2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

RNN and SOM based Classifier to Recognize Assamese Fricative Sounds Designed using Frame based Temporal Feature Sets

Chayashree Patgiri, Mousmita Sarma and Kandarpa Kumar Sarma

Abstract— **In this work, a Recurrent Neural Network (RNN) is trained using cepstral features and a set of difference cepstral feature (DCF) vectors on a frame to frame basis. The DCF vector is formulated to capture the temporal patterns of fricative sounds or phonemes of Assamese language. A hybrid algorithm is developed to recognize these fricative phonemes from certain words containing them. To preserve the temporal information of the speech segment, we here consider a frame-based hybrid approach to recognize fricatives from Assamese speech. A hybrid feature set is developed where simple frame-based feature is combined with differential frame-based feature. Investigation of feature extraction techniques like Linear Predictive Cepstral Coefficient (LPCC) and Mel-Frequency Cepstral Coefficient (MFCC) have been carried out and their performances have been evaluated in comparison to that obtained from the DCF set for Assamese fricative phoneme recognition. Here, speech segment is divided into 20 millisecond frames with overlap of 10 millisecond to extract features. Also, difference of a current frame with its preceding and succeeding frame is considered for forming a more accurate dynamic approach for fricative recognition. The differential processing enables to reduce correlation and retain only the most relevant portion of the input. After obtaining the feature vectors, Self Organizing Map (SOM) has been used to categorize the related features into different classes and remove repeating data. Thus the features obtained from the phoneme signal has been reduced into different sized cluster centres provided by SOM. The reduced feature vector is next applied to the RNN based hybrid classifier for learning the pattern and recognizing any unknown fricative segment.**

I. Introduction

SPEECH consists of sequences of sounds. Phonemes are the smallest distinguishable meaningful unit of the speech signal, which is an abstract representation at some cognitive level. Many numbers of phones may be associated with a particular phoneme, since uttering the same sound repeatedly never produce exactly similar waveform. But they are similar if we consider for a small segment of time, say 20 millisecond. Therefore, if we consider many numbers of phones as a pattern to represent a particular phoneme class, then an Artificial Neural Network (ANN) can be trained on frame by frame basis to learn the pattern of that phoneme. Fricatives are consonant sounds produced with a very narrow constriction in the oral cavity. There is a rapid flow of air through the constriction, creating turbulence in the flow. The random velocity fluctuation in the flow can act as a source of sound. The sound generated in this way is called turbulence noise. Air turbulence produced in this way, by various kinds of constrictions in the vocal tract (the position of which depends on the particular fricative), is the typical sound source for all fricatives [1]-[4]. In Assamese language, fricative forms a major group of speech sounds which has different phonemical characteristics. In Assamese language, voiceless alveolar fricative /s/ and velar fricative /x/ are observed. Further, voiced alveolar fricative /z/ and voiced glottal fricative /fi/ are also identified. Unlike other Indian languages, the presence of voiceless velar fricative /x/ is a specific feature of the language [5]-[7].

We have proposed here an approach of recognizing fricatives based on Recurrent Neural Network (RNN). ANNs are composed of many non-linear computational elements operating in parallel and arranged in the pattern of biological neural network. Hidden Markov Models (HMMs) is the most successful speech recognition technology till date. Unlike HMM, ANN is able to work with model free data and can continuously learn from the surrounding. Moreover, ANNs can retain this learning and use it for subsequent processing. This way ANNs can be helpful for speech processing applications. A recent trend of using ANN for speech processing is observed [8]-[17] so that near human like performance can be achieved.

Among the supervised learning ANNs, the RNNs have the dynamic structure with a capability of learning temporal information and hence are suitable for speech based applications [18] [19]. In our previous work [20], we have described a RNN based algorithm to recognize fricative sounds from Assamese speech, where the feature vector is generated from the specific acoustic-phonetic characteristics i.e. centre of gravity (COG), standard deviation (SD), skewness and kurtosis. Although the model provides acceptable recognition rate, the feature vector used are somewhat static and pre estimated. As a result, the temporal nature of the speech signal which is the primary characteristic of the signal is not considered. This may lead to fall in recognition rate in a different large database. Therefore, the requirement of a model using more robust feature is obvious.

