Int J Speech Technol (2016) 19:9–18
DOI 10.1007/s10772-015-9320-6

Automatic speech segmentation in syllable centric speech recognition system

Soumya Priyadarsini Panda[1] · **Ajit Kumar Nayak**[2]

Received: 13 August 2015 / Accepted: 12 November 2015 / Published online: 21 November 2015

Abstract Speech recognition is the process of understanding the human or natural language speech by a computer. A syllable centric speech recognition system in this aspect identifies the syllable boundaries in the input speech and converts it into the respective written scripts or text units. Appropriate segmentation of the acoustic speech signal into syllabic units is an important task for development of highly accurate speech recognition system. This paper presents an automatic syllable based segmentation technique for segmenting continuous speech signals in Indian languages at syllable boundaries. To analyze the performance of the proposed technique, a set of experiments are carried out on different speech samples in three Indian languages Hindi, Bengali and Odia and are compared with the existing group delay based segmentation technique along with the manual segmentation technique. The results of all our experiments show the effectiveness of the proposed technique in segmenting the syllable units from the original speech samples compared to the existing techniques.

Keywords Speech recognition · Speech segmentation · Syllable · Indian languages · Vowel onset point · Vowel offset point · Zero crossing rate

1 Introduction

In the last few years, there has been a major change in the technology and strong presence of the IT companies in the market leads to the development of a number of sophisticated information processing devices. This increases the demand of human computer interaction via natural languages to enhance the ease of accessibility and user friendliness. Research in the area of speech and language processing enables machines to speak and understand natural languages as like human leading to the development of different essential and luxury products enhancing the quality of life (Mao et al. 2014; Panda et al. 2015). As compared to other approaches, there have been sufficient successes today that suggest that these technologies will continue to be a major area of research and development in creating intelligent systems now and far into the future.

Speech recognition is the ability of a machine to understand and carry out spoken commands by a human. An automatic speech recognition (ASR) system in this aspect converts the spoken speech segments in a language into the respective text units (Kitaoka et al. 2014). The speech recognition systems makes use of various speech and language technologies and are increasingly being used to facilitate and enhance human communication, particularly through their use in human computer interfaces such as in internet search engines and mobile communications (He et al. 2014). The possible application of speech recognition includes voice dialing (Gałka et al. 2014), data entry, designing applications for the elderly or communicatively

✉ Soumya Priyadarsini Panda
sppanda.cse@gmail.com

Ajit Kumar Nayak
ajitnayak@soauniversity.ac.in

[1] Department of CSE, Institute of Technical Education and Research, Siksha 'O' Anusandhan University, Bhubaneswar, Odisha, India

[2] Department of CS&IT, Institute of Technical Education and Research, Siksha 'O' Anusandhan University, Bhubaneswar, Odisha, India

impaired as assistive technology, games, etc. (Lippmann 1997; Wang and Sim 2014).

Recognizing human generated speech by a computer and converting it into respective text units is an inherently complex activity as it requires a set of cognitive and linguistic skills (Koolagudi and Rao 2012; McLoughlin 2014). Several ASR systems has been developed for different languages (Kelly et al. 2013) and the performance of the system may be measured by the overall accuracy obtained in correct word identification which depends on appropriate identification and segmentation of each pronounceable unit and its correct transcriptions in the form of text. Speech segmentation in ASR systems is the process of identifying the boundaries between words, syllables, or a phoneme in any spoken natural languages. Transcription of a continuous speech signal into a sequence of words is a difficult task, as continuous speech does not have any natural pauses in between words. However, most of the existing segmentation methods use manual segmentation techniques. A brief review on the existing works on speech segmentation and transcription are presented in the next section.

The conventional method of building a large vocabulary speech recognizer for any language uses a top-down approach to speech recognition. i.e. these recognizers first hypothesize the sentence, then the words that make up the sentence and ultimately the sub-word units that make up the words requiring a large speech corpus with sentence or phoneme level transcription of the speech utterances (Sakai and Doshita 1963). In addition, it also requires a dictionary with the transcription of the words at phonemic/sub-word unit and extensive language models to perform large vocabulary continuous speech recognition. The recognizer only recognizes the words that exist in the dictionary. However, adapting a new language requires building of a dictionary and extensive language models for the new language along with existing recognizer available for a particular language. Building such a corpus, is a labor-intensive and time consuming process. In a country like India having 22 official and a number of unofficial languages, designing ASR systems requires building huge text and speech databases and transcriptions which is a difficult and time consuming task. Our objective focuses on designing an automatic speech segmentation technique for the Indian languages.

There are fewer research documented for automatic speech segmentation in the Indian languages. As compared to the stress-timed languages like, English, Japanese, French, the syllable-timed Indian languages are highly phonetic in nature. i.e. there is no variation in the written scripts and their pronunciations. Also, most of the Indian languages have similar pronunciation rules making sub word unit identification simpler using the same technique. Whereas, the sub word unit identification is quite complicated for the stress-timed languages due to the presence of stress patterns and varied pronunciation rules. As Indian languages are syllable-centered, the focus of our work is to obtain a vocabulary independent syllable-level transcription of the spoken utterance.

