Recognition of Assamese Spoken Words using a Hybrid Neural Framework and Clustering Aided Apriori Knowledge

PALLABI TALUKDAR, MOUSMITA SARMA and KANDARPA KUMAR SARMA
Gauhati University
Department of Electronics and Communication Technology
Guwahati-781014, Assam
INDIA
(pallabiz95, go4mou, kandarpaks)@gmail.com

Abstract: In this paper, an Artificial Neural Network (ANN) based model is proposed for recognition of discrete Assamese speech using a Self Organizing Map (SOM) based phoneme count determination technique. The phoneme count determination technique takes some initial decision about the possible number of phonemes in the word to be recognized and accordingly the word is presented to some N-phoneme recognition algorithm. In this paper recognition algorithm is designed to recognize three phoneme consonant-vowel-consonant (CVC) type Assamese words. The word recognizer is consisted of another SOM block to provide phoneme boundaries and Probabilistic Neural Network (PNN) and Learning Vector Quantization (LVQ) to identify the SOM segmented phonemes. The recognition of constituent phonemes in turn represents the discrimination between incoming words with a minimum success rate of 90%.

Key–Words: Formant, Phoneme, ANN, KMC, LPC, DWT.

1 Introduction

Spoken word recognition is a distinct subsystem providing the interface between low-level perception and cognitive processes of retrieval, parsing, and interpretation of speech. The process of recognizing a spoken word starts from a string of phonemes, establishes how these phonemes should be grouped to form words, and passes these words onto the next level of processing. There are several theories in the literature which focuses on the discrimination and categorization of speech sounds. One of the earliest theory in this contrast is the Motor theory proposed by Alvin Liberman, Franklin Cooper, Pierre Delattre and other researchers in 1950 [1]. Motor theory postulates that speech is perceived by reference to how it is produced. It means that, when perceiving speech, listeners access their own knowledge of how phonemes are articulated. Analysis by Synthesis Model, proposed by Stevens and Halle, 1960 [2], in turn stated that speech perception is based on auditory matching mediated through speech production. Research on the discrimination and categorization of phonetic segments was the key focus of the works on speech perception before 1970s. The processes and representations responsible for the perception of spoken words became a primary object of scientific inquiry with a curiosity of disclosing the cause and methods of how the listener perceives fluent speech. According to Cohort theory (1980) [3][4], various language sources like lexical, syntactic, semantic etc. interact with each other in complex ways to produce an efficient analysis of spoken language. It suggest that input in the form of a spoken word activates a set of similar items in the memory, which is called word initial cohort. The word initial cohort consisted of all words known to the listener that begin with the initial segment or segments of the input words. In 1994, Marslen-Wilson and Warren revised the Cohort theory. In the original version, words were either in or out of the word cohort. But in the revised version, words candidate vary in the activation level, and so membership of the word initial cohort plays a important role in the recognition [3] [5] . According to the TRACE model of spoken word recognition, proposed by McClelland and Elman, in 1986 [6] [4], speech recognition can be described as a process, in which speech units are arranged into levels which interact with each other. There are three levels: features, phonemes, and words. There are individual processing units or nodes at three different levels. Dennis Norris in 1994, have proposed another model of spoken word recognition called the Shortlist model [4]. According to Norris, a shortlist of word candidates is derived at the first stage of the model. The list consists of lexical items that match the bottom-up speech input. This abbreviated list of lexical items enters into a network of word units in the later stage, where lexical units compete

