Electronics and Electrical Engineering

AIMS Electronics and Electrical Engineering, 4(2): 200–215.
DOI: 10.3934/ElectrEng.2020.2.200
Received: 18 February 2020
Accepted: 14 April 2020
Published: 27 April 2020

http://www.aimspress.com/journal/ElectrEng

Research article

Improving speech recognition using bionic wavelet features

Vani H Y[1,*] and Anusuya M A[2]

[1] Department of Information Science & Engg., JSS Science & Technology University, Mysore, Karnataka, India
[2] Department of Computer Science & Engg., JSS Science & Technology University, Mysore, Karnataka, India

*** Correspondence:** Email: vanihy@jssstuniv.in; Tel: +919964286326.

Abstract: Bionic wavelet transform is a continuous wavelet, based on adaptive time frequency technique. This paper presents a speech recognition system for recognizing isolated words by discretizing the continuous Bionic Wavelet (BW). Conversion from continuous to discrete is achieved by adopting central frequency and thresholding techniques. The BW features of noisy signal are processed through MFCC to obtain the optimal features of the speech signal. SVM, Artificial Neural Network (ANN) and LSTM techniques are used to improve the recognition rate by enhancing the speech signals. The experiments are conducted on FSDD and Kannada data set. The speech feature vector is calculated using the parameters extracted by Bionic wavelet with different central frequencies of Morlet, Daubechies and Bior3.5, coiflet5 mother wavelets. The obtained Bionic-MFCC optimal features are fed to SVM, ANN and LSTM models for the classification and recognition process. The performance of the models is tabulated for correct recognition that varies from 95% to 96% among these models. The models are tested for various SNRs noise levels like 5 dB, 10 dB, 15 dB and the recognition accuracies of these models are presented for convoluted noisy speech data.

Keywords: bionic wavelet transform (BWT); speech recognition; Bionic-MFCC; wavelet transform (WT); support vector machine (SVM); artificial neural network (ANN); long short term memory (LSTM); continuous wavelet transform (CWT); discrete wavelet; Morlet wavelet; adaptive thresholding; center frequency; T-function

1. Introduction

One of the most important branches of speech processing is enhancing the speech recognition for noisy signals i.e. speech enhancement, speech recognition etc.. Reducing noise from a speech signal is very complex process. The main objective of speech enhancement is to find the optimal estimates of speech features. To obtain efficient feature, wavelet transforms are most useful because it is one of the most prominent technique to analyze the non stationary speech signals in both time and frequency domains in a better way.

Using wavelets [1], the noise can be reduced by appropriately selecting the wavelet coefficient threshold. These threshold values are subtracted from the noisy wavelet coefficients to obtain a noise reduced signal. Since features are computed in scalograms the obtained features are more prominent than the features obtained from short term Fourier transform technique.

In wavelet transforms there are two types: Continuous and Discrete wavelet transforms.

Discrete wavelet transform decomposes the signal into approximation and detail components by shifting and scaling the copies of the basic wavelet to a required level. BWT is proposed and used in the present work because, it resembles the auditory model of human cochlea [2–7] and it can be easily correlated with the MFCC feature extraction process. This helps in extracting the prominent features of the noisy speech signal.

CWT is used to obtain simultaneous time frequency analysis. It is preferred because it is based on Auditory model of Human Cochlea [2–7].

In this paper, we propose the optimal feature selection procedure using BWT and MFCC procedures for convoluted noisy speech data for recognizing words. To calculate the optimal features mother wavelet's central frequencies of Morlet [7], Daubechies, Bior, Coiflet wavelets are adapted to BWT with thresholding and central frequency techniques.

Thresholding on BWT is calculated using the following selection methods. They are [8]: i) Stein's unbiased estimate of the risk rule (SURE), ii) heuristic threshold selection rule, iii) fixed selection rule, iv) minimax v) sqtwolog threshold. To handle noise in the signal SURE threshold selection procedure has been adopted to BWT to estimate the recognition accuracy.

The contents of the paper is organized as follows: Section 2 discusses about the works carried out in literature using bionic wavelets. Section 3 provides introduction to continuous bionic wavelet. Section 4 presents the procedure adopted for converting the continuous wavelet to discrete wavelet. Section 5 discusses about the data set used for the experimentation purpose. Proposed system model is discussed in section 6 with results. The performance analysis of different classifier is discussed in section 7. Section 8 presents observations done during the simulation process. Last section discusses about the conclusion and future enhancements.

2. Literature survey

Extracting optimal feature plays a major role in classification and or recognition. However, many studies shows that bionic with Morlet wavelets are used for de-noising the speech signal by enhancing the signal component. At present the features can be extracted at three methods 1) Features from Time Domain, 2) Frequency Domain Features, 3) Features from Raw wave file. MFCC is the most popular method in frequency domain and the last method is now gearing up in the machine learning models. MFCC is well suited for clean speech signal but making it more robust for noisy data is also presented in this paper. In this direction the bionic wavelets are used for de-noising and the MFCC is made robust towards handling convoluted noisy speech data.

Bionic wavelet is made adaptive by applying various methods viz, by changing the 'K' factor, using different hard/soft thresholding methods and applying various base/central frequencies. The following are some of the related work towards the application of bionic wavelets used for denoising the speech data. A. Garg & O. P. Sahu [9] proposed a method to discretize bionic wavelet using CWT and ICWT using Morlet as the mother wavelet.

