Int. j. inf. tecnol. (April 2024) 16(4):2661–2673
https://doi.org/10.1007/s41870-024-01776-3

ORIGINAL RESEARCH

Spoken word recognition using a novel speech boundary segment of voiceless articulatory consonants

Bachchu Paul[1] · Sumita Guchhait[2] · Sandipan Maity[3] · Biswajit Laya[4] · Anudyuti Ghorai[4] · Anish Sarkar[1] · Utpal Nandi[1]

Received: 17 August 2023 / Accepted: 17 November 2023 / Published online: 17 March 2024

Abstract Communication through speech offers the most straightforward channel for man-machine interaction. Nevertheless, it is a barrier for some languages with low data resources. Extracting features and processing silence in a speech signal is an unnecessary extra effort. Noise in the speech signal reduces classification accuracy. Therefore, silence and noise are removed from the signal to improve recognition. Nonetheless, current approaches rely on static Zero-Crossing-Rate (ZCR) and energy values for the detection. Through the analysis of the speech signal, it has been determined that the utilization of fixed ZCR and energy values do not effectively address the delineation of unvoiced consonant boundaries in speech. The use of static values fails to accurately identify the speech boundary during the articulation of these unvoiced consonants. Therefore, in this study, the dynamic value of ZCR and energy has been derived to overcome this problem. Here, roughly a spoken region has first been identified from each speech signal of a non-overlapping frame. In the second step, the dynamic values are derived by two novel algorithms. Two standard datasets, the Free Spoken Digit Dataset (FSDD) and the Bangla 0 to 99 Dataset (Bangla Dataset), spoken words in English and Bengali, respectively, have been used in this study. The Mel Frequency Cepstral Coefficients (MFCC) have been extracted from each raw signal and the proposed pre-processed signal. Subsequently, these features are input into a Bidirectional Long-Short-Term-Memory (BiLSTM) network. The result shows the superiority of the proposed pre-processing methods.

Keywords Isolated word · Manner of articulation · Zero Crossing · Mel Frequency Cepstral Coefficient · Bidirectional-Long-Short-Term-Memory

✉ Bachchu Paul
ableb.paul@gmail.com

Sumita Guchhait
mitaguchhait@gmail.com

Sandipan Maity
maitysandipan88@gmail.com

Biswajit Laya
biswajitlaya007@gmail.com

Anudyuti Ghorai
anudyuti@outlook.com

Anish Sarkar
anishsarkar983@gmail.com

Utpal Nandi
nandi.3utpal@gmail.com

1 Department of Computer Science, Vidyasagar University, Midnapore 721102, West Bengal, India

2 Department of BCA, Belda College, Paschim Medinipur, Belda 721424, West Bengal, India

3 Department of Computer Science, Debra Thana Sahid Kshudiram Smriti Mahavidyalaya, Debra 721126, West Bengal, India

4 Department of Computer Science (BCA), Kharagpur College, Inda, Kharagpur 721305, West Bengal, India

1 Introduction

Speech boundary detection is an essential issue in speech segmentation. A phonetics sentence comprises related words composed of the utterances of phonemes. The declaration of a group of sentences is called continuous speech. Vowels, semi-vowels, diphthongs, and consonants constitute the primary phonemic classes [1, 2]. The phonemes are generated by air pressure flowing through the vibrating vocal cord. Vowels, semi-vowels, and diphthongs are all aspirated sounds; sufficient air flows through the vocal cord. These are

Table 1 The place and manner of articulation of the English consonants

Manner ↓		Voicing	Place						
			Bilabial	Labiodental	Interdental	Alveolar	Palatal	Velar	Glottal
Stop		Voiceless	p			t		k	ʔ
		Voiced	b			d		g	
Fricative		Voiceless		f	θ	s	ʃ		h
		Voiced		v	ð	z	ʒ		
Affricate		Voiceless					tʃ		
		Voiced					dʒ		
Nasal		Voiced	m			n		ŋ	
Liquid	Lateral	Voiced				l			
	Rhotic	Voiced					r(ɹ)		
Glide		Voiced	w				j	(w)	

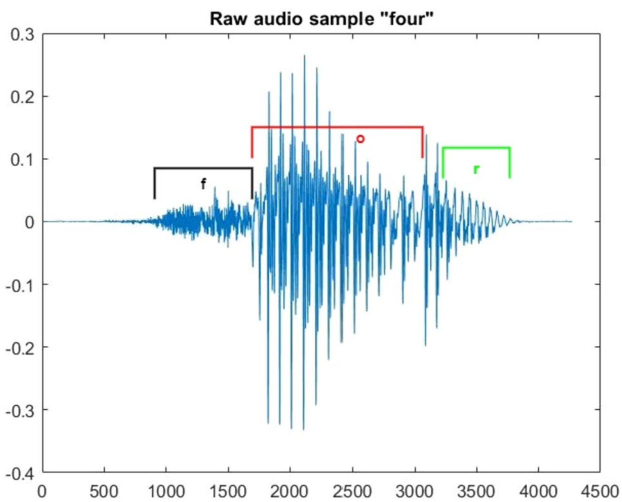


Fig. 1 The audio wave of an utterance of the word “four”

all voiced sounds. However, the pronunciation of consonants can be categorized into voiced and unvoiced. Table 1 shows English consonants’ place and manner of articulation [3, 4]. In English, the unvoiced consonants are /p/, /t/, /k/, /f/, /θ/, /s/, /ʃ/, and /tʃ/. So, the utterance of these consonants generates a shallow air pressure in the vocal cord, resulting in such consonants having very low amplitude (almost zero). For example, the sound wave for audio signal four is presented in Fig.1.

The pronunciation of the word “four” is constructed by the phonemes /f/+/ou/+/r/. So, it begins with the unvoiced consonant /f/ (marked by the color black in Fig. 1), followed by the diphthong /ou/, looks like the utterance /o/ (characterized by the color red in Fig. 1) and ends with the voiced approximant /r/ (drawn by the color green in Fig.1). It is clear that the amplitude of the sound wave for the consonant /f/ is very low, and its utterance is like silent. Similarly, the place and manner of articulation for Bengali (formally known as Bangla) consonants [5, 6] are presented in Table 2.

The consonants /প/, /ফ/,/ত/,/থ/, /ট/, /চ/, /ছ/, /ক/, /খ/, and /স/ are unvoiced. For example, the sound wave of the Bengali word “ছয়” is shown in Fig.2.

The word ‘ছয়’ is constructed by the phonemes /ছ/+/অ/+/য়/. In Fig. 2, it is shown that the air pressure in the vocal cord for /ছ/ is so low that it appears unvoiced (indicated by the color black). The air pressure for the next two phonemes /অ/(a vowel), shown by the red color, and /য়/ (a semi-vowel), led by the pink color, is high enough to generate the high amplitude signal.

In linguistics, a word is formed by one or more syllables, and phonemes form a syllable. Therefore, boundary detection is essential for segmenting continuous speech into syllables, words, and sentences [1, 7].

Eliminating noise and silence has several advantages, such as data reduction and audio signal compression. A long audio signal may contain a single word. An audio signal can be divided into silence, voiced, and noise signals. Extracting the features of silence and noise can increase the feature matrix. Therefore, analyzing and extracting the entire signal will unnecessarily increase the amount of data. Noise and silence removal has several advantages, such as speech segmentation, boundary detection, feature reduction, data compression, vowels, and consonant detection.

