Automatic Segmentation of Kannada Speech Signal into Syllables and Sub-words: Noised and Noiseless Signals

Hemakumar G, Punitha P

Abstract— this paper addresses the problem of Kannada speech signal segmentation. The designed algorithm automatically segments the continuous Kannada speech signal into syllables and sub-words using the dynamic threshold computation by the combination of short time energy and magnitude of signal. The proposed method first pre-processes the original Kannada speech signal then framing is done for every 20 millisecond with an overlapping of 6.5 millisecond. Secondly frames are passed through hamming window using the same size of frame. Thirdly voiced and unvoiced part is detected through computing dynamic threshold using short time energy and magnitude of signal. In this paper 20 unique sentences are used for experiment which can be used as commands to simple mobile sets. Each of these sentences was recorded for 10 times and segmentation testing is done for each signal and computed accuracy rate of segmentation. The automatic segmentation success rate of individually uttered of sentences in experiments is excellent and has reached accuracy rate of 96.69% and miss rate of about 3.31%. Computations are done using Mat lab.

Index Terms— Kannada Language, Magnitude of signal, Pre-emphasize, standardization of signal, Segmentation, Short time energy and Syllables.

1 INTRODUCTION

Speech can be considered as a time-varying signal since the parameters of the signal such as the amplitude, frequency and phase varies in time. Speech signal can be segmented into Words, Sub-words, Syllables and Phonemes. Segmentation is a process of decomposing the speech signal into a set of basic phonetic units. The basic phonetic unit can be a phoneme or a syllable, based on the language. The general idea of segmentation can be described as dividing something continuous into discrete, non-overlapping entities. Several speech recognition systems consider syllable as a basic unit because of its better representational and durational stability relative to the phoneme. Indian languages are syllable-time based languages, syllable is considered as the basic sound-unit in this work.

In normal sentential utterances, the speaker develops a rhythm of stressed and unstressed syllables. Certain languages (e.g., English) have been called stressed-timed because stressed syllables tend to occur at regular time intervals. Other languages (e.g., Most Indian languages like Kannada language) are syllable-timed because each syllable tends to have equal duration. In both cases, the phenomenon is more perceptual than acoustical since physical measurements of duration vary considerably from the proposed isochronizes. The production regularity may exist not at the acoustic level but at the articulatory level, in terms of muscle commands for stressed syllables coming at regular intervals. The acoustic differences between the two types of languages are the stress-timed languages significantly reduce the durations of unstressed syllables compared to stressed ones, while syllable-timed languages do so to a much lesser extent.

This paper discussing the continuous Kannada speech signal segmentation into syllables and sub-words. The Kannada is a syllable-timed based language. The programs written in mat lab for automatic segmentation of continuous speech will segment any Kannada speech signal of any length. The length of signal increases the complicity of segmentation increases, in our experiment if length of signal increases more than 5 seconds then that signal will be divided into 5 seconds with overlapping of 0.5 second and then segmentation is done for each parts.

The remaining part of the paper is organized into five different sections; Section 2 deals with review of speech signal segmentation. Section 3 deals with the signal preprocessing, Section 4 deals with algorithm of proposed model. Section 5 deals with Experimentation results. Section 6 deals with discussion and conclusion.

2 REVIEW OF SPEECH SIGNAL SEGMENTATION

In paper [1] they have experimented to automatically segment and label continuous speech signal into syllable-like units for Indian languages (Tamil and Telugu). In this approach, the continuous speech signal is first automatically segmented into syllable-like units using group delay based algorithm. Similar syllable segments are then grouped together using an unsupervised and incremental training (UIT) technique. Isolated style HMM models are generated for each of the clusters during training. During testing, the speech signal is segmented into syllable-like units which are then tested against the HMMs obtained during training. They showed that performance of segmentation followed by labeling is superior to that of a flat start syllable recognizer. Paper [2] used