Fie Chen [10] proposed adaptive DBWT by changing T-function of BWT and splitting the dyadic tiling map of DWT that uses quadrature-mirror filters, organized as DBWT tiling map for decomposition. M. Talbi [11] proposed entropy technique to BWT to identify the two sub bands having minimal entropy for each coefficient.

Cao Bin-Fang [12] proposed a bionic wavelet method of hierarchical threshold based on PSO. The noisy speech signal is decomposed using bionic wavelet transform. In this Particle Swarm Optimization is proposed for threshold optimization. The noise with high frequency is separated by bionic wavelet transform and this is fed as an input to an adaptive filter. From the experimental work the paper illustrates speech enhancement for various SNR conditions.

A detail analysis is made by Yang Xi, Liu et al. to understand the behavior of bionic wavelet with additive noise for various db's. It clearly explains the usage of bionic with Morlet as a mother wavelet for removing various db level noises from a speech signal. Yao and Zhango proposed an adaptive bionic with a Morlet wavelet base frequency "ωo" of mother wavelet 15165.4 Hz that is suitable for human auditory system.

Mourad used [13] MSS-MAP for wavelet transform and used four different test such as SNR, segmental SNR, Itakura and perceptual evaluation for various types of noises and their levels. A new speech enhancement procedure is proposed by WU Li-ming [14] on improved correlation function processing for Bionic wavelet co-efficient.

Speech recognition for Arabic words is demonstrated in Ben-Nasr [15]. Feature extraction is done by using MFCC with bionic wavelet. To increase the recognition rate Delta-Delta coefficients are used and classification is done by using feedforward back propagation neural network. Zehtabian [16], proposed speech enhancement technique using BWT and singular value decomposition method. The paper illustrates SVD is better than BWT for higher SNR's.

Liu Yan [17], proposed de-noising algorithm on sub band spectrum entropy with bionic wavelet transform. They showed that sub band spectrum is good in detecting the end point of the speech signal. Hence it is used to distinguish speech as well as noise. The experimental work demonstrate sub band entropy de-noising method is superior than Wiener filter algorithm. Pritamdas [18] focus on continuous wavelet transform and thresholding of coefficients for speech enhancements using thresholds and wavelet transform scales in adaptive manner.

From the literature survey, it is observed that a lot of work is reported on Bionic wavelet for speech enhancement with thresholding and rescaling procedures used for converting continuous to discrete wavelet co-efficient's for additive noise only. In this paper, procedure to convert continuous to discrete wavelet based on the central frequency is proposed. New feature extraction technique and the procedure to reduce the noise of convoluted noise is presented.

To the best of our knowledge this work is unique in its own way for de-noising the convoluted noise at various levels. The next section describes the characteristics of Bionic wavelet.

3. Continuous bionic wavelet

Alternative to STFT, is the WT technique [19–22]. When these two are compared visually, The scalograms of WT are better in representing the formant frequencies and structural harmonics of

speech. Hence WT technique is identified as one of the prominent method to handle non stationary signals. CWT is fixed with some base scale [23] that is $2^{1/m}$ where m is an integer greater than 1. Where 'm' is the number of "voices per octave". Different scales are obtained by raising this base scale to positive integer numbers, for example $2^{k/m}$ where k = 1,2, 3…. The translation parameter in the CWT is discretized to integer values, represented by l. The resulting discretized wavelets for the CWT is represented by Eq. 1

$$\frac{1}{2^{\frac{k}{m}}}\Psi\frac{(n-l)}{2^{\frac{k}{m}}} \tag{1}$$

3.1. Bionic wavelet

Bionic wavelet transform (BWT) is an adaptive wavelet transform based on a model of the active biological auditory system [24]. The decomposition of BWT [2] is perceptually scaled and adaptive. It has the following properties:

i) High sensitivity and selectivity
ii) Signal with determined energy distribution
iii) Can be reconstructed

The resolution of bionic wavelet transform can be achieved by adjusting signal frequency and the instantaneous amplitude with its first order differential values.

4. Realization of discrete bionic wavelet from continuous

This section discusses about the mechanism adopted to convert continuous wavelet to discrete wavelet. To convert any continuous to discrete wavelet the discrete thresholding and central or base frequencies of different mother wavelets are adopted.

4.1. Center frequency [25]

Db11, Coif 5 and Bi-ortho3.5 wavelets are considered with central frequencies – 0.67, 0.68, 1.04 Hz respectively. The center frequency is calculated using centfrq of Matlab.

$\omega_m = {}^{\omega_0}/_{(1.1623)^m}$, m varies from 1 to 22 for Morlet. For other wavelets centfrq function of Matlab is used.

All the wavelets possess different characteristics, hence the following four wavelets are considered

Db11: asymmetric, orthogonal, bi-orthogonal.
Coif 5: symmetric, orthogonal, bi-orthogonal.
Bior3.5: symmetric, not-orthogonal ,bi-orthogonal.

22 scales are considered for BW in spite of center frequency .These wavelets are preferred because they mimics the mel-scale mapping of the MFCC [26] procedure and also these are designed to match the basilar membrane spacing i.e. based on nonlinear perceptual model of the auditory system.