Additionally, there is a possibility that noise can decrease classification accuracy, and silence enhances the audio signal’s feature.

However, speech boundary detection is not an easy task. The speech signal exhibits non-stationary characteristics. This means its characteristics change over time and between speakers. Each individual has a different vocal shape [8, 9]. Motivated by this, a novel pre-processing method has been proposed to derive the dynamic threshold ZCR and energy and applied to the isolated word recognition to address the problems in this paper.

Key features of the proposed work are outlined below.

1. We propose a distinctive pre-processing method to derive a dynamic threshold for energy and zero crossing to eliminate silence and noise regions from an audio sample. The approach correctly detects the speech boundary for the phonemes that start or end with the unvoiced consonants.
2. Two algorithms have been designed to derive the dynamic energy and zero-cross threshold. Two different datasets of two separate languages were employed in the experiment.
3. The experiment has been performed in three ways: the MFCC feature has been extracted from the raw, the voiced signal applies the static threshold, and the voiced signal uses the dynamic threshold.
4. The comparative classification accuracy from each feature matrix is analyzed to prove the superiority of the proposed pre-processing technique.

This paper is structured in the following manner: Sect. 2 provides an overview of related research in isolated word recognition. Section 3 outlines the methodology applied. Detailed results and comparative analysis can be found in Sect. 4, while Sect. 5 discusses the conclusion and the future scope of our proposed work.

2 Literature review

Researchers have conducted several studies on audio pre-processing and classifying isolated spoken words in different languages. Some works focused on the features, some on the classifiers, and some on the data pre-processing. This section compiles a selection of recent relevant studies.

Mahalingam et al. (2019) [10] proposed an isolated spoken word recognition task in the English language taken from FSDD. They classified 2000 audio files using a Wavelet Scattering Transform (WST) as a feature and Long-Short Term Memory (LSTM) as a classifier. They obtained 96% of test accuracy. Wu, J et al. (2018) [11] presented a new neural network, namely a spiking neural network (SNN), to categorize 400 Real World Computing Partnership (RWCP) sounds and 4950 audios from the TIDIGITS dataset [12]. They used Short-Term Fourier Transform (STFT) and Log-Auditory Filter Bank as a feature and Self-Organizing Map-SNN for the identification. The SNN addressed the time-warping problem. They got 99.60% accuracy on RWCP and 97.40% on the TIDIGITS dataset. Nayak et al. (2023) [13] proposed a deep learning-based 7090 speech command recognition in the Kui language. MFCC was used as a feature, and several classifiers were used for training. The highest accuracy was incurred at 97% using the attention-LSTM model. A variant of MFCC features, called Bionic Wavelet Transform (BWT), was proposed by Vani et al. (2020) [14]. The experiment used two datasets, FSDD and their own Kannada dataset. They achieved 96% and 90% accuracy on FSDD and Kannada datasets using the LSTM classifier model. Chuchura et al. (2022) [15] focused on spliced audio detection from the spectrogram as a feature and Convolutional Neural Network (CNN) as a classifier to detect forged or original audio. The accuracy obtained was 93.05%. Turab et al. (2022) [16] worked to classify isolated spoken words on three speech corpus of English, Urdu, and Gujarati. They aimed to feature ensembling to increase the recognition rate. The mel-spectrogram, MFCC, and ZCR are fused and classified using a new architecture of NetB0 and obtained the highest accuracy of 99%. The English-isolated word from FSDD was classified by Savitha et al. (2021) [17]. MFCC was used as an audio feature, and simple Recurrent Neural Network (RNN) as a classifier. They obtained 90.31% accuracy and reduced the loss to 0.4391.

Six thousand Bangla audio samples were collected from 120 Bangladeshi Speakers by Shuvo et al. (2019) [18]. They extracted the MFCC feature and fed it into CNN for classification purposes. 93.65% accuracy was achieved by their model for the regional Bangla language. B. Paul et al. (2021) [19] proposed a Bangla speech recognizer model for 1000 isolated spoken Bangla numerals. They used MFCC as a feature and the Gaussian Mixture Model (GMM) as a classifier

Table 2 The place and manner of articulation of Bengali consonants

Place → / ↓ Manner		Bilabial		Dental		Alveolar		Post-Alveolar		Palatal	Velar		Glottal
Stops	Voiceless	প/p/	ফ/ph/	ত/t/	থ/th/	ট/t/	ঠ/th/	চ/c/	ছ/ch/		ক/k/	খ/kh/	
	Voiced	ব/b/	ভ/bh/	দ/d/	ধ/dh/	ড/d/	ঢ/dh/	জ, য/ɟ/	ঝ/ɟh/		গ/g/	ঘ/gh/	
Nasals		ম/m/				ন,ণ/n/					ঙ,ঞ/ŋ/		
Trill						র/r/							
Flap						ড়,ঢ়/ɾ/							
Fricatives						স/s/				শ,ষ,/ʃ/			হ/h/
Lateral						ল/l/							
Approximant										য়/j/			

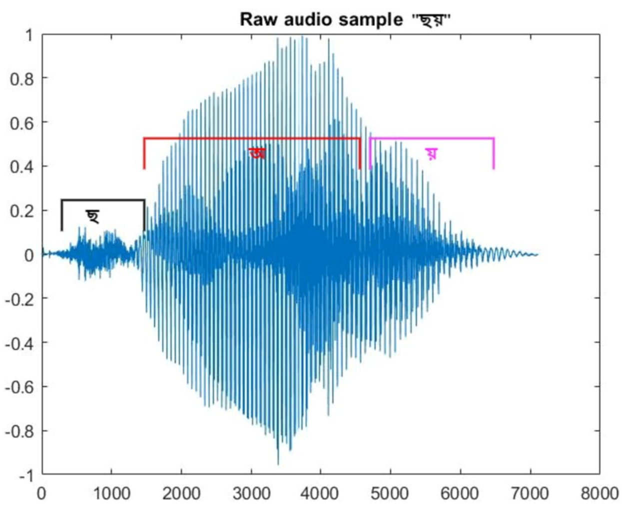


Fig. 2 The audio wave of an utterance of the word "ছয়"

and obtained 91.7% cross-validation accuracy. Four thousand audio samples are classified by Sen et al. (2021) [20]. The audio samples are Bangla spoken numerals recorded by Bangladeshi speakers. They extracted MFCC, ΔMFCC, and ΔΔMFCC and then finally trained by 10-fold cross-validation training using CNN, achieving 96.7% accuracy. Paul et al. (2022) [21] classified Bangla-isolated spoken digits and words using the template-based matching technique Dynamic Time Warping (DTW). They extracted MFCC, ΔMFCC, and ΔΔMFCC as a feature and matched every pattern by DTW. They got 93% test accuracy. An Artificial Neural Network (ANN) based isolated word recognition task was investigated by Noman et al. (2022) [22]. The speaking dialect is Bangladeshi speakers. The Discrete Fourier Transform (DFT) was extracted to feed into ANN for classification and obtained 95.23% accuracy.