with one another via lateral inhibitory links [4]. From all these theories of spoken word recognition we can arrive at the conclusion that the spoken word recognition is a complex multileveled pattern recognition work performed by neural networks of human brain and the related speech perception process can be modeled as a pattern recognition network. Different levels of the language like lexical, semantic, phonemes can be used as the unit of distribution in the model. All the theories proposed that bottom up and top down processes between feature, phoneme, and word level combines to recognize a presented word. In such a situation, Artificial Neural Network (ANN) models has greatest potential, where hypothesis can be performed in a parallel and high computational approach. ANN models are composed of many non-linear computational elements operating in parallel and arranged in the pattern of biological neural network. The brain's impressive superiority while deal with a wide range of cognitive skills like speech recognition, has motivated researchers to explore the possibilities of ANN models in the field of speech recognition in 1980s [7] with a hope that human neural network like models may ultimately lead to human like performance on such a complex tasks. A few ANN based work on speech recognition are described in . But at the later half of 1990, suddenly ANN based speech research remained dormat, at times terminated [7]. Statistical frameworks like Hidden Markov models (HMMs), Gaussian Mixture Models (GMMs) etc. received attention supporting that supports both acoustic and temporal modeling of speech. However, it should be mentioned that the currently available best systems are far from equaling human like performance and many important research issues are still to be explored. Also, ANN is able to work with model free data unlike HMM and can be 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. Therefore, the value of ANN based research is still large and now a days it is considered as the hot field in the domain of speech recognition [8], [9], [10], [11], [12], [13], [14]. This work presents a novel approach to spoken word recognition, where a phoneme count determination technique is used to take some initial decision about the number of phonemes in the word to be recognized and accordingly the word is directed to the 3-phoneme, 4-phoneme or 5-phoneme word recognition model. Here, we have used the 3-phoneme word recognition algorithm explained in [15], which is designed using a Self Organizing Map (SOM) based phoneme segmentation algorithm and Probabilistic Neural Network (PNN) and Learning Vector Quantization (LVQ) based decision algorithm . The phoneme count determination technique is designed using two different approach, initially using K-means clustering (KMC) and then using SOM based clustering along with a Recurrent Neural Network (RNN) block for decision making. The KMC based technique can take around 83% correct decision about the number of phoneme. If the initial decision about the number of phoneme goes wrong, then the success rate of 3-phoneme word recognition suffers. Therefore, the SOM clustering based phoneme count determination technique is adopted and it provides superior performance in terms of percentage of correct decision. In the CVC word recognition method, SOM is trained with different iteration numbers for the same word and thus different weight vectors are obtained, which is considered as different phoneme boundaries. The phoneme segment obtained by training SOM with various iteration numbers are then matches with the PNN patterns. While taking decision about the last phoneme the algorithm is assisted by LVQ codebook which contains a distinct code for every word of the recognition vocabulary. Assamese CVC words are recorded for this work in a noise free environment from five male and five female speakers of varying age. The description included here is organized as below. Section 2 provides a brief description of the phonemical structure of Assamese language. Section 3 explains the proposed word recognition model which consist of two steps - the phoneme count determination and 3-phoneme word recognition block. The results and the related discussions are included in Section 4. Section 5 concludes the description.

2 Phonemical Structure of Assamese Language

Assamese is an Indo-Aryan language originating from the Vedic dialacts, and therefore, is a sister of all the northern Indian languages. Although the exact nature of the origin and growth of the language is yet to be ascertained with certainity, it is supposed that like other Aryan languages, Assamese was also born from $Apabhra\acute{m}\acute{s}a$ dialects developed from $M\tilde{a}gadhi$ Prakrit of the eastern group of Sanskritic languages [16]. Retaining certain features of its parent Indo-European family it has got many unique phonological characteristics which makes Assamese speech unique and hence requires a study exclusively related to the design of a speech recognition / synthesis system in Assamese [17].

The Assamese phoneme tables obtained from [16] are shown in Figure 1 and Figure 2. There are twenty-three consonants and eight vowel phoneme in the stan-

Consonants

	Bilabial		Alveolar		Palatal	Velar		Glottal
	Vl.	Vd.	Vl.	Vd.	Vd.	Vl.	Vd.	Vd.
Unaspirated	p	b	t	d		k	G	
Aspirated	ph	bh	th	dh		kh	gh	
Spirant			s	z		x		fi
Nasals		m		n			ŋ	
Lateral				l				
Trill				r				
Frictionless Continuant		w			j			

Figure 1: Assamese Consonant Phonemes, Vl.-Voiceless, Vd.-Voiced [16]

Vowels

	Front	Central	Back
High	i		u
Higher-mid	e		o
Lower-mid	ɛ		ɔ
Low		a	ɒ

Figure 2: Assamese Vowel Phonemes [16]

dard colloquial Assamese . The consonants may be grouped into two broad divisions: the stops and the continuants. For the stops there are contrast in three points of articulation- the lips, the alveola, and the velum and four-way contrasts in every point as to the presence or otherwise of voice and aspiration. Therefore, a stop may be voiced or voiceless, aspirated or unaspirated. There are continuants- two frictionless, viz, the semivowels /w,j/, four spirants /s z x h/, one lateral /l/, one trill /r/ and three nasals /m, n/ which are stops as well as continuant both at once [16].
The eight vowels present three different types of contrasts [16]. Firstly, eight-way contrasts in closed syllables, and in open syllables when /i u/ do not follow in the next immediate syllable with intervention of a single consonants except the nasal. Again, it shows six-way contrast in open syllables with /i/ occurring in the immediately following syllable with intervention of any single consonant except the nasals, or except with nasalization and finally, five way contrasts in open syllables when /u/ occurs in the immediately following syllable with a single consonant intervening [16].

3 The Proposed Word Recognition Model

The proposed spoken word recognition model can be described by the block diagram of Figure 3. The model consist of two phases-

- Phoneme count determination and

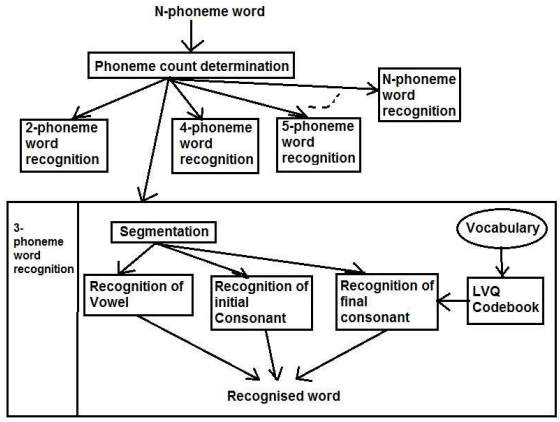

Figure 3: Process logic of the proposed work

- Three phoneme word recognition

The phoneme count determination block determines the number of phonemes in the word and on the basis of that decision, the word can be presented to some N-phoneme word recognition algorithm, where N can be 2, 3, 4, 5 etc. Here, we have described a novel algorithm for recognition of 3-phoneme CVC type words, which uses three different ANN structure, namely SOM, PNN and LVQ. The following section describes both these phases separately.