4.2. Thresholding

This parameter decides about the number of levels used to reduce the redundant information in the CWT towards the discretisation of the wavelet. The following thresholding mechanisms are considered with various levels by trial and error procedure as listed below in the Table 1. Levels are fixed based on the obtained thresholds of the signal. The various ways of calculating the thresholding is as discussed below:

Sqtwolog:

$$thr = \sigma_k \sqrt{2 * \log(p)}$$

where σ is the mean absolute deviation (MAD) and p is the length of the noisy signal. MAD is expressed as

$$\sigma_k = \frac{MAD_k}{0.6745} = \frac{median|\omega|}{0.6745}$$

ω wavelet coefficient and k-scale for wavelet co-efficient

Rigrsure

$$th_k = \sigma_k \sqrt{\omega_c}$$

ω_c is the cth coefficient wavelet square (coefficient at minimal risk) chosen from the vector

ω = [ω1, ω2, .., ωc]

σ is the standard deviation of the noisy signal.

Heursure: Heursure threshold selection rule is a combination of Sqtwolog and Rigrsure methods.

Minimaxi:

$$th_k = \begin{cases} \sigma(0.3936 + 0.10829 \log_2 M, & M > 32 \\ 0, & M < 32 \end{cases}$$

ω: a vector of wavelet coefficients in units scale and M: vector length of the signal.

Table 1. SNR for various thresholding levels.

Sl. No	Thresholding	No. of levels based on the Thresholding	SNR before	SNR after
1	SURE.	2	-12.52	-3.3
2	Heuristic variant	5	-12.52	-11.96
3	Sqtwolog	4	-12.62	-11.96
4	Minimaxi	4	-12.52	-10.24

Algorithm

Steps for Discretizing Bionic Wavelet.

Step 1: Read the speech signal

Step 2: Multiply each value by ‘K’ as shown in Eq. 2

$$BWT_f((a,\tau) = K * WT_f((a,\tau) \quad (2)$$

Step 3: Thresholding function is selected with high SNR using Matlab function thselect.
Step 4: Base/central frequencies of various mother wavelets is applied using centfrq (wname).
Step 5: The modified bionic wavelet coefficients are divided by the 'K' factor to get the coefficients and reconstruction is done by taking its the inverse continuous wavelet transform. Where the 'approximation is done by K'-factor using Eq. 3

$$\frac{1.7772T_0}{\sqrt{T^2+1}} \quad (3)$$

Step 6: Compute the inverse continuous transform
Step 7: Obtain the Mel frequency Cepstral co-efficient for the de-noised signal [26]
Step 8: The Bionic-MFCC features obtained are listed as shown below
Step 9: Classify the same features using SVM, ANN and LSTM classifiers
Following tables shows the sample features are better than bare MFCC features.
Noisy speech Signal: ***The following table presents only the MFCC features without wavelet for noisy signal Co-efficient from MFCC without wavelets.***

3.58	-2	-2.1	-1.88	-1.28	-0.7	0.015	-0.2	-0	-0.3	-0.5	-0.03
3.14	-2.2	-2.7	-2.14	-0.23	0.63	-0.27	-0.7	-0.3	-0.2	-0.2	-0.032
3.21	-0.8	-2.5	-2.1	-0.7	-0.1	-0.35	-0.5	-0.1	-0.2	-0.2	0.0452
3.17	-3	-3.2	-1.22	-1.27	-0.9	-1.86	-0.8	-1	-0.7	-0.2	-0.273

Coefficients after applying Wavelets

4.60	**2.67**	**1.89**	**1.21**	**0.63**	-0.81	**0.25**	-0.24	-0.18	-0.53	**0.26**	-0.53
4.08	**2.74**	**1.87**	**1.11**	**0.74**	-0.59	0.28	-0.19	-0.03	-0.50	0.14	-0.57
4.52	**2.78**	**1.92**	**1.21**	**0.74**	-0.69	0.30	-0.17	-0.11	-0.55	0.19	-0.56
5.94	**6.20**	**0.19**	-2.79	**-1.18**	-0.66	**-1.86**	-0.99	-2.15	-1.15	**0.04**	-0.62

Clean speech Signal

The following table presents only the MFCC features without wavelet for clean signal

2.92	1.1	1.68	1.556	1.36	0.5	0.814	0.16	0.1	-0.6	-0.4	-0.2386
3.33	0.91	0.71	0.943	1.43	0.99	-0.08	-0.6	-0.2	-0.1	-0.1	-0.1655
3.5	0.84	0.43	0.607	1.27	0.79	0.243	-0.5	0.09	0.07	0	-0.2478
-4.9	-1	-2.5	-1.23	-1.8	-0.1	-1.05	-0.9	-0.6	-1	-0.7	-0.5966

Coefficients after applying Wavelets

3.22	0.39	**2.12**	1.349	**1.46**	0.47	**0.826**	**0.17**	0.09	**-0.5**	-0.4	**-0.2157**
3.33	0.55	**0.85**	0.922	1.39	**1.04**	**-0.05**	**-0.6**	-0.2	**-0.1**	-0.1	**-0.1344**
3.55	0.41	**0.62**	0.579	**1.28**	0.86	0.178	**-0.4**	0.05	**0.11**	-0	**-0.2232**
0.01	**0.23**	**-2.1**	-0.71	**-1.5**	0.08	-1.14	-0.9	-0.6	-0.9	-0.7	-0.5968

The above presents the weighted features obtained from step 1 to step 8 of the algorithms. From this it is clear that wavelets weighted feature values are better for both for clean and noisy speech signal.