From the literature, most speech recognition models [11, 13, 15–18, 20, 22] didn't focus on audio signal analysis and pre-processing to improve accuracy. Most existing works [11, 12, 14–18, 20] followed the traditional mechanism of different feature extraction followed by different classification techniques to compare the result in terms of accuracy. However, the existing models fall short in identifying the addition of noise. Few [14, 19, 21, 22] have addressed audio signal pre-processing to enhance accuracy. Also, the computational cost of pre-trained deep learning models is relatively high for these small corpora of isolated words. The cost depends on the duration of the signal and the number of training parameters. However, the entire utterance must be processed to feed it into classifiers. Noise and silence are undesirable components of speech. Thus, if we neglect them, we get more productive results. Even in our previous works [19, 21] of isolated word recognition tasks, the non-voiced consonants were not correctly detected. As a result, the boundary was segmented poorly for the words that started with voiced-less consonants. So, this work focuses on speech recognition to address these issues and enhance classification accuracy.

3 Methodology

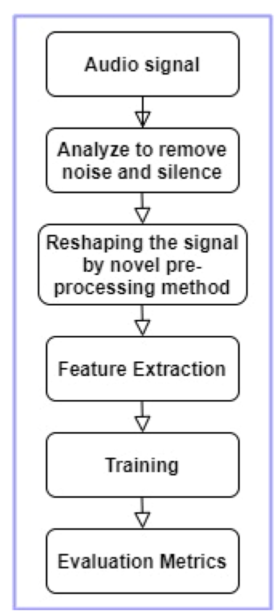
The sequential structure of the proposed method is summarized by a flowchart given in Fig. 3.

The proposed method has been carried out in four main phases: Silence and noise zone detection, derivation of the dynamic threshold for ZCR and energy, MFCC feature extraction, and finally, training & classification. The initial stage involves importing audio samples from the corpora to be analyzed. Section 3.1 covers an in-depth discussion of the speech corpus and its features. Then, a rough estimate of the voiced portion from each clip is obtained using the static values of ZCR and energy. The primary innovation of this approach lies in the creation of an algorithm to derive the dynamic threshold of ZCR and energy from cropping the voiceless part of an audio clip and sharpening the marginal part. Moving on to the subsequent stage, we extracted MFCC features from the pre-processed audio clips. To conclude, the extracted features were input into a BiLSTM classifier to measure and contrast the efficiency of our proposed method.

3.1 Database used

In this research, we utilized two established datasets containing spoken isolated words from two separate languages. The

Fig. 3 The graphical depiction of our proposed methodology



first speech corpus is the Free Spoken Digit Dataset, formerly FSDD [23], containing ten spoken digits in the English dialect. The dataset contains three thousand audio clips recorded by six individuals. The sampling frequency is 8 KHz, and the recording uses the mono channel. The second speech corpus is “Bangla spoken 0–99 number” (Bangla Dataset) [24], spoken by Bangladeshi speakers. The first ten classes have been considered for classification. The sampling frequency is 41.4 KHz, and the stereo channel with the 32-bit resolution was used during the recording. Due to there being 100 samples in every category, we added another 200 audio in each class for the robustness of the proposed method.

3.2 Silence and noise detection

In studying the audio signal, it has been found that the best method for recognizing the speech segment is to calculate the energy and zero crossing. The formula for the energy is given in Eq. 1 [1, 25] and zero-crossing is given in Eq. 2 [26] and Eq. 3 [27, 28], respectively. However, the selection of a threshold for ZCR and the energy of a non-overlapping frame is difficult because the speech signal is non-stationary. Therefore, it is not easy to choose the values of energy and ZCR to form the correct boundary in the speech signal. Again, it is impossible to choose different values from signal to signal. Therefore, a rough boundary is first obtained by selecting the static ZCR and energy values. The proposed method uses 0.3 and 0.1 as initial thresholds for zero-crossing and energy, respectively. Table 3 summarizes the decision for a frame. Algorithm 1 estimates from the utterance signal approximately the voice region, where the voice signal begins and ends when the utterance ends.

$$E_x(m) = \frac{1}{N}\sum_{i=1}^{N}|x(i)|^2 \tag{1}$$

where N is frame length, $E_x(m)$ is the energy of the m^{th} sample.

$$ZCR_x(m) = \frac{1}{2N}\sum_{i=1}^{N}|sign[x(i)] - sign[x(i-1)]| \tag{2}$$

where,

$$sign[x(i)] = \begin{cases} -1, & \text{if } x(i) < 0 \\ 1, & \text{otherwise} \end{cases} \tag{3}$$

Table 3 Frame selection

ZCR> Zero-crossing-threshold	Noise frame
Energy<Energy-threshold	Silence frame
ZCR< Zero-crossing-threshold and Energy>Energy threshold	Voiced frame

3.3 Algorithm for voice zone detection

This algorithm finds an estimated voiced activity zone from an audio signal. It inputs the audio sample x, the energy threshold e_th, and the zero-crossing threshold zc_th.

Algorithm 1 (recorded audio sample x_m , e_th, zc_th)

```
Step 1: frame_len ⟵ 0.025*fs    //fs the sampling
frequency and Set the frame length to 25ms
Step 2:  N ⟵     ⌈length(x_m)/frame_len⌉
           // Count the number of non-overlapping frames
Step 3:   for i = 1 : N       // for each frame i
Step 4:       Px (i) ⟵     energy of the frame (i) using
              formula 1
Step 5:       Zx(i) ⟵    zero crossing of the frame (i) using
              formula 2
Step 6:       MAXX = max(Px(i))
Step 7:       if Px(i) > e_th*MAXX && Zx(i) < z_th
Step 8:       then
Step 9:               v(i) = 1
Step 10:      else
Step 11:              v(i) = 0
Step 12:      endif
Step 13:  end for
Step 14: k ⟵ Find the first index of the first frame for
         which v(i) = 1
Step 15:  l  ⟵ Find the last index of the last frame for
         which v(i) = 1
Step 16:  y ⟵x(k:l)
           // The signal from x(k) to x(l) indicates a rough
           estimation of the actual  word uttered zone
Step 17:   Return (y)
Step 18: End
```

However, this selection can’t address the region where pronunciation begins or ends with unvoiced consonants such as /p/, /t/, /k/, /f/, /θ/, /s/, /ʃ/, and /tʃ/ (in English). Similarly, the pronuntiation of /প/, /ফ/,/ত/,/থ/, /ট/, /চ/, /ছ/, /ক/, /খ/, and /স/ looks silent in Bangla. This affects the recognition of the boundary. The wrong delimiter changes the meaning of the word and decreases the recognition rate.

An example of the use of these static values is - the boundary of the English word “four” shown in Fig. 4.