3.1 Phoneme Count Determination

To develop a discrete speech recognition model for Assamese language, the system should have the capability of dealing with words having different CV combinations. Therefore, in order to enable the model to deal with multiple phoneme case, a block is created, so that the algorithm can take some prior de-

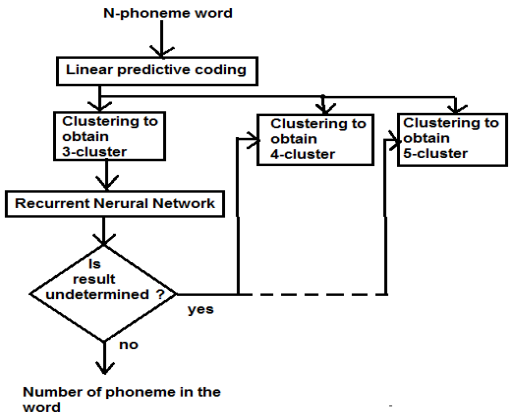

Figure 4: Phoneme Count Determination Block

cision about the number of phonemes in any incoming word. Since, discrete speech signals are consisted of word units separated by some intentional pause or silent part, therefore, it is not very difficult to omit the silent part from the signal in order to separate the words in the speech signal. A simple intensity threshold method can be used for such kind of word- silent separation. The silent part ideally has zero intensity. Thus, a threshold can be put to remove that part of the speech signal which has zero intensity. But the crucial part of discrete speech recognition is to recognize the individual word consisting of multiple phonemes. As a solution to such a problem, here we have proposed an ANN based recognition model where the recognition algorithm initially decides the number of phonemes in the word and then according to that decision the word is presented to recognize the constituent phoneme in some N-phoneme recognition units designed separately. Therefore, a phoneme count determination block is designed which is shown in Figure 4. Two different approaches are carried out to design the phoneme count determination block as explained below.

3.1.1 KMC based Phoneme Count Determination Block

Clustering is the process of partitioning or grouping a given set of patterns into disjoint clusters. This is done such that patterns in the same cluster are alike and patterns belonging to two different clusters are different. Clustering has been a widely studied problem in a variety of application domains including ANNs. The number of clusters k is assumed to be fixed in k-means clustering. In this section, we propose probable technique for generating some apriori knowledge about

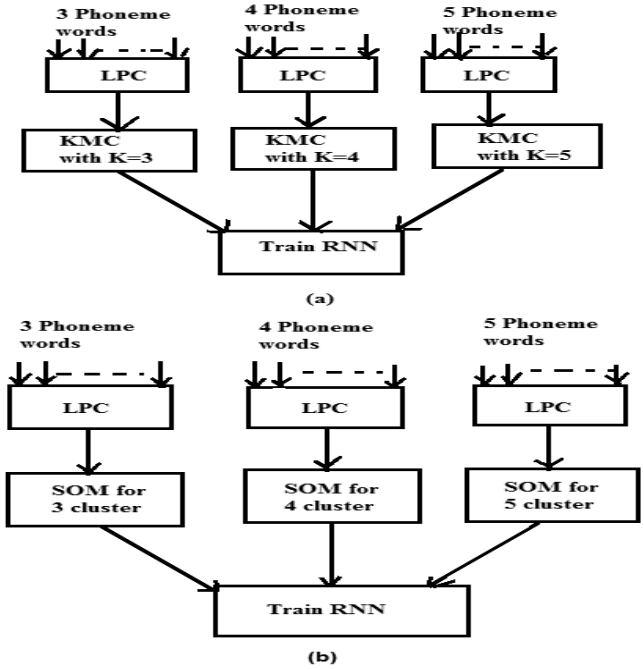

Figure 5: Word Clustering Technique

number of phonemes in an word, without segmentation. The method involves clustering N-phoneme words into N-cluster using KMC. Here, N is the value of k in the KMC. The word clustering technique proposed here, can be visualized from part (a) of Figure 5.

3.1.2 SOM based Phoneme Count Determination Block

The same phoneme count determination block is later redesigned using the SOM. Here, the KMC blocks are replaced by SOM competitive layers. The SOM training algorithm resembles k-means algorithm in the sense that it partitions the input data space into a number of clusters of wining neuron and its neighbors with similar weight vectors. Therefore SOM can be used for clustering data without knowing the class memberships of the input. The SOM algorithm is based on unsupervised, competitive learning, where clusters are formed depending upon a self-organization process of the constituent neurons which groups data depending upon a similarity measure. The similarity measure is decided upon by a euclidian distance between the random connectionist values and the input and finally optimized by a Gaussian spread.The SOM based phoneme count determination technique can be summarized from the part (b) of Figure 5.

