Phonetic Speech Analysis

D.S.Shete [1], S.B. Patil [2]

[1](Department Of Electronics Engineering , D.Y.Patil College of Engg. & Tech., Kolhapur,
Shivaji University Kolhapur,India)
[2](Department Of Electronics Engineering , D.Y.Patil College of Engg. & Tech., Kolhapur,
Shivaji University Kolhapur,India)

ABSTRACT: *This paper presents a description of the work done on phonetic speech analysis. The work aims in generating phonetic codes of the uttered speech in training-less, human independent manner. This work is guided by the working of ear in response to audio signals. The Devnagri script inspires the work presented. The Devnagari script classifies and arranges 46 phonemes in a scientific manner based on the process of its generation. The work at present focuses on identifying the class (varna) of the phoneme as specified by the Devnagari script. More work is needed to identify the variant of the class identified. Phoneme code thus generated can be used in an application specific way.This work also explains and proves the scientific arrangement of the Devnagari script. This work tries to segment speech into phonemes and identify the phoneme using simple operations like differentiation, zero-crossing .*

Keywords *- Devnagari script, Phonetic speech analysis, Phoneme recognition, Speech to Text conversion.*

I. INTRODUCTION

The work discussed here aims in designing training-less,human independent phoneme class recognition system. This work is not speech recognition or speaker recognition system,but a phoneme recognition system. Phonemes are the basic unit of speech of a language. Each language has its own distinctive set of phonemes, typically numbering between 30 and 50;e.g. English can be represented by a set of around 42 phonemes; Hindi is having a set of 46 phonemes. When different phonemes articulate, voice is produced.Irrespective of a human, the way any phoneme uttered is same and this is the principle of arrangement of the Devnagari script. Though the same speech uttered by different persons is felt different to listen, the information (phonemes in this case) that we extract from the signal is same. This can be because the pattern of vibrations produced by air density on eardrum must be similar for the same phoneme. This work uses the same principle in classifying and identifying the uttered phoneme.The work at present does not deal in identifying the exact phoneme rather its class (varna).

$$y(n) = x(n) – x(n+1) \quad (1)$$

The work starts after differentiating input speech signal using Eqn. (1). The work is divided in four parts: - 1) End point detection, 2) Segmenting speech into phonemes 3) Phoneme class identification and 4) Phoneme variant identification in the class identified.Out of these, part (1) is not designed to be very robust and accurate, because already the work has been done satisfactorily. Part (2) is implemented using variation in zero-crossing rates and part (3) is implemented using FFT of the speech segment obtained in part (2). Most of the work is focused on parts (2) and (3). No work has been done on implementing Part (4).

II. THE DEVNAGARI SCRIPT

Devnagari script is a script of phonemes arranged in a well structured scientific manner showing unambiguous classification and grouping of phonemes according to the organs used in producing that sound. The letter order of Devnagari is based on phonetic principles which consider both the manner and place of articulation of the consonants and vowels they represent. Accordingly these letters (Akshar) are grouped into different classes called Varnas ("TulyasyaPrayatnam Savarnam") [1]. Every letter and its pronunciation is unique and can't be represented or pronounced by using any other letter(s).This gives us a unique representation for every word uttered by human irrespective of human and context of speech.

This feature is absent in languages like English in which one representation and pronunciation of a word or letter can be done in more than one way, e.g. bye, buy both are pronounced similarly.The first 25 consonants of Devnagari script, arranged in a 5X5 matrix, form five different groups of phonemes as in Table 1 Each row of five consonants is generated in totally different way. First four rows are classified depending on the touch point of tongue inside the mouth as Kanthhawya (Velar), Talawya (Palatal), Murdhanya (Retroflex) and Dantawya (Dental). The fifth group is called Aushthawya (Labial) because it is generated using lips only. The elements in a single row are generated using the same organs but varying the time period of touch and pressure at the same or near the touch point of group.Different phonemes in these varnas are:

Table 1: Phonemes of Devnagari script

Phone class	Class variant				
	Non-voiced		voiced		Nasal
Kanthwya	ka	Kha	ga	gha	nga
Talwaya	Cha	Chha	ja	jha	nja
Murdhanya	Ta	Tha	da	Dha^	na^
Dantawya	Ta	Tha	da	dha*	na*
Aushthawya	Pa	Pha	ba	ma	ma

