A Segmentation-based Token Identification for Recognition of Audio Mathematical Expression

SBTI Method for Audio Mathematical Expression Recognition

Vaishali A. Kherdekar[1], Sachin A. Naik[2], Prafulla Bafna[3]
Symbiosis Institute of Computer Studies and Research (SICSR), Symbiosis International (Deemed University),
Pune, Maharashtra State, India

***Abstract*—In human-computer interaction, humans can interact with the computer with the help of text, audio, images, speech, etc. Interacting with the computer using speech, speech recognitions in particularly audio segmentation is a challenging task due to accent or way of pronouncing style. To input mathematical symbols, words, functions, and expressions with the help of a keyboard are tedious and time-consuming. Input this with the help of audio, speeds up the input process. In this paper, an SBTI (audio Segmentation Based Token Identification) algorithm is proposed for the recognition of words in an audio mathematical expression. 6 types of audio mathematical expressions are considered for recognition. The proposed algorithm segments the audio file into chunks and from each chunk temporal and spectral characteristics of audio signals are selected to extract the features. The model is trained using a neural network. The proposed algorithm shows a classification accuracy of 100% for the algebraic, quadratic, area, and differentiation expression, 99% for trigonometric expression, and 92% for summation expression.**

Keywords—Audio segmentation; classification; feature extraction; neural network; speech recognition

I. Introduction

In human-computer interaction, various types of data are used such as text, image, audio, and video. Interacting with the computer with the help of audio data increases the speed of interaction. Now a day’s most devices accept audio data as input but input mathematical data with the help of audio is not available. Existing work reveals that few researchers worked on the recognition of audio mathematical expression. To input mathematical symbols, words, functions, and expressions with the help of a keyboard are tedious and time-consuming. In this paper, a novel methodology is designed for the recognition of audio mathematical expressions. Recognition of mathematical expression is a challenging task because it consists of digits, letters, symbols, functions, etc.

Segmenting audio data is an important step in speech processing as well as audio processing applications such as speech recognition [1]. Segmenting audio data is challenging and a need of the day for applications such as speech recognition. Audio segmentation quality [2] affects the performance of recognition of speech. Segmentation of audio data is the process of splitting the audio data into small segments. Continuous audio signals are fragmented into small segments. To segment a sentence into phonemes, words, and syllables, [3] continuous speech segmentation is used. Based on classification techniques, segmentation is categorized into two types [4], [6], [7] classification-dependent segmentation and classification-independent segmentation. In [5] audio is segmented into the voiced and unvoiced parts with the help of ZCR and energy. The purpose of audio segmentation is to find acoustics variations [7] in the audio signal. Audio is segmented into various components of speech such as voiced part, nonvoiced part, noise, silence [8], etc. Segmentation methods are also categorized into three types [6] model-based segmentation, metric-based segmentation, and energy-based segmentation. In energy-based segmentation, silence is detected by thresholding the energy signals. Statistical models are used to design the model-based segmentation [9]. In metric-based segmentation, a distance function is used to determine the boundaries. Speech segmentation is an important step in speech processing [10]. A phoneme or a syllable or a sub-word, is the basic phonetic unit [11] which is based on the language.

Segmentation can be categorized into [18] acoustic audio segmentation and phonetic audio segmentation. Speech segmentation is also categorized into [19] manual segmentation and automatic segmentation. In Manual segmentation, speech signal waveforms are examined and it is segmented but it is very time-consuming and not produced correct results whereas, in automatic segmentation, speech waveforms are automatically segmented into words. Preprocessing of a speech signal is most important in speech recognition. As speech signals are continuous, the first thing is to convert them into digital form with the help of an analog-to-digital converter. In the audio file when silence is present it has no importance because of the lack of information. Silence removal is significant in speech recognition. Silence in the audio file is removed with the help of energy and the threshold value. In an audio file, speech signals are present in the form of frames through which features are extracted by considering the window size and frame rate. Feature extraction is the process of converting an audio signal into a sequence of features called a feature vector. Feature vector consists of the information of audio signals having temporal and spectral characteristics. The selection of feature sets plays a key role to improve the performance of audio segmentation.

In the proposed work, a novel audio segmentation based token identification algorithm is proposed, features are extracted based on various parameters and accuracy for different types of expressions is reported.

