Prosody-Based Sentence Boundary Detection of Spontaneous Speech

Conference Paper · January 2014

DOI: 10.1109/ISMS.2014.59

CITATION
1

READS
76

4 authors, including:

Nursuriati Jamil
Universiti Teknologi MARA
101 PUBLICATIONS **632** CITATIONS
SEE PROFILE

Izzad Ramli
Universiti Teknologi MARA
11 PUBLICATIONS **25** CITATIONS
SEE PROFILE

Noraini Seman
Universiti Teknologi MARA
28 PUBLICATIONS **88** CITATIONS
SEE PROFILE

2014 Fifth International Conference on Intelligent Systems, Modelling and Simulation

Prosody-Based Sentence Boundary Detection of Spontaneous Speech

Nursuriati Jamil, Muhammad Izzad Ramli, Zainab Abu Bakar, Noraini Seman
Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA
Shah Alam, Selangor, Malaysia
liza@tmsk.uitm.edu.my, zadzed89@yahoo.com, zainab@tmsk.uitm.edu.my, aini@tmsk.uitm.edu.my

***Abstract*—Sentence boundary detection (SBD), also known as sentence breaking decides where a sentence begins and ends. This paper describes sentence boundary detection using acoustic and prosodic features for spontaneous Malay language spoken audio. We introduced the addition of volume change rate to 7 prosodic features and rate-of-speech for our preliminary experiment of detecting sentence boundary. Experiments are conducted on a forty-two minutes question-answer (Q/A) session of spontaneous speech comprising 12 adult male speakers and 4 female speakers. The speech datasets are first classified as speech/non-speech segments and only the non-speech segments are further tested as candidates of sentence boundaries. Our proposed rule-based method of boundary detection managed a promising 74.88% accuracy rate. For future work, we are considering to utilize learning algorithm to improve the accuracy rate and reduce false alert.**

Keywords-spontaneous speech; speech/non-speeceh detection; boundary detection; prosody features; Malay language

I. INTRODUCTION

Sentence boundary detection (SBD) for spoken audio has received increasing attention in recent years by many researches in various languages for many purposes [1]. Often, natural language processing tools require spoken audio to be divided into sentences by using sentence boundary detection for number of reasons [3]. The importance of sentence boundary detection was realized by many applications such as subsequent language processing, topic segmentation and summarization. Other than improving readability, SBD can provide structure relevant to language processing, topic segmentation and summarization [2].

The Malay language, has its origin from the ancient Austronesian language, is one of the world most spoken language, being spoken by approximately 180 million people [4]. Unfortunately, unlike English or other languages, speech-related research in Malay language is still at an early stage [7]. Furthermore, sentence boundary detection studies for speech recognition in Malay language are scarce. Several attempts of speech segmentation for spontaneous Malay spoken audio were done [5] [6]. However, they only focused on isolated words not continuous stream of words. Malay language has also been identified as under-resourced language [8] based on the following aspects: limited presence on the web, lack of electronic resources for speech and language processing, such as monolingual corpora, bilingual electronic dictionaries, transcribed speech data, pronunciation dictionaries, and vocabulary lists. Thus, it is empirical that speech-related work in Malay language is been pursued.

Generally, speech boundary detection is done using linguistic approach or acoustic approach or combination of linguistic-acoustic approach [9]. Linguistic-based method used linguistic features in statistical language model to detect the sentence boundary. On the other hand, acoustic approach used prosodic features such as fundamental frequency (*F0*), energy, duration and pause in detecting the sentence boundary. However, combination of linguistic and acoustic methods always produced higher accuracy compared to linguistic and acoustic approach alone. One of the constraints of linguistic approach is the need of a speech recognition component that comprises the language context information and linguistic features for segmenting the sentence [9]. Therefore, the speech recognition component needs to be constructed prior to sentence boundary detection. However, speech recognition often takes processing and higher computational costs. Moreover, speech recognition in Malay language is still at its infancy stage and recognition is limited to several words only [7]. Thus, linguistic approach is an unlikely option for sentence boundary detection for Malay language at the moment.