Hemakumar G. Research Scholar, Bharathiar University, and Asst Professor, Department of Computer Science, Government College for Womens', Mandya. hemakumar7@yahoo.com

Dr. Punitha, Professor & Head, Department of MCA, PESIT, Bangalore.

blind speech segmentation procedure that allowed a speech sample to be segmented into words/sub-word units without the knowledge of any linguistic information (such as, orthographic or phonetic transcription). This was done by using end-point detection technique which detects the proper start and the end points of the speech events. The start and end points are detected by tracing abrupt change of the data sequence, which is greater or less than a given threshold. Here they segmented the continuous Bengali speech signal. Paper [6] worked for automatic Punjabi continuous speech segmentation using Short term energy. Basic units for segmentation are words, phonemes and syllables. They segmented Punjabi words into syllables in Punjabi language and showed that seven types of syllables are recognized in Punjabi. These syllable types are: V, VC, CV, VCC, CVC, CCVC and CVCC; where V and C represent vowel and consonant respectively. Paper [8] has presented acoustic-phonetic analysis of Kannada, spoken in southern and coastal regions of Karnataka. In this work, mid points of the vowels were automatically identified using a trained speech recognition system, and subsequently, formant frequencies corresponding to the mid-point were computed. Preliminary analysis of first and second formant frequencies showed observable difference in the characteristics of vowels /a/, /A/, /i/, /I/, /u/, /U/, /e/, /E/, /o/ and /O/. The frequency domain properties of vowels across phonetic contexts varied less in case of the speakers of coastal Karnataka in comparison with those of speakers from South Karnataka. This is likely due to the fact that Kannada was not the mother tongue of most of the speakers from Coastal Karnataka. Paper [16] has discussed the Speech signal which is basic to study and analysis of speech technology and phonetics. To form meaningful chunks of language, the speech signal needs to have dynamically varying spectral characteristics; sometimes varying within a stretch of a few milliseconds. Phonetics groups these temporally varying spectral chunks into abstract classes roughly called allophones. Distribution of these allophones into higher level classes called phonemes takes us closer to their function in a language. In most of the languages phonemes and letters have varying degrees of correspondence. Since such a relation exists, a major part of a speech technology deals with the correlation of script letters with chunks of time-varying spectral stretches in that language. Indian languages are said to have a more direct correlation between their sounds and letters. Such similarity gives a false impression of similarity of several sets of text-to-sound rules across these languages. A given letter which has parallels across various languages may have different degrees of divergence in its phonetic realization in these languages. They illustrate such differences and point out the grey areas where speech scientists need to pay greater attention in building their systems, especially multilingual systems for Indian languages. In paper [10] they have experimented on the Acoustic segmentation of speech based on landmark detection which is an important stage in keyword spotting based on acoustic matching. In the present work, the class of plosive sounds in continuous speech is considered for detection and classification. Acoustic-phonetic features extracted in the vicinity of landmarks or speech events are shown to be reliable for the detection of unvoiced stops with high temporal accuracy. They tried for Marathi words and sentences by developing their own database for training and testing. In Paper [11] presents simple and novel feature extraction approaches for segmenting continuous Bangla speech sentences into words/sub-words. These methods are based on two simple speech features, namely the time domain features and the frequency-domain features. The time-domain features, such as short-time signal energy, short-time average zero crossing rate and the frequency-domain features, such as spectral centroid and spectral flux features are extracted in this research work. After the feature sequences are extracted, a simple dynamic thresholding criterion is applied in order to detect the word boundaries and label the entire speech sentence into a sequence of words/sub-words. In paper [3], there objective was to develop a method for syllabification of the acoustic signals of Sinhala words and they have achieved considerably high precision in the outcome. It was clear that the wrong pronunciation creates problems in the process of syllabification. In this work the informants were not trained to pronounce the words correctly, and therefore they have collected the pronunciation of words in their natural speech. As an effect, this decreased the accuracy of their methodology to some extent. Further, the two occurrences of the same word generated by the same speaker were behaved differently due to the variation in the pronunciation. Paper [4], they have proposed a method to improve the quality of the segment graph for segment-based speech recognition by attempting to reconstruct segments that are missing due to possible insertion errors. Acoustic discontinuities and manners of articulation are used in evaluating each boundary in the segment graph. The results showed satisfactory phonetic recognition accuracy in Thai continuous speech despite the increase in segment graph size in an intermediate step of the system. However, the algorithm reported in their work only adds new segments into the segment graph. Segment errors due to boundary deletion errors are still remained.