Figure 4a shows the raw audio wave of the word “four,” The phoneme boundaries are separated by black, red, and pink colors for ‘f’, ‘o’, and ‘r’ respectively. It is found that

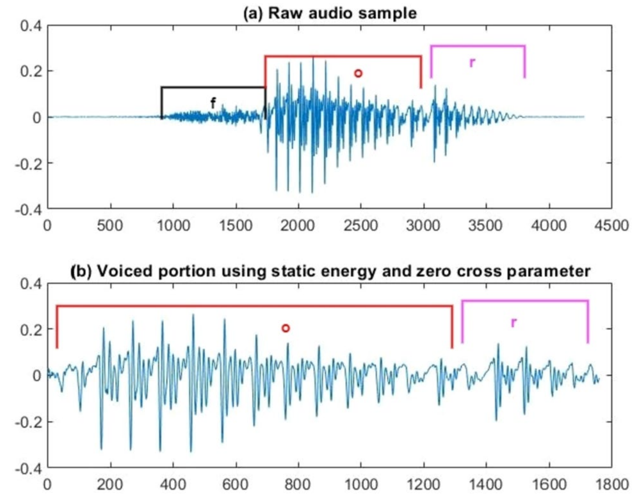


Fig. 4 **a** The raw audio wave of the pronunciation of the word “four”. **b** The voiced section of the corresponding audio wave 4(a) using static zero-crossing and energy

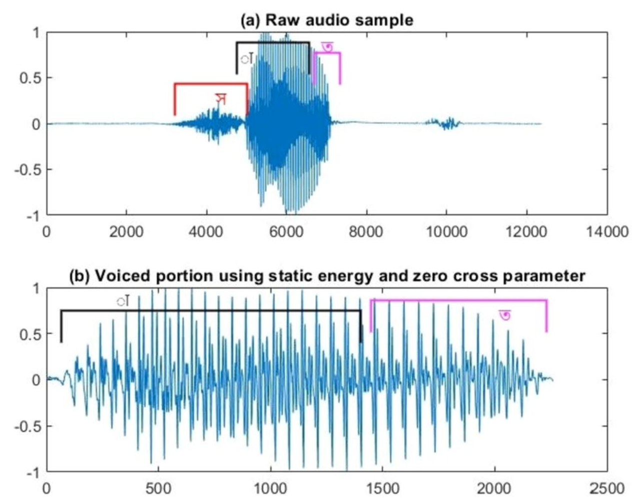


Fig. 5 **a** The raw audio wave of the pronunciation of the word “সাত”. **b** The voiced section of the corresponding audio wave 5(a) using static zero-crossing and energy

when the static zero-crossing and energy are applied, the utterance for the consonant sound “f” is missing in Fig. 4b. Thus, its utterance looks like the word “or”. This changes the meaning of the word since this static value does not properly recognize the boundary. Similarly, the static threshold is used to represent- the boundary of the Bengali word “সাত” in Fig. 5.

Figure 5a shows the raw audio wave of the sound “সাত” and the phoneme boundaries are indicated by the colors red, black, and pink for /স/, /া/, and /ত/, respectively. The voiced part of the sound is shown in Fig. 5b using the static threshold for energy and zero-crossing. The part marked in red (sound wave for /স/) is eliminated in Fig. 5b. However, it carries an “s”, an important fragment of the word “সাত”. This changes the meaning of the word. In Fig.5b, the boundary begins with the vowel “আ”. As a result, the pronunciation of the word সাত is mostly recognized as “আট” by the classifier since the utterance of 5b “আত” mostly matches আট.

3.4 Deriving dynamic threshold value of energy and zero crossing

Here, an algorithm is developed to derive the dynamic value of ZCR and energy threshold. The signal is analyzed in 25ms non-overlapping frame-wise. Here, the threshold range is reshaped by deriving an algorithm. Evaluate the frame count indicating the voice signal using Algorithm 1. Then, the mean energy and mean zero-crossing are determined in the voiced section only. Then count the number of frames with energy greater than the mean energy, called COUNT1. Similarly, the number of frames whose zero-crossing is greater than the mean zero-crossing is called COUNT2. The energy and zero-crossing threshold is then determined from the mean energy, mean zero-crossing, COUNT1, and COUNT2,- according to steps 11 and 12 in Algorithm 2.

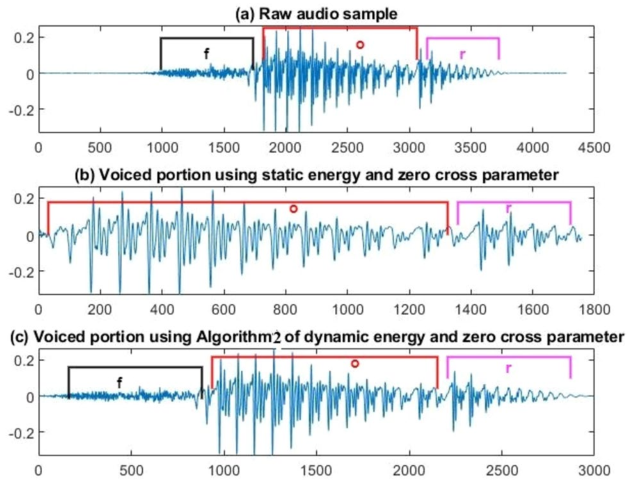


Fig. 6 **a** The raw audio wave of the sound “four”. **b** The boundary of the corresponding 6(a) using the static threshold. **c** The boundary of the corresponding 6(a) using the dynamic threshold

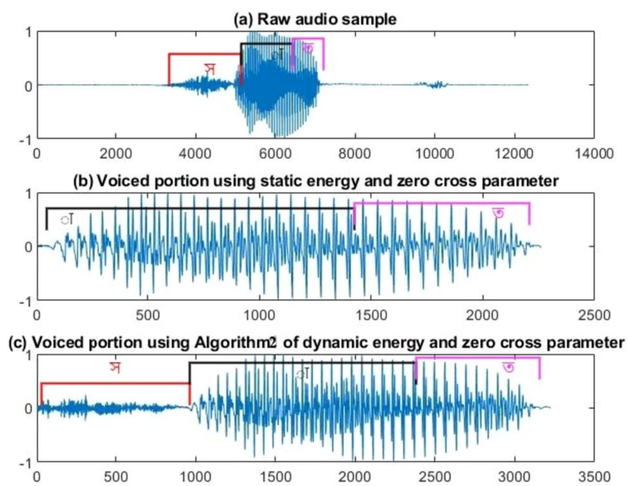


Fig. 7 **a** The raw audio sample of the utterance “সাত”. **b** The effect of selecting static energy and zero-crossing **c** The effect of selecting dynamic energy and zero-crossing by Algorithm 2

3.5 Algorithm for dynamic thresholding

The algorithm for dynamic threshold calculation is described in Algorithm 2:

Algorithm 2 (Utterance Zone y_m)

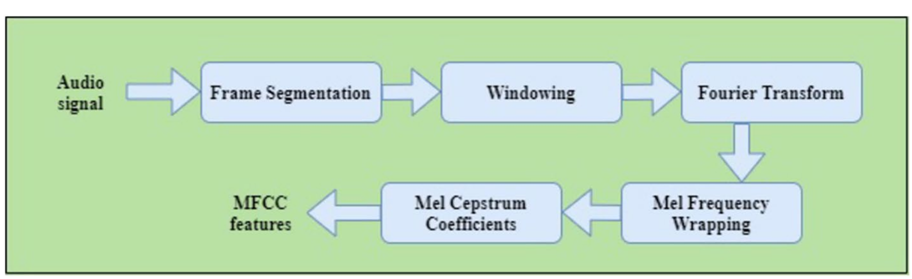
Step 1: frame_len ⟵ 0.025*fs // fs is the sampling Frequency, Set the frame length to 25ms
Step 2: N ⟵ ⌈length(y_m)/frame_len⌉ // Count the number of non-overlapping frames within the voiced region
Step 3: for each frame i
Step 4: Px (i) ⟵ average energy of the frame (i) using formula 1
Step 5: Zx(i) ⟵ average zero crossing of the frame (i) using formula 2
Step 6: end for
Step 7: mean_px ⟵ mean(Px) //Find the mean energy of the audio sample y_m
Step 8: mean_zx ⟵ mean(Zx) //Find the mean zero crossing of the audio sample y_m
Step 9: COUNT1 ⟵ Number of frames having energy > mean_px
Step 10: COUNT2 ⟵ Number of frames having zero-crossing > mean_zx
Step 11: th_px ⟵ mean_px/COUNT1
Step 12: th_zx ⟵ mean_zx + mean_zx/COUNT2
Step 13: Return (th_px, th_zx)
Step 14: End

The application of Algorithm 2 and the effect of the boundary of the audio wave “four” is shown in Fig. 6.

