2008 IEEE Region 10 Colloquium and the Third International Conference on Industrial and Information Systems, Kharagpur, INDIA December 8 -10, 2008.
255

Phonetic Speech Analysis for Speech to Text Conversion

Abhijit V. Bapat, Lalit K. Nagalkar
Dptt. of ECE, Yashwatrao Chavan College of Engineering
R.T.M. Nagpur University
Nagpur, Maharashtra, India
AbhijitBapat@yahoo.com, lalit_nagalkar@yahoo.com

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 calculation and FFT.

Index Terms— Phonetic speech analysis, Phoneme recognition, Devnagari script, 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[2]; 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).

$$y(n) = x(n) - x(n+1) \qquad (1)$$

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 [4, 5]. 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:

IEEE Kharagpur Section & IEEE Sri Lanka Section

TABLE I. FIRST 25 CONSONANTS OF DEVNAGARI SCRIPT:-

Phone Class (he Varna)	Class (Varna) variant				
	Non-voiced		*Voiced*		*Nasal*
	Inaspirated	*Aspirated*	*Inaspirated*	*Aspirated*	
Kanthhwya	ka (cut)	kha	ga	gha	nga
Talwaya	cha(char)	chha	ja	jha	nja
Murdhanya	Ta (tap)	Tha	Da(dog)	Dha^	na^
Dantawya	ta	tha(theme)	da(the)	dha*	na*
Aushthawya	pa (pup)	pha	ba	ma	ma

* These are different than respective upper ones marked by ^.

At present we have focused on these five groups of phonemes only (excluding the nasals in first 2 groups- "nga" and "jna", as none was able to pronounce them properly) and tried to get the details of variations among these groups so as to be able to identify the group (*varna*) first.

III. THE 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 16 different persons (5 females + 11 males) in normal daily use rooms at 8 kHz with 8 bits per sample.

The work being focused mostly on identifying the five classes (*varna*s), 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.

A. 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 zero-crossings 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 difference of 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/, /th/ /h/) is not going to affect the result as long as it is segmented correctly that we discuss next.

B. Speech Segmentation:

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. Zero-crossings 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.

C. Phoneme Recognition:

To identify a phoneme at present we are using a 128 to 512 point FFT depending upon the length of phoneme segment to get the frequency bands of the phoneme. In this part we are computing FFT of the signal segment and taking its threshold with a threshold of 0.1. The density is then computed of this threshold of FFT. Here we observed that often for vowels the density is very less compared to non nasal consonants. Nasals as they are closer to vowels [3] are also often found to produce very low density FFT.

The 0-4 kHz frequency scale is divided into 17 bands each of ~235 Hz. Each band is numbered from 1 to 17. Band 1 corresponds to (0 to 235) Hz, Band 2 to (235 to 470) Hz and so on. The peak densities of stems in FFT are then computed, normalized and corresponding frequency bands are marked. This density is computed with a rectangular window of size 500Hz and in steps of size 200Hz. These frequency values are first converted into corresponding number of points depending upon FFT size. Because size of FFT is a power of 2 we converted step size to nearest higher power of 2 hence 17 bands widening step size to ~235Hz.

The band with less than 50% density is discarded assuming a phoneme will produce its bands with strengths >=50%; this assumption is observed to be a correct decision.

The identification of phoneme class is done by finding the positions of peaks thus produced. The phonemes from same group are found to produce these peaks consistently inside a particular band only or a predictable combination of 2 or 3 bands. Retroflex and Dental phoneme classes are found to produce multiple peaks in different bands. Also the positions of FFT peaks of these two groups are found to be very common with each other as can be seen from the TABLE- I. The Retroflex and Dental are hence, in this work, are grouped together. Till now no concrete distinctions have been found between these two groups.

Because of these observations artifacts producing more than 3 peaks are ignored to be considered to identify. Also non-speech parts are found to be more or less equally distributed on complete frequency range, hence whenever a segment is having more than 9 cells (>50% of total cells) having normalized densities ratio >=60% the maximum that segment is marked as noise.

Vowels and nasals are found to produce densities <=5 most of the times, but this is not a very sharp boundary as vowels also produced densities >10 much like non-nasal consonants considerable number of times. Duration of vowels, nasals and fricatives is one of the features to distinguish between these grouped together and non-nasal consonant. Hence if duration of a segment is >50ms that segment is marked as non-nasal consonant and it can be any of vowel, nasal or fricative. Fricatives (SHA) produce very large zero-crossings (> 80) hence easy to be identified.

All the vowels and phonemes (LA, WA and RA) are occupying band 3 most of the times making them very difficult to be distinguished. FFT peaks for vowels are also producing at most two bands.

Fig (1) shows end points detected and the phone boundaries marked by the algorithm. Out of 13 considerable segments 10 are correctly identified. One phoneme “KA” is lost. Overall system accuracy is expected to be ~75%.

The positions of largest frequency density groups are as tabulated below.