3 SIGNAL PREPROCESSING

In this paper, pulse code modulation with a frequency of 16 KHz, 16-bit mono channel is used. Each sentence signal is separately recorded with a silence region before and after right Signal. These signals were recorded at a little noisy room environment, while Gold Wave Software was used to record with the help of mini microphone of frequency response 50 – 12500Hz. A detail of speech database is described in table 1.

Table 1: Continuous Speech Database Description.

Language	Kannada
Speech type	Read Speech
Number of Sentence Used	20 Sentence
Number of Speaker	1 male speaker
Recording Conditions	Room Environment with little noise
Number of signals	Each sentence recorded 10

used to Testing.	times, total 200 Signals.

4 PROPOSED METHOD

In this paper we have experimented to design the speech segmentation using the combination of time domain feature and frequency domain feature and segmented into syllables and sub-words in 5 stages. Our model works in offline mode, firstly speech signal need to acquire and store in speech database and then pass into program. Each stage is explained using algorithm.

4.1 Preprocessing stage

- Analog signal is digitalized.
- DC Component is removed from digitalized sample values, for i=1 to N s(i) = s(i) – mean(s), where N is length of Signal, s is signal.
- Pre-emphasize phase ŝ(n) = s(n) – ã * s(n-1) , where constant ã value is 0.9955.
- Standardization of amplitude
  S (n) = ŝ (n) / max (| (ŝ (n) |)

4.2 Framing and Windowing

- Frame blocking is done for signal. Here we have framed for every 20 millisecond with an overlapping of 6.5 millisecond.
- Each frame is passed under hamming window by keeping same size of frame.

$$w_m = \begin{cases} 0.54 - 0.46\cos(\frac{2\pi n}{N-1}), & 0 \le n < N \\ 0 & , otherwise \end{cases}$$

4.3 Computation of Short time energy and magnitude

- Short time energy is computed for frame using

$$\boldsymbol{E_n} = \sum_{\boldsymbol{m=-inf}}^{\boldsymbol{m=inf}} [\boldsymbol{x(m)w(n-m)}]^2$$

- Magnitude of frame is computed using 9th order of butter filter with 0.33 low cut-off value.
  1. [B, A] = butter(9,0.33,'low')
  2. Y = filter(B, A, Signal)
  3. Magnitude = abs(sum(Y))
- Check for end of signal, if end of signal comes out of the loop; otherwise repeat the steps.

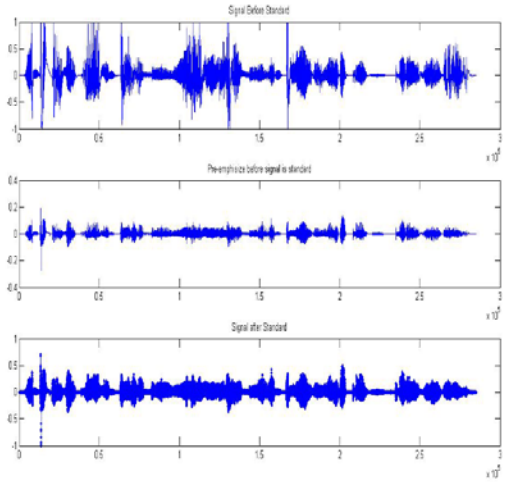
4.4 Dynamic Threshold detection

Magnitude of frame

Thr_msf = [((∑ msf)/ n) – min (msf) * 0.6] + min (msf)
If (msf > Thr_msf) mark V_msf =1