A syllable centric ASR system identifies the syllable boundaries in the continuous speech signal and segments the speech into syllable units for converting into the respective written scripts or text (Wang 2000; Lin et al. 1996). One of the major reasons for considering syllable as a basic unit for ASR systems is its better representational and durational stability relative to the phonemes. A syllable is typically made up of a syllable nucleus, a vowel (V) with optional initial and final margins, consonants(C) (Origlia et al. 2014). The syllables thus may encompass both CV (consonant–vowel) and VC (vowel-consonant) transitions, including most co-articulatory and other phonological effects within its boundaries which makes syllable boundary identification easier (Li et al. 2013). While, the stress-timed languages like English suffers from issue like co-articulation effect (phoneme-to-phoneme transitions) between phonemes due to the stress used in pronunciations making phone boundary identification difficult, the boundary identification process is quite simpler for the syllable-timed languages. For syllable based speech segmentation, the syllable end points are needed to be obtained. The proposed technique uses a vowel offset point identification technique to segment the speech units at syllable boundaries. As the onset of vowel is an important event which makes the transitions from the consonant part to the vowel part, it helps in identifying the anchor points for different important acoustic events like aspiration, burst, transitions, etc. playing an important role in different speech segment identification. To analyze the performance of the model a number of test samples in the three Indian languages, Hindi, Bengali and Odia are considered and the model is compared with the existing group delay based and manual segmentation techniques. A subjective evaluation test is also performed to analyze the performance of the proposed technique for appropriate speech segmentation compared to the existing techniques.

The remainder of the paper is organized into the following sections. Section 2 describes a brief overview of the existing methods for speech segmentation. Section 3 discussed about the proposed segmentation technique for automatic speech segmentation at syllable boundaries. Section 4 discusses about the result analysis of the proposed model showing the efficiency of the technique in producing segmented syllable units from a set of continuous speech samples in the three Indian languages, Hindi, Bengali and Odia. Section 5 concludes the discussion showing the summary of the presented work and the future direction of the work where further work may be carried out.

2 Related work

Developing automatic speech recognition systems capable of obtaining high accuracy in different Indian languages is a difficult and ongoing process. There are a number of techniques available for the stress-timed languages like, English, French, Japanese, etc.; however fewer researches have been documented for the syllable-timed languages like, the Indian languages. Speech segmentation is an important problem in ASR systems as segmentation of the continuous speech signals into smallest pronounceable units helps in proper identification of the units by the ASR system. For extracting the syllable boundary information from continuous speech signal a temporal flow model (TFM) network has been discussed in (Shastri et al. 1999), where the time varying properties of the speech are captured by the TFM. The TFM is a neural network architecture that supports arbitrary connectivity across different layers, provides for feed-forward as well as recurrent links, and allows variable propagation delays along links. However, the approaches require analysis of the speech signals with respect to various aspects for training the model.

Different studies have investigated the speech segmentation problem of ASR systems in different cultures, most of which have focused on an ANN based approach (Sirigos et al. 2002) that needs training of the model with a speech corpus having information on end point in the respective languages. There are fewer studies that investigated syllable end point identification dynamically. Prasad et al., in (2004) presented a new algorithm for automatic segmentation of speech signals into syllable-like units based on short-term energy function. A syllable level segmentation technique is proposed in (Ziolko et al. 2006) for Japanese language based on a common syllable model where, the segment boundaries of the units are detected by finding the optimal HMM sequence. Zhao and Shaughnessy in (2008) proposes a hybrid automatic segmentation method that utilizes silence detection, convex hull energy analysis, and spectral variation analysis for of syllable units in Mandarin speech. In (Obin et al. 2013) a novel paradigms for syllable-based segmentation is proposed that uses the time–frequency representation, and the fusion of intensity and voicing measures through the frequency regions for selecting the pertinent information for the segmentation.

In (Musfir et al. 2014) a discussion on a modified group delay based approach is presented for reducing error proportions for the syllable based segmentation particularly for the fricatives, nasals and unvoiced stop type of units. For segmentation of the speech signal, the ratio of energy in the high frequency bands to the low frequency bands is used. However, the issues arise for the semivowels are not addressed. In (Besacier et al. 2014) a discussion focusing on automatic speech recognition for under resourced languages is presented which addresses the lacking of automatic unsupervised techniques for syllable based speech segmentation. Many models have been developed for decades and some progress has been achieved in speech segmentation, nevertheless the quality in terms of the accuracy still presents gaps, particularly regarding the adaptations in Indian languages.

3 Automatic speech segmentation technique

Segmenting continuous speech signal into syllable units is not a single phase conversion rules, instead may be carried out through different phases. The phases of automatic speech segmentation at syllable boundaries are shown in Fig. 1 and the details of the phases are discussed next.

3.1 Time domain representation

The input to the model is.wav files recorded at 8000 Hz, mono, which may contain speech samples for words or sentences in the considered Indian languages (Hindi, Bengali and Odia). For obtaining the syllable end points for automatic speech segmentation, the speech samples in.wav files are needed to be represented in the time domain first. However, the silent gaps (if present) in the input speech sample are removed first before further processing. The time varying spectral characteristics of the speech signal are represented graphically as wave patterns. The correct number of points to represent the wave file are obtained by considering the sampling frequency ($f_s = 1/t_s$) or the number of samples per second in the wav file. Figure 2 shows the time domain representation of the vowel "aa"

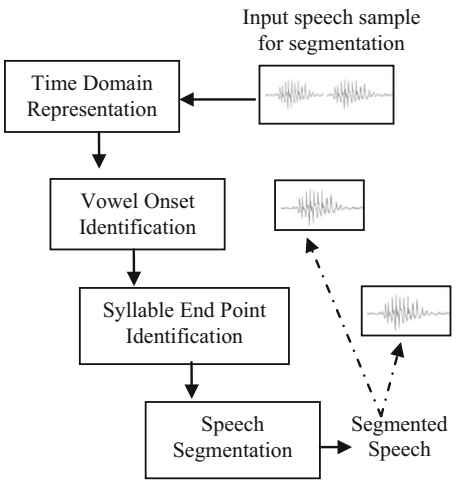

Fig. 1 Phases of speech segmentation

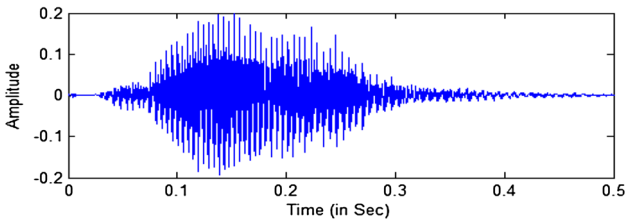

Fig. 2 Time domain representation of the vowel "aa"