II. Literature Review

There are various challenges in audio segmentation and speech recognition such as separating the audio data into regions, multiple classes will be used for different types of segmentation boundaries [1]. To segment audio data various features are considered such as BIC, Boundary confidence, Warping factor variance, and word length [1] where the MAP decoder framework is presented which is applicable for segment features. It has been observed that many researchers have used various audio segmentation techniques [6, 7, 8, 9,10,14] which are based on either acoustic information or word level timing information [2]. In [7, 8,10] author proposed audio segmentation methods for classifying audio components into speech, non-speech, noise, music, and silence. By combining model-based and metric-based algorithms, a hybrid algorithm is proposed where results indicate that the model-based algorithm gives high precision and moderate recall whereas the metric-based algorithm indicates high recall and moderate precision. Gish distance function is used which improves the performance. Researchers have used various feature extraction techniques for segmentation such as MFCC [1], zero-crossing rate, and energy [5] where authors concluded that ZCR is low and energy is high for the voiced part and ZCR is high and energy is low for the unvoiced part. In [9] author proposed a speech segmentation method for streaming end-to-end automatic speech recognition. In [11] 20 Kannada audio sentences are used where speech signals are processed and framed into 20 milliseconds, the model shows an accuracy of 87.76% using HMM. In [12] 150 Punjabi-connected words are used for the recognition of speech in noisy and noise-free environments. In [13] Voice Input Speech Output calculator is developed for the recognition of Bangla numerals. In [14] weather news data in Myanmar language is segmented and computed the recognition accuracy. In [15] threshold and energy value are used for the segmentation of recognition of English audio. To select voiced and unvoiced parts from audio signal Voice Activity Detectors [16] are used. In [20] author developed the model based on handwritten recognition and speech recognition for 74 mathematical symbols, 39 features were extracted from each frame. The model showed a 50.09% recognition rate for speech recognition and 81.55% recognition rate for handwritten recognition which was developed using SVM classifier. In [21] author proposed a mathematical expression recognition based on handwritten recognition, speech recognition, and fusion method. They used CROHME dataset for handwritten recognition and HAMEX dataset for speech recognition. For automatic speech recognition, each speech signal is filtered, re-sampled and then reframed back. It is segmented into 25 ms frames with an overlap of 10 ms. Results showed that recognition rate of 80% for handwritten recognition, 50% for speech recognition and 98% after fusion. They conclude that bimodal processing gives better results. In [22] author discussed the ICT related needs to study mathematics for disabled persons. They have reported that Human-Computer Interaction in the domain of mathematics is still lagging, creating & editing electronic mathematical content still remains difficult for both able and disabled students, and particularly for people studying "at a distance" and those relying on mobile computing devices such as Smartphones and tablet PCs. In [23] author developed a speech recognition system for speaker-independent isolated words. They have created a database of 200 samples by recording Urdu digits 1 to 20 from 10 speakers at a sampling rate of 16KHz. DWT is used to extract the features and the model is trained using FFANN (Feed Forward Artificial Neural Network).

In [24] author developed a speaker-independent model for the Kannada language. PRAAT software was used to record 10 Kannada words (from 1 to 10) 10 times from one female speaker at an 8KHz sampling rate with some background noise. A statistical method is used to remove the silence from the speech signal. Speech signals are segmented into frames of 160 samples in length with overlapping of 80 samples. Each frame is multiplied by a 160-point Hamming window. Features are extracted using the MFCC technique. They trained the model using two clustering algorithms of the vector quantization method. VQ1 was developed by Juang which is based on a binary splitting algorithm i.e. splitting every cluster into two clusters, and VQ2 is developed by Lipeika, based on splitting a cluster with the largest average distortions into two clusters. The result shows that when the silence removal algorithm is used error rate has decreased from 2.59 to 1.56 in the case of VQ1 clustering algorithm and 2.5 to 1.45 for the VQ2 algorithm. In [25] author proposed a speech recognition model for isolated speaker-independent 300 Gujarathi numerals from age group of 5 to 40. Result shows an accuracy of 78.13% with MFCC features and KNN classifier. To overcome the problem of low accuracy and high computational complexity, [26] proposed a method for speech segmentation.

Based on the above literature, it shows that audio segmentation is challenging [13], and segmenting an audio mathematical expression is more challenging, hence there is scope to work on it. Researchers reported that the syllable counting method can be used in the segmentation [14] process to improve the result. In this paper, we have proposed a segmentation algorithm based on tokens for the recognition of audio mathematical expressions.

Authors have experimented to segment the speech into phonemes with the help of phonemes specialists, results in known as tedious, expensive and subjective [27].