3.1.3 Role of RNN in Decision Making of the Proposed Technique

RNN is a special structure of supervised ANN known for data recurrency and is suited for capturing the temporal nature of speech signals [7] . Therefore it has been chosen for taking decision about the number of cluster in a clustered speech data. An RNN is trained with the clustered data to learn the number of cluster so that it can classify any unknown clustered word according to the number of cluster. The sample clustered data set is obtained from KMC or SOM, where 3-phoneme words has 3 clusters, 4-phoneme words has 4 clusters and 5-phoneme words has 5 clusters. Accordingly, the RNN classifies the data into 3, 4 and 5-cluster groups and suppose they are stored as DK3, DK4 and DK5 respectively. Here, we have considered only certain variations like 3, 4 and 5- phonemes. Now, any N-phoneme word obtained from the discrete speech signal, is first clustered with K=3, 4, 5. Next, at first DK3 is presented to the trained RNN to classify into any of the three classes. If RNN fails to classify, then DK4 and DK5 is presented consequently. If any of DK3, DK4 and DK5 do not come under any of the classes defined by RNN, then that word is discarded. The decision making process of the RNN is depicted in Figure 4. The same decision making method is used in case of clustered data set obtained from both KMC and SOM based clustering technique. Comparing the success rate of correct decision for both the techniques it is obtained that the SOM based technique can provide more correct decisions (around 7 $\%$ improvement) and therefore it is used with the 3-phoneme word recognition technique in the proposed model. The result of the both the techniques are shown in Section 4.3

3.2 Recognition of Three Phoneme Words

SOM has a special property of effectively creating spatially organized "internal representations" of various features of input signals and their abstraction [18]. SOMs can be considered as a data visualization technique i.e. it provides some underlying structure of the data [18]. This idea is used in our phoneme segmentation technique, process logic of which is given in Figure 2. The weight vector obtained by training an one dimensional SOM with the LPC features of a word containing the phoneme to be segmented, is used to represent the phoneme. Training the same SOM with various iteration numbers, we get different weight vectors, each of which is considered as a segment of different phonemes constituting the word. The weight vectors thus obtained are classified into vowel, initial consonant and final consonant. The work only covers the unaspirated phoneme family, which have phonemes like /p/, /b/, /t/, /d/, /k/ and /g/ and all the eight vowel of Assamese language [19]. The classification is done by PNNs trained with these clean unaspirated phonemes and vowel phonemes. The PNN is based on statistical principles derived from Bayes decision strategy and non-parametric kernal based estimators of probability density function (pdf)s. It finds pdf of features of each class from the provided training samples using Gaussian Kernel. These estimated densities are then used in a Bayes decision rule to perform the classification. Advantage of the PNN is that it is guaranteed to approach the Bayes optimal decision [20]. The proposed three phoneme word recognition algorithm can be stated in two distinct parts-

- SOM based Segmentation Algorithm and
- PNN and LVQ based Recognition algorithm

The word recognition algorithm is well described in [15]. Although the following sections provides a brief description of the two parts of the algorithm.

3.2.1 SOM based Segmentation Algorithm

The SOM weight vector extraction algorithm can be visualize from the block diagram of Figure 6. The SOM weight vectors thus extracted are stored as SW1, SW2, SW3, SW4, SW5 and SW6. SOMs role is to provide segmentation boundaries for the phonemes. Here six different segmentation boundaries are obtained for six separate sets of weight vectors. The segmented phonemes are next applied to the PNN and LVQ based decision algorithm which performs the role of decision making for constituent vowel and consonant phoneme.

3.2.2 PNN and LVQ based Recognition Algorithm

Here we have performed some two class PNN problem, where three PNNs are trained with two clean unaspirated consonant phonemes and are named as PNN1, PNN2 and PNN3. Similarly, four other PNNs are trained with clean vowel phonemes and are named as PNN4, PNN5, PNN6 and PNN7. That means the output classes of PNN1 are /p/ and /b/, the output classes of PNN2 are /t/ and /d/, the output classes of PNN3 are /k/ and /g/, the output classes of PNN4 are /i/ and /u/ etc. and this way we obtain total seven PNNs which are considered to know the patterns of all the unaspirated consonant phonemes and vowel phonemes of Assamese.