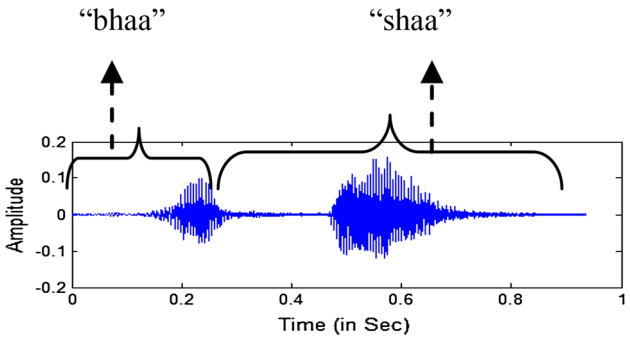

Fig. 3 Time domain representation of the word "bhaa-shaa"

and Fig. 3 shows the wave pattern of the word "bhaa-shaa" in Odia language (English meaning: language) on time axis with two syllable units "bhaa" and "shaa".

3.2 Vowel onset identification

For identifying the syllable end points the event of vowel onset and offset concepts in speech production are used. While, the event of occurrence of a vowel in a speech signal is the vowel onset point, the event of end of the vowel section is called the offset point. Significant changes occurred in the speech signal may be noticed in the energies of excitation source, spectral peaks, and modulation spectrum at the VOP instances. Also, the speech signals for the vowels are produced with a higher energy compared to the consonants and the variation may easily be detected by analyzing the spectral peaks in the speech signal. Therefore, the use of vowel onset and offset events helps in obtaining the syllable boundaries in the speech signals.

The syllable units in the Indian languages may be of the form: V, CV, CCV, CCVC and CVCC, where C and V represent consonant and vowel, respectively. In Indian languages, more than 90 % of the syllables are of CV type (Prasanna et al. 2009). Our model focuses on segmentation of the three forms of syllables: V, CV and CCV from continuous speech samples in the considered three languages. Segmenting a syllable into vowel and consonant regions can be performed by determining the onset and offset points of the vowels as in a CV unit, speech segment

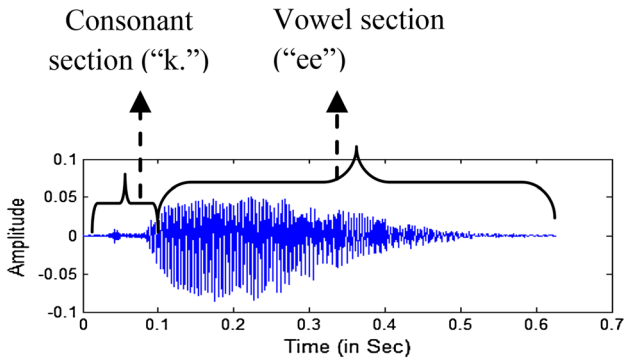

Fig. 4 Consonant and vowel regions in an utterance "kee" (CV)

before the vowel onset point is the consonant region, and after the vowel onset point is the vowel region. The variation in the consonant and vowel regions and the offset and onset points are shown in Fig. 4.

The pronunciation of Indian language consonants contains an inherent vowel "a" within it. The syllable units must have an ending vowel "a" or some other vowels. Therefore the speech sample is segmented into the number of segments equals to the available number of vowels in the input sample. The vowel sections may easily be identified in the wave representation as the vowel sounds are produced with higher energy compared to the consonants. The high energy portions in the wave patterns shows the vowels, the low energy sections shows the consonant sections and zero axis values shows the silent gap.

The vowel onset points (VOP) plays an important role not even for the number of segment identification but also used for the syllable end point identification in the next phase. For obtaining the VOPs, the spectral peaks method (Prasanna et al. 2009) is used. As the instance at which the event of onset of a vowel takes place must have high energy. The peak point in the energy spectrum shows the instance of the VOP over the time axis. However, to show the accuracy of the time instance obtained for VOPs, first the VOPs are manually labeled through wave analysis process as discussed in (Prasanna et al. 2009). The instance of the VOP obtained by manual wave analysis process for the syllable unit "ka" is shown in Fig. 5a.

For obtaining the VOPs dynamically using the spectral peaks method, the speech signal needs to be processed in blocks of 20 ms with a 10 ms shift. i.e. the time domain speech signal is processed in overlapping blocks. The Discrete Fourier Transform (DFT) is used for calculating the frequency spectrum of the speech signal for examining the information encoded in the frequency, phase, and amplitude. A 256-point DFT is computed on each block of speech signal and the sum of ten largest peak points is computed from the first 128 points. The sum of ten peak