TABLE II. EXPERIMENTAL OBSERVATIONS FOR PHONEME CLASS CLASSIFICATION

Class (*varna*)	Zero-Cross range	Prominent Freq. Band positions
Velar	25- 49	(6, 7,); 5 also seen with ZC>=35
Palatal	45- 90	(11 to 15)
Retroflex	30- 60	(3, 4); (8, 9, 10); (12, 13, 14)
Dental	20- 55	(3, 4); (8, 9, 10); (12, 13, 14)
Labial	10- 55	(2, 3, 4, 5); 5 seen with ZC<35

For Retroflex and Dental the main identity group is (8, 9, 10). Groups (3, 4) and (12, 13, 14) appear in combination with each other and/ or (8, 9, 10). The peak appears at any one of position of these groups. Phonemes of Velar group appear sometime at 5 also, but this is found to be very rare hence Velar is considered to occupy bands (6, 7) only.

Because 0-4kHz range is divided into 17 bands each band of ~235Hz the numbers in last column in Table-1 indicates the highest frequency of that band (e.g. 7x235=1645Hz). The lower bound is the upper bound of previous numbered band. From this table it can be seen that the zero-crossing ranges are highly overlapping hence this feature is given least weight for identification purpose. FFT is first used to mark likely phoneme groups and then ZC is used to finalize the group if more than one group is found to be considered. Once the group (Varna) or row is finalized then the variations in the neighborhoods of the segment will be used to identify the column in the phoneme table to identify the phoneme.

IV. ASSUMPTIONS

Though the data used in designing the system is not completely noiseless, the system do not deals in removing noise rather it operates on the input data directly assuming its noise to be already filtered. The system do eliminates the DC offset before further processing so that it will not affect calculations in zero crossings.

V. 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 and FFT of the speech segments is verified. At present the accuracy of system is not very high (~75%) and we are not able to separate Retroflex and Dental groups. Also, though IDs of vowels and other consonants of the script overlap with these five groups, work is advancing in positive direction to solve these problems and identify the variant of the group also. Also it is expected the duration of silence before a plosive utterance will be a good parameter to consider as in Fig. 1.

Also as suggested by Devnagari script, because of the use of same organs for a *Varna,* patterns from same *varna* will be highly correlated. Hence it'll be easier determining first the Varna and then the phoneme (*Akshar*).

ACKNOWLEDGMENT

I would like to express my sincere thanks towards my guide Prof. Mr. Abhijit V. Bapat for his invaluable continuous guidance directing the course of action in the necessary and new direction and imparting me the knowledge to be able to write and present this paper.

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.

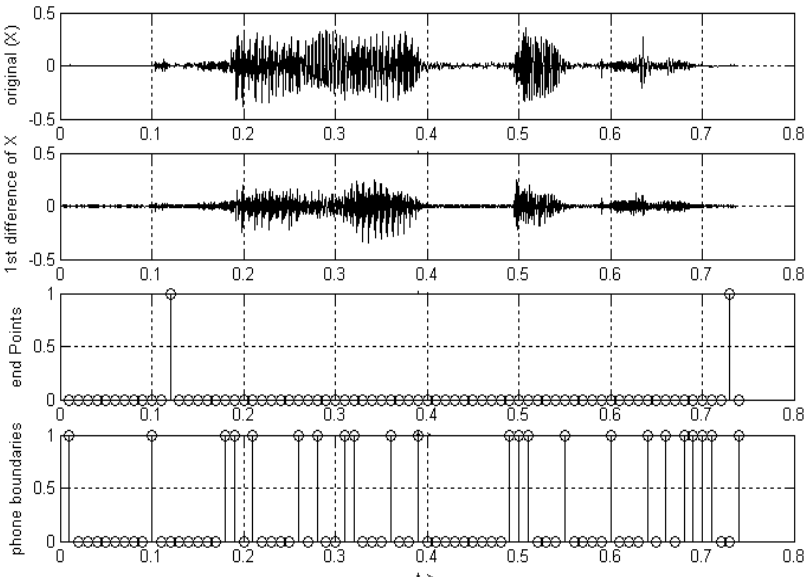

Fig 1: Speech Segmentation of utterance "PHONETICS" by a male speaker. From top to bottom:
Fig (a) is original signal. Time axis is in ms
Fig (b) is 1st difference of Fig (a)
Fig (c) shows end points as detected (12 to 73) 10s of ms
Fig (d). Shows phone boundaries these are (1, 10, 18, 19, 21, 26, 28, 31, 32, 36, 39, 49, 50, 51, 55, 60, 64, 66, 68, 69, 70, 71, 74) in 10s of ms

Groups Identified As:- (11-21) is PA; (20-26) is Vo; (25-28) is NA; (27-31) is Vo (here (25-31) is nasal NA); (30-32) is Vo; (31-36) is CH; (35-39) is Vo; (39-49) is silence; (49-51) is CH; (50-55) is Vo; (55-59) is silence; (59-73)is SA. Here phone "KA" is lost.

Vo: Vowel.
Group (31-36) will be identified as vowel if no silence before it is considered thus avoiding wrong identification as "CH".
Every Consonant written is first variant of the one of the five groups identified except SA.

NOTE: To reduce space needed here, some segments with same identification and occupying same phoneme are grouped above; e.g. Segments (10-18-19-21) are manually grouped into one segment (11-21); identification of these segments is same i.e. PA.