To develop such an algorithm which can be used in various practical scenarios, a different approach is considered here. A hybrid classifier is designed where cepstral feature and difference cepstral feature is used. Self Organizing Map (SOM) based clustering technique is used to group the similar feature vectors into one cluster. Here, two hybrid RNN fricative recognizers are designed using Linear Predictive Cepstral Coefficients (LPCC) and Mel-Frequency Cepstral Coefficients (MFCC) feature extractors respectively. For each of the hybrid classifiers, feature vector is formed using frame-

Chayashree Patgiri, Mousmita Sarma and Kandarpa Kumar Sarma are with the Department of Electronics Communication Engineering, Gauhati University, Guwahati-14, Assam, India. (email: {chayashreepatgiri21, go4mou, kandarpaks}@gmail.com).

978-1-4799-1484-5/14/$31.00 ©2014 IEEE

based LPCC/MFCC vectors and frame-based differential LPCC/MFCC vectors of the speech signal. After obtaining the features vectors, to categorize the related features into different classes and remove repeating and correlated feature vectors, a SOM based clustering has been used. SOM is a method of data analysis used for clustering and projecting multi-dimensional data into a lower-dimensional space. SOM is used here to cluster the feature vectors extracted in that way since adjacent frames may possess less variation. SOM's role is to simply bring similar frames into one cluster. The cluster provided by SOM is used as pattern vector for the RNN classifier. This way original feature vectors obtained from 20 millisecond frames are reduced into the difference sized cluster centers. SOM is used as a data compaction unit since it reduces large numbers of frames of speech segment into small numbers of clustered frames. The reduced feature vector is fed to the RNN based hybrid phoneme recognizer for performing the discrimination between different fricative classes.

The rest of this paper is organized as follows. Section II describes about the fricative sounds of the Assamese language. Section III describes the theoretical consideration required for this work. Here, a brief depiction about the feature extractors considered for the work are described. A brief account about the relevant details of the speech database collected for the work is described in Section IV. The experimental details and results are included in Section V. Finally, Section VI concludes the description.

II. Fricative Sounds of Assamese language

Assamese is a major language spoken in the North-Eastern part of India. It is the official language of state of Assam. It is an Indo-Aryan language originated from Vedic dialects but the exact nature of the origin and growth of the language is not very clear as yet [5]. It is supposed that like other Aryan languages, Assamese was also born from *Apabhraṁśa* dialects developed from *Mãgadhi* Prakrit of the eastern group of Sanskritic language [5]. Assamese phonemic inventory consists of eight vowels and twenty-one consonants. The consonants may be grouped into broad divisions: the stops and the continuants. There are eleven continuants out of which four spirants or fricatives /s/, /z/, /x/, /ɦ/ are identified [5] [21] [22]. These four Assamese fricative as shown in Table I are described below [5]-

1) Voiceless alveolar sibilant, /s/: It is one of the most common sound cross linguistically. Its manner of articulation is sibilant fricative, which means it is generally produced by channelling air flow along a groove in the back of the tongue up to the place of articulation, at which point it is focused against the sharp edges of the nearly clenches teeth, causing high frequency turbulence. Its place of articulation is alveolar means it is articulated with tongue at the alveolar ridge. Its phonation is voiceless, which means it is produced without vibrations of the vocal cords.
2) Voiced alveolar fricative, /z/: Its manner of articulation is also sibilant. But its phonation is voiced, which means the vocal cords vibrate during the articulation.
3) Voiceless velar fricative, /x/: Its place of articulation is velar, which means it is articulated with the back of the tongue at the soft palate. Its phonation is voiceless. Assamese is unusual among eastern Indo-Aryan language for the presence of voiceless velar fricatives. It is similar to the velar sound in German of Europe. Phonetically, this /x/ sound is pronounced somewhat in between the sounds /s/, /kh/ and /h/ and is similar to the German sound /ch/ as pronounced in the word 'Bach' or the Scottish sound as found in the word 'Loch'. It may be an Indo- European feature, which has been preserved by ɑxɑmija [23] [24]. It is an important phoneme in the language.
4) Voiced glottal fricative, /ɦ/: Its phonation is voiced, which means the vocal cords vibrate during the articulation.