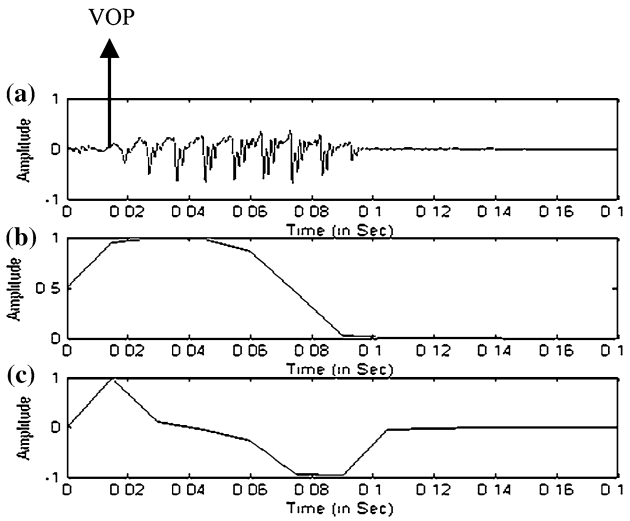

Fig. 5 VOP instance identification **a** Original waveform of "ka" CV type unit with manually labeled VOP instance **b** sum of 10 peak points in each block **c** enhanced values using FOD showing VOP evidence

points in each block plotted as a function of time is shown in Fig. 5b. The change at the VOP available in the spectral peaks energy is further enhanced by computing its slope using First Order Difference (FOD). These enhanced normalized values show the VOP evidence. The VOP evidence plot using spectral peaks method for the speech signal "/ka/" is presented in Fig. 5c.

3.3 Syllable end point identification

The instance at which a vowel section ends is known as the vowel offset points (VOF) and offset identifications leads to the identification of the syllable end points as Indian language syllables ends with a vowel. To identify the syllable end points, the zero crossing rate (Kay and Sudhaker 1986; Sreenivas and Niederjohn 1992) identification method on the speech signal within two identified VOPs or after the VOP instance is used. The zero-crossing rate (ZCR) is the rate of sign-changes along a signal (Lau and Chan 1985), i.e., the rate at which the signal changes from positive to negative or back. The zero crossing count is an indicator of the frequency at which the energy is concentrated in the signal spectrum. Voiced speech is produced because of excitation of vocal tract by the periodic flow of air at the glottis and usually shows a low zero-crossing count, whereas the unvoiced speech is produced by the constriction of the vocal tract narrow enough to cause turbulent airflow which results in noise and shows high zero-crossing count.

A zero-crossing is a point where the sign of a mathematical function changes (e.g. from positive to negative),

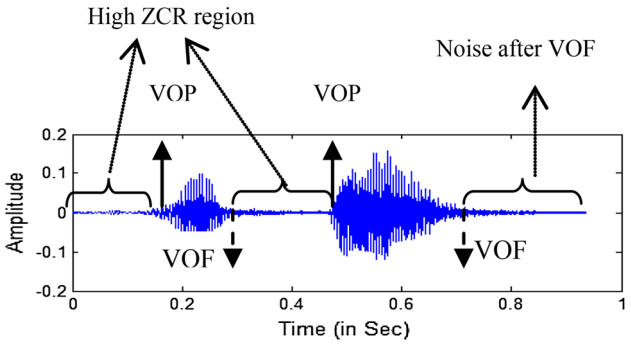

Fig. 6 Onset and offset point identification on speech sample "/bhaa-shaa/" with high ZCR regions showing consonant sections

represented by a crossing of the axis (zero value) in the graph of the function. The speech segments between two VOPs are processed to obtain the ZCR values and the higher ZCR after the VOP instance show the offset points for the vowel units and starting point of the next syllables as shown in Fig. 6. The rate at which zero crossings occur is a simple measure of the frequency content of a signal. The ZCR may be defined as given in (Lau and Chan 1985) as:

$$Z_n = \sum_{m=-\infty}^{\infty} sgn[x(m)] - sgn[x(m-1)]|w(n-m) \quad (1)$$

where,

$$sgn[x(n)] = \begin{cases} 1, & x(n) \geq 0 \\ -1, & x(n) < 0 \end{cases}$$

and w(n) is the windowing function with a window size of N samples as given in Eq. (2).

$$W = \begin{cases} 1/2N, & 0 \leq n \leq N-1 \\ 0, & otherwise \end{cases} \quad (2)$$

3.4 Speech segmentation

The proposed algorithm used for syllable based speech segmentation is given in algorithm 1. The input to the algorithm is.wav files containing pronunciation of different words in the considered languages and the algorithm produces segmented syllable units in separate.wav files as outputs. The wave data in the input speech samples (words) are copied into an array and are represented in the time domain, where t is the total time duration of the input speech as discussed in Sect. 3.1. The time instances of the VOPs are obtained by using the spectral peaks method (Sect. 3.2). For obtaining the syllable end points the ZCR values are computed and the VOFs are identified as discussed in Sect. 3.3. The input speech sample is then segmented into p samples based on the identified VOPs and

VOFs using the function *f(t)* as presented in step 5 of the proposed algorithm, where S_1 represents wave data between 0 and VOF_1, S_2 represents the wave data between VOF_i to VOF_{i+1} $(1 < i < n)$ and S_3 is the wave data between VOF_n to end.

For example, for the wave pattern presented in Fig. 6 discussed above, the number of VOPs identified is two; therefore the input speech segment is segmented into two syllable units. The syllable boundaries may be identified by the VOFs. For the first segment, $f(t) = S_1$ i.e. the data values from 0 time instance to VOF_1 time instance are segmented as first syllable segment and for the 2nd VOF, $f(t) = S_3$. i.e. the wave data between VOF_1 to end are segmented for the second syllable. Therefore, the two segmented speech units for the input word sample "bhaa-shaa" are "bhaa" and "shaa" as shown in Fig. 7.

