A Lexical Knowledge Driven Manner Based Speech Recognition Model

A.K. Dutta, Advisor, C-DAC, Kolkata

1. INTRODUCTION

The emergence of speech as a more direct means of interaction with the computers, promises the possibilities of leap-frogging into development of tomorrow's interfaces in Indian Languages for IT to Mass. Speech recognition is one of the most challenging speech technologies, for the development of speech mode man-to-machine communication. Speech recognition is the process of converting an acoustic signal, captured by a microphone or a telephone, to a set of words. The recognized words can be the final results, as for applications such as Commands & Control, Data Entry, Knowledge retrieval and Document Preparation. They can also serve as the input to further linguistic processing in order to achieve speech understanding.

Use of spectral domain parameters in speech recognition is a well-studied phenomenon. In spectral domain Mel-Frequency Filter Coefficients – their temporal derivatives and the Formant Frequency etc. – are used as parameters. Various statistical classification techniques including HMM are used to recognize the basic units of the speech signal. But due to overlapping nature of the spectral parameter for different speech units, it is difficult to increase the recognition rate. Furthermore, extractions of frequency domain parameters are worthy in terms of time, but lack necessary robustness. Recent studies reveal efficacy of certain shape parameters in speech recognition. The extraction of these has been seen to be quite robust with low time complexity as well as provide competitive recognition rate at least for vowels.

However acoustic parameters alone cannot achieve speech recognition as well as the human beings do. It is known that only about 40 – 60% intelligence is contained in the acoustics of speech signal. Hence, the Speech Recognition Systems focused on acoustics alone may not be the final solution. The recognition efficiency of such acoustics-based recognition system can be vastly improved by harnessing the associated lexical knowledge. Integrating lexical knowledge in speech recognition is one of the most effective ways to achieve good recognition efficiency. A hierarchical approach is intended for the purpose.

In general speech recognition systems use word as the intermediate units for final understanding of the content. Unfortunately for speech the words are not clearly separated which one sees in the textual languages. Thus word boundary detection becomes a significant problem in continuous speech. The next problem is the recognition of words. One could use the signal constituting words as the units of recognition. Such systems are known to be practical and efficient only for very small vocabularies usually not exceeding a few hundred words. The other approach is to use phonemes as basic unit of recognition. While phonemes exhibits good potential of individual recognition (around 80%-85%) its usefulness in word recognition becomes doubtful for words consisting of multiple phonemes as the error rate gets multiplied.

One way to circumvent this is to use lexical knowledge to guide phoneme recognition so that the best possible strategy for each word can be achieved. The present paper proposes such an approach. In this approach the speech signal is first segmented into words. The signal for a word is then further segmented into zones of selected manners using phase space approach. This segmentation produces a pseudo representation of the actual word. The phonetic representations of a very large vocabulary of words (of the order of 50,000) are converted into corresponding pseudo-words. These pseudo-words break up the actual vocabulary into cohorts of small cardinality. Each of these cohorts are analysed using confusion matrices for phoneme recognition to establish the strategy, which gives the best recognition for all the words in the cohort. This constitutes the lexical expertise. This expertise is used to guide the recognition module when the input signal for a word is analysed into its manner based pseudo-word configuration.

2. Methodology

In the Manner Based Model of speech recognition a multilevel lexically driven hierarchical approach is used. In the first phase of recognition, a manner-based criterion for pseudo word formation is used. The total vocabulary is split into cohorts of words corresponding to different pseudo words. In the second phase, we use shape parameters for phoneme recognition to separate words in each cohort. These manner based labeling allows lexical knowledge to be used in recognition stage. The figure1 represents the basic block diagram of the above model.

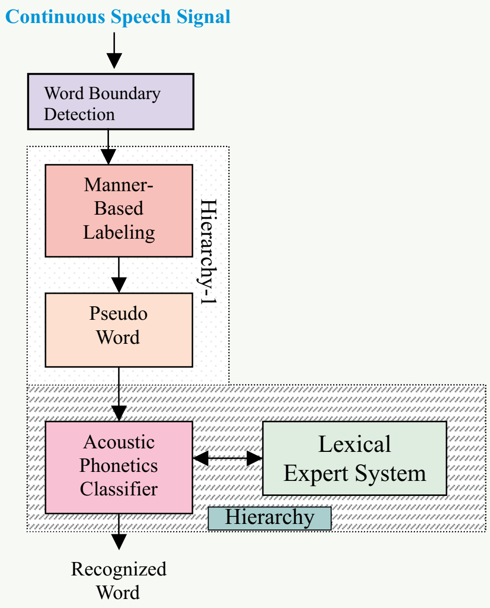

Figure 1 Basic block diagram for speech recognition

Hierarchical System for the development of the recognition engine:

- Word boundary detection
- Manner-base labeling
- Phonetic classification
- Lexical Expert System

2.1 Word Boundary Detection

Continuous speech does not offer any clear signature of the word boundary, like pause or silence. In fact the phonemes at the end of a word exhibits similar co-articulatory phenomena with the initial phoneme of the next word as they do inside a word. Word boundary detection is, therefore, a problem of great importance and has different connotations in different languages. One needs to identify clues based on speech signal which is largely influenced by the characteristics of the particular language. Prosodic features such as pitch, amplitude, duration and pause can be used as speech clues to detect word boundaries, while language features such as phoneme sequence constraints and syntactic markers can also be exploited for placing word boundaries. In most of the speech recognition systems, the word boundaries are detected as a part of the word hypothesisation process in the lexical analyzer. But this is a tedious and complex procedure. Standard colloquial Bengali (SCB) is a bound-stress language, with stress on the first syllable, which introduces some information in the suprasegmental parameters of the first syllable in the speech signal and may help to detect the word boundary in the continuous speech signal. The parameters used in the present system are: 1) Difference of nucleus vowel duration across the syllable boundary, 2) Difference of normalized nucleus vowel power across the syllable boundary, 3) Normalized pitch difference across the syllable boundary, 4) Difference of average normalized pitch across the syllable boundary, 5) Difference of normalized maximum periodic power of the nucleus vowel across the syllable boundary, 6) Onset duration of the nucleus vowel.

Altogether 225 sentences spoken by five native Bangla informants of both sexes, in the age group of 20-50 years are used in the study. These sentences contain 2735 syllables and 1105 words, the sentence terminal word being excluded. A recognition score of 87.8% with a classifier based on a distance function weighted by inverse of variance is studies. Tables 1 and 2 respectively present the confusion matrices for speaker independent and speaker dependent cases respectively.

TABLE 1

Confusion matrix for SPEAKER NDEPENDENT case

		Classified as	
		Boundary	Non-boundary
Original class	Boundary	968 (87.8%)	135
	Non-boundary	303	1328 (81%)

TABLE 2

Confusion matrix for SPEAKER NDEPENDENT case

		Classified as	
		Boundary	Non-boundary
Original class	Boundary	874 (79.2%)	229
	Non-boundary	303	1422 (87%)

2.2. Manner Based Labeling

Manner based labeling of a speech signal consists of segmenting the continuous speech signal according to some manner of production and then labeling them appropriately. Though all different manners of production of speech leaves their tell-tale marks on the signal the robust extraction of all of them may not always be possible. The approach is to use only those manners which can be robustly as well as economically differentiated through a particular signal processing approach. In phase-space approach, some simple manipulation of the high dimensional trajectory matrix generated from the one dimensional time series representing the continuous speech signal provide a low dimensional parametric representation of the signal. Such parametric representations have been recently used for voicing detection algorithms (VDA) as well as Pitch detection algorithms (PDA)[2]. Such an approach has also been used to obtain certain low-level classification of the quasi-periodic region into different manner groups like vowels other vocalic segments. This algorithm has been tested on a large database of continuous speech. VDA is used for the manner based labeling of the signal into the voiced, unvoiced, sibilant, affricate and nasal murmurs. This classification is mainly on the basis of three parameters, namely, no of minima, standard deviation of minima values and flat count, which are obtained from deviation vs delay plot. This procedure labels the signal into the five selected set of manners namely, plosive, vowel, sibilant, affricate and vocalic non-vowel segments. This produces the corresponding pseudo-word for the input word-signal. After the manner based labeling, we are now ready for pseudo word formation. Pseudo words are formed by assigning P for plosive, V for vowel, S for sibilant, A for affricate and L for vocalic non-vowel segments. As for example if the word is [illegible] then the corresponding pseudo word is AVL (Figure-2). We have to do this for all twenty thousand words we have taken for this experiment

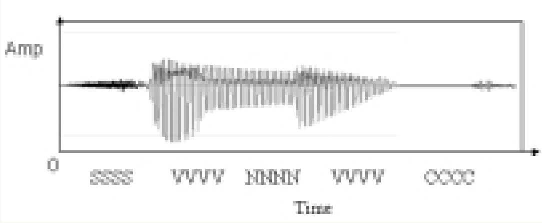

Figure 2 Manner based Labeling of a speech signal

2.3. Lexical expert system

Bangla has a syllabic script. The script for a word has to be converted into strings of phonetic symbol before forming the pseudo words. This was done using a phonological converter. The pseudo words are formed from the phonetic words using a transformation table, which gives the selected manner representation of each phoneme. The dictionary containing two columns, one for the actual phonetic words and the other for the pseudo words is now created. Sorting of the table by the pseudo word column gives the groups, which have same pseudo word. The groups so formed are known as cohorts. Some of these cohorts may contain only single word, which are called unique cohorts.