III. SPEECH PROCESSING WORK

The work aims in designing training-less, human independent phonetic speech analysis system to generate phonetic codes of uttered speech. This work is guided by the working of ear in response to audio signals.Speech signals are composed of a sequence of sounds. These sounds and the transitions between them serve as a symbolic representation of information. Though the arrangement of these sounds is governed by the rules of the language, the elemental sounds called phonemes (Akshars) remains the same. Also the way different human produce these phonemes are also same because the difference remains in the parameters of the signal produced like pitch, energy etc. It also is known that the spectral properties of speech waveform such as energy, zero crossings and correlation can be assumed fixed over time intervals on the orders of 10 to 30 ms [2].When same speech is uttered by different humans, the information that we extract i.e. phonemes is same irrespective of the speaker. This gave us the question where might this information be present. Hence as first step before starting to work with the speech signal we tried to find some of these parameters that store information of speech. We first converted speech signal into a series of (+1, 0, -1) by changing all values above a +ve threshold to +1, below a –ve threshold to -1 and in between values to 0. The threshold was selected by manual inspection. The speech was still understandable though was heavy in noise. Next only the points which fell on zero-crossings were marked +/- 1 according to its sign. The result was same as previous. Next we differentiated the signal using (1) effectively high pass filtering it, as was expected the speech was still preserved. With these results we concluded these two simple parameters viz. zero-crossing and magnitude variation are holding much of the information. Hence these two parameters are always used to further process the signal.We are using complete list of Devnagari alphabets uttered by 10 different persons (5 females +5 males) in normal daily use rooms at 8 kHz with 8 bits per sample.The work being focused mostly on identifying the five classes (varnas), accuracy around 75% is obtained when speech is composed of consonants from these groups only. The work identifies distinct patterns produced by these five classes. Next we discuss the three parts implemented.

3.1 End- Point Detection:

In a same phoneme class it is observed that the first variant is unique in a sense that it is repeated in other three of the four plosive variant except the nasal and is coupled with aspiration and/ or voice bar. Hence even in the absence of this aspiration or voice bar within the marked end points, we can correctly segment and identify the class of the phoneme.Here the technique used to identify endpoints is not very robust and hence low energy nasals and other low energy plosives as specified by [2, pg 132] are tried to be avoided.This limitation is used because already many techniques have already been developed [4, 5] and this part is not main focus of the work. Use of one of these techniques is advised.The method used here is inspired by [1, 4] and is magnitude based only. The average magnitude envelop of 1st difference of input speech signal is obtained using a rectangular 15ms window moved forward at 10ms steps. A threshold equal to 15% of the maximum of this envelop is used to partition this envelop in two parts abvTh and blwTh.In the 1st difference of this magnitude envelope its zerocrossings are marked. Zero-crossings nearest to start and end of abvTh is taken as tentative start and end points. As done in [4] we move backward and forward at start and end respectively. The distance of nearest previous/ next ZC to start/ end is computed. If this width is >3 start/ end is moved to this point and again we move forward or backward. If we get three consecutive zero-crossings in this 1st difference of amplitude envelop which are not wider than 3 durations of steps and within a threshold of 1 step above or below the width of zero crossing that we started with we finalize the point that we started with as start/ end point. If the differenceof widths between current two zero-crossings and the width we started with is >1, we move start/ end point to this point and again start to find three consecutive zero-crossings matching above condition. Because the work focuses on identifying the phone class of first 25 phonemes and presence of first phoneme of class in other variants of same class loss of friction (aspiration) before present in weak fricatives (/f/, /the/ /h/) is not going to affect the result as long as it is segmented correctly that we discuss next

3.2 Speech Segmentation

Speech segmentation can be described as a process of identifying boundaries in the speech signal and labeling each of the speech segments between two adjacent boundaries with a symbol. This process of identifying boundaries and labeling, can be addressed at various levels of details such as Speech / non-speech segmentation. This is a task of detecting the begin and end of speech in the audio signal. Apart from speech, an audio signal may also contain non-speech data such as music, noise, silence etc. It is important to segment the audio data into speech / non-speech segments, as it acts as pre-processing step for many speech systems such as speech recognition,speaker recognition, etc.,. Output of this process is the boundaries between the speech and non-speech segments in an audio signal.