Figure 6a represents the raw audio wave of an utterance of the word “four”, and the phonemes boundaries are marked by the colors black, red, and pink for /f/, /o/, and /r/, respectively. Fig. 6b shows the voiced part of corresponding audio 6(a) using Algorithm 1, and Fig. 6c shows the voiced part using Algorithm 2. A clear distinction between 6b and 6c in that /f/ is cropped in Fig.6b but is visible in Fig. 6c.

Similarly, from the Bangla dataset after applying Algorithm 2, the boundary of the utterance সাত is shown in Fig. 7. The raw audio sample of the utterance “সাত” is presented in Fig. 7a. The voiced part of the corresponding audio

Fig. 8 Steps for MFCC feature extraction



is displayed in Fig. 7b by selecting the static energy and zero-crossing threshold. Finally, the voiced part of the audio is depicted in Fig. 7c by applying Algorithm 2, deriving the dynamic energy and zero-crossing threshold.

3.6 Feature extraction

An extensively adopted speech feature, MFCC, incorporating twenty-six (26) dimensions has been extracted here. Figure 8 shows how the various steps are performed to determine the MFCC feature.

Framing: Each audio sample is truncated with a duration of 25ms and an overlap of 60%, resulting in 99 frames/sec. Each frame is analyzed to find 26 MFCC features.

Windowing: Here, each frame is convolved with a Hamming window. Equation 4 [29] shows the formula for a Hamming window.

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) \tag{4}$$

In this context, w(n) denotes the signal after windowing at the n^{th} point.

Fast Fourier Transform: The conversion from the signal's time-domain to its frequency-domain components is established by Fourier transform using a fast algorithm of Discrete Fourier Transform (DFT) [30]. Equation 5 [30, 31] is used to determine the DFT of a windowed signal.

$$S(k) = \sum_{i=1}^{N} w(n)e^{-\frac{j2\pi ki}{N}} \; 1 \le k \le K \tag{5}$$

Here, K represents the DFT length, which is the nearest power of 2 greater than the window length.

We apply the formula to calculate the frame energy using the DFT values given in Eq. 6 [29, 31] for the next level of analysis.

$$P_i(k) = \frac{1}{N}|S(k)|^2 \tag{6}$$

Mel scale filter bank: In the MFCC calculation process at this point, we employ 26 overlapping triangular filter banks to transform the spectrum into the Mel scale using Eq. 7 [32, 33].

$$m = 2595\log_{10}\left(1 + \frac{f}{700}\right) \tag{7}$$

Here, m signifies the Mel frequency, and f denotes the frequency in Hz.

Discrete cosine transform (DCT): In the ultimate step, we revert the Mel frequency spectrum to the time domain by performing the DCT on the log Mel power spectrum of frame i. The mathematical expression for finding the DCT is Equation 8 [34, 35].

$$C_m = \sum_{k=1}^{M} \cos\left[m\left(k - \frac{1}{2}\right)\frac{\pi}{M}\right]E_k \tag{8}$$

For this purpose, M is set to 26, indicating the number of filter banks, and $1 \le m \le L$ denotes the allowed range for the number of MFCC coefficients.

The 26 coefficients of Cm are considered MFCC for a single frame. The dimension of the feature matrix is {number_of_frames x 26} for variable length audio. The next phase feeds this feature into a single Bidirectional Long Short Term Memory Network (BiLSTM).

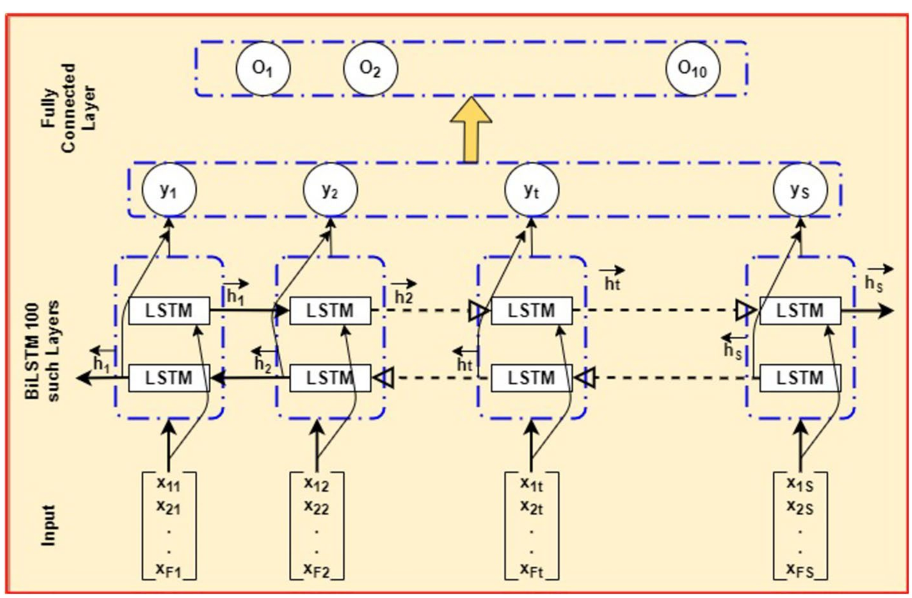


Fig. 9 Proposed architecture for the sequence to label architecture

3.7 Classifier model

When it comes to classification, out of the available classifiers, the most popular sequence-to-label classifier, a BiLSTM [36, 37, 38.], is used in the experiment with hyper-tuned parameters. The datasets are split into a training set and a test set with 80% and 20%, respectively. The proposed architecture for this classification is shown in Fig. 9. This classifier is chosen because only LSTM can classify the variable length input sequence. Other classifiers, both for machine learning and deep learning, require feature shaping. This means the feature matrix or vector must be transformed into a uniform shape before being fed into the classifiers. Uniform feature design is implemented either by a zero pad in the feature (post-process) or by splitting audio into the same utterance duration (pre-process), which requires considerable effort.

In Fig.9, the time series input (X) of the MFCC feature is fed into the input layer. Here 'F' and 'S' represent the dimension of the MFCC feature, the length of the input sequence, which is 26, and the number of audio frames, respectively. The prediction accuracy of BiLSTM is better than LSTM [36, 38]. It predicts the output from both directions of the input. BiLSTM is nothing but the combination of two LSTMs. At time 't', the output of the t^{th} BiLSTM unit 'y_t' is generated by Eqs. 9, 10, and 11 [37, 39].

$$\overrightarrow{h}_t = \sigma(\overrightarrow{W}_{xh} \cdot x_t + \overrightarrow{W}_{hh} \cdot \overrightarrow{h}_{t-1} + \overrightarrow{b}_h) \quad (9)$$

$$\overleftarrow{h}_t = \sigma(\overleftarrow{W}_{xh} \cdot x_t + \overleftarrow{W}_{hh} \cdot \overleftarrow{h}_{t+1} + \overleftarrow{b}_h) \quad (10)$$

$$y_t = \overrightarrow{W}_{hy} \cdot \overrightarrow{h}_t + \overleftarrow{W}_{hy} \cdot \overleftarrow{h}_t + b_y \quad (11)$$

In Eq. (9) $\overrightarrow{W}_{xh}$ forward input-hidden weight, $\overrightarrow{W}_{hh}$ is forward hidden-hidden weight, and $\overrightarrow{b}_h$ is forward bias. Similarly, in Eq. (10), all these are identical for the backward direction. Finally, $\overrightarrow{h}$ *and* $\overleftarrow{h}$ are combined to obtain the output 'y_t' given by Eq. (11) [40, 41].