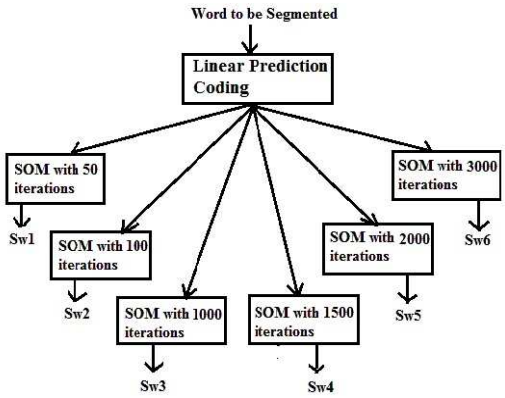

Figure 6: SOM Segmentation Block [**?**] [15]

The reason behind the use of two class PNNs is to increase the success rate by performing a binary level decision. Although it increases the computational complexity and memory requirements, it is tolerable for the sake of increasing success rate. Since PNN is the fastest, use of more than one PNN does not harm much the speed of the algorithm. PNNs handle data that has spikes and points outside the norm better than other ANNs. Therefore, PNN is suitable for problems like phoneme classification [20]

As mentioned earlier, the segmented vowel and consonant phonemes are identified with the help of seven PNNs trained with clean phonemes. An LVQ Codebook is used along with the PNN based recognition algorithm while deciding the last phoneme of the incoming word. The vocabulary used for this work contains words having the last phoneme any of /t/, /k/, /r/ and /n/. One four class PNN is trained with clean phonemes to learn all these four phoneme patterns separately. This PNN is used to recognize the last phoneme from the SOM segmented phonemes of the incoming word. While taking decision about the last phoneme the algorithm is assisted by the LVQ codebook. Learning vector quantization (LVQ) is a method for training competitive layers in a supervised manner. In an LVQ network competitive layer learn to classify input vectors into target classes chosen by the user unlike strictly competitive layer possessed by SOM [21]. The codebook designed by LVQ has a distinct code for every word of the vocabulary. Suppose, the PNN decides about a phoneme, but if no words of the vocabulary end with that phoneme then the decision is discarded. Thus, the codebook assistance assures that most likely decision about the last phoneme can be obtained.

Table 1: Table representing phoneme values for C_i, V_j and C_k

i/j/k	C_i	V_j	C_k
1	$/p/$	$/i/$	$/p/$
2	$/b/$	$/u/$	$/b/$
3	$/t/$	$/e/$	$/t/$
4	$/d/$	$/o/$	$/d/$
5	$/k/$	$/\varepsilon/$	$/k/$
6	$/g/$	$/ao/$	$/g/$
7	$/ph/$	$/a/$	$/ph/$
8	$/bh/$	$/oo/$	$/bh/$
9	$/th/$		$/th/$
10	$/dh/$		$/dh/$
11	$/kh/$		$/kh/$
12	$/gh/$		$/gh/$
13	$/s/$		$/s/$
14	$/z/$		$/z/$
15	$/x/$		$/x/$
16	$/h/$		$/h/$
17	$/m/$		$/m/$
18	$/n/$		$/n/$
19	$/l/$		$/\eta/$
20	$/r/$		$/l/$
21			$/r/$
22			$/w/$
23			$/j/$

4 Experimental Result and Discussion

The work is carried out as per the flow diagram of Figure 3. Mathematically, the problem in hand can be stated as-

Suppose, C_i is the initial consonant phoneme which can be vary within all the phoneme families, V_j be the vowel phoneme and C_k be the last consonant phoneme.
Then any incoming word may have the form
X=$C_iV_jC_k$
where, i = 1 to 20 (excluding /η, w, j/)
j = 1 to 8 and
k = 1 to 23.
Identify, C_i, V_j and C_k.
The values of C_i, V_j and C_k are given in Table 1.

4.1 Experimental Speech Signals

The experimental speech samples are recorded in three phase. At first clean consonant and vowel

Table 2: CVC Type Word List of for Spoken Word Recognition Model

Sl No	Vowel (V)	Initial Consonant(C)	Last Consonant(C)	CVC Word
1	/i/	/x/	/t/	/xit/
2			/r/	/xir/
3			/k/	/xik/
4			/s/	/xis/
5	/u/	/d/	/kh/	/dukh/
6			/r/	/dur/
7			/t/	/dut/
8			/b/	/dub/
9	/e/	/b/	/x/	/bex/
10			/d/	/bed/
11			/s/	/bes/
12			/l/	/bel/
13	/o/	/m/	/n/	/mon/
14			/k/	/mok/
15			/r/	/mor/
16			/t/	/mot/

Table 3: CVC Type Word List of for Spoken Word Recognition Model

Sl No	Vowel (V)	Initial Consonant (C)	Last Consonant (C)	CVC Word
17	/ɛ/	/kh/	/l/	/khɛl/
18			/p/	/khɛp/
19			/d/	/khɛd/
20			/r/	/khɛr/
21	/ao/	/r/	/n/	/raon/
22			/kh/	/raokh/
23			/th/	/raoth/
24			/x/	/raox/
25	/a/	/l/	/bh/	/labh/
26			/z/	/laz/
27			/th/	/lath/
28			/kh/	/lakh/
29	/oo/	/n/	/l/	/nool/
30			/d/	/nood/
31			/kh/	/nookh/
32			/m/	/noom/