TABLE I
Assamese Fricative Phonemes

PHONATION	PLACE OF ARTICULATION		
	Alveolar	Velar	Glottal
Voiceless	/s/	/x/	
Voiced	/z/		/ɦ/

III. Theoretical Consideration

Here we briefly describe the related theoretical aspects.

A. LPCC feature extraction method

In the linear prediction analysis of speech, each sample is predicted as a linear weighted sum of the past p samples, where p represents the order of prediction [25]. If $s(n)$ is the present sample, then it is predicted by the past p samples as

$$\hat{s}(n) = -\sum_{k=1}^{p} a_k s(n-k) \tag{1}$$

The difference between the actual and predicted sample value is termed as the prediction error or residual, which is given by

$$e(n) = s(n) - \hat{s}(n) = s(n) + \sum_{k=1}^{p} a_k s(n-k) \tag{2}$$

where $\{a_k\}$ are the linear prediction coefficients. The linear prediction coefficients are typically determined by minimizing the mean squared error over an analysis frame. The coefficients can be obtained by solving the set of p normal equations,

$$\sum_{k=1}^{p} a_k R(n-k) = -R(n), n = 1, 2, \ldots, p \tag{3}$$

where

$$R(k) = \sum_{n=0}^{N-(p-1)} s(n)s(n-k), k = 0, 1, 2, \ldots, p \tag{4}$$

and $\{s(n)\}$ are the speech samples and N is the numbers of samples in one analysis frame.
In the frequency domain, the eq. (2) can be represented as,

$$E(z) = S(z) + \sum_{k=1}^{p} a_k S(z) z^{-k} \tag{5}$$

i.e.

$$A(z) = \frac{E(z)}{S(z)} = 1 + \sum_{k=1}^{p} a_k z^{-k} \tag{6}$$

A cepstrum is the result of taking the Inverse Fourier transform (IFT) of the logarithm of the estimated spectrum of a signal.The concept of cepstrum was defined in a 1963 paper by Bogert et al [26]. A short-time cepstrum analysis was proposed by Schroeder and Noll for application to pitch determination of human speech [27] [28]. Cepstral parameter extraction in speech recognizers system is based on converting LPC parameters to cepstral coefficients by utilizing the recursion relationship.
Cepstral coefficients of $A(z)$ is given by,

$$c(k) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log A(e^{j\omega}) e^{jn\omega} d\omega \tag{7}$$

The cepstrum parameters may be computed directly from the LP parameters using the following recursion.

$$c(k) = a(k) - \sum_{m=1}^{k-1} \frac{m}{k} c(k) a(k-m), 1 \leq k \leq p \tag{8}$$

In speech recognition systems, the cepstrum also plays a significant role. Specifically, the cepstral coefficients have been found empirically to be a more robust, reliable feature set for speech recognition than linear predictive coding (LPC) coefficients or other equivalent parameter sets [29]. Thus, the cepstral coefficient of the LPC obtained are applied to SOM for clustering which will form the feature vector for the RNN classifier in this paper.

B. MFCC feature extraction method

An alternative use of the cepstrum in speech recognition is the mel-frequency cepstrum [30]. The mel-frequency cepstrum is based on calculating the cepstrum from the logarithm of the spectrum obtained from a filter bank with center frequencies and bandwidths determined by a constant mel-frequency interval. Mel-frequency cepstral coefficients (MFCCs) are set of values that collectively make up an MFC. The difference between the cepstrum and the mel-frequency cepstrum is that in the MFC, the frequency bands are equally spaced on the mel scale, which approximates the human auditory system's response more closely than the linearly-spaced frequency bands used in the normal cepstrum. This frequency warping can allow for better representation of sound. Human perception of the frequency content of sounds follows a subjectively defined nonlinear scale called the Mel scale [2]. This is defined as,

$$f_{mel} = 2595 \log_{10}(1 + \frac{f}{700}) \tag{9}$$

If $\{y(n)\}$ represent a frame of speech that is pre-emphasized and Hamming-windowed then in frequency domain, $y(n)$ is converted to M-point DFT which leads to the energy spectrum,

$$\mid Y(k) \mid^2 = \mid \sum_{n=1}^{M} y(n) e^{-\frac{j2\pi nk}{M}} \mid^2, 1 \leq k \leq M \tag{10}$$