Short time energy

Thr_STE = [((∑ STE) / n) – min(STE) * 0.5] + min(STE)
If (STE ≥ Thr_STE) mark V_STE = 1



. Fig 1. Show the details of preprocessing stage result for speech signal. Command is '9480770862ge kare maaDu'. Command meaning is that 'make a call to 9480770862'. Figure represented has time verse amplitude

4.5 Detection of voiced part / Segmentation part

If V_STE * V_MSF equal to 1 then
Mark has voiced part
Otherwise mark has unvoiced part.

5 EXPRIMENTATION

Figure 1 shows the details of how the original Kannada speech signal will be and how it will be after the preprocessed signal. This preprocessed stage will makes the signal in standard format which leads at increasing the segmentation accuracy rate. The figure-1 showing the original signal, pre-emphasize signal without standard the signal and pre-emphasized signal after DC component removed and standardized the signal. This stage helps more clear to segment the signal and identify the noise part. The figure 3 shows the Rabiner technique and proposed model for preprocessing stage. The proposed model increases the speech signals decibel, amplitude of signal and more smothering the signal leads better to segmentation.

Table 2 shows the details of time taken for the execution of 200 signals and there segmentation accuracy rate and error rate occurs in short time energy, magnitude of signal and combination of both the techniques. This experiment shows that combination of both time domain and frequency domain will reduce the segmentation error rate. The short time energy shows very less time of execution but in this case there will be a problem of insertions and deletions of segment part in which signal occurs more than magnitude technique. The calculation of accuracy rate of segmentation is done by counting the

segmentation of signal and actual number of syllables and sub-words are present in the original sentence. The results of segmented parts of each speech signal is verified manually and computed the accuracy rate and error rate of each signal. Finally the averages of those have been taken.

Table 3 shows the sentences used in this experiment and in each sentences how many words and syllables are present. In this experiment there are totally 123 words and 295 syllables and each sentence is recorded for 10 times and segmentation testing has done for each signal. Table 4 showing the average accuracy rate for each sentence in each technique and showing that combination of short time energy and magnitude of signal will increases the accuracy rate and decreases the error rate and also increases the execution time.

Figure 2 showing the detection of voiced and unvoiced part in speech signal. Here figure shows how the short time energy and magnitude of signal will looks for same signal and how the voiced part is detected and marked. Here voiced part is detected using the dynamic threshold computation. It is also showing the missing part or unidentified voiced part in signal. The table 4 showing the details of accuracy rate and error rate of each sentence.

Table 2: showing the accuracy rate, error rate and time taken to execute the 200 continuous Kannada speech signal segmentation into syllables and sub-words. Program executed in Intel core i5 CPU @ 2.67GHz and RAM 3GBytes. Segmentation done using dynamic threshold computation for following methods.

Sl. No	Methods	Time taken in Second	accuracy Rate	Error Rate
1	Short Time Energy	33.9573	93.66	6.34
2	Magnitude of signal	112.7415	95.14	4.86
3	Combination of short time energy and magnitude	135.9312	96.69	3.31

Table 3: Showing the details of sentence used in experiment and number of words and syllables contained in sentences.