Several common acoustic features used for detecting endpoints and sentence boundaries are pause [9][12][14][16], zero-crossing rates [10][11], fundamental frequency [12][13][16], pitch [12][13][16], and energy [9][12]. In one of the earliest sentence boundary study, Gotoh and Renal [14] proposed pause audio segmentation without speech recognition using only pause feature. However, this study did not present an efficient method for automatic pause detection in various contexts. Furthermore, Wang et al. [15] argued using pause is not enough for practical segmentation usage. In their work, acoustics features such as frame energy, zero-crossing rate, pitch, pause, rate-of-speech and prosodic features are extracted from audio speech broadcast news for speech segmentation. The final accuracy achieved 82.3% for sentence boundary and 86.7% for non-boundary. In a more recent work of endpoint detection, Seman et al. [10] combined short-term energy, zero-crossing rates, frame-based Teager's energy and energy entropy features to detect endpoints for isolated words in Malay language. They tested 1,250 isolated Malay word utterances for recognition using Discrete-Hidden Markov Model and results showed that the highest average recognition rate of 80.76% are achieved using energy entropy feature. Wang et al. [12] utilized pause

2166-0662/14 $31.00 © 2014 IEEE
DOI 10.1109/ISMS.2014.59

duration, average F0 ratio and energy ratio to predict sentence boundary and punctuation from TDT3 English broadcast news corpora. They combined lexical, prosodic and modified n-gram score features into a dynamic conditional random fields framework and managed to reduce 20% relative recognition error.

During acoustical analysis of Malay language spontaneous speech done by [23], Hamzah et al. discovered that the last segment at the end of a Malay language speech sentence is an unvoiced segment. There are two types of speech segment that is voiced segment and unvoiced segment. Voiced segment is represented by a vowel because pronunciation of vowel is louder than consonant [19]. On the other hand, unvoiced segment is represented by a consonant [20]. In Japanese language, a vowel such as 'a', 'e', 'i', 'o', 'u' is presented as voiced segment and consonant is an unvoiced (voiceless) segment [21]. In general, vowel in Malay language is also known as voiced segment and consonant is unvoiced segment. However, there are certain consonants in Malay language that is categorized as voiced segments particularly consonants spoken at the end of a sentence [22] [24]. Based on these characteristics, we predict that the volume feature used by Jang et al. [17] for endpoint detection can also be used in sentence boundary detection to increase rate of sentence boundary detection and reduce false alert. In our study, we revisit the work by [15] and improved their methods by incorporating fundamental frequency and volume features at different stages of sentence boundary detection. This paper is composed of several sections. Section 1 identifies the motivation of the research. Section 2 explains the Malay spoken speeches as the data set. Details of methods and algorithms are presented in Section 3. Section 4 discusses the results of the experiments and finally, Section 5 concludes with recommendations for further work.

II. SPEECH DATASET

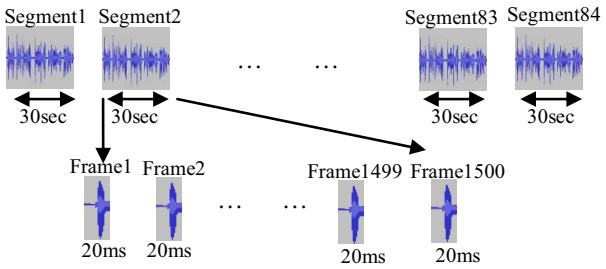
Our proposed methods are tested on Malaysia Parliamentary Hansard Document (MPHD) audio data (.wav) gathered from Malaysia Parliamentary debates dated 28 August, 2008 [18]. The Hansard documents contains spontaneous and formal speeches of parliamentary sessions surrounded with medium noise condition or environment (≥ 30 dB), disfluencies such as "um", repeat and self-repair [1], speakers interruption (Malay, Chinese and Indian races) and different speaking styles (low, medium and high intonation or shouting). Apart from that, the audio data also contains noises such as claps, laughter, whispers, and arguments. For our experiments, 185 minutes of one parliamentary session document was selected as our dataset. The selected Hansard document consists of two sessions. After analyzing the audio data of both sessions, the first session is omitted as it consists of formal speeches with read text prepared before the session. Only the second session of the debate is used as they are from the unplanned questions and answer (*Q/A*) session spontaneously answered during the parliamentary debate. The duration of the second session is 88 minutes. This 88-minutes audio data is further segmented into 176 non-overlapping segments of 30 seconds for faster processing. However, only 84 segments totaling to 2,520 seconds (i.e. 42 minutes) of audio data comprising 4 females and 12 males are used in our sentence boundary detection experiments. The purpose of selection is to allow variety of speakers that speak a minimum of two continuous sentences with minimum total duration of speech of at least 30 seconds. In the 42-minutes dataset, there are a total of 227 sentence boundaries.