This is followed by the construction of a filter bank with Q unity height triangular filters, uniformly spaced in the Mel scale. Next, this filter bank is imposed on the spectrum calculated in eq. (10). The final MFCC coefficients C_m can be written as [31],

$$C_m = \sqrt{\frac{2}{Q}} \sum_{l=0}^{Q-1} \log[e(i+1)] cos[m(l-0.5)\frac{\pi}{Q}] \tag{11}$$

where $0 \leq m \leq (R-1)$, R is the desired number of cepstral features and $e(i)$ are the outputs of the Mel-scaled band-pass filters can be calculated by a weighted summation between respective filter response $\Psi_i(k)$ and the energy spectrum $\mid Y(k) \mid^2$ as,

$$e(i) = \sum_{k=1}^{\frac{M}{2}} \mid Y(k) \mid^2 \Psi_i(k) \tag{12}$$

where i is the number of filters in the mel filter bank. The MFCC coefficients analysis is performed for every frame (20 ms) and result is used to generate the feature vector for RNN classifier.

C. SOM

The SOM is a method of data analysis used for clustering and projecting multi-dimensional data into a lower-dimensional space to reveal hidden structures of the data. The SOM [32] is a class of ANN capable of recognizing the main features of the data they are trained on. Kohonen proposed SOM architecture which can automatically generate self organization properties during unsupervised learning process.
Kohonen SOM is unsupervised system which is based on the competitive learning. It means that a competition process takes place before each cycle of learning. In the competition process a winning processing element is chosen by some criteria. Usually this criteria is to minimize an Euclidean distance between the input vector and the weight vector. After the winning processing element is chosen, its weight vector is adapted according to the learning law used [18]. The learning procedure of Kohonen feature maps is similar to that of competitive learning networks. A similarity (dissimilarity) measure is selected and the winning unit is considered to be the one with the largest (smallest) activation. For Kohonen feature maps, the winning unit's weights and also all of the weights in a neighborhood around the winning units are updated [33].

CVC		
WORD	IPA	MEANING IN ENGLISH
চাহ	[saɦ]	'tea'
জহ	[zɒɦ]	'heat; warmth'
জাহ	[zaɦ]	'digested'
সজ	[xɒz]	'honest'
সাহ	[xaɦ]	'courage'
শহ	[xɒɦ]	'crop'
শাহ	[xaɦ]	'the kernel of a fruit'
শীহ	[xiɦ]	'ear of corn'
হাস্	[ɦax]	'to laugh'
চঁছ	[sɒs]	'smooth'
শেষ	[xex]	'end'
হাঁহ	[ɦaɦ]	'a web-footed bird'

VCV		
WORD	IPA	MEANING IN ENGLISH
আহা	[aɦa]	'an exclamation of pleasure'
আশা	[axa]	'hope'
ইসি	[ixi]	'this and that'
অহা	[ɒɦa]	'present'
অহি	[ɒɦi]	'snake'
অহো	[ɒɦɔ]	'alas'
আজি	[azi]	'today'
আশী	[axi]	'eighty'

Fig. 1. Wordlist Prepared for Recognition purpose

IV. Speech Database

The speech database is created from speakers of four different dialects of Assamese language. A wordlist as shown in fig. 1 is prepared containing fricative-vowel-fricative (C_iVC_i and C_iVC_j) and vowel-fricative-vowel (V_iCV_i and V_iCV_j) syllables and are recorded by the trained speakers in a noise free environment. Each CVC and VCV token is repeated five times, yielding a total of 245 tokens per speaker (49 syllables $\times$ 5 repetitions). For recording, the speech analysis software Wavesurfer [34] and a PC headset is used with the following specification-

- Sampling frequency: 48000 Hz and
- Bit resolution: 16 bit per sample

According to [21], Assamese has four different dialects namely Eastern, Central, Kamrupi and Goalpariya groups. From every dialect there are 3 speakers. After recording fricatives are annotated and segmented in the speech analysis software PRAAT [35].