5. Data set

Two different datasets considered are free spoken digit dataset (FSDD) [27] and Kannada dataset (Table 2) with recordings of spoken digits and words sampled at 8 kHz and 16 kHz respectively. The recordings are trimmed, so that they have near minimal silence at the beginnings and ends. It consists of English pronunciation words of numbers from one to nine from four different speakers. Totally 900 signals with 100 signals of each digit is collected. The second data set is isolated words Kannada data set. The words considered are as shown in Table 3. These signals are sampled at 16 KHz frequency consisting of 30 speakers with 20 male and 10 female speakers. 1000 word samples are collected from both genders for Kannada data set. The signals are artificially convoluted with street noise [28] with the SNR of 5, 10 and 15db to create convoluted noisy speech signals.

Table 2. Kannada dataset.

Slno	Kannada Word	Slno	Kannada Word
1	Kannada	21	Habba
2	Namaskara	22	Alli
3	Pusthaka	23	Utsava
4	Oota	24	Hogu
5	Nale	25	Niru
6	Nanu	26	Tarakari
7	Neeru	27	Uppu
8	Naalku	28	Balagade
9	Aaaru	29	Edagade
10	Aidu	30	Munde
11	Olage	31	Edina
12	Purva	32	Habba
13	Raathri	33	Horage
14	Samvidhana	34	Pustaka
15	Tarakari	35	Halu
16	Udda	36	Hinde
17	Ondu	37	Dayavittu
18	Eradu	38	Paschima
19	Muru	39	Daxina
20	Nalku	40	Uttara

Table 3. English dataset.

Sl. No.	English
1	One
2	Two
3	Three
4	Four
5	Five
6	Six
7	Seven
8	Eight
9	Nine

6. System model for the proposed approach

The obtained features are modeled for classification and recognition using machine learning models like SVM [29–32] ANN [15,33] and LSTM [34,35] in the proposed work. The overall data flow diagram of adopting all the models is as shown below Figure 1.

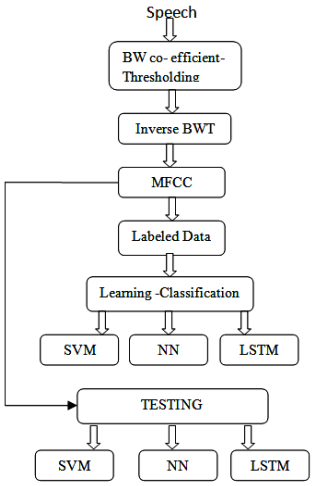

Figure 1. Flowchart of the proposed work.

General experimental setup:

The obtained features of all the signals are grouped into training and testing samples. These signals are convoluted with 5 db, 10 db, 15 db street noise [28]. The same data set is used by all the models for testing and training purpose to evaluate the recognition accuracies performance of all the models. The results are discussed at two levels namely, i) signal to noise ratio before and after the application of bionic wavelet ii) Recognition accuracies of the models compared with the existing models if any.

6.1. Signal to Noise Ratio (SNR) [36]

It is a best indicator for identifying noise interference in a given signal. SNR is computed using the following formulas.

$$snr_{before} = \frac{mean(orgsignal)^2}{mean(transpose(noise).^2}$$

$$snr_{before_{db}} = 10 * log10(snr_{before}) \quad \% \text{ in dB}$$

$$snr_{after} = \frac{mean(enhancesp.^2)}{mean(redidual\_noise.^2)}$$

$$snr_{after_{db}} = 10 * log10(snr_{after}) \quad \% \text{ in dB}$$

The Table 4 presents the application of different central frequencies to bionic wavelet to reduce the noise levels. It is clear that an average of 2db of noise is reduced.

Table 4. SNR for various central frequency with their mother wavelets.

Slno	dB	SNR Before	SNR after applying			
			Morlet	Db11	Coif5	Bior3.5
			sqtwolog			
1	5	-24.34	2.047	3.314	2.9	2.14
2	10	-19.3604	2.035	3.32	2.88	2.15
3	15	-14.24	2.06	3.35	2.90	2.16
			heursure			
4	5	-24.34	2.10	3.71	2.9	2.14
5	10	-19.3604	2.04	3.343	2.8	2.15
6	15	-14.24	2.06	3.345	2.9	2.16
			minimaxi			
7	5	-24.34	1.88	4.05	3.25	1.79
8	10	-19.3604	1.86	4.04	3.21	1.76
9	15	-14.24	1.83	4.025	3.16	1.74

Table 5 depicts the application of bionic wavelets for convoluted noise considering the 22 scales as mentioned in the literature.

Table 5. SNR for convoluted noise using bionic 22 scales.

Slno	dB	SNR Before	SNR After
1	5	-24.34	0.004379
2	10	-19.25	0.0117
3	15	-14.40	0.0363

Comparing Table 4 and 5 SNR level is better for convoluted noise. Hence noise reduction is better in Table 4 than in Table 5.

