International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:3, No:4, 2009

Assamese Numeral Corpus for Speech Recognition using Cooperative ANN Architecture

Mousmita Sarma, Krishna Dutta and Kandarpa Kumar Sarma

Abstract—Speech corpus is one of the major components in a Speech Processing System where one of the primary requirements is to recognize an input sample. The quality and details captured in speech corpus directly affects the precision of recognition. The current work proposes a platform for speech corpus generation using an adaptive LMS filter and LPC cepstrum, as a part of an ANN based Speech Recognition System which is exclusively designed to recognize isolated numerals of Assamese language- a major language in the North Eastern part of India. The work focuses on designing an optimal feature extraction block and a few ANN based cooperative architectures so that the performance of the Speech Recognition System can be improved.

Keywords—Filter, Feature, LMS, LPC, Cepstrum, ANN.

I. Introduction

Speech recognition is a method that uses an audio input for data entry to a computer or a digital system in place of a keyboard. In simple terms it can be a data entry process carried out by speaking into a microphone so that the system is able to interpret the meaning out of it for further modification, processing or storage. Speech corpus is the basis for both analyzing the characteristics of speech signal and developing speech synthesis and can be used to create acoustic model for an Automatic Speech Recognition (ASR) system. Speech corpus can be categorized according to its content, speaking style, channel property, phonetic coverage, dialectal accent or application area and can be generated by extracting carefully chosen features from the speech signal [1]. Feature extraction involves transforming the input data into the set of values that best describes the input under consideration. Over the years various speech features like intensity, pitch, short-time spectrum, LPC cepstrum, formats, nasal co-articulation, spectral correlation harmonic features and cepstral measures have been investigated [2]. This work focuses on the generation of a LPC cepstrum based optimal feature extraction block so that the performance of an ANN-based Speech Recognition System can be improved. The raw samples are the speech-recordings captured with male-female gender and mood variations uttering isolated numerals of Assamese language. Speech is a time varying signal. As such, the recognition system used in an ASR should have the ability to capture time-varying characteristics. Among the supervised learning ANNs, the Recurrent Neural Network (RNN)s for their time-varying nature of learning are suitable for applications like ASR and speech synthesis [3]. This work explains the extraction of speech corpus of Assamese numerals using an adaptive pre-emphasis filter block for use with an RNN as part of an ASR for numerals in Assamese- a major language in the North Eastern part of India. The results derived from the RNN based ASR block show a success rate of around 96% which is improved further to about 98% by using a cooperative block formed by two RNN sections to deal separately with gender based recognition. The description included here is organized as below. Section II provides a brief account of the distinct phonological features of Assamese speech and it uniqueness. A brief account about the importance of speech features and the relevant details of the raw speech samples collected for the work is described in Section III. The experimental details are included in Section IV. The results and the related discussion is included in Section V. Section VI concludes the description.

II. Distinctive Phonological Features of Assamese Language

The Assamese is a major language in the North - Eastern part of India with its own unique identity, language and culture though its origins root back to the Indo-European family of languages. These languages are spoken by more than a billion people, chiefly in Afghanistan, Bangladesh, India, Iran, Nepal, Pakistan, and Sri Lanka. It also is related to the Indo-Iranian subfamily. This class can be subdivided into three groups of languages: the Dardic(or *Pisacha*), the Indic (or Indo-Aryan), and the Iranian. Assamese is the easternmost member of this New Indo-Aryan (NIA) subfamily spoken in the Brahmaputra Valley of Assam [5] [6] [7]. Retaining certain features of its parent Indo-European family it has got many unique phonological characteristics. Some of those may be cited as below:

- A unique feature of the Assamese language is a total absence of any retroflex sounds. Instead the language has a whole series of alveolar sounds, which include oral and nasal stops, fricatives, laterals, approximants, flaps and trills, unlike other Indo-Aryan and Dravidian languages [5].
- Another striking phonological feature of the Assamese language is the extensive use of velar nasal / η/. In other New Indo Aryan languages this / η/ is always attached to a homorganic sound like /g/. In contrast it is always used singly in Assamese.
- The voiceless velar fricative / x / is a distinct characteristic of Assamese language which is not to be found in any language in the entire country. 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 'Asomiya'. It is an important phoneme in the language [5] [7].

There are other phonological uniqueness of Assamese pronunciation which shows minor variations when spoken by people of different regions of the state. This makes Assamese speech unique and hence requires a study exclusively directly to develop a speech recognition / synthesis system in Assamese.

III. Raw speech signals

Voice (or vocalization) is the sound generated by humans, animals and other vertebrates using a combination of the lungs and the vocal folds in the larynx, or voice box. All air entering and leaving the lungs passes through the vocal folds, which form a valve across the top of the trachea (windpipe). When brought together and appropriately tensed, air passing between the vocal folds causes them to vibrate, or more accurately, they open and close in quick succession. The frequency of vibration or fundamental frequency (henceforth F_0) is tightly correlated with the perceived pitch of the voice [10]. Speech is a bit different. Speech contains emotions and feelings and is generated by precisely coordinated muscle actions in the head, neck, chest, and abdomen. Speech results after a gradual process involving years of learning and practice [8]. In an ASR, speech recorded with gender, mood and condition variation can be distinguished. The system requires an appropriate set of features which can be used to achieve the objectives of recognition. Feature is a set of values extracted from an input speech that uniquely represents the key attributes of the sample. Speech signal is the output of a time varying vocal tract stimulated by a time-varying excitation signal. The vocal tract system is approximately described in terms of the acoustic features such as the frequency response of the resonances (formants) and anti- resonances (anti-formants) of the systems. These features are easier to extract from the signal than the articulatory parameters. The excitation of the vocal tract consists of broadly three categories [9] [11].