V. Experimental Details and Results

Fig. 2 shows the block diagram of the proposed fricative recognition system using hybrid RNN classifier. The steps involved on the proposed work can be summarized below-

1) Recording of 245 number of fricative-vowel-fricative (C_iVC_i and C_iVC_j) and vowel-fricative-vowel (V_iCV_i and V_iCV_j) syllables from 12 speakers covering all the four dialects of Assamese language.
2) Extraction of features from recorded fricative sounds using frame-based LPCC and frame-based differential LPCC method and repeat the steps for MFCC method.
3) Generation of hybrid feature vectors for both LPCC and MFCC methods those will be used for training and testing of said classifiers.
4) Training each of the RNN with 80 fricative samples using those feature vectors.
5) Testing of the algorithm with 30 samples per fricative.

A. Results of RNN classifier using frame-based LPCC features

In the first step of the fricative recognition model, RNN classifier is used. The feature vectors generated from the frame-based LPCC of the fricative speech are presented to the SOM for clustering the large dimensional data. Here, SOM is used to reduce the size of the feature vector which will form the pattern layer of the RNN classifier. A 12^{th} order linear prediction analysis is performed for every frame of 20 millisecond speech with overlap of 10 millisecond. So, the size of the vector after taking LPCC will be $N \times 12$, where N will be the number of frames present in the speech segment. SOM is used here to cluster the feature vectors extracted in that way since adjacent frames may possess less variation. SOM's role is to simply bring similar frames into one cluster. Thus vectors obtained after taking LPCC are fed to a SOM network with different cluster size, M $(M < N)$ for grouping the similar data. The cluster size, M used here are 8 and 10. The cluster provided by SOM is used pattern vector for the RNN classifier. This way original feature vectors obtained from 20 millisecond frames are reduced into the difference sized cluster centers.

Initially different training algorithm like gradient descent with adaptive learning rate backpropagation algorithm Resilient Backpropagation (RBP) and Levenberg-Marquardt (LM) are used to train the RNN with 3 hidden layer and 40 feature vectors. But recognition rate was observed to be somewhat lower and requires more time. Finally, Scaled Conjugate Gradient (SCG) algorithms is used with 80 feature

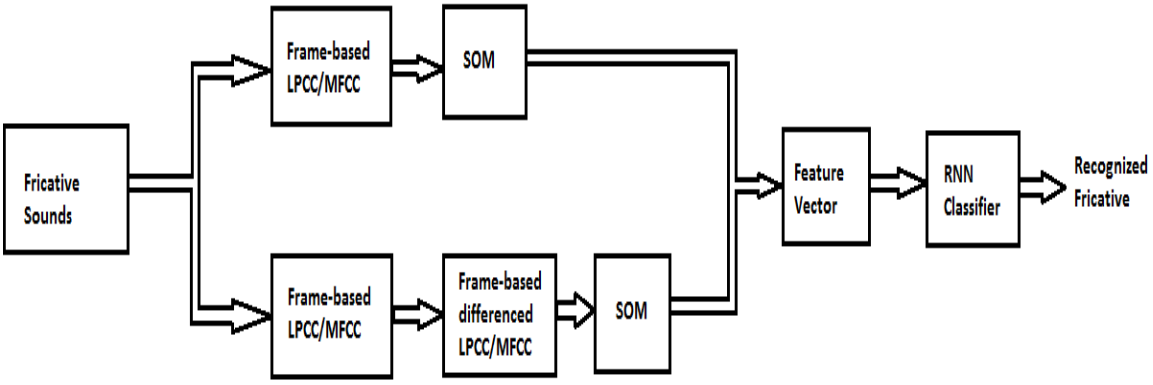

Fig. 2. System block diagram

TABLE II
SUCCESS RATE OF ASSAMESE FRICATIVES USING FRAME-BASED LPCC (M=8)

Fricative	Recognition rate(%)
/s/	86.67
/z/	70
/x/	70
/fi/	80

TABLE III
SUCCESS RATE OF ASSAMESE FRICATIVES USING FRAME-BASED LPCC (M=10)

Fricative	Recognition rate(%)
/s/	83.33
/z/	76.67
/x/	70
/fi/	86.67

vectors, which increases the success rate to an acceptable mark and is found to be the best among all the four algorithms in terms of recognition rate. Table II shows the percentage of correct recognition using frame-based LPCC with cluster size, M of value 8. It is observed that the overall recognition rate is 77% considering LPCC and RNN with three hidden layer and 80 feature vector trained with SCG learning algorithm where $M = 8$. The same RNN is now trained with $M = 10$ and Table III shows the result which indicates an improvement of overall recognition rate for fricative upto 80%. The primary reason behind selection of different values of M (namely 8 and 10) is to ascertain the effect of cluster size on recognition. With $M = 10$, more information is extracted than is the case with $M = 8$, hence better success rate is obtained with the former.