7. Performance analysis of various classification methods

In our earlier works [36,37], the experiments were carried no clean and noisy speech data set with normal MFCC features. The current feature extraction procedure applies bionic wavelets for extracting better features for the dataset specified in section 3. Hence, In this paper the new Bionic-MFCC features are used for the recognition purpose by reducing the noise using discrete bionic wavelets. Experiments are performed on standard benchmark dataset (FSDD) and Kannada dataset. The various models and their parameters used are as follows:

7.1. Support vector machine

Since the speech features are non-linear in nature, the features need to be mapped to high dimensional space. The basic idea is that the input space need to be mapped into a high dimensional feature space by nonlinear transformation and the optimal hyper plane is found in the new space. The optimal hyper plane not only needs to ensure that different categories can be discriminated correctly, but also the maximum categorization interval between them should be promised. Thus, the generalization capability of the support vector machine is stronger. The target function corresponding to the nonlinear separable support vector machine is given by:

$$\min\left(\frac{1}{2}\boldsymbol{\omega}^T\boldsymbol{\omega}+C\sum_{i=1}^{N}\xi_i\right)$$

$$\text{s.t.}\quad y_i\left(\boldsymbol{\omega}^T\boldsymbol{x}_i+b\right)\geq 1-\xi_i,\xi_i\geq 0,i=1,2,\ldots,N$$

where ω represents the weight coefficient vector, and b is a constant. C denotes the penalty coefficient to control the penalty degree for misclassified samples and balance the complexity of the model and loss error. ξi represents the relaxation factor to adjust the number of misclassified samples that allowed exit in the process of classification.

When SVM is used to solve the classification problems, two strategies can be adopted. One is ONE-TO-ALL, and ONE-TO-ONE. In this paper ONE-To-ALL method is applied for multi-classification. Kernel functions are also the key functions for SVM. Hence, polynomial and radial basis kernel functions are considered.

Table 6 and Figure 2 depict the recognition performance using SVM model. To implement SVM RBF(r) and polynomial kernel (p) functions are used. It is observed that Bionic-MFCC features, well classifies the noisy signal compared to clean speech proposed using Bionic-MFCC features [30]. SVM performs better with RBF kernel function for standard data set. Whereas, as it fails for

Kannada data set. Polynomial function performs better for Kannada data set as shown in Figure 2. From this it identifies that the kernel performance depends on the data set.

Table 6. Classification accuracy of SVM.

Slno	Center/Base frequency	SVM:Recognition accuracy with various SNR for											
		English						Kannada					
		5dB		10dB		15dB		5dB		10dB		15dB	
		r	p	r	p	r	p	r	p	r	p	r	P
1	Morlet(0.7958)	95	94	96	95	96	95	88	87	89	87	90	89
2	Db11(0.67)	96	95	96	95	96	95	89	88	90	88	91	90
3	Coif5(0.689)	96	95	96	95	96	96	89	88	90	88	91	90
4	Bior3.5(1.004)	91	90	91	90	91	91	80	88	80	89	89	89
5	MFCC –car noise[39]	77.58		82.27		87.72		Not available in the literature					

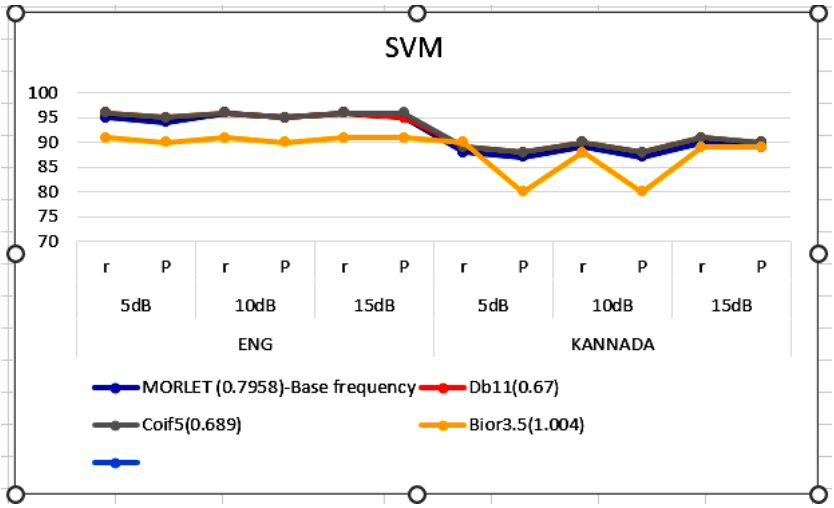

Figure 2. Classification accuracy of SVM.

7.2. Neural network

In the literature bionic wavelets are applied with Morlet base frequency for additive noisy Arabic speech recognition system [15,34,38] using NN. Hence in this paper bionic wavelets are tried for convoluted noisy speech data to identify the level of noise reduction and feature weights for recognition accuracy. Standard dataset has good recognition rate compared to Kannada data set. Less performance is due to the variable word length and existence of ambiguity in the utterance of the speaker. The Table 7 and Figure 3 show the recognition accuracies obtained.

NN Implemented Procedure:

Neural network model has 9 nodes, each with 12 bionic MFCC features at the input layer. Two hidden layers are considered with 9 nodes at the output layer representing each word. The output

layer has 9 nodes with one node for each digit. Softmax activation function is applied on the top of the network to get output class label probabilities. The model is optimized by adam-delta optimizer that adapts learning rate by moving window.

Learning is continued and network is learnt for all updates. The model is constructed and categorical cross entropy is used for multi classification.

Table 7. Recognition accuracy of NN.