The output of all BiLSTM units is concatenated and passed to ten fully connected layers for ten output classes of softmax activation.

4 Result and discussion

To demonstrate the novelty of the proposed method, the experiment is conducted in three different ways. First, the MFCC features are extracted from the raw audio signal.

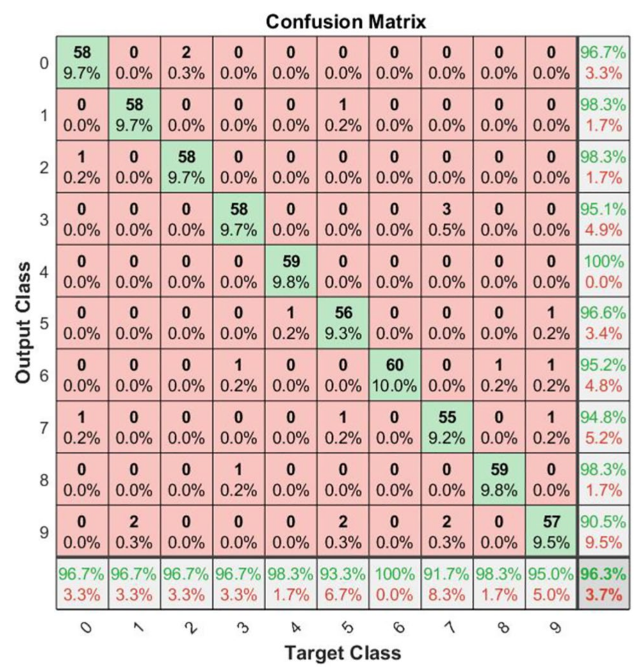


Fig. 11 Confusion matrix by applying the pre-processed MFCC on FSDD

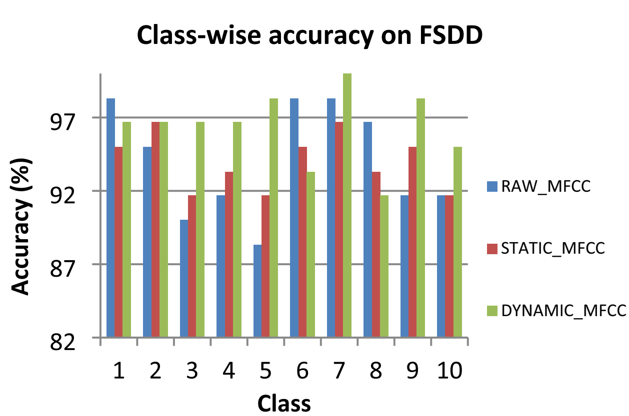


Fig. 10 Classification accuracy of recorded, cropped, and proposed pre-processed audio on FSDD

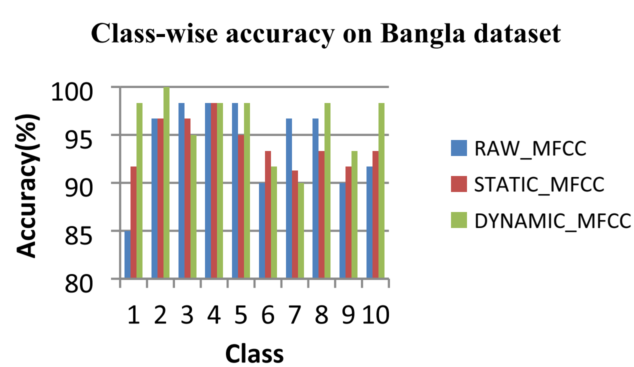


Fig. 12 Classification accuracy of recorded, cropped, and proposed pre-processed audio in the Bangla dataset

Second, the MFCC features are extracted from the Voiced Zone of the audio clip with Algorithm 1. Finally, the MFCC features are extracted from the proposed pre-processed audio using Algorithm 2. We named these different audio features RAW_MFCC, STATIC_MFCC, and DYNAMIC_MFCC, respectively. Now, these three different features have been fed to the proposed BiLSTM network with a hyper-tuned model for each dataset for classification. The experiment is performed with an Intel i5 CPU processor using MATLAB 2018a and its associated libraries. The model is trained for 100 epochs using the 'adam' optimizer; the initial learning rate is 0.001, the minibatch size is 128, and L2 regularization is used.

The average accuracy of the RAW_MFCC, STATIC_MFCC, and DYNAMIC_MFCC features on FSDD is 94%, 94.01%, and 96.3%, respectively and the class-wise accuracy is shown in Fig.10.

From Fig.10, the x-axis represents the ten output classes for the ten numeric digits from zero to nine. The y-axis indicates the percentage of accuracy for the corresponding class. Within each class, you'll find three color bars denoting accuracy—blue for RAW_MFCC, brown for MFCC of the clipped audio, and green for MFCC of the proposed pre-processed audio clips. The confusion matrix for the best result obtained by applying the proposed pre-processed technique is shown in Fig. 11.

Similarly, the average accuracy of the RAW_MFCC, STATIC_MFCC, and DYNAMIC_MFCC features on the Bangla Dataset is 94.17%, 94.13%, and 96.2%, respectively, and Fig. 12 shows the class-wise accuracy obtained with the different extracted MFCC features for the Bangla dataset.

In Fig.12, the x-axis represents the ten output classes of spoken Bangla digits শূন্য to নয়, and the y-axis indicates the percentage of accuracy for the corresponding class. For each class, there are three colored bars: blue, brown, and green, representing the accuracy obtained by entering the RAW_MFCC, the MFCC of the truncated audio, and the MFCC of the proposed pre-processed audio clips. The confusion matrix for the best result obtained by applying the proposed pre-processed technique to the Bangla dataset is shown in Fig. 13.

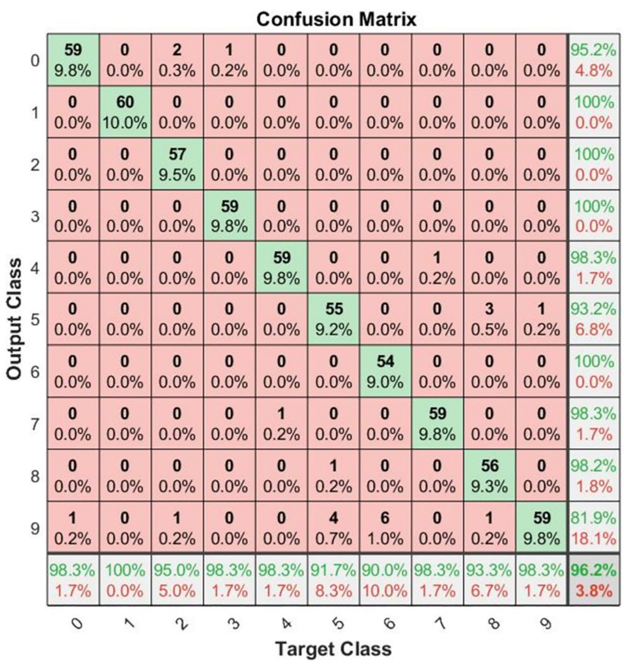


Fig. 13 Confusion matrix by applying the pre-processed MFCC on the Bangla dataset