- Voiced source (due to vibrating vocal folds)
- Unvoiced source (turbulent air flow at narrow constriction in the vocal tract) and
- Plosive source (abrupt release).

In general the short-time characteristics of the speech signal are represented by the short-time (10-20 mS) spectral features of the vocal tract system as well as the nature of excitation in the short-time segment. These are called segmental features [9] [11].

Selecting appropriate features from a speech signal is an important issue to achieve high accuracy in speech recognition. Feature extraction involves analysis of speech signal. Broadly the feature extraction methods are classified as temporal analysis and spectral analysis techniques. In temporal analysis the speech waveform itself is used for analysis. In spectral analysis spectral representation of speech signal is used for analysis. The extracted feature vector must contain information that is useful to identify and differentiate speech sounds insensitive external noise and other irrelevant factors [4]. As mentioned earlier, since the work is a part of Assamese Speech

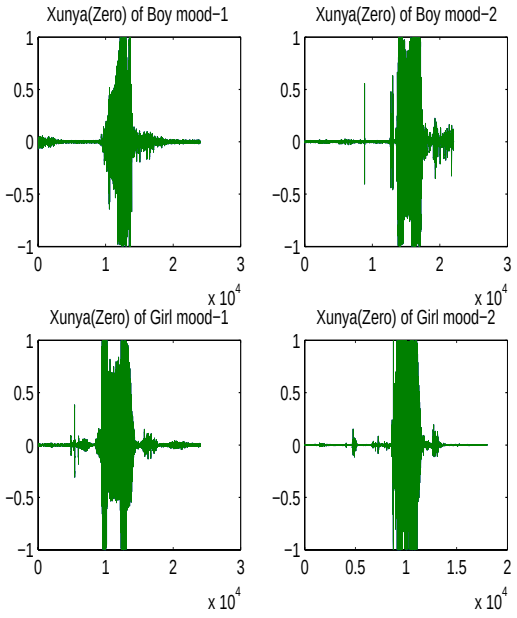

Fig. 1. Waveform of *xunya*(Zero in Assamese)with gender and mood variations

Recognition System, speech signals with gender and mood variation, uttering Assamese numerals from zero to nine,are captured. This results in a total of following broad sets of speech signals:

1) *Girl mood1*
2) *Girl mood2*
3) *Boy mood1*
4) *Boy mood2*
5) *Girl Reference*
6) *Boy Reference*

The waveforms of speech signal uttering *Xunya* (Zero in Assamese)in various mood and gender is shown in Figure 1 The gender distinction is brought into the samples to ascertain the ability of the proposed model to deal with such variations. In reality male and female voice and speech samples are different. Male vocal folds tend to be longer and thicker than female vocal folds causing them to vibrate more slowly. Male speakers have an average F_0 upto 200 Hertz (for speakers of German and English language it is about 100120 Hz while for females this values is twice. Female vocal folds are shorter and lighter and vibrate at approximately twice the male frequency [10].

The spectrum of a speech signal uttering *Xunya* (Zero in Assamese) is shown in Figure 2 .

For recording the speech signal, a PC headset and a sound recording software, Gold Wave, is used. GoldWave's Monitor recording option helps to adjust the volume level before recording. While recording, the sampling rate taken is 8000 Hz in mono channel mode.

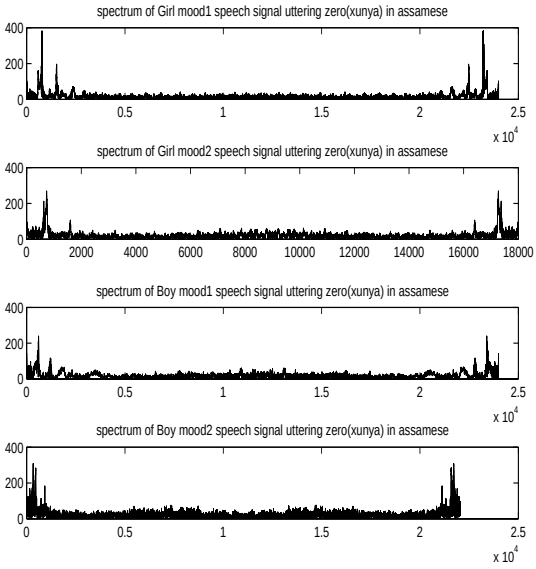

Fig. 2. Spectral representation of *xunya*(Zero in Assamese)with gender and mood variations

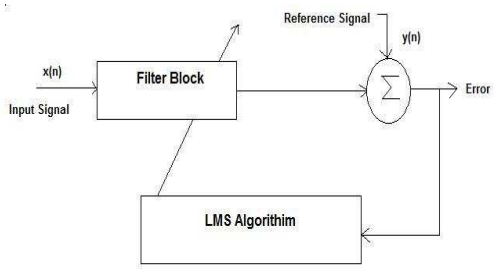

Fig. 3. Adaptive Pre-emphasis filter

IV. Experimental Details

The work is carried out as per the process diagram of a generic recognition system. The captured signals are grouped into six broad categories. Further a few more samples are generated from *Girl Reference* and *Boy Reference* mixing noise with them. The sample set thus generated consists of over 100 sets of data covering all the numerals. Of these 40 are categorized as training set and the rest taken for testing the recognizer. The following section provide a brief description of each of the constituent blocks.