An optimised parameters of automatic speech segmentation has been done into syllable units. This can be developed using time domain energy-based features and static threshold is used to detect the syllable boundary [28].

Speech Recognition technology is challenging because of sensitivity to the environment (background noise) or the weak representation of grammatical and semantic knowledge [29].

The proposed research study is unique because the dataset is generated as it is not available and samples are collected by varied age group. To segment long utterances is challenging. Here, Complex mathematical expressions are executed using the proposed approach, and recognition accuracy is reported.In this experiment the eight different categories of mathematical expressions are considered because of their different spatial and structural representation.

III. Methodology SBTI (Segmentation-based Token Identification)

Fig. 1 shows the methodology used for implementation of the proposed algorithm of segmentation.

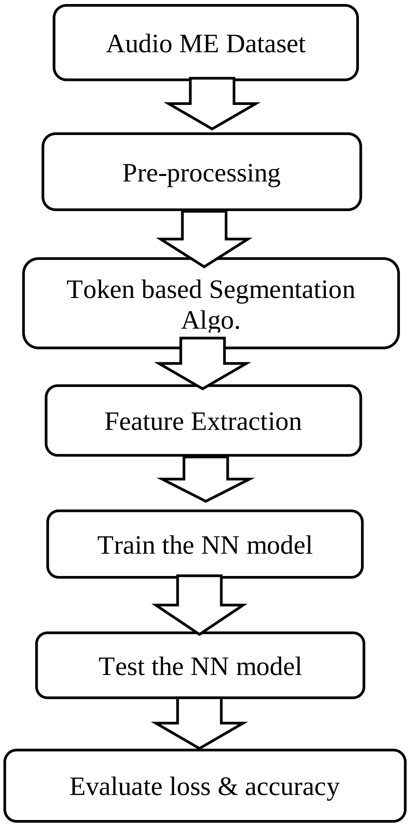


Fig. 1. Methodology

A. Audio Mathematical Expression (AME) Dataset

To perform the experiment we have selected 6 categories of mathematical expression which include algebraic expression, quadratic expression, area formulae, trigonometric expression, differentiation, and summation. To cover the samples from different age groups, people from the age group of 10 to 55 are used to create the dataset. All these expressions are recorded using the Audacity tool and taken in .wav. Table I shows the number of samples used for this study.

B. Pre-processing

In this step, the voiced part from the audio file is selected with the help of ZCR.

C. Segmentation

The proposed segmentation algorithm segments audio mathematical expression by considering the parameters such as the number of tokens, frame rate, and duration of audio mathematical expression in milliseconds. Fig. 3 shows the signal of $x^2 + y^2 = z^2$ audio mathematical expression in .wav format which is given as input to the proposed algorithm.

Proposed Segmentation Algorithm 1: SBTI

```
Input: Audio Mathematical Expression (AME) in .wav format
Output: Segments of  AME
  Start
    D←S/ Fr     // Duration of AME
    Fr ← // Number of samples per second
    T←  // Number of tokens in AME
    Ch← D/T   // Number of chunks
    St←0 //Starting pint
    Lt←0 //Ending point
    For I←1 to T
        Ch (I) ← Seg(St. . Lt )
        St ←  Ch   + I
        Lt ← Lt + Ch
    End loop
    End
```

D. Feature Extraction

From the audio mathematical expression, we have selected the features such as chroma_stft, rmse, spectral_centroid, spectral_bandwidth, roll-off, zero_crossing_rate, and 21 MFCC coefficients. Chromogram features and root mean square energy are features calculated in chroma_stft and RMSE. Root Mean Square Energy for each frame is calculated in RMSE.

Spectral centroid is a frequency-based feature that computes the “Location of mass”. Equation (1) is used to calculate spectral centroid.

$$f_C = \frac{\Sigma_k s(k) f(k)}{\Sigma_k s(k)} \quad (1)$$

where s(k) is the spectral magnitude at frequency bin k, f(k) is the frequency at bin k.

A variation between higher and lower frequencies is calculated using spectral_bandwidth. Spectral_rolloff computes the shape of the signal. It shows the frequency at which high frequencies deteriorate to 0. ZCR is used to check the voiced and unvoiced parts of speech. It is the rate at which the signal changes from positive to negative or vice-versa. It is computed as

$$ZCR = \sum_{n=1}^{N-1} |S[x(n+1)] - S[x(n)]| \,/\, 2(N-1) \quad (2)$$

In MFCC, the first audio signals are divided into short frames. Calculate the Fourier transform for each frame. Find the log of all filter bank energies. Compute the discrete cosine transform of each Mel log power. MFCC coefficients are calculated as

$$\text{Mel (f)} = 2595 * \log_{10}(1 + \frac{f}{700}) \quad (3)$$

There are various applications of neural network present such as Machine translation, speech recognition, sentiment analysis, chatbots, named entity recognition, etc. To study such areas machine learning’s sequence model is very useful.