Algorithm 1: Speech Segmentation Algorithm

1. For each sample $word_i$ read .wav file contents into W_i , where l be the length of W_i.
2. Represent the wav data in time domain with time axis from 0 to t, where t be the total time duration of the speech sample.
3. Identify all VOPs in the speech signal using spectral peaks which gives the number of syllables (n) present in the input speech sample
4. Obtain the syllable endpoints by identifying the vowel offset points using the higher ZCR values.
5. Segment the input speech samples into *p* segments based on the following rule:

$$f(t) = \begin{cases} S_1, & VOF=VOF_1 \\ S_2, & VOF_1<VOF<VOF_n \\ S_3, & \text{Otherwise} \end{cases}$$

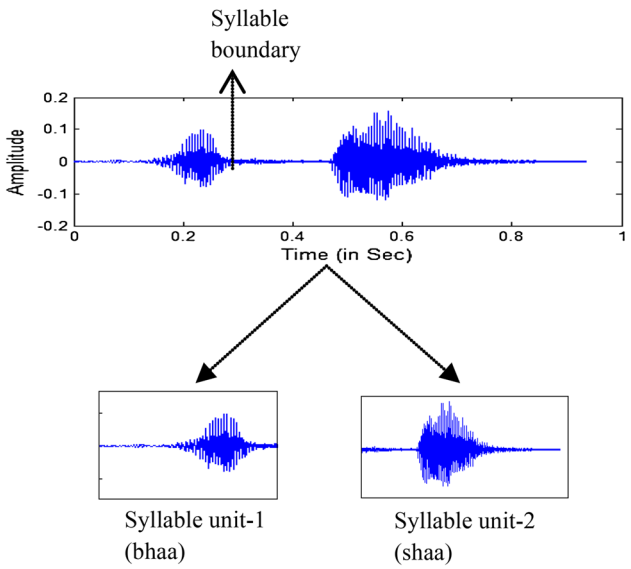

Fig. 7 Syllable based segmentation based on the proposed algorithm for the speech sample "bhaashaa" in Odia language

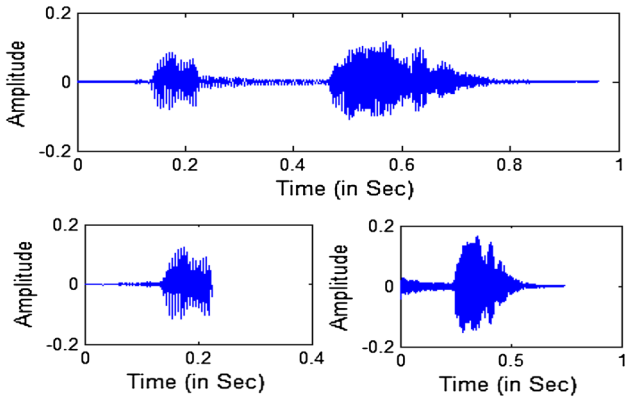

Fig. 8 Wave pattern of input speech "sa-bda" in Odia language (*top one*) and segmented output syllable units "/sa/" and "/bda/" (*bottom ones*)

4 Result analysis

For analyzing the performance of the segmentation algorithm in Indian languages, a set of random test samples (words) in Hindi, Bengali and Odia language are considered with the three syllable forms (V, CV, CCV) and the algorithm is run for segmenting the speech at syllable boundaries. The output of the speech segmentation algorithm for the input speech sample "sabda" in Odia language is shown in Fig. 8, where the speech unit is segmented into two syllable units "sa" (CV) and "bda" (CCV). Figure 9 shows the output of the segmentation algorithm for the word "hindi" in Hindi language, where the input speech is segmented into two syllable units "hi" (CV) and "ndi" (CCV) and Fig. 10 shows the speech segmentation process on a Bengali word "baanglaa" with three syllable units "baa", "ng" and "laa". To show the efficiency of the technique to perform syllable based speech segmentation on sentences, a set of experiments are also performed on different sentences in the considered languages. An example sentence is presented in Fig. 11 with the wave patterns of the segmented syllable units. For all the experiments performed, the same set of words and sentences are testes on the existing group delay based technique (Musfir et al. 2014). Even though, the group delay based approach produces poor quality results for the fricatives stop consonants or nasal sounds, few results are presented showing the duration of identified syllable units in the next section along with the proposed and manual segmentation technique.

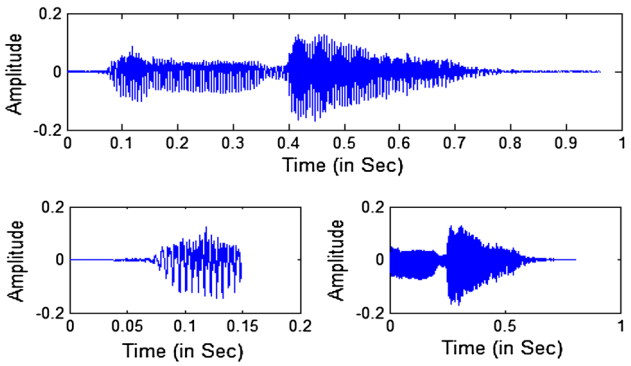

Fig. 9 Wave pattern of input speech "hindi" in Hindi language (*top one*) and segmented output syllable units "/hi/" and "/ndi/" (*bottom ones*)