A. Pre-emphasis Filter

A pre-emphasis filter emphasizes the high-frequency components of an input audio signal. Pre-emphasis filter is a digital filter, with adjusting components changing the filter coefficient so as to update the frequency characteristics [13]. The pre-emphasis filter plays a critical role in capturing the features of the input speech samples. The pre-emphasis block first is designed following the considerations as described below-

TABLE III
LMS-Algorithm

for known filter length M and step size parameter μ such that $0 < \mu < \frac{2}{MS_{max}}$
Given u[n]= M -by- 1 tap input at time n $u[n] = [u(n), u(n-1), ...u(n-M+1)]^T$
d[n] the desired response at time n,
Compute for n=0, 1, 2..
e = d(n) - $w^H(n)u(n)$
w(n+1)=w(n) + $\mu u(n)e^*(n)$

*1) **Filter block**:* Digital filters are required for two broad purposes- separation of signals that have been combined and restoration of signals that have been distorted in some way. Signal separation is needed when a signal has been contaminated with interference, noise, or other signals. Similarly signal restoration is used when a signal has been distorted in some way [14] [15]. Our work focuses first on removing noise from the input samples which help in proper extraction of the features. For noise removal operation the following digital filter structures have been tested.

1) Finite Impulse Response (FIR) filter
   a) Direct
   b) Transposed
   c) Cascade
2) Infinite Impulse Response (IIR) filter
   a) Direct form 1
   b) Direct form 2
   c) Lattice

All the FIR structures are implemented using Kaiser window, Rectangular window, Hamming window, Equi-ripple and Least Square methods. Similarly all the IIR structures are implemented using Chebychev, Elliptic and Butterworth methods. Then a comparative study is carried out using MSE values of all the filters. As can be seen from Tables I and II, among all the structures the Transposed Equi-ripple FIR filter and Elliptic Lattice IIR filter works best for all sets of speech signals.

But the IIR filtered signals have more errors in comparison to the FIR filtered signals. Thus the transposed Equi-ripple FIR filter is selected to be best suited for the work. The Transposed FIR Equi-ripple filter which is selected to be most errorless uses Remez / Parks-McClellan algorithm. The Remez/Parks McLellan method produces a filter which just meets the specification without over performing. These filters are optimal in the sense that the maximum error between the desired frequency response and the actual frequency response is minimized [16]. Yet to obtain the best performance the above filter structures must dynamically minimize the error in the corrupted signal upto the desired mark. Hence an adaptive architecture is required. The pre-emphasis filter is modified using an adaptive block involving the Least Mean Square (LMS) algorithm. The block diagram of this modified design is as in Figure 3. The renown LMS algorithm is used for the design of the adaptive filter is given in Table III [12].

2) Equalization: The noise corrupted signals of the captured speech samples require equalization after the pre-emphasis filtering. This is carried out using an LMS al-

TABLE I
AVERAGE MSE OF VARIOUS FIR FILTER STRUCTURE

Sl. No	Speech Signal	Finite Impulse Response Filter		
		Direct (Equiripple)	Transposed (Equiripple)	Cascade (Least Square)
1	*Girl Mood1*-1	0.02857	0.02123	0.024785
2	*Girl Mood2*-2	0.0235	0.0037	0.0223
3	*Boy Mood1*-1	0.01451	0.01124	0.01368
4	*Boy Mood2*-2	0.02389	0.01425	0.01495

TABLE II
AVERAGE MSE OF VARIOUS IIR FILTER STRUCTURE

Sl No	Speech Signal	Infinite Impulse Response Filter			
		Direct-1 (Elliptic)	Direct-2 (Elliptic)	Lattice (Elliptic)	Cascade (Elliptic)
1	*Girl Mood1*-1	0.05125	0.05132	0.05013	0.05128
2	*Girl Mood2*-2	0.06013	0.0543	0.0345	0.0545
3	*Boy Mood1*-1	0.02827	0.02821	0.02814	0.02820
4	*Boy Mood2*-2	0.04523	0.04516	0.04510	0.04525

gorithm based equalizer. LMS-algorithm produces the least mean squares of the error signal i.e. difference between the desired and the actual signal [17].The MSE value fixed during equalization is 10×10^{-4}, which LMS equalizer attains within the first 10 iterations for each of the sample considered.

B. *Windowing*

Windowing is used to minimize the signal discontinuities at the borders. A window is a function that is zero-valued outside the chosen interval. When another function or a signal data is multiplied by a window function, the product is also zero-valued outside the interval. If the window is defined as w[n], $0 < n \leq N-1$, then the windowed signal is

$$\hat{x_l}[n] = x_l[n]w[n] \tag{1}$$

where $0 < n \leq N-1$. In our work we used a Hamming Window, the most common window used for speech analysis. The coefficients of a Hamming window are computed from the following equation

$$w[k+1] = 0.54 - 0.46cos(2\Pi\frac{k}{n-1}) \tag{2}$$

where k=0,...,n-1.