TABLE IV
SUCCESS RATE OF ASSAMESE FRICATIVES USING FRAME-BASED LPCC AND DIFFERENTIAL LPCC (M=10)

Fricative	Recognition rate(%)
/s/	86.67
/z/	76.67
/x/	76.67
/fi/	86.67

B. Results of RNN classifier using frame-based LPCC and Differential LPCC features

To increase the recognition rate further we design a hybrid feature set to train the RNN classifier. For that purpose, difference of LPCC from one frame to another is used to create another feature vector. This provides another set of feature vectors which is clustered using SOM. A 2^{nd} RNN classifier is trained using differential LPCC, which add knowledge to the main classifier. This way recognition rate improves. The difference coefficients for frame n are the difference between the coefficients of frame $n+\delta$ and $n-\delta$. In our implementation, a differential coefficient is computed every frame, with $\delta = 1$ frames. Frame-based differential LPCC is combined with simple frame-based LPCC which forms the hybrid feature vector for the RNN. Table IV shows the result of recognition rate of hybrid RNN classifier with three hidden layer and $M = 10$ using hybrid LPCC features. With these parameters overall recognition rate of Assamese fricatives found to be approximately 82%.

C. Results of RNN classifier using frame-based MFCC features

The recognition rate of Assamese fricative is also evaluated for the RNN classifier using MFCC features. Here

TABLE V

SUCCESS RATE OF ASSAMESE FRICATIVES USING FRAME-BASED MFCC (M=8)

Fricative	Recognition rate(%)
/s/	70
/z/	80
/x/	80
/fi/	70

TABLE VI

SUCCESS RATE OF ASSAMESE FRICATIVES USING FRAME-BASED MFCC (M=10)

Fricative	Recognition rate(%)
/s/	73.33
/z/	83.33
/x/	80
/fi/	70

TABLE VII

SUCCESS RATE OF ASSAMESE FRICATIVES USING FRAME-BASED MFCC AND DIFFERENTIAL MFCC (M=10)

Fricative	Recognition rate(%)
/s/	86.67
/z/	86.67
/x/	86.67
/fi/	83.33

TABLE VIII

NETWORK PARAMETERS FOR LPCC AND MFCC FEATURES

Parameters	LPCC	MFCC
Number of hidden layers	3	3
Number of epoch	1000	1000
Training Algorithm	SCG	SCG
Training Time	80.45 sec	63.76 sec

TABLE IX

COMPARISON OF OVERALL SUCCESS RATE OF ASSAMESE FRICATIVES WITH LPCC AND MFCC FEATURES

Feature type	LPCC	MFCC
With frame-based feature(M=8)	77%	75%
With frame-based feature(M=10)	80%	77%
With frame-based hybrid feature(M=10)	82%	86%

also, MFCC coefficients are calculated for frame size of 20 millisecond and with 13 cepstral features. Tables V and VI show the recognition rate using MFCC features with $M = 8$ and $M = 10$ respectively.

D. Results of RNN classifier using frame-based MFCC and Differential MFCC features

Table VII shows the results of recognition rate of hybrid RNN classifier with three hidden layer and $M = 10$ using MFCC hybrid features. As done previously, here also MFCC hybrid feature set is generated using frame-based MFCC combining with differential MFCC like that of LPCC. With these parameters overall recognition rate of Assamese fricatives found to be approximately 86%.