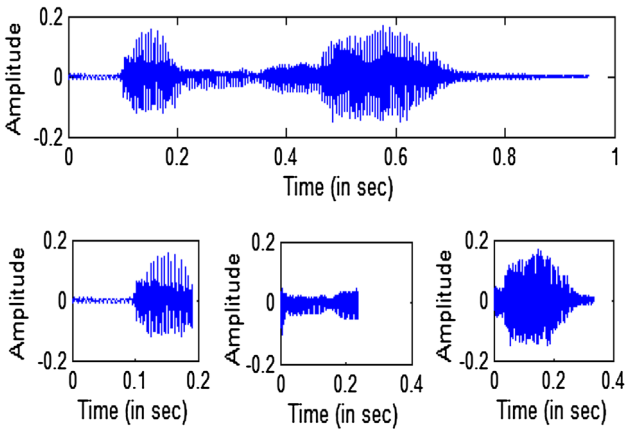

Fig. 10 Wave pattern of input speech "Baanglaa" in Bengali language (*top one*) and segmented output syllable units "/baa/", "ng" and "/laa/" (*bottom ones*)

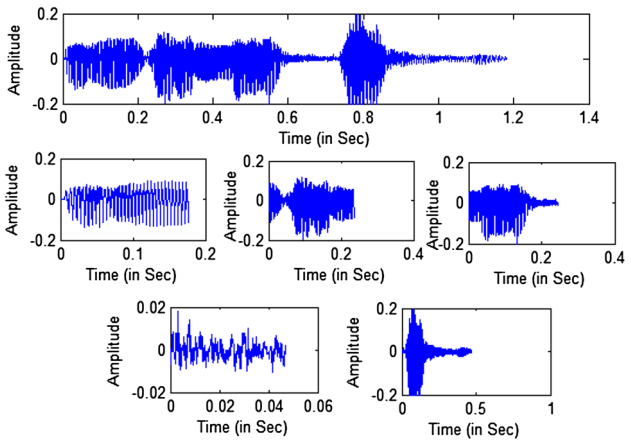

Fig. 11 Wave pattern of input speech "Hindi me sabd" in Hindi language (*top one*) and segmented output syllable units "hi", "ndi", "me", "sa", "bd" (*bottom ones*)

4.1 Variation in segmented syllable durations

Experiments are performed for sixty different units from each category of considered syllable types out of which six units are selected from each category to show the effectiveness of the technique in producing appropriate syllable segments. The speech samples in both male and female voices recorded at 8000 Hz, mono,.wav format are considered for all the experiments. The duration data for these six different units from the three considered syllable forms are presented in this section in male and female voice. For obtaining the predicted duration values of all the considered syllable units, a manual labeling approach is used. The same sets of words are used for segmenting the speech at syllable boundaries using the proposed algorithm as well as by the group delay based technique. All the experiments on syllable duration identification are performed on same set of speech data in male and female voice. The difference in the manually predicted and segmented speech durations for the V type units are presented in Figs. 12 and 13 in male and female voice respectively. Figures 14 and 15 shows the syllable duration values for the test results obtained for the CV type units. The duration values for the CCV type units are shown in Figs. 16 and 17 respectively. In all the experiments performed it may be observed from the results, the proposed segmentation technique achieves close results compared to the predicted syllable durations, whereas the group delay based technique shows a high degree of variation in syllable durations compared to the predicted durations.

The error percentage in the predicted syllable duration (actual duration) and the segmented duration is computed by the formula given in Eq. (3). Where, E is the percentage of error, D_S is the duration of segmented speech and D_A is the actual duration of the syllable obtained by manual segmentation. The average % of error in the three forms of syllable units (V, CV and CCV) for the 6 considered sample units are obtained separately for the proposed and existing segmentation technique. The proportion of error obtained in obtaining the exact duration syllable for the proposed and existing group delay based technique are

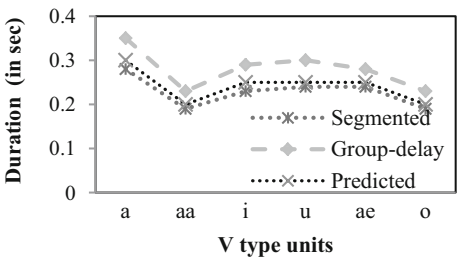

Fig. 12 Duration of V type units obtained by manual analysis and segmentation algorithm in male voice

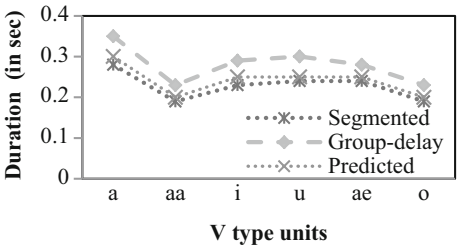

Fig. 13 Duration of V type units obtained by manual analysis and segmentation algorithm in female voice

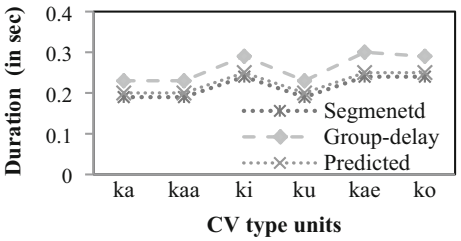

Fig. 14 Duration of CV type units obtained by manual analysis and segmentation algorithm in male voice

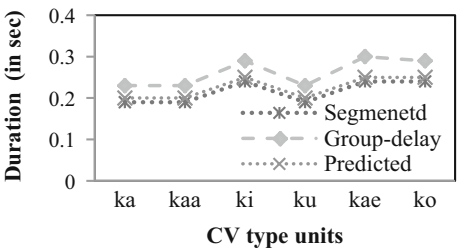

Fig. 15 Duration of CV type units obtained by manual analysis and segmentation algorithm in female voice