Sequence modeling consists of data in the form of sequences, a data structure.

It is used in supervised learning algorithms. Sequence models are categorized based on the input and output which may be in the form of scalar, audio, text, image, and video form.

The performance of the model is measured using various parameters such as accuracy, loss, F1-score, precision, recall, etc. The accuracy of the model depends on predicted and observed values. It is computed as

$$\mathrm{Acc(M)} = \frac{TrP+TrN}{Tot_P+Tot_N} \tag{4}$$

Where TrP and TrN are true positive and true negative respectively. Tot_P is the combination of true positive and false positive samples whereas Tot_N is the inclusion of false positive and false negative.The model's performance is visualized using accuracy-loss graphs.

E. Training and Testing

To recognize the speech for audio mathematical expression, the proposed algorithm is used to segment an audio mathematical expression. Audio mathematical expression in .wav format is given as input as shown in Fig. 2. The frame rate and the number of samples present in each expression are computed. Based on the number of samples and frame rate duration of audio mathematical expression is find out. The number of tokens and duration of audio mathematical expressions are used to segment the audio mathematical expression.

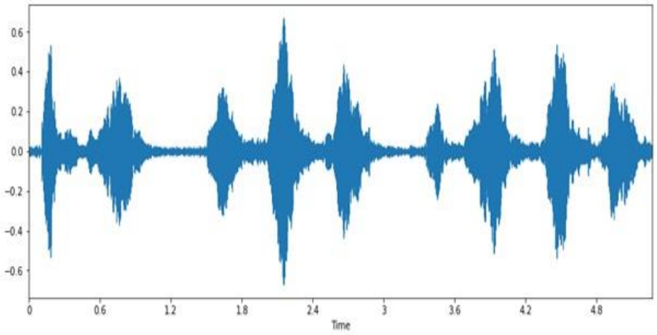


Fig. 2. An Audio signal of the mathematical expression "$x^2 + y^2 = z^2$".

Features such as chroma_stft, rmse, spectral_centroid, spectral bandwidth, roll-off, zero_crossing_rate, and 21 MFCC coefficients are extracted from each segment. The parameters used to train and test the NN (Neural Network) model are RElu and Softmax activation function, sparse_categorical_crossentropy loss function, adam optimizer, 50 epochs, and 128 batch size. The total number of samples for each expression is split into Training 80% and testing set 20%.

IV. Results and Discussion

Table I, represents the samples used for segmentations alog with its total number of segments and number of labels used. Table II indicates that training accuracy for algebraic expression, quadratic expression, trigonometric expression, and area formulae whereas it is less for summation expression. Testing accuracy is also less for summation expression as compared to another category because of more number of labels. Table II shows testing accuracy for each category of audio mathematical expression. Fig. 3 shows the graph of training accuracy of summation and differentiation expression. It indicates that initially training accuracy is low but gradually it goes on increasing.

Table III shows the comparative result of the proposed model with the existing result for the recognition of words in continuous speech.

TABLE I. No. of Samples used for Segmentation

Expression Type	No. of Samples	No. of Segments	No. of Labels
Algebraic (x2 + y2 = z2)	160	1280	8
Quadratic (ax2 + bx + c)	237	1896	6
Area (πr2)	225	675	3
Trigonometric (Sin2θ + cos2 θ =1)	102	918	7
Differentiation (d/dx x2 =2x)	100	600	6
Summation (∑n = n(n+1)/2)	100	1000	10

TABLE II. Testing Accuracy

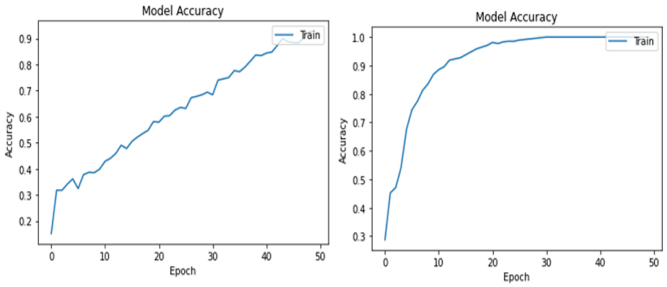
Expression Type		Testing Accuracy
Algebraic	$x^2 + y^2 = z^2$	0.86
Quadratic	$ax^2 + bx + c$	0.91
Area	πr^2	0.91
Trigonometric	$\mathrm{Sin}^2\theta + \mathrm{Cos}^2\theta = 1$	0.85
Differentiation	$d/dx\ x^2 = 2x$	0.89
Summation	$\sum n = n(n+1)/2$	0.60