Sl.no	Central frequency of different wavelets **(Proposed method)**	**NN: Recognition accuracy with various snr with**						**MLP**		
		English			**KANNADA**			**Arabic SR by BWT**		
		5dB	**10dB**	**15dB**	**5dB**	**10dB**	**15dB**	**5dB**	**10dB**	**15dB**
1	MORLET (0.7958)-Base frequency	94	95	95	81	83	89	78.7	93.9	95
2	Db11(0.67)	95	94	95	83	85	90	No information available		
3	Coif5(0.689)	94	94	95	83	85	90			
4	Bior3.5(1.004)	90	90	91	79	82	85			

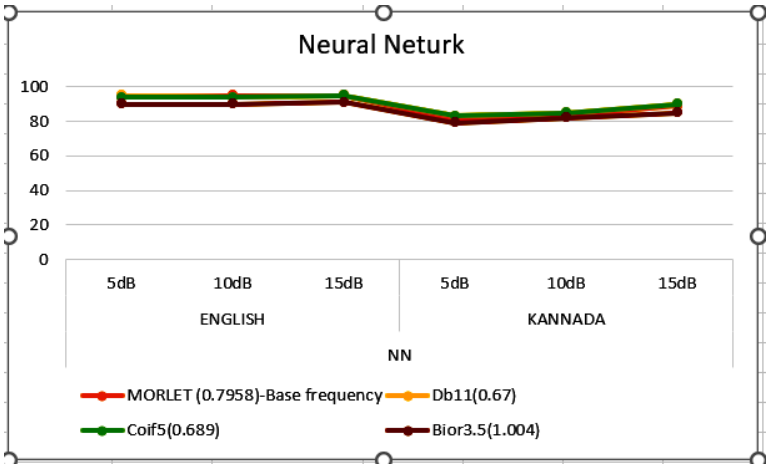

Figure 3. Recognition accuracy of NN.

7.3. LSTM

Procedure: The MFCC features are fed to the input-layer to build basic LSTM Cell. Wrapping of each layer in a dropout layer is considered with 0.5 probability value, for learning in each iteration. A group of dropout wrapped LSTMs are fed to a MultiRnn cell to group the layer together.

The CTC model helps to learn for labeling a variable –length sequence when the input-output arrangement is not known. Consider the features m = (m_1, m_2,….m_T) and the label n = (n_1, n_2,….n_U). The CTC is trained based on maximum probability. The loss function of the CTC model is computed as

$$l_{CTC} =\triangleq -lnP(n|m) \approx= ln \sum_{\pi \epsilon \Phi} \prod_{t=1}^{T} P(k = \pi_t|m)$$

The label sequence π is all expanded possible CTC path alignments Φ having length T
P(k = πt|m) is a label distribution at time step t.

Finally, stacked LSTM layers are embedded. The CTC [39,40] loss function and Adam-delta optimizer functions are used to define the model to create a single fully connected layer with

SoftMax activation function to get the labeled predictions. The activation function is as given below:

$$P_t(k|m) = \frac{exp(h_t^L(k))}{\sum_{i=1}^{K+1} exp(h_t^L(i))}$$

The Ada-delta optimizer is considered to minimize the loss by feeding the predictions to mean squared error loss function. Accuracy metric is used for training and testing process. The predicted values minimized with errors using mean squared error and Adam-delta optimizer .Then at the end accuracy metric is used for training and testing.

In the literature, works are carried out using Bi-LSTM and LSTM model for speech classification [34] with 95% and 96.58% of accuracy for clean speech signal. T. Goehring, et al. [41] uses recurrent neural network model for feature extraction for Babble noise for 5 dB and 10 dB with a recognition accuracy of 78% and 82% as illustrated in Table 8.

Whereas in the proposed work LSTM model is applied to convoluted noisy speech data and the performance of the model is shown in Table 8 and Figure 4, demonstrates better results than identified in the literature. Using Bionic-MFCC features recognition accuracy is improved by 1% compared to Bi-LSTM model for speech data. Among SVM ANN, and LSTM models, LSTM is better in modeling the convoluted speech data using db11 mother wavelet.

Table 8. Classification accuracy of LSTM.

Slno	Center frequency	LSTM:Recognition accuracy with various SNR for					
		English			Kannada		
		5dB	10dB	15dB	5dB	10dB	15dB
1	Morlet(0.7958)	91	95	95	81	83	89
2	Db11(0.67)	92	96	96	83	85	90
3	Coif5(0.689)	92	96	96	83	85	90
4	Bior3.5(1.004)	92	96	96	83	85	90
5	MFCC–Clean[40]	Clean -96.58			Not Availbable in the literature		
6	RNN(Feature Extraction)[41] Babble Noise	78	82	-			

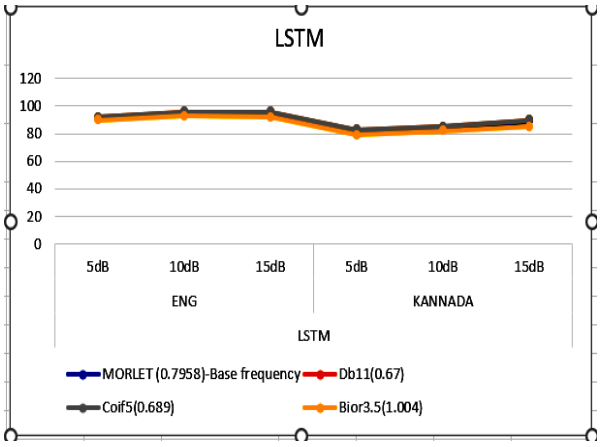

Figure 4. Classification accuracy of LSTM.