Each cohort is converted into a binary tree structure where each node dictates the details of the classification process. Here the nodes keep the information of the best division of the word list into two subgroups. As vowel classification is reported to give highest recognition rate the vowels are used for designing the expert system. The different vowel positions in a cohort are detected first. The idea is to successively examine the vowels in different positions to find the best possible binary classification of the word list in terms of recognition rate. E.g., the three words kobi, khokaa, paakaa in a cohort can be separated into two groups kobi, khokaa and paakaa by the first vowel position. Again kobi and khokaa can be separated by the second vowel position.

Let us consider the arrangement of Bangla vowels in the sequence u, o, a, aa, ae, e, i. It happens that in such a sequence mis-recognition occurs primarily between contiguous vowels. If corresponding to one position of vowels in a cohort all vowels in the aforesaid sequence are present except, say, 'a' then we may say that a hole exists there. In that case we may make a binary division of the list such that in one group all the words with vowels 'o' and 'u' goes to one group and the rest to the other group. In such a case the classification generally yields high recognition rate. Various such knowledge are put into the formation of the tree structure. The process of such binary partitioning continues until no further partitioning can be done. Thus we get a binary tree structure where the leaf nodes contain either a single word or such words, which can only be differentiated, through a classification of consonants.

The detailed algorithm for obtaining a tree for a cohort is as follows:

Step1. If there is single word in cohort then go to Step8 else Step2.

Step 2. Search for a hole in any vowel position for all the words in that cohort. If hole is found then go to Step3 else go to Step4

Step 3. Break up the cohort into two sub cohorts such that the words (or word) having vowels, which lie before, the hole goes to cohort one and others to cohort two. Else Go to Step 7

Step4. Check the first vowel position of words in the cohort. If the first vowel position of those words contains same vowel then go to Step5 else Step6

Step5. The cohort is not broken further and the recognition rate is distributed equally among all the words in that cohort. Go to Step 7

Step6. At this point we can say that all the vowels at a particular position of all the words in that cohort is contiguous to one another (i.e. no hole can be found between them). Therefore the cohort is broken into two sub cohorts at that vowel position where the recognition rate is very high. This vowel position can be obtained from table 1, which has been calculated previously.

Step 7. Check all the nodes sequentially to see whether there is a single word or such words where vowels in each position are same. If the condition is satisfied then go to Step8 else take the node as new word list and go to Step2.

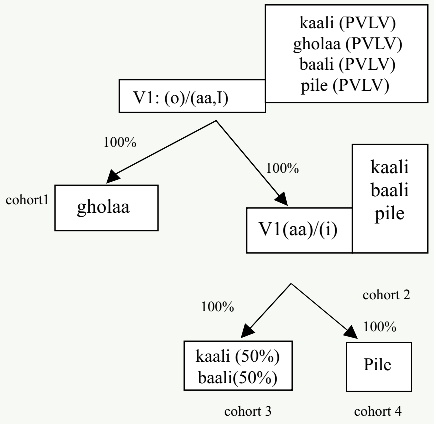

Figure 7 Binary tree for the cohort PVLV

Step 8: End.

Figure 7 describes how the cohort is broken using lexical knowledge

Let us consider a cohort PVLV consisting of four Bangla word, like /*kaali*/, /*gholaa*/, /*baali*/ and /*pile*/. On examining the first vowel position of these words we can see that there is hole between the vowel "o" and the vowel "aa". Therefore we can break this cohort into 2 sub cohorts. Thus the word gholaa goes to cohort1 and the rest of the words like kaali , baali and hire goers to cohort 2. Thus the recognition rate of each of the cohort is 100%. Again examining the first vowel position of the words in the cohort2 we can see that there is hole between the vowel "aa" and "i". Thus cohort2 is broken down to two sub cohorts namely cohort3 and chort4. Hence the words kaali and baali goes to cohort3 and hire to cohort 4. On careful examination of the words of cohort 4 we find that they all have the same vowels at all positions. Thus we cannot divide this cohort into another two sub cohorts. Therefore the recognition rate for each of the word comes out to be 50%

Now the total recognition rate of the cohort is calculated by summing up the error rate at the leaf nodes and dividing it by the number of leaf nodes. This will give us the total error rate. In this case the recognition rate of the cohort comes out to be75%.