TABLE III. Comparative Results of the Proposed Model with Existing Results

Ref. No.	Dataset	Language	Recognition Rate
[11]	20 unique sentences	Kannada	87.76%
[12]	150 distinct Words	Punjabi	86.05%
[17]	25 samples of 0- 9 digits, some operators	English	80%
[20]	74 Mathematical Symbols	French	50%
[21]	HAMEX	French	50%
Proposed Algo.	AME	English	85% to 91%

V. Conclusion

In Human-computer interaction, speech recognition plays a vital role, but developing systems for speech recognition is challenging. In this paper, a segmentation-based token identification algorithm is proposed to segment an audio mathematical expression and classify it into symbols, letters, digits, Greek letters, and operators. Chroma_stft, Rmse, Spectral_centroid, Spectral_bandwidth, Rolloff, Zero_crossing_rate, and 21 MFCC coefficients were used to extract the features. Results indicated that the proposed algorithm works accurately for the segmentation of audio mathematical expressions. This study will be useful to speed up the entry of mathematics in documents for normal people as well as for blind persons. It will help the researchers to work on the recognition of speech for expressions used in mathematics.



(a) Summation (b) Differentiation

Fig. 3. Training accuracy of summation and differentiation.

In the future, we would like to expand the proposed segmentation algorithm for audio mathematical expressions having a large number of tokens.

Acknowledgment

Researchers thank the faculty of Symbiosis Institute of Computer Studies & Research Centre for providing support and guidance.

References

[1] D. Rybach, C. Gollan, R. Schluter, and H. Ney, "Audio segmentation for speech recognition using segment features, " In 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4197-4200. IEEE, 2009.

[2] S. E. Tranter, K. Yu, G. Evermann, and P. C. Woodland, "Generating and evaluating segmentations for automatic speech recognition of conversational telephone speech," In 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 1, pp. I-753. IEEE, 2004.

[3] S. N. Endah, N. Fadlilah, R. Kusumaningrum and S. Adhy, "Continuous Speech Segmentation Using Dynamic Thresholding of Short-term Features," In Journal of Engineering Science and Technology, 17(4) 2022, 2919-2935.

[4] J. X. Zhang, J. Whalley, and S. Brooks, "A two phase method for general audio segmentation," In 2009 IEEE International Conference on Multimedia and Expo, pp. 626-629. IEEE, 2009.

[5] R.G. Bachu, S. Kopparthi, B. Adapa, and B. D. Barkana, "Separation of voiced and unvoiced using zero crossing rate and energy of the speech signal", In American Society for Engineering Education (ASEE) zone conference proceedings, pp. 1-7. American Society for Engineering Education, 2008.

[6] G. M. Bhandari, "Different audio feature extraction using segmentation," International Journal for Innovative Research in Science Technology 2, no. 9 (2016): 1-5.

[7] G. M. Bhandari, R. S. Kawitkar, and M. C. Borawake, "Audio segmentation for speech recognition using segment features," International Journal of Computer Technology and Applications 4, no. 2 (2013): 182.

[8] Panagiotakis, Costas, and G. Tziritas, "A speech/music discriminator using RMS and zero-crossings," In 2002 11th European Signal Processing Conference, pp. 1-4. IEEE, 2002.

[9] Y. Shu, H. Luo, S. Zhang, L. Wang and J. Dang, "A CIF-Based Speech Segmentation Method for Streaming E2E ASR," in IEEE Signal Processing Letters, vol. 30, pp. 344-348, 2023, doi: 10.1109/LSP.2023.3261662.

[10] S. Zahid, F. Hussain, M. Rashid, M. H. Yousaf, and H.A. Habib, "Optimized audio classification and segmentation algorithm by using ensemble methods," Mathematical Problems in Engineering 2015 (2015).

[11] P. Punitha, and G. Hemakumar, "Speaker dependent continuous Kannada speech recognition using HMM." In 2014 International Conference on Intelligent Computing Applications, pp. 402-405. IEEE, 2014.