Sl. No	Sentence	Total Number	
		Words	Syllables
1	naMjuMDayyanige kare maaDu	3	10
2	naMjuMDanige kare maaDu	3	9
3	naMjuMDappanige phoonu maaDu	3	10
4	naMjuMDeeShanige kaal maaDu	3	9
5	kumaarage kare maaDu	3	8
6	kumaarige kare maaDu	3	8
7	aaytu phoona iDuteene	3	8
8	kare radu maaDu	3	6
9	9480770862ge kare maaDu	13	25
10	5971368026ge kare maaDu	13	27
11	9448073552 kare maaDu	13	26
12	8970560279 kare maaDu	12	27
13	7259776816 kare maaDu	12	26
14	123 kare maaDu	5	11
15	55066 kare maaDu	7	14
16	naMjuMDayyanige SMS maaDu	3	11
17	naMjuMDappanige SMS maaDu	3	11
18	8971368026 SMS maaDu	12	28
19	naMjuMDanige MMS maaDu	3	10
20	naMjuMDeeShanige MMS maaDu	3	11
	Total =	123	295

Table 4: Showing the details of segmentation accuracy rate and error rate. In this experiment each sentence is recorded 10 times and tested for all 10 signals. Results showing are of average of these signals. Acc= Accuracy rate, Err = Error Rate.

Sl. No. sentence	Short Time Energy		Magnitude of signal		Combination	
	Acc	Err	Acc	Err	Acc	Err
1	95.00	5.00	100.00	0.00	99.00	1.00
2	98.89	1.11	95.56	4.44	98.89	1.11
3	98.00	2.00	97.00	3.00	99.00	1.00
4	81.12	18.88	86.67	13.33	85.56	14.44
5	95.00	5.00	95.00	5.00	100.00	0.00
6	93.75	6.25	97.50	2.50	93.75	6.25
7	85.00	15.00	91.25	8.75	91.25	8.75
8	96.67	3.33	100.00	0.00	100.00	0.00
9	95.60	4.40	96.80	3.20	95.60	4.40
10	97.41	2.59	96.67	3.33	100.00	0.00
11	92.70	7.30	96.16	3.84	94.24	5.76
12	94.08	5.92	96.30	3.70	97.04	2.96
13	92.31	7.69	97.70	2.30	93.85	6.15
14	92.73	7.27	98.19	1.81	94.55	5.45
15	91.43	8.57	96.43	3.57	94.29	5.71
16	96.37	3.63	90.91	9.09	100.00	0.00
17	90.91	9.09	93.64	6.36	99.10	0.90

18	95.72	4.28	87.50	12.50	99.65	0.35
19	96.00	4.00	94.00	6.00	99.00	1.00
20	94.55	5.45	95.46	4.54	99.10	0.90
To-tal	93.66	6.34	95.14	4.86	96.69	3.31

6 DISCUSSION AND CONCLUSION

In this paper automatic Kannada speech segmentation into syllables and sub-words is done for the speech recognition. The recognition of syllables or sub-words is done for the selected voiced part of signal and not on the frame of the signal. The recognition part is not discussed in this paper. If the accuracy rate of segmentation is good then recognition accuracy rate will be good. In our experiment for identifying the voiced part or unvoiced part in the signal we have used the combination of short time energy and magnitude of signal and showed the increases of accuracy rate in segmentation. We have tested this method on isolated Kannada words signal, continuous Kannada read type of signal and audio Kannada dialogues file downloaded from internet of different time duration. In our experiment the error occurs only in the segmenting the noised speech signal and speaker having more breathing air-pressure noise occurring during the time of speaking or reading. To the continuous speech signal segmentation the noise from breathing is more problem than the noise from external environment.

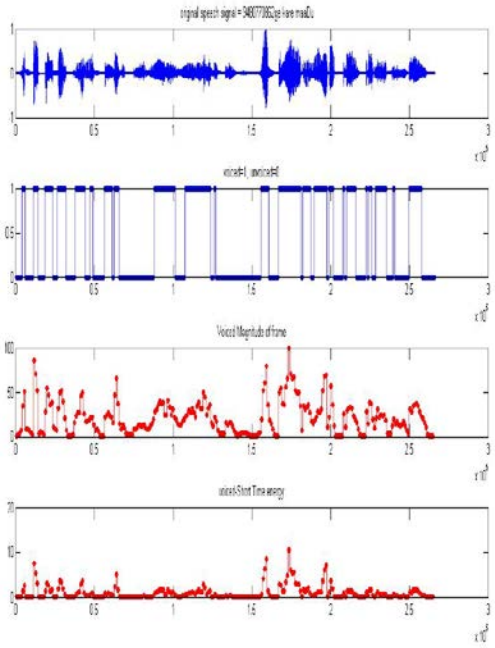


Fig 2: Showing segmentation of speech signal using short time energy and magnitude of frame.