III. METHODOLOGY

There are four stages involved in our experiments of sentence boundary detection: 1) audio segmentation 2) feature extraction 3) speech/non-speech classification and 4) boundary detection.

A. Audio Segmentation

Prior to feature extraction, the 42-minutes audio data which comprises 84 segments of 30-seconds spontaneous speech are further divided into 20 milliseconds (0.02 sec) non-overlapping frames. Fig. 1 illustrates the audio segmentation procedure into a total of 126,000 frames. These smaller frames are used in feature extraction for classification of speech/non-speech segments discussed later.



Number of frames per 30-second segment = $\frac{30sec}{0.02sec} = 1{,}500\ frames$

Total frames = 1,500frames x 84 segments = 126,000 frames

Figure 1. Audio segmentation of the speech dataset

B. Feature Extraction

There are two stages of feature extraction in our experiment. The first stage of feature extraction is used for speech/non-speech classification while the second stage is for sentence boundary detection. Three acoustic features that are fundamental frequency (F0), energy and zero-crossing rate (ZCR) are extracted from each 126,000 frames individually to classify speech/non-speech fragments. In the second stage of feature extraction, energy and F0 features are combined with seven prosodic features to detect sentence boundaries. These features are rate-of-speech, volume change rate, pause, succeeding and preceding sentence duration, succeeding and preceding pause duration, and rate-of-speech duration. Therefore, a total of ten features are extracted from the audio sentence boundary candidates. These candidates are non-speech fragments identified during speech/non-speech classification.

1) Fundamental frequency (F0): F0 is defined as the lowest frequency of a periodic waveform. A period of the waveform is the shortest possible time after which the waveform repeats itself. This single period is the smallest

repeating unit and it will describe the signal completely. F0 is calculated using (1).

$$F0 = \frac{1}{T} \tag{1}$$

where *F0* is the fundamental frequency and *T* s the fundamental period. F0 is used to classify the speech dataset into speech/non-speech segments. F0 before and after non-speech segments are also extracted to be used in sentence boundary detection.

2) Energy: Energy is very much related to the amplitude. It is a way of representing the amplitude changes in speech signal. Energy (E_k) for *k*-th segment is defined in (2), where $g(t)$ is the amplitude for *t*-th frame and *N* is the number of frames.

$$E_k = \sum_{n=1}^{N-1} |g(t)|^2 \tag{2}$$

Energy of the speech dataset is one of the features used to classify it to speech/non-speech segments. The energy preceding and succeeding the non-speech segments are also used to detect sentence boundary detection.

3) Zero-Crossing Rate: Zero crossing rate (ZCR) is measured based on the number of times the audio signal crosses the zero amplitude line by transition from a positive to negative or vice versa. The zero crossing value (Z_k) for the k-th segment is computed using (3), where $sgn[x_i(n)]$ can be three possible value that is +1, 0, -1 depending on whether the sample is positive, zero or negative.

$$Z_k = \sum_{n=1}^{N-1} |sgn[x_i(n)] - sgn[x_i(n-1)]| \tag{3}$$

where

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

In classification of the speech dataset, ZCR is used as a threshold to determine the speech/non-speech segments.

4) Rate-of-Speech (ROS): ROS often has two main definitions based on word/minute (WPM) and syllable/second (SPS) [25]. In real implementation, we will take vowel as syllable [15]. The total vowels that exist in a certain duration will represent the value of rate-of-speech. ROS is calculated using (4).

$$ROS = \frac{n}{\sum d_i} \tag{4}$$

where *n* is the vowel count, and d_i is the *i-th* vowel duration. In this paper, two ROS-related features are extracted to be used for detecting sentence boundaries. They are duration of ROS and ROS preceding and succeeding non-speech segments.

5) Volume Change Rate: Volume at endpoints normally descends gradually in a longer duration. This characteristic is similar to sentence boundary. On the other hand, pause's volume caused by disfluency or inter-word pause showed an abrupt change in a shorter duration. Fig. 2 illustrates an example of volume changes occuring at a disfluency and sentence boundary.

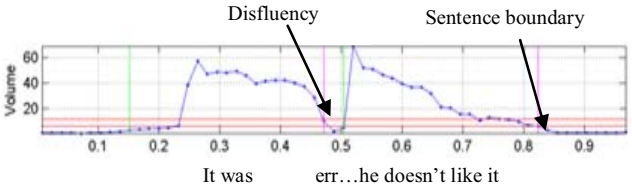


Figure 2. Voume slope at occurrence of disfluency and end of sentence

As disfluencies and pause are commonly mistaken as sentence boundary, we proposed to extract volume change rate to decrease false alert during sentence boundary detection. Equation 5 shows the equation we used to extract volume change rate.

$$Volume\ change\ rate = \frac{d_i - d_j}{t_i - t_i} \tag{5}$$

where represents the volume at time and represent the volume at time . is the time at frame and is the time at frame .