C. Linear Predictive coding (LPC)

Linear predictive analysis of speech has become the predominant technique for estimating the basic parameters of speech. Linear predictive analysis provides an accurate estimate of the speech parameters and also an efficient computational model of speech. The basic idea behind linear predictive analysis is that a specific speech sample at the current time can be approximated as a linear combination of past speech samples. Through minimizing the sum of squared differences, over a finite interval between the actual speech samples and linear predicted values a unique set of parameters or predictor coefficients can be determined [18]. The next section describes the processing step involved in the LPC analysis using the autocorrelation method of order p. In matrix form

$$\mathbf{Ra} = \mathbf{r} \tag{3}$$

where $\mathbf{r} = [r(1)r(2)...r(3)]^T$ is the autocorrelation vector, $\mathbf{a} = [r_1 r_2 ... r_p]^T$ is the filter co-efficient vector, and

$$R = \begin{pmatrix} r(0) & r(1) & ... & r(\text{p-1}) \\ r(1) & r(0) & ... & r(\text{p-2}) \\ r(2) & r(1) & ... & r(\text{p-3}) \\ . & . & . & . \\ r(\text{p-1}) & r(\text{p-2}) & ... & r(0) \end{pmatrix} \tag{4}$$

is the Toeplitz autocorrelation matrix. This matrix is nonsingular and gives the solution

$$a = R^{-1}r \tag{5}$$

These coefficients form the basis for linear predictive analysis of speech. The coefficients of a p^{th}-order linear predictor (FIR filter) that predicts the current value of the real-valued time series x is based on past samples as shown in eq. 6

$$\hat{x(n)} = -a(2)x(n-1) - a(3)x(n-2) - .. - a(p+1)x(n-p) \tag{6}$$

p is the order of the prediction filter polynomial, a = [1 a(2) ... a(p+1)].

D. Cepstral co-efficient

The actual predictor coefficients obtained as mentioned above can never be used in speech recognition, since they typically show high variance. Hence it is required to transform the predictor coefficient to a more robust set of parameters known as cepstral coefficients. They can be directly derived from the set of LPC co-efficients using the recursion
c_0 = r(0),

$$c_m = a_m + \sum_{k=1}^{M-1} \frac{k}{m} c_k a_{m-k}, \tag{7}$$

where $1 < m < p$, and

$$c_m = \sum_{k=1}^{M-1} \frac{k}{m} c_k a_{m-k}, \tag{8}$$

where $m > p$. The cepstral co-efficient are the coefficients of the Fourier transform representation of the log magnitude of

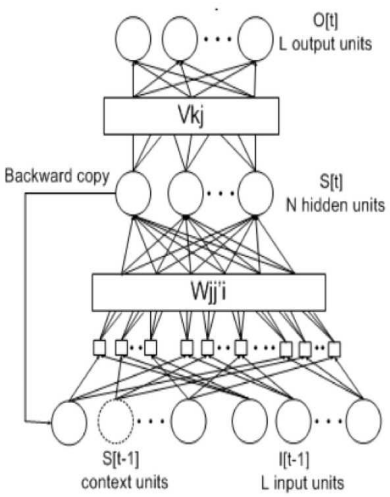

Fig. 4. Recurrent Neural Network

the spectrum [19]. The size of the LPC feature vector take is 20. The considerations upon which this size of the cepstral co-efficient are fixed is given in Section IV-E and the related results are provided in Table IV.

E. ANN configuration and training consideration

Artificial Neural Network (ANN)s are non- parametric prediction tools that can be used for a host of pattern classification/application including speech recognition [20] [21] [22]. One of the most commonly used ANN structures for pattern classification is the Multi-Layer Perceptron (MLP) which have found application in a host of work related to speech synthesis and recognition.[20] [21] [22]. The MLPs work very well as an effective classifier for vowel sounds with stationary spectra, while their phoneme discriminating power deteriorates for consonants characterized by variations of short-time spectra [23]. Feed forward MLPs are unable to deal with time varying information as seen in the speech spectra. The Recurrent Neural Networks (RNNs) have the ability to deal with time varying nature of the inputs for which these are found to be suitable for application like speech recognition [3]. A RNN is constituted for this work with one hidden layer and tan-sigmoid activation functions (Figure 4). The input layer size is equal to the length of the feature vector and the output layer is equal to the number of classes which in this case is ten. The Table IV shows its performance with a number of hidden neurons when trained upto 2000 epochs with ten numeral samples of *Girl Reference*. The RNN training is carried out using (error) back propagation with momentum(BPM) and Levenberg-Marquardth back propagation algorithms. The precision generated by both training methods are comparable but the time taken by the first one is more. But memory requirement of the second is higher. Hence, the results derived are the average values of both the methods. But for testing the RNN trained with BPM is adopted. The RNN configuration

TABLE V
VARIATION OF THE AVERAGE TRAINING TIME OF A MLP

Sl Num	Epoch	Time is sec.s
1	500	53.2
2	1000	89.1
3	1500	124.1
4	2000	153.4

TABLE VI
VARIATION OF THE AVERAGE TRAINING TIME OF A RNN

Sl Num	Epoch	Time is sec.s
1	500	38.2
2	1000	67.2
3	1500	82.1
4	2000	94.8

with 40 hidden neurons gives the best time and success-rate combination. It requires the least amount of time to attain a success rate of around 96% within 2000 epochs while taking several sets of *Girl Reference* sample which included SNR variations between 1 to 25 dB. Hence, this RNN configuration is used for performing the speech recognition.