phonemes are recorded from five girl speakers and five boy speakers, which results in a total of following broad sets of speech signals: *Girl1*, *Girl2*, *Girl3*, *Girl4*, *Girl5*, *Boy1*, *Boy2*, *Boy3*, *Boy4* and *Boy5*. These are used to train PNNs. The description included in Section 2 simply revels the fact that Assamese consonant phonemes can be classified into six distinct phoneme families as *unaspirated*, *aspirated*, *spirant*, *nasal*, *lateral* and *trill*, excluding the semivowels /w/ and /j/. The sample words selected for this work covers variation in the initial phoneme from different families as well as variation of all the vowel phonemes. For the second phase of collecting samples, a few samples collected are shown in Table 2 and Table 3, which are recorded from the same girl and boy speakers used in the first phase. The third phase covers some more samples recorded with mood variations, for testing the ANNs. Further the third phase samples are extended by adding $\pm 3db$ white gaussian noise to it.For recording the speech signal, a PC headset and a sound recording software, Gold Wave, is used. The recorded speech sample has a duration of 2 seconds, sampling rate of 8000 samples/second and bit resolution of 16 bits/sample.

4.2 Preprocessing

The pre-processing of the speech signal consist of two operations namely-smoothing of the signal by median filter and removal of the silent part by threshold method. Although the speech signals are recorded in a noise free environment, presence of some low frequency distortion is observed. Therefore, a median filtering operation is performed on the raw speech signals, so that the phoneme segmentation does not suffer due to any type of unwanted frequency component [22] [23].

The smoothed signal S_{smooth} contains both speech and non speech part. The non-speech or silent part occurs in a speech signal due to the time spend by the speaker before and after uttering the speech and this time information is considered to be redundant for phoneme segmentation purpose. The silent part ideally have zero intensity. But in practical cases it is observed that even after smoothing, the silent part of the speech signal have intensity about 0.02 in the scale 1. Our silent removing algorithm considers this intensity value as a threshold. Thus, a pure signal containing only the necessary speech part is obtained. These preprocessing operations are applied to every speech signal used in the work.

4.3 Phoneme Count Determination Result

Any word coming to the recognizer is first clustered with KMC algorithm for k-value 3. Then LPC features of the clustered data is extracted and presented to the RNN for classification. The RNN is trained to learn number of clusters in the data. If RNN fails to classify the data into any of the defined class, then the

Table 4: Word List Prepared for Clustering Technique

Sl No	3-phoneme Words	4-phoneme Words	5-phoneme Words
1	/xit/	/xiki/	/xital/
2	/xir/	/xakhi/	/xapon/
3	/xik/	/xati/	/xijal/
4	/xis/	/xari/	/xiphal/
5	/dukh/	/dili/	/datal/
6	/dur/	/dukhi/	/dupat/
7	/dut/	/duni/	/duphal/
8	/dub/	/dora/	/dojal/
9	/bex/	/besi/	/besan/
10	/bed/	/beli/	/betal/
11	/bes/	/burhi/	/betan/
12	/bel/	/bora/	/bixal/
13	/mon/	/mona/	/mohar/
14	/mok/	/mora/	/methon/
15	/mor/	/mula/	/mukut/
16	/mot/	/muthi/	/muazar/

Table 5: Word list Prepared for Clustering Technique

Sl No	3-phoneme words	4-phoneme words	5-phoneme words
17	/khɛl/	/kheda/	/khabar/
18	/khɛp/	/bhada/	/khangal/
19	/khɛd/	/dhara/	/khorak/
20	/khɛr/	/thaka/	/khorang/
21	/raon/	/roza/	/ratan/
22	/raoth/	/ronga/	/rabar/
23	/raokh/	/rati/	/ragar/
24	/raox/	/roa/	/razan/
25	/labh/	/lopha/	/lagan/
26	/laz/	/loa/	/labar/
27	/lath/	/lora/	/lasit/
28	/lakh/	/lani/	/lagar/
29	/nool/	/noga/	/nazar/
30	/nood/	/nila/	/natun/
31	/nookh/	/nura/	/nadan/
32	/noom/	/nija/	/nagar/

word is clustered with KMC algorithm for k-value 4 and the process is repeated. If the RNN again fails to classify the data then the word is clustered with k-value 5 and the same process is repeated again. In this way, the proposed logic is used to determine the possible number of phonemes in the word for 3-, 4- and 5-phoneme words. If the RNN fails to take any decision, then that particular word is discarded.

Selection of proper set of sample words as a representative of all the N-phoneme words, where (N=2, 3 or 4) is an important factor for better success rate of the proposed technique. The RNN should be trained with enough data, so that it can learn the difference between 3, 4 and 5-phoneme words. The work described in this paper experimented the proposed technique only for 3-phoneme, 4-phoneme and 5-phoneme words. Accordingly a word list is prepared. The word list is shown in Table 4 and Table 5. Here, we have considered words having all the Assamese vowel variations. Words with all the possible vowel and consonant combination of Assamese vocabulary if are used for the purpose shall provide better results. The words are recorded from selected speakers with 70 % of the samples used for training and 30 % for testing the KMC technique.