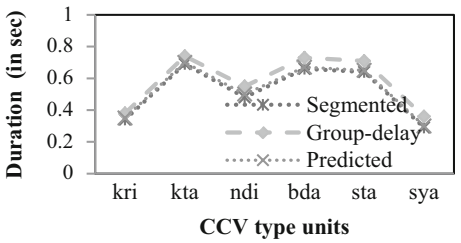

Fig. 16 Duration of CCV type units obtained by manual analysis and segmentation algorithm in male voice

presented in Figs. 18 and 19 respectively. In an average while 5, 4 and 2 % error occurred for the V, CV and CCV type units respectively by the proposed technique, the existing group delay based technique shows 14, 16 and 10 % of errors respectively for the syllable types in average. In other words, while the proposed segmentation algorithm gives 96 % accurate segmentation results (with

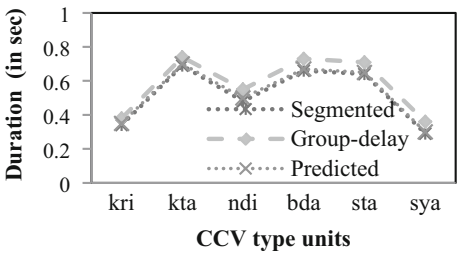

Fig. 17 Duration of CCV type units obtained by manual analysis and segmentation algorithm in female voice

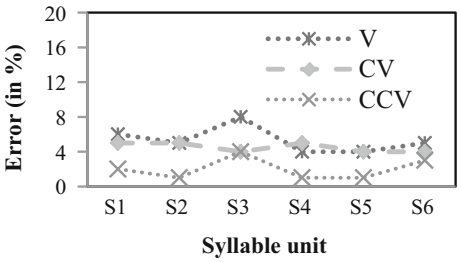

Fig. 18 Percentage of error in duration values for proposed technique

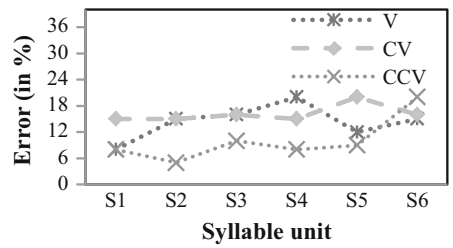

Fig. 19 Percentage of error in duration values for group-delay based technique

4 % error in average for all type of units), the existing group delay based technique achieves 86 % accuracy (with 14 % error in average for all type of units) for the considered speech samples.

$$E = \frac{D_S - \mathrm{D}_A}{D_A} \times 100 \tag{3}$$

4.2 Subjective evaluation

To analyze the quality of the segmented speech, a subjective evaluation test is performed, where a listeners test is performed by five different listeners to evaluate the quality of the segmented speech by the proposed and group delay based technique compared to the manually segmented speech. In these tests, the MOS (mean opinion score) (Panda and Nayak 2015) for a set of words in Hindi, Bengali and Odia language with the three syllable forms are considered. The speech units for the selected words for

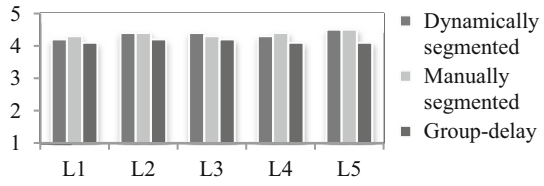

Fig. 20 Average MOS score for 10 samples by 5 listeners for Hindi language speech units

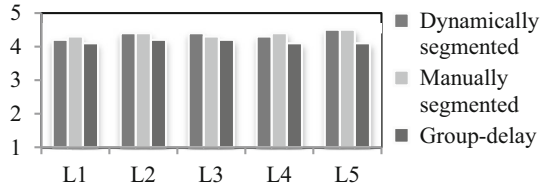

Fig. 21 Average MOS score for 10 samples by 5 listeners for Bengali language speech units

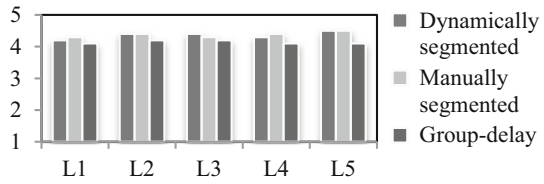

Fig. 22 Average MOS score for 10 samples by 5 listeners for Odia language speech units

the manually segmented and dynamically segmented by the algorithms and the output speech are played one by one. The listeners have to measure the speech quality in a five point scale (1: very low, 2: low, 3: average, 4: high, 5: very high) on the basis of their feeling to the speech for understandability and clarity. All the tests were performed with a headphone set and the only information the listeners are provided with is, they have to compare the three speech segments for the sample words. Figures 20, 21 and 22 shows the average MOS test results for ten words by each listener for our experiments on Hindi, Bengali and Odia language speech respectively. The results show comparable results of the proposed segmentation technique with the manually segmented speech. While, the group delay based technique shows relatively poor results due to the improper segmentations.

5 Conclusions

In this work, an automatic speech segmentation algorithm is proposed for dynamically segmenting Indian language speech at syllable boundaries. The use of the VOP identification techniques along with the ZCR technique performs segmentation of the continuous speech signals dynamically at syllable boundaries overcoming the time requirement and database limitation of manual segmentation technique. Experiments were performed to analyze the overall performance of the model in segmenting different type of syllable units in the three Indian languages, Hindi, Bengali and Odia. The presented results show very less proportion of error on segmented syllable durations compared to the existing technique on the actual durations of the speech samples. Also, this method saves a lot of time as needed by the manually labeling process for speech segmentation by giving the segmented output in few mille seconds only.