F. Cooperative Heterogeneous ANN Architecture- configuration and training consideration

The best training and testing time performance recorded by the RNN based ASR is around 96 % which need to be improved. The work load given to the RNN block is next segregated into two sections configured to hand gender specific inputs. For that a Cooperative Heterogenous ANN (CHANN) architecture is formulated as shown in Figure 5. The entire system has two distinct parts for dealing with two classes of input classified into male and female clusters. The first block is formed by a Multi-Layer Perceptron (MLP) which acts like a class mapper network. It categories the inputs into two gender based clusters. It contains one hidden layer with tan-sigmoid activation function and one input and the other the output layer each fitted with log-sigmoid activation functions and is trained by the back propagation algorithm with Levenberg-Marquardt (LM) optimization. The classification is just in terms of two classes male and female. The decision of this network is placed as a class code into the input sample vector and then passed on to two RNN - blocks which performs the gender specific recognition. The MLP training time is as in Table V. After about 2000 epochs the MLP produces a class - mapping performance of around 98 % which is acceptable for taking its predicted output and the class codes generated for doing a gender specific clustering of the samples used for training. The next stage is to feed the gender specific features to the RNNs to train. The RNNs are faster but take different times to reach the desired goal. The average time taken by the RNNs to train upto 2000 epochs is shown in Table VI. The training of each of the RNN blocks are monitored separately. If one RNN trains faster and reaches the desired goal earlier, it is stopped while the other continues to train. The two criteria followed in the training of the RNNs are the mean square error (MSE) convergence and the classification rate. The RNNs are provided with individual training goals though these ultimately

TABLE IV
PERFORMANCE OF RNN WITH NUMBER OF HIDDEN NEURONS

Sl No	Predictor size	ANN training time in sec	% of successful recognition
1	20	28.3	82
2	25	32.4	84
3	30	36.1	86
4	35	48.3	91
5	40	54.8	96
6	45	88.2	93.4
7	50	99.3	93.2
8	55	108.2	93.8

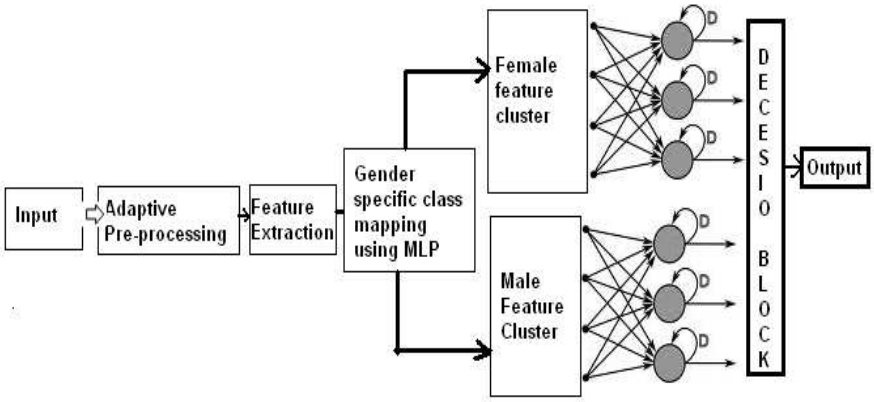

Fig. 5. Cooperative Heterogenous ANN (CHANN) architecture

dictate the global goal. The RNNs should be over-trained as there is always a chance of the ANNs will lose the ability to generalize.

V. PERFORMANCE EVALUATION

To evaluate the performance of the various filter structures, several sets of noise corrupted tests signals are created. The main objective is to make the recognition system compatible with noisy environment. The noisy inputs include SNR variation of 1 to 25 dB. After successfully filtering the noisy data the MSE of the filtered and unfiltered signals are computed. The results in Tables VII and VIII show MSE value evaluation for the speech signal sets with Transposed equi-ripple FIR structure. This filter shows the best performance. Some of the results derived using Transposed Equi-ripple FIR filter is as in Tables I and II.

Yet the performance is not dynamic. The adaptive filter design is, therefore, adopted and applied with the Transposed equi-ripple FIR block. The adaptive filter block is designed to generate a MSE value in the range of 0.09 to 0.0001 in case of an input sample compared to the the reference signal as shown in the Table IX. The adaptive filter block takes less than 10 iterations on average to generate the desired MSE value for the 40 training samples. The discontinuities from the filtered signal are removed using an Hamming window. The LPC parameters are evaluated from the windowed signal, using which the final cepstrum co-efficients are generated. Figure 6 and Figure 7, represents the corpus and their minimum phase reconstructed version of speech signals of *Girl Mood1* and *Boy Mood1*. The output of the corpus generation process can be represented as in Figures 8, which is for a particular speech signal *Xunya* (Assamese zero) from the set *Girl Mood1*. Finally the correlation between the noise free and noise mixed signal, noise mixed and recovered signal, noise free and recovered signal are observed. Such a correlation of speech signals from the sample set *Girl Mood1* and *Boy Mood1* is shown in Figure 9 and Figure 10.

The adaptive filter generates MSE values in 10^{-4} range which improves the recognition considerably. This is shown by the Table X.