The probability of correct decision depends on several factors. Varying LPC predictor length significantly affects the success rates and training time. Table 6 shows overall success rate for 3-, 4- and 5-phoneme words with varying predictor length and corresponding RNN training time. As can be seen from Table 6, with 50 predictor size, we obtain minimum

Table 6: Performance of RNN v/s Predictor Size for Clustering Technique

Sl No	RNN training Time in sec	Predictor Size	Success Rate
1	2545.012	50	63.2%
2	1546.037	30	75.4%
2	1546.037	10	81%

success rate and with 10 we obtain maximum success rate. Similarly, Table 7 shows success rate for 3-, 4- and 5-phoneme words with fixed predictor size 30.

From the experimental results, it is seen that the KMC based method can give around 83 % success rate while determining number of phonemes in a word using KMC aided apriori knowledge. Obviously, in order to use the logic in the proposed spoken word recognition method, the success rate must be improved; otherwise the whole recognition system shall suffer. Therefore, the SOM based phoneme count determination block is designed as explained in Section 3.1.2. The success rate phoneme count determination using SOM is shown in Table 8 and Table 9 for noise free and noisy signal. The noisy signal set is created by adding 3 dB white Gaussian noise with the noise free signals. It can be seen that the noise free signal set shows around 7 % improvement in comparison to the KMC based method. Further increasing the RNN training signals the success rate of the noisy signal set can also be improved.

Table 7: Success Rate of Number of Phoneme Determination with KMC based method

Sl No	Value of N in N-phoneme word	Success Rate without noise	Success Rate with noise
1	3	83%	76.2%
2	4	67.3%	62.1%
3	5	78.3%	71%

Table 8: Phoneme count determination Success Rate for noise free signals using SOM based method

Sl No	Word	Correct Decision	False Decision
1	3-phoneme	92%	6%
2	4-phoneme	89%	8%
3	5-phoneme	90%	8%

4.4 Vowel Segmentation and Classification Results

The segmented phonemes are then checked one by one to find which particular segment represents the vowel phoneme by matching pattern with the trained PNNs. It is observed that the recognition success rate for various vowel is around 95% with the SOM based segmentation technique. Table 10 shows success rate for various vowels.

4.5 Consonant Phoneme Segmentation and Classification Results

The SOM is trained for six different iterations which provides identical number of decision boundaries. For 50 sessions a segment named SW1 is obtained (Figure 1). Similarly, for 100, 1000, 1500, 2000 and 3000 session other five segments are obtained. This is somewhat similar to the segmentation carried out using DWT. The DWT provides various levels of decomposition with faster processing time than the SOM. In case of SOM, some vital time slots are lost in training. But the segments provided by SOM provide better resolution which helps the PNN to provide improved discrimination capability subsequently.Similarly Table 11 shows success rate for various consonant phoneme as initial phoneme and last phoneme. Table 12 summarizes this performance.
Overall 3-phoneme word recognition success rate using the phoneme count determination block and without using phoneme count determination block can be summarized by the Table 13.
The experimental results shows that the success rate of phoneme count determination block has to be improved further, so that the word recognition model

Table 9: Phoneme count determination Success Rate for noisy signals using SOM based method

Sl No	Word	Correct Decision	False Decision
1	3-phoneme	84%	10%
2	4-phoneme	86.2%	8.5%
3	5-phoneme	87%	6.2%

Table 10: Success Rate of Vowel Phonemes

Sl No	Word	Success rate of SOM
1	/i/	98%
2	/u/	95%
3	/e/	94%
4	/o/	92.5%
5	/ɛ/	97%
6	/ao/	96.7%
7	/a/	92%
8	/oo/	94.5%

can show better success rate. But the novelty of the work is that it successfully integrates a phoneme count block as part of a discrete speech recognition system using a hybrid neural framework with a minimum success rate of 90%.

5 Conclusion and Future Direction

In this work, we have described a word recognition model with a phoneme count determination block, which can take some initial decision about the number of phonemes in any incoming word, so that the algorithm control can move from 2-phoneme to 3-phoneme word recognition algorithm, 3-phoneme to 4-phoneme word recognition algorithm and so on. The novelty of the work is that it acts as a self sustaining, fully automated mechanism for spoken word recognition with 3-, 4- and 5-phoneme variation with no manual intervention. By developing similar kind of word recognition method for 2-phoneme, 4 phoneme or any N- phoneme word a complete discrete speech recognition model for Assamese language can be developed.

References:

[1] R. L. Diehl, Lori L. Holt, “Speech Perception”, Annu. Rev. Psychol. 2004.