Even though, the proposed technique works well for the three considered syllable forms (V, CV, and CCV) in Hindi, Bengali and Odia language speech, the algorithm may occasionally produce un-natural results when multiple vowel regions fused together (e.g.: the words “aa-i” and “u-ee” in Odia language of type V–V). Therefore, the proposed algorithm may further be enhanced to get appropriate results for V–V pairs. This algorithm may also be incorporated in the Indian language speech recognition systems for segmenting the continuous speech signals into syllable units for further processing avoiding the need of a large manually labeled database for syllable duration information.

References

Besacier, L., Barnard, E., Karpov, A., & Schultz, T. (2014). Automatic speech recognition for under-resourced languages: A survey. *Speech Communication, 56*, 85–100.

Gałka, J., Masior, M., & Salasa, M. (2014). Voice authentication embedded solution for secured access control. *IEEE Transactions on Consumer Electronics, 60*(4), 653–661.

He, Y., Han, J., Zheng, T., & Sun, G. (2014). A new framework for robust speech recognition in complex channel environments. *Digital Signal Processing, 32*, 109–123.

Kay, S. M., & Sudhaker, R. (1986). A zero crossing-based spectrum analyzer. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 34*(1), 96–104.

Kelly, F., Drygajlo, A., & Harte, N. (2013). Speaker verification in score-ageing-quality classification space. *Computer Speech & Language, 27*(5), 1068–1084.

Kitaoka, N., Enami, D., & Nakagawa, S. (2014). Effect of acoustic and linguistic contexts on human and machine speech recognition. *Computer Speech & Language, 28*(3), 769–787.

Koolagudi, S. G., & Rao, K. S. (2012). Emotion recognition from speech using source, system, and prosodic features. *International Journal of Speech Technology, 15*(2), 265–289.

Lau, Y. K., & Chan, C. K. (1985). Speech recognition based on zero crossing rate and energy. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 33*(1), 320–323.

Li, M., Han, K. J., & Narayanan, S. (2013). Automatic speaker age and gender recognition using acoustic and prosodic level information fusion. *Computer Speech & Language, 27*(1), 151–167.

Lin, C. H., Wu, C. H., Ting, P. Y., & Wang, H. M. (1996). Frameworks for recognition of Mandarin syllables with tones using sub-syllabic units. *Speech Communication, 18*(2), 175–190.

Lippmann, R. P. (1997). Speech recognition by machines and humans. *Speech Communication, 22*(1), 1–15.

Mao, Q., Dong, M., Huang, Z., & Zhan, Y. (2014). Learning Salient Features for Speech Emotion Recognition Using Convolutional Neural Networks. *IEEE Transactions on Multimedia, 16*(8), 2203–2213.

McLoughlin, I. V. (2014). Super-audible voice activity detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 22*(9), 1424–1433.

Musfir, M., Krishnan, K. R., & Murthy, H. (2014). Analysis of fricatives, stop consonants and nasals in the automatic segmentation of speech using the group delay algorithm. In Twentieth National Conference on Communications (NCC) (pp. 1–6).

Obin, N., Lamare, F., & Roebel, A. (2013). Syll-O-Matic: an adaptive time-frequency representation for the automatic segmentation of speech into syllables. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), (pp. 6699–6703).

Origlia, A., Cutugno, F., & Galatà, V. (2014). Continuous emotion recognition with phonetic syllables. *Speech Communication, 57*, 155–169.

Panda, S. P., & Nayak, A. K. (2015). An efficient model for text-to-speech synthesis in Indian languages. *International Journal of Speech Technology, 18*(3), 305–315.

Panda, S. P., Nayak, A. K., & Patnaik, S. (2015). Text-to-speech synthesis with an Indian language perspective. *International Journal of Grid and Utility Computing, 6*(3–4), 170–178.

Prasad, V. K., Nagarajan, T., & Murthy, H. A. (2004). Automatic segmentation of continuous speech using minimum phase group delay functions. *Speech Communication, 42*(3), 429–446.

Prasanna, S., Reddy, B. V. S., & Krishnamoorthy, P. (2009). Vowel onset point detection using source, spectral peaks, and modulation spectrum energies. *IEEE Transactions on Audio, Speech, and Language Processing, 17*(4), 556–565.

Sakai, T., & Doshita, S. (1963). The automatic speech recognition system for conversational sound. *IEEE Transactions on Electronic Computers, 6*, 835–846.

Shastri, L., Chang, S., & Greenberg, S. (1999). Syllable detection and segmentation using temporal flow neural networks. In International Congress of Phonetic Sciences (pp. 1721–1724).

Sirigos, J., Fakotakis, N., & Kokkinakis, G. (2002). A hybrid syllable recognition system based on vowel spotting. *Speech Communication, 38*(3), 427–440.

Sreenivas, T. V., & Niederjohn, R. J. (1992). Zero-crossing based spectral analysis and SVD spectral analysis for formant frequency estimation in noise. *IEEE Transactions on Signal Processing, 40*(2), 282–293.

Wang, H. M. (2000). Experiments in syllable-based retrieval of broadcast news speech in Mandarin Chinese. *Speech Communication, 32*(1), 49–60.

Wang, G., & Sim, K. C. (2014). Regression-based context-dependent modeling of deep neural networks for speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 22*(11), 1660–1669.

Zhao, X., & Shaughnessy, D. O. (2008). A new hybrid approach for automatic speech signal segmentation using silence signal detection, energy convex hull, and spectral variation. In Canadian Conference on Electrical and Computer Engineering (pp. 145–148).

Ziolko, B., Manandhar, S., Wilson, R. C., & Ziolko, M. (2006). Wavelet method of speech segmentation. In 14th European Signal Processing Conference (pp. 1–5).