6) Duration: Duration of a sentence is important in determining whether a speech segment is a sentence or otherwise. In this paper, segment with longer duration is classified as a solid sentence. On the other hand, short segment is classified either as a part of sentence or a short sentence. This classification is crucial as sentence boundary exists at the end of a complete sentence. Therefore, speech segments with longer duration indicated a higher potential of occurrence of sentence boundary at the end of it. Togashi et al. [26] in summarizing spoken lectures deleted short sentences such as "no" and "yes" as they are regarded as low importance. However, we deemed short sentences as important because it may be related to preceding or suceeding sentences. Therefore, we retained the audio segments of short sentences. For the purpose of speech boundary detection, we extracted duration of the speech preceding and succeding the non-speech segments and duration of the pause preceding and succeding the non-speech segments.

7) Pause: Pause feature is known to play a critical role in sentence boundary detection and disfluency detection [27]. A study by [28] related to pause behaviour in spontaneous speech showed that fluent pause or sentence boundary's duration is observed to be statistically longer than disfluency pause. Therefore, pause is used in our work to discriminate between sentence boundary and disfluency pause based on pause's duration.

C. Speech/Non-speech Classification

The purpose of speech/non-speech classification is to categorize the 42-minutes (i.e. 84 segments) speech dataset into speech and non-speech segments. Before the experiment is conducted, a groundtruth dataset is

constructed by manually labeling the speech/non-speech segments of the speech datasets using Audacity 1.3.12-beta. An example of a manually annotated short sentence "Terima kasih" is illustrated in Fig. 3. *Label 1* shows the "Terima kasih" text aligns with its waveform and *Label 2* depicts the labeled speech (*S*) and non-speech (*NS*) segments of "Terima kasih" aligned with its corresponding waveform. A total of 6,413 segments are annotated from 84 segments consisting of 3,206 speech segments and 3,207 non-speech segments.

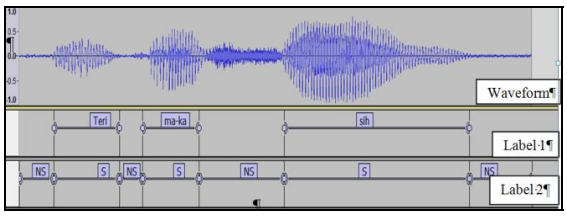


Figure 3. Manual annotation of sentence "Terima kasih"

Due to hardware constraint, the 84 segments are further divided into 20 milliseconds non-overlapping frames totaling to 126,000 frames. Fundamental frequency (F0), energy and zero-crossing rates (ZCR) are extracted from each of these frames to classify them into speech and non-speech segments. Frames that have high ZCR are categorized as speech segments and frames with low ZCR are categorized as non-speech segments. A ZCR threshold value (Thr_{ZCR}) is calculated to determine the speech/non-speech segments. Frames that have very low value of F0 are categorized as non-speech segments and frames with high F0 are categorized as speech segment. Energy feature is used to discriminate between speech and non-speech segments with selected set of threshold. A non-speech segment has much lower amplitude than the speech segment, resulting to non-speech segment to have lower energy. In our audio data, speech segment energy is higher than 30db, making it easier to discriminate from pause/silence.

Speech and non-speech classifications are done using the vowel/consonant/pause (V/C/P) classification rules adapted from [15]. However, we improved the rule by adding fundamental frequency feature as described in our earlier work [29]. The improved classification rules are presented in Fig. 4. Once all the frames are classified as vowel, consonants or pause, the final step is to merge vowel and consonant frames as speech segments and classify pause frames as non-speech segments. The non-speech segments are used to detect sentence boundary as speech segments are regarded as non-boundaries.

```
If   Frame_ZCR < Thr_ZCR  then
        Frame_Type = Consonant
Else if  Frame_F0 = 0,  then
        Frame_Type = Pause
Else if  Frame_Energy < Noise_Level  then
        Frame_Type = Pause
Else    Frame_Type = Vowel
```

Figure 4. Improved V/C/P classification rules

D. *Speech Boundary Detection*

For speech boundary detection, we only consider non-speech segments in our experiment as boundary candidates as possible boundaries existed only in these segments. From a total of 3,207 boundary candidates, we removed 935 boundary candidates as they have duration of less than 0.12 seconds. This is because upon closer analysis of the boundary candidates, we discovered that the minimum length of pause duration for our speech dataset is 0.12 seconds. Therefore, boundary candidates that are shorter than 0.12 seconds are not considered as potential sentence boundaries. After omitting shorter non-speech segments, we are left with 2,272 sentence boundary candidates.

