmenting the input speech signal based on deep learning architectures. Here the segmentation is carried out for processed signal using Kernelized deep belief networks (KDBN). The k-means clustering strategy is then employed to identify the k most discriminant frames, also known as key fragments, from all extracted feature vectors of one voice signal.A set of basic classifiers is then built for the ensemble after all training speech signals' speech features have been extracted.The experimental results shown in perspective of accuracy, signal error rate, Signal to noise ratio, minimum variance distortion.

References

[1] Arias-Vergara, T., Klumpp, P., Vasquez-Correa, J. C., Noeth, E., Orozco-Arroyave, J. R., & Schuster, M. (2021). Multi-channel spectrograms for speech processing applications using deep learning methods. *Pattern Analysis and Applications*, *24*(2), 423-431.

[2] Krecichwost, M., Mocko, N., & Badura, P. (2021). Automated detection of sigmatism using deep learning applied to multichannel speech signal. *Biomedical Signal Processing and Control*, *68*, 102612.

[3] Yang, X. K., Qu, D., Zhang, W. L., & Zhang, W. Q. (2018). An adapted data selection for deep learning-based audio segmentation in multi-genre broadcast channel. *Digital Signal Processing*, *81*, 8-15.

[4] Santhanavijayan, A., Naresh Kumar, D., & Deepak, G. (2021). A semantic-aware strategy for automatic speech recognition incorporating deep learning models. In *Intelligent system design* (pp. 247-254). Springer, Singapore.

[5] Papakostas, M., & Giannakopoulos, T. (2018). Speech-music discrimination using deep visual feature extractors. *Expert Systems with Applications*, *114*, 334-344.

[6] França, R. P., Monteiro, A. C. B., Arthur, R., & Iano, Y. (2021). An overview of deep learning in big data, image, and signal processing in the modern digital age. *Trends in Deep Learning Methodologies*, 63-87.

[7] Yao, Z., Wang, Z., Liu, W., Liu, Y., & Pan, J. (2020). Speech emotion recognition using fusion of three multi-task learning-based classifers: HSF-DNN, MS-CNN and LLD-RNN. Speech Communication, 120, 11–19.

[8] Sharmadha, S., Shivani, K., Shruthi, K., Bharathi, B., and Kavitha, S. (2020) "Automatic speech recognition using Deep Neural Network". International Conference on Soft Computing and Signal Processing, Hyderabad, India, pp. 353–361

[9] Korkmaz, Y., & Boyacı, A. (2022). milVAD: A bag-level MNIST modelling of voice activity detection using deep multiple instance learning. *Biomedical Signal Processing and Control*, *74*, 103520.

[10] Li, X., Ma, D., & Yin, B. (2021). Advance research in agricultural text-to-speech: the word segmentation of analytic language and the deep learning-based end-to-end system. Computers and Electronics in Agriculture, 180, 105908.