E. Comparison Between LPCC and MFCC feature sets

A comparative analysis is carried out between the LPCC and MFCC features for fricative recognition in a RNN based learning environment. From Table VIII, we can conclude that with same network parameters, training time required for the network with LPCC feature set is more than that of MFCC feature set. From Table IX, we observed that hybrid RNN classifier with MFCC feature gives better recognition rate for Assamese fricatives than LPCC feature. From experimental results, it can be concluded that RNN based hybrid classifier can recognize fricatives in Assamese language. Overall success rate for the proposed model with frame-based hybrid MFCC feature is found to be 86%. This validates the effectiveness of the proposed approach. Compared to the works described in [36] [20], the advantage of this frame based feature set, is its ability to learn the temporal patterns from fricative phones and use it subsequently to recognize the phoneme.

VI. CONCLUSIONS

Here, we have used frame-based LPCC and MFCC along with frame-based differential LPCC and MFCC features for designing of the hybrid RNN classifier. We observed that while combining differential frame-based LPCC/MFCC with the original frame-based LPCC/MFCC features then hybrid RNN classifier shows improvement in terms of success rate. The frame based hybrid MFCC feature set with SOM clustering gives better result than that of frame based hybrid LPCC feature set. Thus, the hybrid approach based on RNN classifier provides better performance while applying for Assamese fricative phoneme recognition. This is a prototype work done to recognize the fricative phonemes. The work can be later extended to a complete phoneme recognition system combining this temporal feature set with the proposed RNN based neural classifier.

REFERENCES

[1] A. Jongman, R. Wayland, and S. Wong, "Acoustic characteristics of English fricatives," *Journal of Acoustical Society of America*, vol. 108, no. 3, Sepember, 2000.

[2] D. O'Shaughnessy, *Speech Communication Human and Machine,* 2^{nd} Edition, IEEE Press, New York, 2000.

[3] Kenneth N. Stevens, *Acoustic Phonetics*, 1^{st} MIT Press paperback Edition, The MIT Press, Cambridge, Massachusetts, London, England, 2000.

[4] P. Ladefoged, S. F. Disner, *Vowels and Consonants*, 3^{rd} Edition, Wiley-Blackwell Publishing Ltd., West Sussex, UK, 2012.

[5] G. C. Goswami, *Structure of Assamese,* 1^{st} Edition, Department of Publication, Gauhati University, Guwahati, Assam, India, 1982.

[6] U. N. Goswami, *An Introduction to Assamese,* Mani-Manik Prakash, Guwahati, Assam, India, 1978.

[7] G. C. Goswami and J. P. Tamuli, "Asamiya", in G. Cardona and D. Jain (eds.), *The Indo-Aryan Languages,* London: Routledge, pp. 391-443, 2003.

[8] M. Sarma and K. K. Sarma, "Speaker identification model for Assamese language using a neural framework," *In Proceedings of International Joint Conference on Neural Networks (IJCNN)*, pp. 1-7, Dallas, TX, USA, August, 2013.

[9] M. Sarma and K. K. Sarma, "An ANN based approach to recognize initial phonemes of spoken words of Assamese language," *Applied Soft Computing,* vol. 13, no. 5, pp. 2281-2291, May, 2013.

[10] G. Dede, M. H. Sazli, "Speech recognition with artificial neural networks," *Digital Signal Processing,* vol. 20, no. 3, pp. 763-768, May, 2010.

[11] T. L. Kumar, T. K. Kumar, K. S. Rajan, "Speech Recognition Using Neural Networks," *In Proceedings of International Conference on Signal Processing Systems,* pp. 248-252, Singapore, May, 2009.

[12] M. Oprea, D. Schiopu, "An artificial neural network-based isolated word speech recognition system for the Romanian language," *In Proceedings of* 16^{th} *International Conference on System Theory, Control and Computing (ICSTCC)*, pp. 1-6, Sinaia, October, 2012.

[13] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano and K. J. Lang, "Phoneme recognition using time-delay neural networks," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, no. 3, pp. 328-339, March, 1989.

[14] D. Paul, R. Parekh, " Automated speech recognition of isolated words using neural networks," *International Journal of Engineering Science and Technology(IJEST)*, vol. 3, no. 6, pp. 4993-5000, 2011.

[15] K. Elenius and G. Takacs, "Phoneme Recognition with an Artificial Neural Network," *In Proceeding of* 2^{nd} *European Conference on Speech Communication and Technology (EUROSPEECH)*, pp. 121-124, Italy, September, 1991.