A groundtruth dataset is constructed prior to conducting sentence boundary detection.. A speech transcript of the 42-minutes spoken speech dataset is acquired from the parliament. The speech transcript also annotates laughter, claps and noises as non-speech segments. Sentence boundary is manually labeled [SB] based on the symbol period, ' . ' and question mark, ' ? ' as shown in the example in Fig. 5. There are a total of 227 sentence boundaries in the speech dataset. The groundtruth is later used to evaluate the sentence boundary detection's performance.

No	Sentence	SB
6	.. pa peratus jenayah yang berjaya diselesaikan setakat ini? [SB] Kita tidak mahu angka-angka atau statistic daripada luar negara tetapi apakah statistik di dalam negara kita sendiri.[SB] Daripada jumlah yang diambil DNA ini, berapa peratus yang dapat diselesaikan – kes-kes pembunuhan, kes rogol, kes-kes orang yang hilang dan sebagainya. [SB] Berapa banyak kes jenayah ulangan yang…	3

Figure 5. Groundtruth dataset for speech boundary [SB] detection

After the construction of groundtruth dataset, a total of 10 audio features consisting of 7 prosodic features, 2 rate-of-speeches (*ROS*) and a volume feature are extracted from the 2,272 boundary candidates. The features as discussed in previous section are illustrated in Fig. 6, their summary depicted in Table 1.

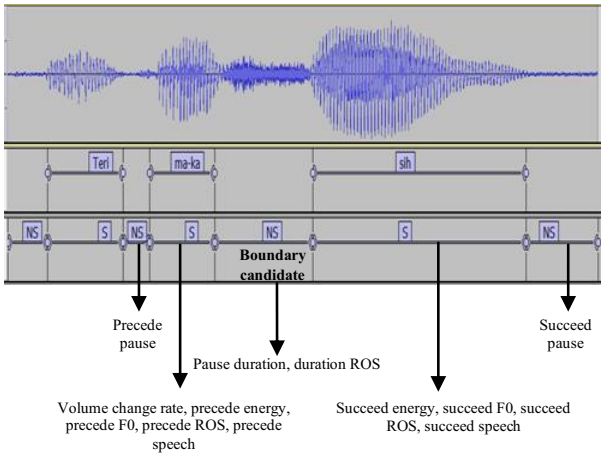


Figure 6. Extracted features of a boundary candidate

TABLE I. DESCRIPTION OF FEATURES USED FOR SBD

No.	Features	Description
1.	Succeed speech	Duration of the speech succeeding boundary candidate
2.	Precede speech	Duration of the speech preceding boundary candidate
3.	Succeed pause	Duration of the pause succeeding boundary candidate
4.	Precede pause	Duration of the pause preceding boundary candidate
5.	Pause duration	Duration of boundary candidate
6.	Fundamental frequency	Difference between preceding and succeeding fundamental frequency
7.	Energy	Difference between preceding and succeeding energy
8.	Duration rate-of-speech	Rate of boundary candidate duration and rate-of-speech
9.	Rate-of-speech	Difference between preceding and succeeding rate-of-speech
10.	Volume change rate	Volume change of rate preceding boundary candidate

In our initial experiment of sentence boundary detection, a basic classification method is used that is rule-based. Mean of each 10 features is calculated using (6) and used as a threshold for determining the sentence boundary.

$$Mean_{features} = \frac{1}{s}\sum_{k=1}^{s} F_k \quad (4)$$

where s is the total number of sentence boundary and is k-th feature value. The rules for detecting sentence boundary are shown in Fig. 7.

$$\begin{aligned} &If\ Feature_{Candidate} \geq Feature_{Thr}\ then \\ &\ Score = 1 \quad \%\ boundary\ hit \\ &else \\ &\ Score = 0 \quad \%\ boundary\ missed \end{aligned}$$

Figure 7. Rule-based sentence boundary detection

Each feature has its own threshold value and if a boundary candidate's feature evaluated to TRUE, a hit score is assigned to the boundary candidate indicating a sentence boundary. Meanwhile, if a boundary candidate's feature evaluated to a FALSE, a missed is assigned to the sentence boundary score. Boundary candidates that have a high score of boundary hits are classified as true sentence boundary.