[2] R. Mannell, “Speech Perception Background and Some Classic Theories”, Department of

Table 11: Success rate for Consonant Phoneme Recognition

Sl .No	Phoneme	Success Rate of Initial Phoneme	Success Rate of Last Phoneme
1	/p/	×	85%
2	/b/	90%	100%
3	/t/	×	90%
4	/d/	97%	90%
5	/k/	×	85%
6	/g/	×	×
7	/ph/	×	×
8	/bh/	×	89%
9	/th/	×	100%
10	/dh/	×	×
11	/kh/	91%	100%
12	/gh/	×	
13	/s/	×	90%
14	/z/	×	85%
15	/x/	91%	95%
16	/h/	×	×
17	/m/	93%	87%
18	/n/	91%	100%
19	/l/	93%	95%
20	/r/	93%	85%

Table 12: Overall Success rate of Phoneme Recognition

Sl No	Phoneme	Success Rate
1	$Vowel$	95.25%
2	$InitialConsonant$	94%
3	$LastConsonant$	93.2%

Table 13: Overall Success Rate for 3 Phoneme Words

Sl No.	System	Success Rate
1	With Phoneme Count Determination Block	90%
2	Without Phoneme Count Determination Block	98%

Linguistics, Feculty of Human Sciences, Macquarie University, Downloaded on September, 2010.

[3] M. W. Eysenck; Psychology-an international persepective; books.google.co.in; 2004;

[4] P. W. Jusczyk, P. A. Luce, Speech Perception and Spoken Word Recognition: Past and Present, Departments of Psychology and Cognitive Science, Johns Hopkins University, Baltimore, Maryland; and Department of Psychology and Center for Cognitive Science, University at Buffalo, Buffalo, New York; 2002.

[5] B. Bergen; Linguistics 431/631: Connectionist language modeling; Meeting 10: Speech perception; 2006;

[6] J. L. McClelland, D. Mirman and L. L. Holt; Are there interactive processes in speech perception?;TRENDS in Cognitive Sciences Vol.10 No.8;www.sciencedirect.com; 2006;

[7] Y. H. Hu and J. N. Hwang, Handbook of Neural Network Signal Processing, The Electrical Engineering and Applied Signal Processing Series; CRC Press, USA; 2002.

[8] M. Sarma, K.K. Sarma, "Vowel Phoneme Segmentation for Speaker Identification Using an ANN-Based Framework", Journal of Intelligent Systems, vol.22 , Issue.2; pp.111-130; 2013;

[9] M.Sarma, K.K. Sarma, "An ANN based approach to Recognize Initial Phonemes of Spoken Words of Assamese Language", Elsevier's Applied Soft Computing; Vol. 13, Issue 5, pp. 22812291, 2013.

[10] M.Sarma, K.K. Sarma, "Segmentation of Assamese phonemes using hybrid soft-computational approach", International Journal of Parallel, Emergent and Distributed Systems; Vol. 28, Issue 4, pp.370-382 ; 2013.

[11] W. AL-Sawalmeh, K. Daqrouq and O. Daoud, "The Use of Wavelet Entropy in Conjuction with Neural Network for Arabic Vowels Recognition", WSEAS Transactions on Signal Processing, Issue 3, volume 7, 2011.

[12] J. Karam, "BiorthoganalWavelet Packets and Mel Scale Analysis for Automatic Recognition of Arabic Speech via Radial Basis Functions", WSEAS Transactions on Signal Processing, Issue 1, volume 7, 2011.

[13] R. Thangarajan, A. M. Natarajan and M. Selvam, "Word and Triphone Based Approaches in Continuous Speech Recognition for Tamil Language", WSEAS Transactions on Signal Processing, Issue 3, volume 4, 2008.

[14] F. Neri, Cooperative evolutive concept learning: an empirical study, WSEAS Transaction on Information Science and Applications; vol. 2; issue 5; pp. 559-563; 2005.

[15] M. Sarma and K. K. Sarma, "Recognition of Assamese Phonemes using Three Different ANN Structures", Proceedings of CUBE International IT Conference, Pune, India, 2012.

[16] G. C. Goswami: Structure of Assamese, First Edition, Department of publication, Gauhati University, Guwahati, Assam, India, 1982.

[17] Courtesy: Prof. Gautam Baruah, tdil.mit.gov.in / assamesecodechartoct02.pdf, Dept. of CSE, IIT Guwahati, India.

[18] T. Kohonen,"The Self-Organizing Map", Proceedings of the IEEE, vol 78, no. 9, September, 1990.

[19] G. C. Goswami: "Structure of Assamese", First Edition, Department of publication, Gauhati University, 1982.

[20] D. F. Specht, "Probabilistic Neural Networks; Neural Networks, vol. 3.,no. 7, pp. 109-118, 1990

[21] T. Kohonen, J. Hynninen, J. Kangas, J. Laaksonen and K. Torkkola: LVQ PAK,The Learning Vector Quantization Program Package, Programming Team of the Helsinki University of Technology, Laboratory of Computer and Information Science, Version 3.1, Finland, 1995.

[22] B. T. Tang, R. Lang, H. Schroder, A. Spray and P. Dermody, "Applying Wavelet analysis to speech Segmentation and Classification. H. H. Szu, editor, Wavelet Applications, volume Proc. SPIE pp. 2242:750-761, 1994.

[23] L. R. Rabiner and R. W. Schafer, Digital Processing of Speech signals, Third Impression, Pearson Education, New Delhi, 2009.