[12] A. Kaur, and A. Singh. "Optimizing feature extraction techniques constituting phone based modelling on connected words for Punjabi automatic speech recognition." In 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 2104-2108. IEEE, 2016.

[13] T. Ahmed, Md. F. Wahid and Md. A. Habib, "Implementation of Bangla Speech Recognition in Voice Input Speech Output (VISO) Calculator," International Conference on Bangla Speech and Language Processing (ICBSLP), 21-22 September, 2018.

[14] Y.W. Chit, and S.S. Khaing, "Myanmar continuous speech recognition system using fuzzy logic classification in speech segmentation," In Proceedings of the 2018 International Conference on Intelligent Information Technology, pp. 14-17. 2018.

[15] L. Kun, Y. Zhang, J. Li, and W. Dong, "A Study on English Audio Segmentation Methods Based on Threshold Value and Energy Sequence," In MATEC Web of Conferences, vol. 22, p. 02017. EDP Sciences, 2015.

[16] V. Kanabur, S.S. Harakannanavar and D. Torse, "An Extensive Review of Feature Extraction Techniques, Challenges and Trends in Automatic Speech Recognition," International Journal of Image, Graphics and Signal Processing 10, no. 5 (2019).

[17] U. Shrawankar, and A. Mahajan. "Speech: a challenge to digital signal processing technology for human-to-computer interaction." arXiv preprint arXiv:1305.1925 (2013).

[18] H. Frihia and H. Bahi, "HMM/SVM segmentation and labelling of Arabic speech for speech recognition applications," International Journal of Speech Technology 20, no. 3 (2017): 563-573.

[19] M. Kaur and A. Kaur, "A review: Different methods of segmenting a continuous speech signal into basic units," International Journal of Engineering and Computer Science 2, no. 11 (2013).

[20] S. Medjkoune, H. Mouchère, S. Petitrenaud, and V. Christian, "Handwritten and audio information fusion for mathematical symbol recognition," In 2011 International Conference on Document Analysis and Recognition, pp. 379-383. IEEE, 2011.

[21] S. Medjkoune, H. Mouchère, S. Petitrenaud, and V. Christian, "Using Speech for Handwritten Mathematical Expression Recognition Disambiguation," In 2012 International Conference on Frontiers in Handwriting Recognition, 2012.

[22] D. Attanayake, G. Hunter, J. Denholm-Price, E. Pfluegel, "Novel Multi-Modal Tools to Enhance Disabled and Distance Learners' Experience of Mathematics," International Journal on Advances in ICT for Emerging Regions (ICTER), 6(1), 2013.

[23] B. Rehman, Z. Halim, G. Abbas, T. Muhammad, "Artificial Neural Network-based Speech Recognition using DWT Analysis Applied on Isolated words from Oriental Languages," Malaysian Journal of Computer Science 28, no. 3, pp. 242-262,1015.

[24] M.A. Anusuya, and S.K. Katti, "Speaker Independent Kannada Speech Recognition using Vector quantization," Proceedings published by International Journal of Computer Applications (IJCA), pp. 316-319 ISSN: 0975 – 8887, 2012.

[25] B. C. Patel and A. A. Desai, "Recognition of Spoken Gujarati Numeral and Its Conversion into Electronic Form", International Journal of Engineering Research & Technology (IJERT) IJERT ISSN: 2278-0181 Vol. 3 Issue 9, September- 2014.

[26] S. Niveditha, S. Shreyanth, V. Kathiroli, P. Agarwal and S. Ram Abishek, "Kernelized Deep Networks for Speech Signal Segmentation Using Clustering and Artificial Intelligence in Neural Networks," 2023 IEEE 12th International Conference on Communication Systems and Network Technologies (CSNT), Bhopal, India, 2023, pp. 667-674, doi: 10.1109/CSNT57126.2023.10134609.

[27] Sharma,Meenakshi, and Vijay Kumar. "Importance of Artificial Intelligence in Neural Network: Speech Signal segmentation using K-means clustering with Kernelized deep belief networks." Eur. Chem. Bull. 2023 , 12 (Special Issue 7), 2061-2065

[28] Riksa Meidy Karim, Suyanto, "Optimizing Parameters of Automatic Speech Segmentation into Syllable Units", International Journal of Intelligent Systems and Applications (IJISA), Vol.11, No.5, pp.9-17 2019. DOI: 10.5815/ijisa.2019.05.02

[29] Chen, L. Special Issue on Automatic Speech Recognition. Applied Sciences, 13(9), 5389, 2023.