Performance measures:

Word classification error rate is computed by

$$\text{Classification Accuracy} = \frac{\text{No. of correctly classified audio}}{\text{Total No. of audio files files}}$$

$$\text{Classification Error rate} = \frac{\text{No. of incorrectly classified audio}}{\text{Total No. of audio files files}}$$

8. Observations and discussions

This section discusses about the observations done on the models used for the classification and recognition purposes.

SVM:

- In SVM the classification rate can be improved by applying different normalization methods.
- SVM performance varies with the choice of kernel function
- Non-linear SVM kernels are well suited for classification of speech data

ANN:

- Recognition accuracy can be increased by using large data set and the selection of appropriate optimizer function
- Increasing the number of hidden nodes improves the learning phase

LSTM:

It works on par with ANN, except the proper choice of the CTC loss function. The suitable selection of cost function will also help us to yield the good recognition rate. LSTM requires less features than SVM and ANN to model the data.

In general, SVM and ANN equally perform well compared to other model but not as good as LSTM. This is due to the optimality of the features obtained by the weighted values from Bionic-MFCC features. The results of LSTM model on FSDD dataset is better with db11 compared to other models because of fine-tuned dataset of FSDD. The results for db11 wavelet for 15 db is better because of high signal to noise ratio of noisy data.

9. Conclusion and future enhancements

In this work the discretization procedure of continuous bionic wavelet has been proposed for convoluted noisy speech recognition. The obtained bionic wavelet features are used for reducing the noise level in the speech data. These features are also used in MFCC to obtain the Bionic-MFCC speech features. It also presents the improvement of MFCC features using continuous wavelets. From the obtained results of the models it is clear that LSTM with DB11 wavelet at 15dB SNR outperforms. It is also observed that the recognition accuracies depend on the nature of dataset also.

It is a unique work of applying continuous bionic wavelet for feature extraction using the central frequencies of DB11, coif5 and Bior-3.5 wavelets for convoluted noisy speech data. This work also demonstrates that, even basic mother wavelets features can also be adopted in converting the continuous to discrete wavelets. It is very tedious to handle convoluted noisy speech data because of overlapping and the identification of the frequency of noise with the original data (convolution of

signal and noise). According to our study, the additive noise can be completely removed by using filters but not convoluted noise. Hence this approach is towards reducing the noise using continuous wavelet for the isolated word recognition. As per the study, LSTM model better classifies and improves the recognition accuracy up to 96% with 4% of word error rate than other models. Hence bionic wavelet well sustains and it can be made adaptive in nature, by applying various thresholding concept. From this study it is also observed that central frequency and the thresholding concept plays a major role in noise reduction as well as in the conversion of continuous to discrete wavelet. For Kannada dataset, word error rate is high because of variation in speaker's pronunciations. Whereas FSDD has good recognition rate because of its fine-tuned dataset.

Future enhancements:

In spite of thresholding, genetic algorithms can be adopted for feature reduction. Other wavelets central frequencies can also be tried for discretization of the wavelets. The performance of the above models can be verified for different types of noises for various noise levels to identify SNR. The model performances can also extend to sentence level recognition. The DWT trees can also be used for speech enhancement by noise reduction.

Acknowledgment

We are thankful to all the persons who helped in formulating this paper. The authors remain grateful to Dr. S.K. Katti for all his support.

Conflict of interest

All authors declare no conflicts of interest in this paper.

References

1. Donoho DL (1995) De-noising by soft-thresholding. *IEEE T Inform Theory* 41: 613–627.
2. Yao J, Zhang YT (2001) Bionic wavelet transform: A new time–frequency method based on an auditory model. *IEEE T Biomed Eng* 48: 856–863.
3. Yao J, Zhang YT (2002) The application of bionic wavelet transforms to speech signal processing in cochlear implants using neural network simulations. *IEEE T Biomed Eng* 49: 1299–1309.
4. Yuan XL (2003) Auditory Model-based Bionic Wavelet Transform for Speech Enhancement. A thesis submitted to the graduate school in partial fulfillment.
5. Gold T (1948) Hearing II: The physical basis of the action of the cochlea. *Proc Roy Soc B-Biol Sci* 135: 492–498.
6. Xi Y, Bing-wu L, Fang Y (2010) Speech enhancement using bionic wavelet transform and adaptive threshold function. *Second International Conference on Computational Intelligence and Natural Computing* 1: 265–268.
7. Cohen MX (2019) A better way to define and describe Morlet wavelets for time-frequency analysis. *Neuroimage* 199: 81–86.
8. Valencia D, Orejuela D, Salazar J, et al. (2016) Comparison analysis between rigrsure, sqtwolog, heursure and minimaxi techniques using hard and soft thresholding methods. *XXI Symposium on Signal Processing, Images and Artificial Vision (STSIVA)*, 1–5.