IV. RESULTS AND DISCUSSIONS

This section provides the experimental results and further discussions of our proposed sentence boundary detection. An example result for each stage of speech/non-speech classification is illustrated in Fig. 8 and Fig 9. The identified non-speech segments known as boundary candidates are employed in sentence boundary detection experiment and 10 features are extracted. Then, the rule-based method is applied on all 2,272 boundary candidates. Table 2 showed samples of the 7 prosody features results as hit scores. Based on the hit scores, the boundary candidates are then labeled as sentence boundary <SB>. Fig. 10 shows an example of a boundary candidate detected as sentence boundary. Even though there are 3 boundary candidates in the speech segment, only one is confirmed as sentence boundary based on the accumulated hit scores.

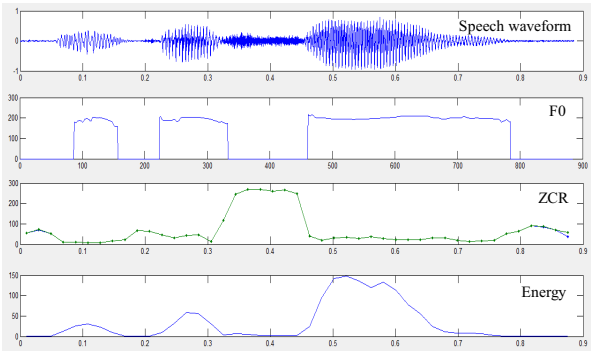


Figure 8. Acoustic features of 30-second speech segment

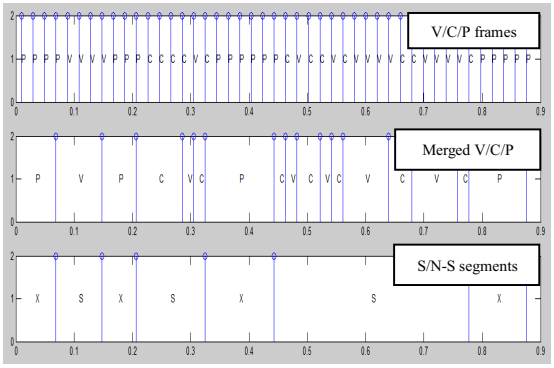


Figure 9. Vowel (V) and consonants (C) merged as speech segments (S); pause (P) merged as non-speech segments (X)

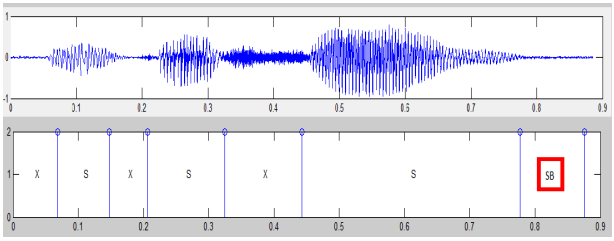


Figure 10. Detected sentence boundary

Performance of the proposed rule-based method is evaluated using accuracy rate (5), while total error rate (8) is calculated as a sum of false alert (6) and missing alert (7) [15].

$$Accuracy = \frac{Total\ correct\ sentence\ boundary}{Total\ sentence\ boundary} \quad (5)$$

$$False\ Alert = \frac{False\ detection\ of\ sentence\ boundary}{Total\ candidates\ of\ sentence\ boundary} \quad (6)$$

$$Missing\ Alert = \frac{Missing\ sentence\ boundary}{Total\ sentence\ boundary} \quad (7)$$

$$Total\ error = \frac{False\ detection + Missing\ detection}{Total\ candidates\ of\ sentence\ boundary} \quad (8)$$

Table II shows the results of sentence boundary detection using 10 features discussed earlier. The highest detection rate of sentence boundary at 74.88% is achieved by candidates with at least 4 total hits. However, it produced the highest false alert that is 80.63%. All the results showed that as the detection rate of sentence boundary increases, the false alert rate also increases. Even though the results seemed impractical due to the high false alert, the experiments showed that acoustic/prosodic features are deemed to be reliable properties for sentence boundary detection.