The adaptive equalizer subsequently smoothens off other remaining distortions in the sample which now becomes suitable for feature extraction. The size of the feature set is an important criteria. If the feature set is selected without any logic, the results vary. Experiments are carried out to fix the length of the feature vector for generating the corpus. Tables XI shows the effect of predictor length in generating acceptable recognition performance by an ANN. The LPC predictor size of 20 gives the best results from the RNN during

TABLE VII
MSE FOR VARIOUS SNR VALUES PRODUCED BY THE TRANSPOSED FIR EQUI-RIPPLE FILTER FOR *Girl Mood1*

Sl No	Input	MSE in Various SNR for *Girl mood*-1					
		1 db	5 db	10 db	15 db	20 db	25 db
1	0	0.1988	0.0851	0.0345	0.0179	0.0128	0.0113
2	1	0.2188	0.1028	0.0515	0.0395	0.0297	0.0281
3	2	0.2057	0.0952	0.0440	0.0274	0.0226	0.0210
4	3	0.1920	0.0776	0.0266	0.0104	0.0052	0.0035
5	4	0.1936	0.0819	0.0328	0.0161	0.0111	0.0095
6	5	0.2066	0.0915	0.0423	0.0261	0.0210	0.0192
7	6	0.2059	0.0921	0.0423	0.0255	0.0207	0.0190
8	7	0.2172	0.1074	0.0561	0.0398	0.0343	0.0332
9	8	0.2254	0.1141	0.0626	0.0467	0.0413	0.0395
10	9	0.2141	0.1019	0.0504	0.0343	0.0293	0.0279

TABLE VIII
MSE FOR VARIOUS SNR VALUES PRODUCED BY THE TRANSPOSED EQUI-RIPPLE FIR STRUCTURE FOR *Boy Mood1*

Sl No	Input	MSE in Various AWGN Noise for *Boy mood*-1					
		1 db	5 db	10 db	15 db	20 db	25 db
1	0	0.2022	0.0877	0.0347	0.0192	0.0141	0.0124
2	1	0.1965	0.0809	0.0292	0.0135	0.0085	0.0067
3	2	0.1938	0.0888	0.0361	0.0205	0.0150	0.0136
4	3	0.2017	0.0882	0.0358	0.0204	0.0152	0.0136
5	4	0.1987	0.0842	0.0327	0.0203	0.0113	0.0109
6	5	0.1980	0.0893	0.0373	0.0208	0.0158	0.0142
7	6	0.1980	0.0888	0.0362	0.0203	0.0152	0.0135
8	7	0.2059	0.0906	0.0407	0.0243	0.0192	0.0175
9	8	0.1985	0.0810	0.0382	0.0166	0.0112	0.0096
10	9	0.1975	0.0843	0.0325	0.0164	0.0113	0.0097

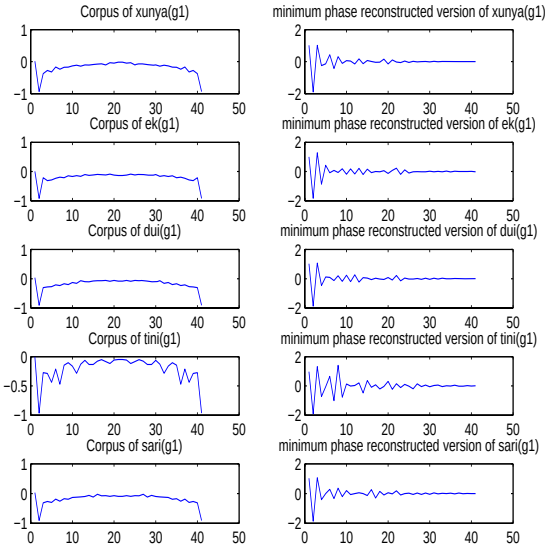

Fig. 6. Corpus set of *xunya* (Assamese zero) to *sari* (Assamese four) of girl mood1

TABLE IX
AVERAGE MSE OF VARIOUS SAMPLE INPUTS USING ADAPTIVE LMS FILTER BLOCK

Sl No	Speech Signal	MSE for the Speech Signal			
		Girl Mood-1	*Girl Mood*-2	*Boy Mood*-1	*Boy Mood*-2
1	0 (*Xunya*)	0.000059	0.000025	0.000057	0.000054
2	1 (*Ek*)	0.000002	0.000073	0.000014	0.000018
3	2 (*Dui*)	0.000035	0.000064	0.000095	0.000049
4	3 (*Tini*)	0.000025	0.000070	0.000016	0.000029
5	4 (*Sari*)	0.000044	0.000044	0.000017	0.000092
6	5 (*Pass*)	0.000092	0.000038	0.000073	0.000041
7	6 (*Soy*)	0.000039	0.000005	0.000003	0.000004
8	7 (*Saat*)	0.000015	0.000040	0.000060	0.000068
9	8 (*Aath*)	0.000007	0.000062	0.000009	0.000026
10	9 (*Nau*)	0.000081	0.000045	0.000046	0.000029