For this experiment we have taken 20,000 commonly spoken Bangla words. Phonetic labeling of these words is done at first and then corresponding pseudo word is generated based on the manner based labeling. And then finally we have used our lexical knowledge system over these words for the estimation of recognition rate.

The result of the classification based on the manner based labeling in the first hierarchy is given in table 3. In this table all the figures are in percentage of total number windows which is slightly greater than 5000. In table 3 C represents consonant class, S represents Sibilant class, a represents affricate class, L represents lateral, nasal and murmur class and V represents vowel.

The overall recognition rate obtained during manner based labeling is 98%, which is considered to be quite good.

		Classified as				
		V	S	A	L	C
Original class	V	95			5	
	S		99	1		
	A		2	98		
	L	8			92	
	C	1				99

Table 3 Confusion matrix for manner based labeling (values are percentage wrt row total)

Table 4 presents the confusion matrix for all samples in the speaker independent mode. This means that in the training set each vowel class is represented by only one representative for all the informants. The percentage of correct classification for the entire seven Bangla vowels is 81.8%. This may be considered as encouraging since this represent a multi-speaker environment. The main source of error is in a mix up between i and u.

Table 4. Confusion matrix for vowel recognition in speaker independent mode (values are percentage wrt row total)

Table 5 a, b, c gives us the information about the class pair and the highest rate of recognition for the corresponding sequence of contiguous vowel. This piece of information is very important for breaking up the cohort into two sub cohorts when all the vowels are contiguous to each other.

		Classified as							
		u	o	a	aa	ae	e	i	#
Original class	u	85						15	
	o	6	88			2	4		
	a		6	80	2	10			2
	aa				10	80	9		1
	e			1	4	6	84	4	1
	i	20				1	4	70	5

Table 4. Confusion matrix for vowel recognition in speaker independent mode (values are percentage wrt row total)

R represents the highest recognition rate for the class pair

Sequence	Class pair	R
u o a aa ae e i	(u o)/(a aa ae e i)	97%
o a aa ae e i	(o a aa ae e)/ i	95%
u o a aa ae e	(u o a aa ae)/e	97%
a aa ae e i	(a aa ae e)/i	94%
u o a aa ae	u/(o a aa ae)	97%
o a aa ae e	(o a)/(aa ae e)	88%
Aa ae e i	(aa ae e) /i	96%

Table 5 (a) Best possible class pair and highest recognition rate for contiguous vowel sequence.

Sequence	Class pair	R
u o a aa	u /(o a aa)	97%
o a aa ae	o/(a aa ae)	98%
a aa ae e	a/(aa ae)	95%
ae e i	(ae e)/ i	96%
u o a	u/(o a)	98%
o a aa	o/ a aa)	98%
a aa ae	a/(aa ae)	94%

Table 5 (b) Best possible class pair and highest recognition rate for contiguous vowel sequence.

Sequence	Class pair	R
aa ae e	a /(ae e)	93%
e i	e / i	96%
u o	u /o	97%
o a	o /a	97%
a aa	a /aa	93%
aa ae	aa /ae	91%
ae e	ae /e	91%

Table 5 (b) Best possible class pair and highest recognition rate for contiguous vowel sequence.

After using the lexical expert system over 20,000 Bangla word following results are obtained:

Total number of cohort generated is 2479

Total number of unique cohort generated is 1097

Maximum size of the cohort is 624 for the pseudo word PVPV.

Average size of each cohort is 12.65

Maximum size of ambiguity is 50

No of unique leaf node is 7000

Overall recognition rate is 95%.

This recognition rate does not reflect the true recognition rate, as we have not considered the recognition rate obtained during manner based labeling.

Thus the actual recognition rate comes out to be 93%

2.4 Phonetic Labeling

The phoneme level classification would be attempted using different shape parameters as (a) Peak to Peak Distance, (b) Extrema Rate, (c) Zero Crossing Rate and (d) Perturbation Area.

(a) Peak-to-Peak Distance (PPD):It is the distance between the largest maxima and smallest minima for a given window (figure 3). This is likely to capture the location of the frequency of the lowest resonant band F_1.