Table 4 Evaluating accuracy in comparison to the current approach

Dataset	Author	Feature	Classifier	Result
FSDD	Mahalingam et al. [11]	MFCC	LSTM	96% testing accuracy on 2000 samples.
	Vani et al. [14]	MFCC	SVM, ANN, LSTM	96% testing accuracy for LSTM.
	A Chuchra et al. [15]	Spectrogram	CNN	93.05% testing accuracy
	M. Turab et al. [16]	Mel Spectrogram, MFCC, and ZCR	EfficientNetB0 (Pre-trained model)	99.00%
	Savitha et al. [17]	MFCC	LSTM	90.31% for training accuracy.
	Proposed Method	MFCC	BiLSTM	96.3% test accuracy
Bangla Spoken Digit Dataset	Shuvo et al. [18]	MFCC	CNN	93.65%
	Paul et al. [19]	MFCC	GMM	91.7%
	O. Sen et al. [20]	MFCC	CNN	96.7% on Cross-validation accuracy
	B.Paul et al. [21]	MFCC+ΔMFCC+ΔΔMFCC	DTW	93%
	Noman et al. [22]	DFT	ANN	95.2%
	Proposed Method	MFCC	BiLSTM	96.2%

4.1 Discussion

Based on the results, it can be analyzed that the application of the proposed pre-processing technique enhances classification accuracy. In Fig. 10 and 12, the green color bars are higher than the blue and brown bars in most classes. This means the proposed word boundary selection using dynamic thresholding selects the most accurate boundary. The average accuracy of almost three percent is improved using the proposed noise and silence zone suppression technique.

The proposed pre-processing technique has several practical advantages: audio data reduction, reduced feature extraction, noise reduction, boundary detection, etc. It saves time and storage space for classification. Another advantage of the proposed BiLSTM classifier is that the feature matrix does not need to be converted into a uniform dimension. The architecture shows how the feature matrix can be fed into the classifier with variable length (as the number of frames varies from audio to audio). Other classifiers, such as Convolutional Neural Network- Long Short Term Memory (CNN-LSTM) classifiers, significantly improve recognition rates. However, the feature must be post-processed using zero-padding or other feature-shaping techniques.

Comparative Study: In two different ways, the superiority of the proposed method has been established. First, with the novel pre-processing technique the accuracy of the pre-processed audio is enhanced compared to the raw audio signal on both datasets as given in Fig. 10 and Fig. 12. Second we have compared the isolated word recognition on two datasets: FSDD and Bangla spoken digits dataset with this study. The result is compared with some recent works (see Table 4). The works [11, 14–17] are focused on the FSDD. Although the study [16] shows high classification accuracy, however, they used EfficientNetB0 pre-trained classifier and feature-ensembling method that requires much computational cost. On the Bangla spoken digits dataset, the study [20] showed a little higher classification accuracy, however, it is cross-validation accuracy. But this study shows test accuracy.

5 Conclusion and future scope of work

A novel speech signal pre-processing is developed by deriving the dynamic thresholds of ZCR and energy. The dynamic thresholds of ZCR and energy provide correct discrimination of noise and silence in an audio signal. This increases the classification accuracy. Two algorithms are developed to determine the dynamic values. The effect of unvoiced consonants in the articulation zone is presented. The developed algorithms detect the boundary of the voiced part. The proposed pre-processing technique has been implemented in isolated word recognition in two different datasets of two different languages. In this isolated word recognition study, the MFCC features have been extracted from the raw audio samples and the modified clipped audio sample using the proposed pre-processing technique. The average classification accuracies are 94% and 96.3% on the raw and pre-processed audio samples for the FSDD; also, 94.2% and 96.2% for the Bangla dataset. The result shows the superiority of the proposed pre-processing method. The result has been compared with some recent existing works.

Although the proposed pre-processing is good enough to detect whether a frame is noise, voice, or silence. This mechanism only detects if noise is present in the signal interval. However, the technique cannot eliminate the background and random noise in the signal. Further investigation is needed to eliminate the background and random noise using novel filtering techniques. There is also a possibility of using the proposed pre-processing approach for speech segmentation in the future. So, the proposed algorithms can be applied to the continuous speech signal in the future for speech separation, segmentation, vowel onset point detection, etc.

Acknowledgments The authors would like to thank the Department of Computer Science, Vidyasagar University, for the facility of the laboratory to conduct the experiment. We would also thank the volunteers who helped with the audio data recording.

Author contributions Conceptualization, Problem Statement analysis, Methodology, and Experimental implementation: Bachchu Paul. Manuscript preparation, Language editing, Figures, Charts: Bachchu Paul, Sumita Guchhait, and Anish Sarkar. Proofreading, typesetting, Drafting: Bachchu Paul, Sandipan Maity, Biswajit Laya, Anudyuti Ghorai. Responses to reviewer's comments: Bachchu Paul, and Utpal Nandi.

Funding The authors did not receive support from any organization for the submitted work. No funding was received to assist with the preparation of this manuscript. No funding was received for conducting this study. No funds, grants, or other support were received.

Data availability The FSDD is available in the Kaggle repository from the web link: https://www.kaggle.com/datasets/joserzapata/free-spoken-digit-dataset-fsdd. The "Bangla spoken 0-99 number" dataset generated during and/or analyzed during the current study is available in the Kaggle repository from the web link: https://www.kaggle.com/datasets/piasroy/bangla-spoken-099-numbers.

Declarations

Conflict of interest The authors have no conflict of interest regarding this manuscript's preparation and submission. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethical approval Not applicable.