9. Garg A, Sahu OP (2019) A hybrid approach for speech enhancement using Bionic wavelet transform and Butterworth filter. *International Journal of Computers and Applications*, 1–11.
10. Chen F, Zhang YT (2006) A new implementation of discrete bionic wavelet transform: Adaptive tiling. *Digit Signal Process* 16: 233–246.
11. Mourad T, Lotfi S, Sabeur A, et al. (2009) Speech Enhancement with Bionic Wavelet Transform and Recurrent Neural Network. *5th International Conference: Sciences of Electronic, Technologies of Information and Telecommunications SETIT,* 22–26.
12. Bin-Fang C, Jian-Qi L, Peixin Q, et al. (2014) An Optimization Adaptive BWT Speech Enhancement Method. *Information Technology Journal* 13: 1730–1736.
13. Mourad T (2016) Speech enhancement based on stationary bionic wavelet transform and maximum a posterior estimator of magnitude-squared spectrum. *International Journal of Speech Technology* 20: 75–88.
14. Wu LM, Li YF, Li FJ, et al. (2014) Speech Enhancement Based on Bionic Wavelet Transform and Correlation Denoising. *Advanced Materials Research,* 1386–1390.
15. Ben-Nasr M, Talbi M, Cherif A (2012) Arabic Speech Recognition by MFCC and Bionic Wavelet Transform using a Multi-Layer Perceptron for Voice Control. *International Journal of Software Engineering and Technology*.
16. Zehtabian A, Hassanpour H, Zehtabian S, et al. (2010) A novel speech enhancement approach based on Singular Value Decomposition and Genetic Algorithm. *2010 International Conference of Soft Computing and Pattern Recognition.*
17. LIU Y, NI W (2015) Speech enhancement based on bionic wavelet transform of subband spectrum entropy. *Journal of Computer Applications* 3: 58.
18. Singh RA, Pritamdas K (2015) Enhancement of Speech Signal by Transform Method and Various threshold Techniques: A Literature Review. *Advanced Research in Electrical and Electronic Engineering* 2: 5–10.
19. Tan BT, Fu M, Spray A, et al. (1996) The use of wavelet Transforms in Phoneme Recognition. *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96* 4: 2431–2434.
20. Rioul O, Vetterli M (1991) Wavelets and Signal Processing. IEEE SP Magazine.
21. Kobayashi M and Sakamoto M (1993) Wavelets analysis of acoustic signals. *Japan SIAM Wavelet Seminars II*.
22. Jones DL, Baraniuk RG (1991) Efficient approximation of continuous wavelet transform. *Electron Lett* 27: 748–750.
23. Swami PD, Sharma R, Jain A, et al. (2015) Speech enhancement by noise driven adaptation of perceptual scales and thresholds of continuous wavelet transform coefficients. *Speech Communication* 70: 1–12.
24. Johnson MT, Yuan X, Ren Y (2007) Speech signal enhancement through adaptive wavelet thresholding. *Speech Communication* 49: 123–133.
25. Wavelet center frequency, MATLAB centfrq, MathWorks. Available from: https://in.mathworks.com/help/wavelet/ref/centfrq.html.
26. Anusuya MA and Katti SK (2011) Front end Analysis of Speech signal processing: A Review. *International Journal of speech technology,* Springer.
27. GitHub. Available from: https://github.com/Jakobovski/decoupled-multimodal-learning.
28. Hu Y and Loizou P (2007) Subjective evaluation and comparison of speech enhancement algorithms. *Speech Communication* 49: 588–601.
29. Aida-zade K, Xocayev A, Rustamov S (2016) Speech recognition using Support Vector Machines. *2016 IEEE 10th International Conference on Application of Information and Communication Technologies (AICT),* 1–4.

30. Mini PP, Thomas T, Gopikakumari R (2018) Feature Vector Selection of Fusion of MFCC and SMRT Coefficients for SVM Classifier Based Speech Recognition System. *2018 8th International Symposium on Embedded Computing and System Design,* 153–157.

31. Thiruvengatanadhan R (2018) Speech Recognition using SVM. *International Research Journal of Engineering and Technology (IRJET)* 5: 918–921.

32. Zou YX, Zheng WQ, Shi W, et al. (2014) Improved Voice Activity Detection based on support vector machine with high separable speech feature vectors. *2014 19th International Conference on Digital Signal Processing.*

33. Gupta A, Joshi A (2018) Speech Recognition using Artificial Neural Network. International Conference *on Communication and Signal Processing*, India.

34. Al-Rababah MAA, Al-Marghilani A, Hamarshi AA (2018) Automatic Detection Technique for Speech Recognition based on Neural Networks Inter-Disciplinary. *(IJACSA) International Journal of Advanced Computer Science and Applications* 9: 179–184.

35. Swedia ER, Mutiara AB, Subali M (2018) Deep Learning Long-Short Term Memory (LSTM) for Indonesian Speech Digit Recognition using LPC and MFCC Feature. Third International Conference on Informatics and Computing (ICIC).

36. Vani HY, Anusuya MA (2015) Isolated Speech recognition using K-means and FCM Technique. *International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT).*

37. Vani HY, Anusuya MA (2017) Noisy speech recognition using KFCM. *International Conference on Cognitive Computing Information Processing.*

38. ICSI Speech FAQ. Available from: https://www1.icsi.berkeley.edu/Speech/faq/speechSNR.html.

39. Yangyang S, Mei-Yuh H, Lei X (2019) END-TO-END SPEECH RECOGNITION USING A HIGH RANK LSTM-CTC BASED MODEL. *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).*

40. Graves A, Jaitly N (2014) Towards End-to-End Speech Recognition with Recurrent Neural Networks. *Proceedings of the 31 st International Conference on Machine Learning*, Beijing, China.

41. Goehring T, Keshavarzi M, Carlyon RP, et al. (2019) Recurrent neural networks to improve the perception of speech in non-stationary noise by people with cochlear implants. *J Acoust Soc Am* 146: 705–718.

AIMS Press