TABLE III. SENTENCE BOUNDARY DETECTION RESULTS BASED ON HIT SCORES

Hit Scores	Total error	False alert	Missing alert	Detection
≥ 4	**83.14**	**80.63**	25.11	**74.88**
≥ 5	70.99	65.88	51.10	48.89
≥ 6	52.37	44.54	78.41	21.59
≥ 7	31.77	22.71	90.75	9.25
≥ 8	17.82	7.83	100	0
≥ 9	11.26	1.27	100	0
10	10.16	0.17	100	0

V. CONCLUSION

This paper presents sentence boundary detection of Malay language spontaneous speech using acoustic/prosodic features only. Even though advanced method of speech boundary detection incorporate linguistic component, we managed to achieve satisfactory results using acoustic/prosodic features. The introduction of volume change rate as one of the prosody feature seemed appropriate as Malay language has some unique properties of unvoiced segment. Even though results of sentence boundary detection showed a high false alert, we believe that acoustic/prosody features only can be utilized. Rule-based classification method is a primitive method as classification results are based on mean threshold of the features. Our research direction in the future is to reduce the false alert drastically while maintaining a high detection rate. We further proposed to employ learning technique for classifying the boundary candidates to achieve better results.

ACKNOWLEDGMENT

The research was funded by Research Cluster Fund, Universiti Teknologi MARA, Shah Alam, Malaysia, no: 600-RMI/DANA 5/3/CG (5/2012).

REFERENCES

[1] S. Yildirim and S. Narayanan, "Automatic detection of disfluency boundaries in spontaneous speech of children using audio-visual information," IEEE Transactions of Audio, Speech & Language Processing, pp. 2-12, 2009.

[2] D. Hillard, M. Ostendorf and A. Stolcke, "Improving automatic sentence boundary detection with confusion networks," Proc. *HLT-NAACL 2004: Short Papers* (HLT-NAACL-Short '04). Association for Computational Linguistics, 2004, pp. 62-72.

[3] Y. Liu, E. Shriberg, A. Stolcke and M.P. Harper, "Using machine learning to cope with imbalanced classes in natural speech: Evidence from sentence boundary and disfluency detection," Proc. INTERSPEECH 2004, Oct 2004, pp. 2-5.

[4] AH. Omar AH, The Encyclopedia of Malaysia: Languages and Literature. Didier Millet, Singapore Editions, 2005.

[5] N. Seman, Z.A. Bakar and N.A. Bakar, "The optimization of artificial neural networks connection weights using genetic algorithms for isolated spoken Malay parliamentary speeches," Proc. 2010 International Conference on Computer and Information Application (ICCIA 2010), Dec 2010, pp. 162-166.

[6] M.S. Salam and M. Dzulkifli, "Segmentation of Malay syllables in connected digit speech using statistical approach," Journal of Computer Science and Security, vol. 2, pp. 23-33, 2008.

[7] C.Y. Fook, M. Hariharan, S. Yaacob and A. Adom, "A review: Malay speech recognition and audio visual speech recognition," Proc. 2012 International Conference on Biomedical Engineering (ICoBE 2012), Feb 2012, pp. 479-484.

[8] L. Besacier, E. Barnard, A. Karpov and T. Schultz, "Automatic speech recognition for under-resourced languages: A survey," Speech Communication, vol. 56, pp. 85-100, Jan 2014.

[9] A. Srivastava and F. Kubala, F, "Sentence boundary detection in Arabic speech," Proc. European Conference on Speech Communication and Technology (EUROSPEECH 2003), Sept 2003, pp. 949-952.

[10] N. Seman, Z.A. Bakar and N. Bakar, "An evaluation of endpoint detection measures for Malay speech recognition of an isolated words," Proc. 2010 International Conference of Information Retrieval and Knowledge Management (CAMP '10), Mac 2010, pp. 1628-1635.

[11] M.R.M. Zainal, A. Hussain and S.A. Samad, Segmentation of Audio Visual Malay Digit Utterances Using Endpoint Detection, Recent Progress in Data Engineering and Internet Technology, Lecture Notes in Electrical Engineering, vol. 156, Springer Berlin Heidelberg New York, 2013, pp. 281-286.

[12] X. Wang, H. T. Ng, and K.C. Sim, "Dynamic conditional random fields for joint sentence boundary and punctuation prediction," Proc. 13th Annual Conference of the International Speech Communication Association, (INTERSPEECH, ISCA 2012), 2012, pp. 281-286.

[13] R. Hamzah, N. Jamil and N. Seman, Acoustical Analysis of Filled Pause in Malay Spontaneous Speech: Computer Applications for Communication, Networking, and Digital Contents, Springer Berlin Heidelberg New York, 2012, pp. 251-259.