References

1. Paul B, Phadikar S (2023) A novel pre-processing technique of amplitude interpolation for enhancing the classification accuracy of Bengali phonemes. Multimed Tools Appl 82:7735–7755. https://doi.org/10.1007/s11042-022-13594-5
2. Koutchadé IS, Adjibi SS (2021) Explaining the english consonant sounds to efl learners: more attention on voicing dimension/l'explication des sons consonantiques anglais aux apprenants de l'anglais langue etrangere: plus d'attention au voisement. Eur J Appl Linguist Stud 3(1):12
3. https://sandiegovoiceandaccent.com/american-english-consonants/place-manner-and-voicing-of-the-american-english-consonants (Last access: 07-MAR-2023)
4. Bhowmik T, Mandal SKD (2018) Manner of articulation based Bengali phoneme classification. Int J Speech Technol 21:233–250. https://doi.org/10.1007/s10772-018-9498-5
5. Hamooni H, Mueen A, Neel A (2016) Phoneme sequence recognition via DTW-based classification. Knowl Inf Syst 48:253–275. https://doi.org/10.1007/s10115-015-0885-9
6. Hasan MR, Hasan MM, Hossain MZ (2022) Effect of vocal tract dynamics on neural network-based speech recognition: A Bengali language-based study. Expert Syst 39(9):e13045
7. Moulin-Frier C, Nguyen SM, Oudeyer P-Y (2013) Self-Organization of Early Vocal Development in Infants and Machines: The Role of Intrinsic Motivation. Front Psychol 4:1006. https://doi.org/10.3389/fpsyg.2013.01006
8. Mohanty P, Nayak AK (2022) CNN based keyword spotting: An application for context based voiced Odia words. Int. j. inf. tecnol. 14:3647–3658. https://doi.org/10.1007/s41870-022-00992-z
9. Aldarmaki H, Ullah A, Ram S, Zaki N (2022) Unsupervised automatic speech recognition: A review. Speech Commun 1398:76
10. Mahalingam H, Rajakumar M (2019) Speech recognition using multiscale scattering of audio signals and long short-term memory of neural networks. Int. J. Adv. Comput. Sci. Cloud Comput 7:12–16
11. Wu J, Chua Y, Zhang M, Li H, Tan KC (2018) A spiking neural network framework for robust sound classification. Front Neurosci 12:836
12. R Gary Leonard (1993) George Doddington. TIDIGITS LDC93S10. Web Download. Philadelphia: Linguistic Data Consortium.
13. Nayak SK, Nayak AK, Mishra S, Mohanty P (2023) Deep learning approaches for speech command recognition in a low resource KUI language. Int J Intell Syst Appl Eng 11(2):377–386. https://ijisae.org/index.php/IJISAE/article/view/2641
14. Vani HY, Anusuya MA (2020) Improving speech recognition using bionic wavelet features. AIMS Electron Electr Eng 4(2):200–215
15. Chuchra A, Kaur M, Gupta S (2022) A Deep Learning Approach for Splicing Detection in Digital Audios. In: Saraswat M, Sharma H, Balachandran K, Kim JH, Bansal JC (eds) Congress on Intelligent Systems Lecture Notes on Data Engineering and Communications Technologies. Springer, Singapore, p 543
16. Turab, M., Kumar, T., Bendechache, M., Saber, T. (2022). Investigating multi-feature selection and ensembling for audio classification. arXiv preprint arXiv:2206.07511.
17. Savitha G (2021) Deep Recurrent Neural Network Based Audio Speech Recognition System. Inform Technol Ind 9(2):941–949
18. M. Shuvo, S. A. Shahriyar, and M. Akhand, "Bangla numeral recognition from speech signal using convolutional neural network." In 2019 International Conference on Bangla Speech and Language Processing (ICBSLP). IEEE, 2019, pp. 1–4.
19. Paul B, Bera S, Paul R, Phadikar S (2021) Bengali Spoken Numerals Recognition by MFCC and GMM Technique. In: Mallick PK, Bhoi AK, Chae GS, Kalita K (eds) Advances in Electronics Communication and Computing ETAEERE 2020 Lecture Notes in Electrical Engineering. Springer, Singapore, p 85
20. Sen, O., & Roy, P. (2021, September). A convolutional neural network based approach to recognize bangla spoken digits from speech signal. In 2021 International Conference on Electronics, Communications and Information Technology (ICECIT) (pp. 1–4). IEEE.
21. Paul B, Paul R, Bera S, Phadikar S (2023) Isolated Bangla Spoken Digit and Word Recognition Using MFCC and DTW. In: Gyei-Kark P, Jana DK, Panja P, Abd Wahab MH (eds) Engineering Mathematics and Computing Studies in Computational Intelligence. Springer, Singapore, p 1
22. Noman A, Cheng X. (2022). Bengali Isolated Speech Recognition Using Artificial Neural Network. In Mechatronics and Automation Technology (pp. 14-23). IOS Press.
23. https://github.com/Jakobovski/free-spoken-digit-dataset/tree/v1.0.8 DOI https://doi.org/10.5281/zenedo.1342401
24. https://www.kaggle.com/datasets/piasroy/bangla-spoken-099-numbers
25. Ying M, Kaiyong L, Jiayu H, Zangjia G (2019) Analysis of Tibetan folk music style based on audio signal processing. J Electr Electron Eng 7(6):151–154
26. Jothimani S, Premalatha K (2022) MFF-SAug: Multi feature fusion with spectrogram augmentation of speech emotion recognition using convolution neural network. Chaos, Solitons Fractals 162:112512
27. Sasmal S, Saring Y (2023) A zero-resourced indigenous language phones occurrence and durations analysis for an automatic speech recognition system. Int J Inf Tecnol. https://doi.org/10.1007/s41870-023-01451-z
28. Biswas M, Rahaman S, Ahmadian A, Subari K, Singh PK (2023) Automatic spoken language identification using MFCC based time series features. Multimedia Tools Appl 82(7):9565–9595
29. Sasmal S, Saring Y (2023) Isolated words recognition of Adi, a low-resource indigenous language of Arunachal Pradesh. Int. J. Inf. Tecnol. 15:3079–3092. https://doi.org/10.1007/s41870-023-01339-y
30. Ai OC, Hariharan M, Yaacob S, Chee LS (2012) Classification of speech dysfluencies with MFCC and LPCC features. Expert Syst Appl 39(2):2157–2165
31. Li Qin, Yang Yuze, Lan Tianxiang, Zhu Huifeng, Wei Qi, Qiao Fei, Liu Xinjun, Yang Huazhong (2020) MSP-MFCC: Energy-efficient MFCC feature extraction method with mixed-signal processing architecture for wearable speech recognition applications. IEEE Access 8:48720–48730
32. Choudakkanavar G, Mangai JA, Bansal M (2022) MFCC based ensemble learning method for multiple fault diagnosis of roller bearing. Int. J. Inf. Tecnol. 14:2741–2751. https://doi.org/10.1007/s41870-022-00932-x
33. Koduru A, Valiveti HB, Budati AK (2020) Feature extraction algorithms to improve the speech emotion recognition rate. Int J Speech Technol 23(1):45–55
34. Sahidullah M, Saha G (2012) Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition. Speech commun 54(4):543–565
35. Paul B, Mukherjee H, Phadikar S, Roy K (2020) MFCC-Based Bangla Vowel Phoneme Recognition from Micro Clips. In: Bhateja V, Satapathy S, Zhang YD, Aradhya V (eds) Intelligent Computing and Communication ICICC 2019 Advances in Intelligent Systems and Computing. Springer, Singapore, pp 511–519
36. Shashidhar R, Patilkulkarni S, Puneeth SB (2022) Combining audio and visual speech recognition using LSTM and deep convolutional neural network. Int. J. Inf. Tecnol. 14:3425–3436. https://doi.org/10.1007/s41870-022-00907-y

37. Ihianle IK, Nwajana AO, Ebenuwa SH, Otuka RI, Owa K, Orisatoki MO (2020) A deep learning approach for human activities recognition from multimodal sensing devices. IEEE Access 8:179028–179038
38. Shah SRB, Chadha GS, Schwung A, Ding SX (2021) A sequence-to-sequence approach for remaining useful lifetime estimation using attention-augmented bidirectional lstm. Intell Syst Appl 10:200049
39. Thakur A, Dhull SK (2022) Language-independent hyperparameter optimization based speech emotion recognition system. Int. J. Inf. Tecnol. 14:3691–3699. https://doi.org/10.1007/s41870-022-00996-9
40. Girirajan S, Pandian A (2022) Acoustic model with hybrid Deep Bidirectional Single Gated Unit (DBSGU) for low resource speech recognition. Multimedia Tools Appl 81(12):17169–17184
41. Oruh J, Viriri S, Adegun A (2022) Long short-term memory recurrent neural network for automatic speech recognition. IEEE Access 10:30069–30079