[16] M. Sarma and K. K. Sarma, "Vowel Phoneme Segmentation for Speaker Identification Using an ANN-Based Framework," *Journal of Intelligent Systems*, vol. 22, no. 2, pp. 111-130, April, 2013.

[17] Md Salam, D. Mohamad and S. Salleh, "Malay Isolated Speech Recognition Using Neural Network: A Work in Finding Number of Hidden Nodes and Learning Parameters," *The International Arab Journal of Information Technology*, vol. 8, no. 4, pp. 364-371,October, 2011

[18] S. Haykin, *Neural Networks:A Comprehensive Foundation,* 2^{nd} Edition, Prentice-Hall of India Pvt. Ltd., Delhi, India, 2005.

[19] S. Kumar, *Neural Networks: A Classroom Approach,* 3^{rd} Edition, Tata McGraw-Hill Education, New Delhi, India, 2004.

[20] C. Patgiri, M. Sarma, and K. K. Sarma,"Recurrent Neural Network based Approach to Recognize Assamese Fricatives using Experimentally Derived Acoustic-Phonetic Features," *In Proceedings of* 1^{st} *IEEE International Conference on Emerging Trends and Applications in Computer Science (ICETACS-2013)*, pp. 33-37, Shillong, India, September, 2013.

[21] Resource Center for Indian Language Technology Solutions, IIT Guwahati, Available via
$http://www.iitg.ernet.in/rcilts/phaseI/languages/asamiya.htm$

[22] Sarma B D, Sarma M, Sarma M and Prasanna S R M, "Development of Assamese Phonetic Engine: Some Issues," *In Proceedings of INDICON-2013*, IIT Bombay, Mumbai, India, 2013.

[23] Rajen Barua, *The X sound in Assamese language,*The Assam Tribune, Guwahati, Sunday, March 5, 2006.

[24] Prof. Gautam Baruah, Dept. of CSE, IIT Guwahati, Available via
$tdil.mit.gov.in/assamesecodechartoct02.pdf$

[25] J. Makhoul, "Linear prediction: A tutorial review," *In Proceedings of IEEE,* vol. 63, pp. 561-580, 1975.

[26] B. P. Bogert, M. J. R. Healy, and J. W. Tukey,"The Quefrency Alanysis of Time Series for Echoes: Cepstrum, Pseudo Autocovariance, Cross-Cepstrum and Saphe Cracking," *Proceedings of the Symposium on Time Series Analysis,* Chapter 15, pp. 209-243, 1963.

[27] A. M. Noll and M. R. Schroeder, "Short-Time 'Cepstrum' Pitch Detection," *Journal of the Acoustical Society of America,* vol. 36, no. 5, pp. 1030-1036, 1964.

[28] A. M. Noll , "Short-Time Spectrum and Cepstrum Techniques for Vocal-Pitch Detection," *Journal of the Acoustical Society of America,* vol. 36, no. 2, pp. 296-302, 1964.

[29] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing,* Englewood Cliffs, NJ:Prentice-Hall, 1975.

[30] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. 28, no. 4, pp. 357-366, August, 1980.

[31] S. Chakroborty, A. Roy and G. Saha, "Improved Closed Set Text-Independent Speaker Identification by combining MFCC with Evidence from Flipped Filter Banks," *International Journal of Signal Processing,* vol. 4, no. 2, pp. 114-121, November, 2006.

[32] K. Haese, "Self-organizing feature maps with self-adjusting learning parameters", *IEEE Transactions on Neural Networks,* vol. 9, pp. 1270-1278, 1998.

[33] J. S. R. Jang, C. T. Sun and E. Mizutani, *Neuro-Fuzzy and Soft-Computing,* 1^{st} Edition, Prentice-Hall of India Pvt. Ltd., Delhi, India, 2011.

[34] WaveSurfer. Available via
$http://www.speech.kth.se/wavesurfer/man.html$

[35] P. Boersma and D. Weenink, *Praat: doing phonetics by computer.* Available via
$http://www.fon.hum.uva.nl/praat/$

[36] A. M. A. Ali , J. V. Spiegel and P. Mueller, "Acoustic-phonetic Features for the Automatic Classification of Fricatives", *Journal of Acoustical Society of America,* vol. 109, no. 5, pp. 2217-2235, May, 2001.