As discussed earlier about experimental observations, it is found that variations in amplitude and the rate of zero- crossings are two of very important parameters that represent the information content of the speech signal. Hence these two parameters are used to segment the speech into its constituent phonemes.The zero-crossings are computed on 1st difference of the input speech signal with a rectangular window of 15 ms duration at steps of 10 ms duration. Similarly the magnitude envelop is computed with same window. After this we marked boundaries with following steps. 1) The positions where the zero-crossing rates in two neighboring cells vary more than 13.30% are marked. Zerocrossings of nasals and liquids are not found to produce any appreciable change above this threshold so as to get separated from the neighboring phonemes. But magnitude variation is found to show sufficient separation.Also this segmentation produced multiple segments of longer phonemes like vowels, nasals and fricatives and mixed nasals and liquids with neighboring phonemes. Hence these segments are first mixed using ZC rate of each segment averaged over 10ms and if difference of this ZC per 10ms with neighboring segment is <=15% the two segments are mixed.This procedure grouped all the multiple segments of longer phonemes. But these groups still included nasals and liquids.These are separated using amplitude envelop.2) As first step in amplitude based segmentation all the peaks in magnitude envelop that differ from its neighboring valley by > 15% of maximum of magnitude profile are marked. This is done because when nasals and liquids got mixed with their neighboring phonemes their magnitudes are found to vary appreciably than the neighboring phonemes.Next the parts above and below a threshold of 15% of maximum of magnitude profile are marked. These segments are then mixed with segments obtained form above step (1).If any of the marked peaks occurs between these segments of speech, this segment is partitioned into above and below the 15% of average value of this segment. Segments <= 50ms or having zero-crossing >= 80 per 10ms is not segmented and is accepted as it is.Once the speech is segmented into phonemes next step is to recognize it.The present work focuses on identifying zero crossing rates of five phoneme classes .The following table shows zero crossing rates of five classes.

Table 2: Experimental Observations for Phoneme Class

Class	Zero Crossing Range
Velar	25-48
Palatal	45-90
Retroflex	30-60
Dental	20-55
Labial	10-55

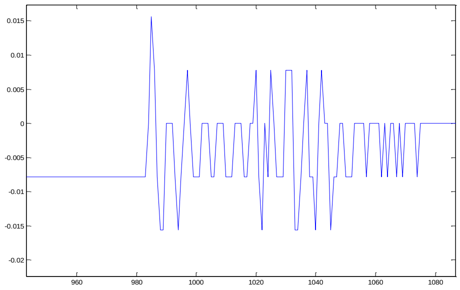

Fig.1 Sample of speech signal of velar class

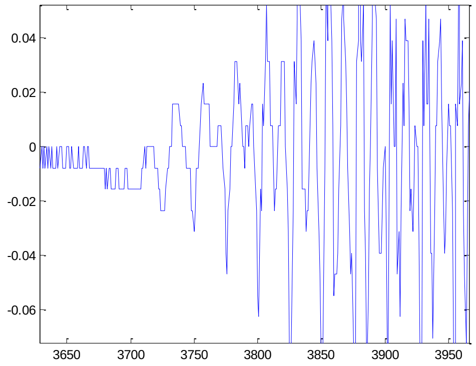

Fig.2 Sample of speech signal of palatal class

IV. CONCLUSION

As was expected in the beginning of the work that a Devnagari script based phoneme recognition system can be designed by considering simple parameters like zero-crossings magnitude. Number of samples of speech are taken and generalization is that if zero crossing rate is high, speech signal is unvoiced while if zero crossing rate is low, speech signal is voiced .At present, whatever zero crossing rate, we are getting, its accuracy is not high and some ranges of zero crossings are also overlapping.There are some drawbacks.The manual intervention in Phoneme recognition system is extremely time consuming and tedious which drastically increases the time required to obtain the phonetic boundaries. If a new database have to be segmented, then all the processes of this approach have to be repeated.

REFERENCES

[1] *AshtadhyaeeBhashyam ; Swami Dayanand Saraswati*

[2] *Digital Processing of Speech Signals; Rabiner, Schafer; Pearson Education.*

[3] *Discrete Time Speech Signal Processing; Quatieri; Pearson Education.*

[4] *A Speaker independent digit recognition system, L. Rabiner , M.Sambur.*

[5] *Robust entropy based endpoint detection, Jia-lin Shen, Jei-Weih Hung,Lin-Shan Lee.*

[6] *Phonetic Speech Analysis for Speech to Text Conversion, 2008 IEEE Region 10 Colloquium and the Third International Conference on Industrial and Information Systems, Kharagpur,INDIA December 8 -10, 2008.*

[7] *A new approach for phoneme segementation of speech signals,Douglas O'Shaughnessy Ladan Golipour, in Proceedings of Interspeech ,Antwerp,Belgium,August 2007.*

[8] *Manual segmentation and labelling of speech, A. Van Erp and L. Boves,in In Proceedings of Speech-88, 1988, pp. pp. 1131–1138. A. Van Erp and L. Boves,in In Proceedings of Speech-88, 1988, pp. pp. 1131–1138.*