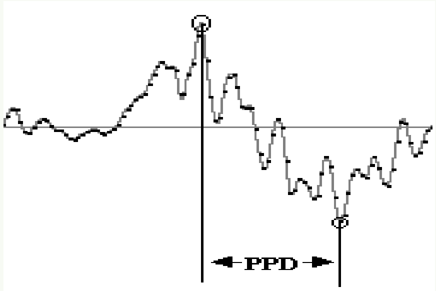

Figure 3 Peak-to-peak distance

(b) Extrema Rate (ER): Total number of peaks and troughs in a given window are the extrema points of the signal. This is likely to be influenced by the frequency of the higher resonant bands like $F_{2.}$ The extrema points are indicated by open circles in figure 4.

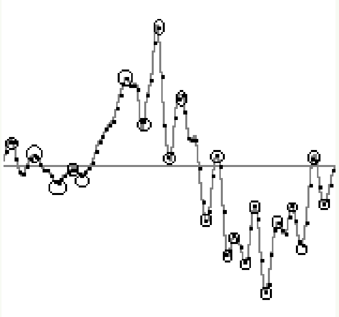

Figure 4. Extrema points

(c) Zero Crossing Rate (ZR): The number of times the signal has crossed the zero-line for a given window. The zero crossing points are indicated by the open circle in figure 5.

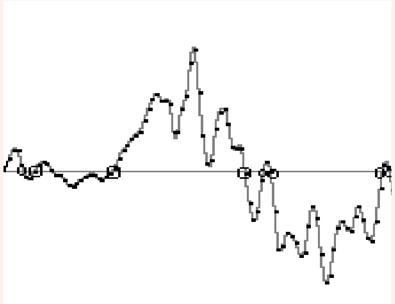

Figure 5 Zero Crossing points

(d) Perturbation Area (PA): The area enclosed between the curves joining the positive and negative high frequency peaks. This is likely to represent the intensity of the higher resonant band F_2. It may be noticed that all extrema points are not used in determining the area of perturbation. The perturbation area is indicated in figure 6.

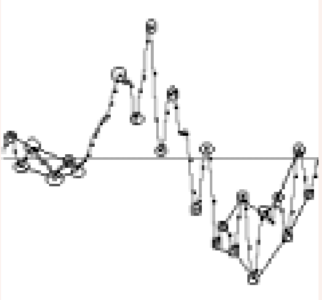

Figure 6 Perturbation Area

3. Conclusion

Integrating lexical knowledge in speech recognition is one of the most effective ways to achieve good recognition efficiency. The proposed method can work as very good isolated word recognition system for large vocabularies say 50000 words. But in case of continuous speech the detection of word boundary is not so much satisfactory. Then we can think about the slow dictation mode of continuous speech where the word boundary more or less defines. A lexical expert system, developed off-line, examines each cohort with respect to a previously determined confusion matrix for vowel recognition and builds up an expert system through successive binary splitting up of the each cohort into classes, which gives best splitting in terms of recognition score.

4. References

1. Datta A.K., Gunguly N.R and S. Roy, "*Recognition of unaspirated plosive: a statistical approach*", IEEE Trans ASSP, Vol. ASSP-28, No. 1, 85-91, 1990.

2. Shyamal Kr. Das Mandal, A. K. Datta, B. Gupta, "*Word Boundary Detection of Continuous Speech Signal for Standard Colloquial Bengali (SCB) using Suprasegmental Features*"FRSM-2002,IIT Kanpur.

3. Dan T. K and Datta A. K. "*Speech synthesis: a time domain model*". Proc. International Workshop on recent trends in speech, music and allied signal processing, 9-11 Dec, 1991, Delhi, India, pp. 39-45.

4. Datta A. K, Gunguly N R, Mukherjee B. "*Intonation in segment-concatenated-speech*". Proc. ESCA Workshop on speech synthesis, Sep 1990, France, pp. 153-156.

5. Soumen Chowdhury, A.K.Datta and B.B.Chaudhuri, "*Pitch detection algorithm using state phase analysis*" J. Acous. Soc. Ind, Vol.28 No.1-4, 2000,pp247-250.

6. Soumen Chowdhury, A.K.Datta and B.B.Chaudhuri "*Pitch detection algorithm using state phase analysis*", J. Acous. Soc. Ind, Vol.28 No.1-4, 2000,pp247-250.

7. S. Rajendran, B. Yegnanarayana. "*Word boundary hypothesization for continuous speech in Hindi based on F0 patterns*" Speech Communication 18(1996) 21-46.