ACKNOWLEDGMENT

The authors would like to thank friends, reviewers and Editorial staff for their help during preparation of this paper.

REFERENCES

[1] G Lakshmi Sarada et al., "Automatic transcription of continuous speech into syllable-like units for Indian languages", S¯adhan¯a Vol. 34, Part 2, April 2009, pp. 221–233. © Printed in India.

[2] Md. Mijanur Rahman et al., "Continuous Bangla Speech Segmentation, Classification and Feature Extraction", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 2, No 1, March 2012, ISSN (Online): 1694-0814.

[3] P. G. N. Priyadarshani and N. G. J. Dias, "Automatic Segmentation of Separately Pronounced Sinhala Words into Syllables", Journal of Science - Volume 6, University of Kelaniya, (2011) : 35-44.

[4] Krerksak Likitsupin, et al., "Improving Segment-based Speech Recognition by Recovering Missing Segments in Segment Graphs – A Thai Case Study".

[5] R. ÈMEJLA and P. SOVKA, "Speech Segmentation Using Bayesian Autoregressive Change point Detector", Radio-Engineering, Vol. 7, No. 4, December 1998.

[6] Er. Amanpreet Kaur and Er. Tarandeep Singh, "Segmentation of Continuous Punjabi Speech Signal into Syllables", Proceedings of the World Congress on Engineering and Computer Science 2010 Vol I, WCECS 2010, October 20-22, 2010, San Francisco, USA.

[7] Kiruthiga S1 and Krishnamoorthy K, "Annotating Speech Corpus for Prosody Modeling in Indian Language Text to Speech Systems", International Journal of Computer Science Issues, Vol. 9, Issue 1, No 1, January 2012.

[8] Nagesha, K Samudravijaya and G Hemantha Kumar, "ACOUSTIC-PHONETIC ANALYSIS OF KANNADA ACCENTS".

[9] A.P.Henry Charles & G.Devaraj, "Alaigal-A Tamil Speech Recognition", Tamil Internet 2004, Singapore.

[10] Veena Karjigi et al., "Identification of stop consonants for acoustic keyword spotting in continuous speech", Proc. of Wireless Personal Multimedia Communications (WPMC), September 2007, Jaipur, India.

[11] Md. Mijanur Rahman and Md. Al-Amin Bhuiyan, "Continuous Bangla Speech Segmentation using Short-term Speech Features Extraction Approaches", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 3, No. 11, 2012.

[12] Hioka Y and Namada N, "Voice activity detection with array signal processing in the wavelet domain", IEICE TRANSACTIONS on Fundamentals of Electronics, Communications and Computer Sciences, 86(11):2802-2811, 2003.

[13] Beritelli F and Casale S, "Robust voiced/unvoiced classification using fuzzy rules", In 1997 IEEE workshop on speech coding for telecommunications proceeding, pages5-6, 1997.

[14] Qi Y and Hunt B, "Voiced-unvoiced-silence classification of speech using hybrid features and a network classifier", IEEE Transactions on Speech and Audio Processing, I(2):250-255, 1993.

[15] Florian Pausch, "Automatic Segmentation of Speech into Sentences Using Prosodic Features", submitted to the Institute of Electronic Music and Acoustics of the University of Music and Performing Arts Graz in November 2011.

[16] Perı Bhaskararao, "Salient phonetic features of Indian languages for Speech Technology", *Tokyo University of Foreign Studies*, Tokyo, Japan.