[14] Y. Gotoh and S. Renals, "Sentence boundary detection in broadcast speech transcripts," Proc. Automatic Speech Recognition: Challenges for the new Millennium (ASR-2000), 2000, pp. 228-235.

[15] D. Wang, L. Lu and H. Zhang, "Speech segmentation without speech recognition," Proc. Acoustics, Speech, and Signal Processing (ICASSP '03), April 2003, vol. 1, pp. 468-471.

[16] Y. Liu, E. Shriberg, A. Stolcke and M. Harper, "Using machine learning to cope with imbalanced classes in natural speech: Evidence from sentence boundary and disfluency detection." Proc. Empirical Methods in Natural Language Processing (EMNLP 2004), July 2004, vol 1, pp. 2-5.

[17] C-Y. Lin, J-S.R. Jang and K-T. Chen, "Automatic segmentation and labeling for Mandarin Chinese speech corpora for concatenation-based TTS," International Journal of Computational Linguistics and Chinese Language Processing, vol 10, no.2, pp. 145–166, 2005.

[18] N. Seman, Z.A. Bakar and N. Bakar, "An evaluation of endpoint detection measures for Malay speech recognition of isolated words," Proc. 2010 International Symposium in Information Technology, Kuala Lumpur (ITSim 2010), Jun 2010, pp. 1628-1635.

[19] M.B. Mustafa, R.N. Ainon and R. Zainuddin, "EM-HTS: Real-time HMM-based Malay emotional speech synthesis," Proc. Seventh ISCA Workshop on Speech Synthesis (ISCA 2010), Sept 2010, pp. 240-244.

[20] R.O. Tachibana, T. Kitamura and M. Fujimoto, "Differences in articulatory movement between voiced and voiceless stop consonants," Acoustical Science and Technology, vol 33, no.6, 2012, pp. 391-393.

[21] A. Cutler, T. Otake and J.M. McQueen, "Vowel devoicing and the perception of spoken Japanese words," The Journal of the Acoustical Society of America, 2009, pp.125-135.

[22] J S. R. S. Jaafar, "Nasal substitution in Sarawak Malay dialect," Asian Social Science, vol 9, no. 4, 2013, pp. 92-109.

[23] H. Hamzah, J. Fletcher and J. Hajek, "Durational correlates of word-initial voiceless geminate stops: The case of Kelantan Malay." Proc. 17th International Conference of Phonetic Sciences (ICPhS XVII), August 2011. pp. 89-92.

[24] H. Hamzah, J. Fletcher and J. Hajek, "A taste of prosody: Possible effects of the word-initial singleton-geminate contrast on post-consonantal vowel duration in Kelantan Malay," Proc. 6th International Conference on Speech Prosody, May 2012, pp. 490-493.

[25] N. Morgan and E. Fosler-Lussier, "Combining multiple estimators of speaking rate," Proc. 1998 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP 1998), May 1998, vol. 2, pp. 729–732.

[26] S. Togashi, M. Yamaguchi and S. Nakagawa, "Summarization of spoken lectures based on linguistic surface and prosodic information," Proc. 2006 IEEE Spoken Language Technology Workshop, 2006, pp. 34-37.

[27] D. Wang and S. Narayanan, "A multi-pass linear fold algorithm for sentence boundary detection using prosodic cues," Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2004), May 2004, pp. 525-528.

[28] C.H. Nakatani and J. Hirschberg, "A corpus-based study of repair cues in spontaneous speech," Journal of the Acoustical Society of America, vol 95,no. 3, pp. 1603–1616, 1994.

[29] M. Izzad, N. Jamil and Z.A. Bakar, "Speech/non-speech detection in Malay language spontaneous speech," Proc. 2013 International Conference on Computing, Management and Telecommunications (COMMANTEL 2013), Jan 2013, pp. 219-224.

.

TABLE II. RESULTS OF RULE-BASED SENTENCE BOUNDARY DETECTION

Candidate no.	Pause duration	Energy	F0	Precede pause	Succeed Pause	Precede sentence	Succeed sentence	Total hits
1	0	0	0	1	0	1	1	3
2	1	0	1	1	1	1	1	6
3	0	0	1	0	1	1	1	4
4	0	0	0	1	0	1	1	3
5	1	0	1	1	1	1	1	6
6	0	0	1	0	1	1	1	4
7	1	0	1	1	1	1	0	5
:	:	:	:	:	:	:	:	:
2,272	1	1	0	0	1	0	1	4