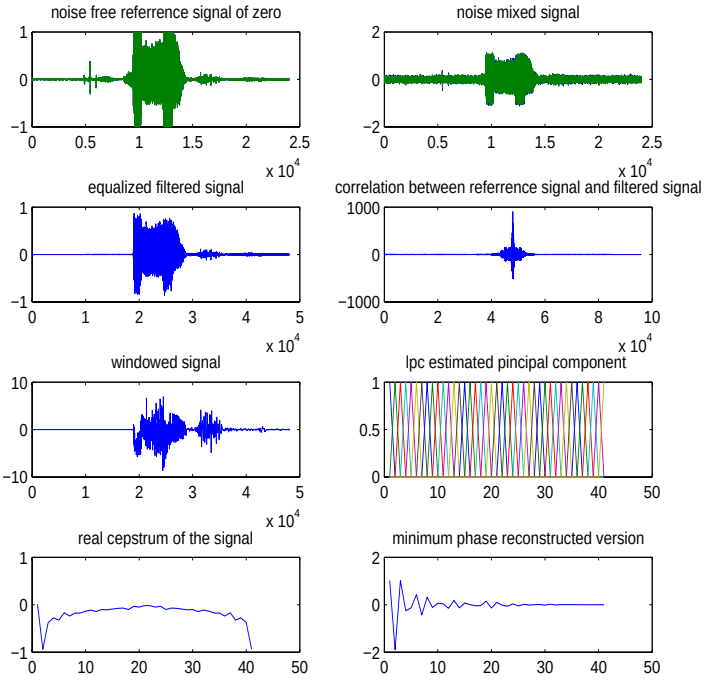

Fig. 8. Various Stages in corpus generation

training. Its successful recognition rate and the time required to reach the desire MSE value is the best among the eight cases considered. The time-precision combination of the 20^{th} order LPC prediction when applied to an RNN generates the best precision level in recognizing the input samples. If the prediction size is too long, the training time becomes extended which is not desirable. The results generated are the average values for 40 training samples applied to an RNN with 40 hidden neurons.

Under test conditions, the RNN shows results which approximately match the results provided in Table XI for a 40 hidden neuron case and the 20 predictor size input sample as included in Table IX. A similar set of results have also been derived for the MLP which took less time to train compared to the RNN but the success rate was around 93 % for the set of speech signals. With a feature size of 25, on an average for 50 samples of each of the 10 sets of noise-free, noise mixed, stress free and stressed samples with gender specific recognition, the success rates of the singular RNN block is around 96 % but using the CHANN this rate improves to about 97.5 %. The average performance of the CHANN block for 50 samples of each of the 10 sets of noise-free, noise mixed, stress free and stressed samples are as in Table XII. The results clearly indicate that the RNN as a singular block at times though fails to make correct recognition and its performance suffers while dealing with a composite input, the CHANN with gender specific feature cluster is better equipped to deal with such variations. The effectiveness of the corpus set generated using

TABLE X
VARIATION OF ANN RECOGNITION RATE WITH TWO DIFFERENT FILTER STRUCTURES

Sl No	Filter Type	Average MSE	percentage of recognition
1	Transposed Equiripple FIR filter	0.013	92
2	Adaptive FIR Design	0.00013	95.6

TABLE XI
PERFORMANCE OF RNN V/S PREDICTOR SIZE

Sl No	Predictor size	RNN training time in sec	% of successful recognition
1	5	24.3	76
2	10	29.4	79
3	15	35.2	84
4	20	43.8	95
5	25	44.8	94
6	30	48.2	93.4
7	35	59.1	93.6
8	40	88.2	94.2

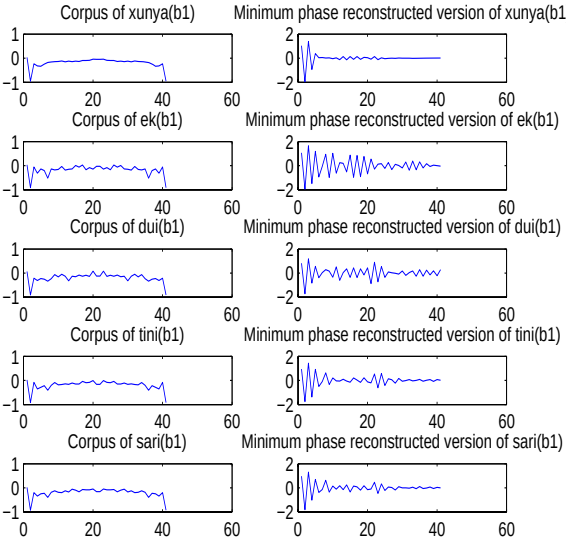

Fig. 7. Corpus set of *xunya* (Assamese zero) to *sari* (Assamese four) of boy mood1

TABLE XII
AVERAGE PERFORMANCE OF THE CHANN ARCHITECTURE SHOWING RESULTS OF MALE AND FEMALE VOICES UTTERING *sunya* (ZERO IN ASSAMESE)

Code book Size	Input Sample	Recognition in %
25	Male- noise-less	97.1
	Male- noise-mixed	95.1
	Male- stressed	95.8
	Male- stress free	97.6
	Female- noise less	97.2
	Female- noise-mixed	95.1
	Female- stressed	94.4
	Female- stress free	97.7

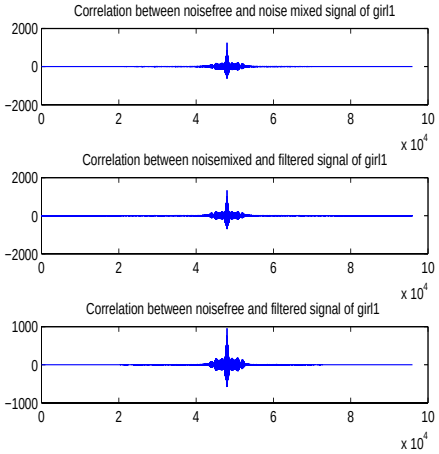

Fig. 9. Correlation for *Girl Mood1*

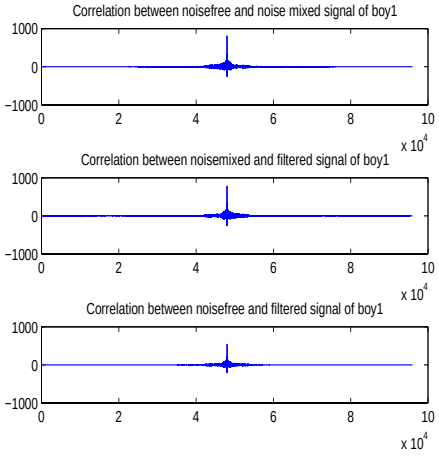

Fig. 10. Correlation for *Boy Mood1*

the adaptive filter block for speech recognition of Assamese numerals using an RNN block and a CHANN block is thus apparent.

VI. Conclusion and Further Direction

The work shows the role played by an adaptive pre-emphasis filter in extracting features of Assamese numerals captured with gender, mood and recording environment variations. The result also demonstrate the superiority of a RNN compared to a feed-forward network like the MLP in speech processing applications. The RNN is more suitable for time -varying signal inputs. If the RNN is used as a part of a heterogenous cooperative block, the performance improves further. The work can include vowels and consonants as well to extent it as an effective tool for Automatic Speech Recognizer in Assamese language. An extended form of the work can include a speech to text converter in Assamese.

References

[1] A. Okatan1, N. Ayanolu, S. Senycel, *Voice Recognition by Cepstrum Method*, Baheehir University, Faculty of Engineering, Department of Computer Eng., Istanbul, Turkey International Intelligent Knowledge Systems Society (IKS), Istanbul, Turkey.
[2] Wikipedia, the free encyclopedia*"Speech corpus"*, en.wikipedia.org/wiki/Speechcorpus.
[3] S. Haykin, *Neural Networks A Comprehensive Foundation*,2^{nd} . Pearson Education, New Delhi, 2003.
[4] K. K. Paliwal and W. B. Kleijn, *Quantization of LPC Parameters*,
[5] Prof. Gautam Baruah, Dept. of CSE, IIT Guwahati, *tdil.mit.gov.in / assamesecodechartoct02.pdf*,
[6] *"The X sound in Assamese language"*, The Assam Tribune Editorial, March 5, 2006.
[7] *Indo-Iranian*. http://www.questia.com/library / encyclopedia/ indo-iranian.jsp
[8] "National Institute on Deafness and Other Communication Disorders", (www.nidcd.nih.gov/health/voice/whatisvsl.htm),
[9] B. Yegnanarayana, *Artificial Neural Networks*, 1^{st} Ed., PHI, New Delhi, 2003.
[10] A. P. Simpson, "Phonetic differences between male and female speech", Language and Linguistics Compass 3/2, 621640, 2009.
[11] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, 1^{st} Ed., Prentice Hall, 1978.
[12] S. Haykins,*Adaptive Filter Theory*, 4^{th} Ed., Pearson Education, New Delhi, 2002.
[13] K. Hisashi, F. T. Mano, *Patent application title: Filter Circuit*, mi.eng.cam.ac.uk / ajr / SA95/ $node43.html$.
[14] S. W. Smith, *The Scientist and Engineer's Guide to Digital Signal Processing,* 2^{nd} *ed.*, Available at www.healthcare.analog.com / static /imported-files /tech... / dsp-book-frontmat.pdf.
[15] *Introduction to Digital Filters*, www.dsptutor.freeuk.com / digfilt.pdf .
[16] *Introduction to DSP - filtering: design by equiripple method*, www.bores.com / courses / intro/ filters/4_equi.htm.
[17] Wikipedia, the free encyclopedia, *Least Mean Square Filter*, www.bores.com / courses / intro/ filters/4_equi.htm.
[18] *Feature Extraction*, cslu.cse.ogi.edu /toolkit /old /old /version $2.0a$ /.../ $node5.html$.
[19] M. P. Kesarkar, *Feature Extraction for Speech Recogntion*, M.Tech. Credit Seminar Report, Electronic Systems Group, EE. Dept, IIT Bombay, November, 2003.
[20] Jurafsky, Daniel and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, (1^{st} ed.). Prentice Hall, 2000.
[21] A. K. Paul, D. Das, and Md. M. Kamal, *Bangla Speech Recognition System Using LPC and ANN*, 3rd ed. Proccedings of Seventh International Conference on Advances in Pattern Recognition,04-06, February, 2009.
[22] G. Dede and M. H. Sazl, *Speech recognition with artificial neural networks*, Digital Signal Processing), Volume 20, Issue 3, Pages 763-768, May 2010.
[23] A. M. Ahmad, S. Ismail, D. F. Samaon, *Recurrent Neural Network with Backpropagation through Time for Speech Recognition*, Proccedings of Intemational Symposium on Communications and Information Technologies 2004 (ISCIT 2004) Sapporo, Japan, October 26- 29, 2004. Harlow, England: Addison-Wesley, 1999.

Mousmita Sarma , completed BSc in Electronics from Gauhati University, Assam, India in 2008. She is presently pursuing MSc in Electronics and Communication Technology at Department of Electronics and Communication Technology, Gauhati University, Assam, India. Her areas of interest include Adaptive Filters, Speech Processing and Applications of ANN.

Krishna Dutta , completed BSc in Electronics from Gauhati University, Assam, India in 2008. He is presently pursuing MSc in Electronics and Communication Technology at Department of Electronics and Communication Technology, Gauhati University, Assam, India. His areas of interest include Adaptive Filters, Speech Processing and Applications of ANN.

Kandarpa Kumar Sarma , presently with the Department of Electronics and Communication Technology, Gauhati University, Assam, India, completed MSc in Electronics from Gauhati University in 1997 and MTech in Digital Signal Processing from IIT Guwahati, Guwahati, India in 2005 where he further continued his research work. His areas of interest include Applications of ANNs, Neuro-Computing, Document Image Analysis, 3-G Mobile Communication and Smart Antenna.