Nursuriati Jamil[1*], Muhammad Izzad Ramli[2], Noraini Seman[3]

J. Electrical Systems 11-3 (2015): 308-318

Regular paper

Journal of Electrical Systems

Sentence boundary detection without speech recognition: A case of an under-resourced language

Sentence boundary detection (SBD), also known as sentence segmentation decides where a sentence begins and ends. Previous method of SBD is either done by linguistic approach or acoustic approach; or combination of both approaches. Even though linguistic approach generally performed better than acoustic approach, it requires the need of a speech recognition component. This is a constraint for Under Resource Languages such as the Malay language. This paper describes the SBD for spontaneous Malay language spoken audio. Experiments are conducted on a forty-two minutes question-answer (Q/A) Malaysia parliamentary session 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. Seven prosodic features, rate-of-speech and volume are then extracted from the boundary candidates for classification. Our proposed SBD method using supervised Adaboost classifier managed a promising100% accuracy rate with 19.44% error rate. For future work, we intend to reduce the error rate by implementing end-point detection on the boundary candidates.

Keywords: Sentence boundary detection; spontaneous speech; prosody features, AdaBoost.

1. Introduction

In most languages, a written sentence is defined as the largest independent unit of grammar; typically begins with a capital letter and ends with a period, question mark or exclamation point. A formal written sentence normally has a subject as well as a predicate. In spoken audio, we define a sentence as a word or group of words that expresses a complete idea. Naturally, it displays recognizable intonation patterns and is often marked by preceding and following pauses [1]. Sentence boundary detection (SBD) task is deemed to be important as it acts as an initial processing part of most natural language processing (NLP) applications. Errors caused during SBD formidably affect the subsequent processes of NLP applications such as speech recognition, topic segmentation and speech summarization.

Generally, SBD is done using linguistic approach or acoustic approach or combination of linguistic-acoustic approach [2]. 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 [2]. Therefore, the speech recognition component needs to be constructed prior to sentence boundary detection. However, speech recognition often takes

[*] Corresponding author: N. Jamil, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia, E-mail: lizajamil@computer.org

[1,2,3] Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia

processing and higher computational costs. For an under-resourced language such as Malay language [3], this requirement poses a problem. Speech recognition in Malay language is still at its infancy stage and recognition is limited to several words only [4]. Thus, linguistic approach is an unlikely option for sentence boundary detection for Malay language at the moment.

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 [5]. Unfortunately, the Malay language is categorized as an under-resourced language [6] due to its limited presence on the web and lack of electronic resources for speech and language processing such as monolingual corpora, bilingual electronic dictionaries, transcribed speech data and pronunciation dictionaries [3]. Speech-related research in Malay language is still at an early stage [4] and sentence boundary detection studies for speech recognition in Malay language are scarce. Several attempts of speech segmentation for spontaneous Malay spoken audio were done [7] [8]. However, they only focused on isolated words not continuous stream of words. Thus, it is empirical that speech-related work in Malay language is been pursued

2. Related work

In [9], SBD systems for text are categorized into two approaches that are rule-based approach and machine learning approach. Similarly, these two approaches are also popular for SBD of spoken audio. Rules are encoded in rule-based SBD according to the acoustic features extracted from the speech segments [10]. 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. Rule-based approach however, is not exhaustive and not robust to conflicts.

Machine learning methods remained the focus of SBD work in recent research work. In speech processing, Hidden Markov Model [11] is the de facto learning methods followed by others such as neural network [13], Bayesian network [14], nearest-neighbour algorithms and decision trees [12]. These methods treats detection as a classification problem and in general performs better compared to rule-based approach. In [15], different classification methods for sentence segmentation of English and Mandarin broadcast news were presented. Among the classifiers tested, boosting-based classifier performed better compared to hidden-event language model, maximum entropy and decision trees. A popular adaptive boosting method known as AdaBoost combines weak-based classifiers to building up a strong classifier. At each iteration of the learning procedure, a new weak learner, h_t is conjured through resampling and reweighting of the previous learners. A different weighting over the training examples is used to give more emphasis to examples that are often misclassified by the preceding weak classifiers. Finally, all the weak learners used in each iteration, t are linearly combined to form the classification function in equation (1).

$$f(x,l) = \sum_{t=1}^{T} \alpha_t h_t(x,l) \tag{1}$$

where α_t is the weight of the weak learner h_t and T is the number of iterations. A lengthy explanation of AdaBoost can be found in [16].

3. 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 [17]. 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 [14], 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 as shown in Figure 1

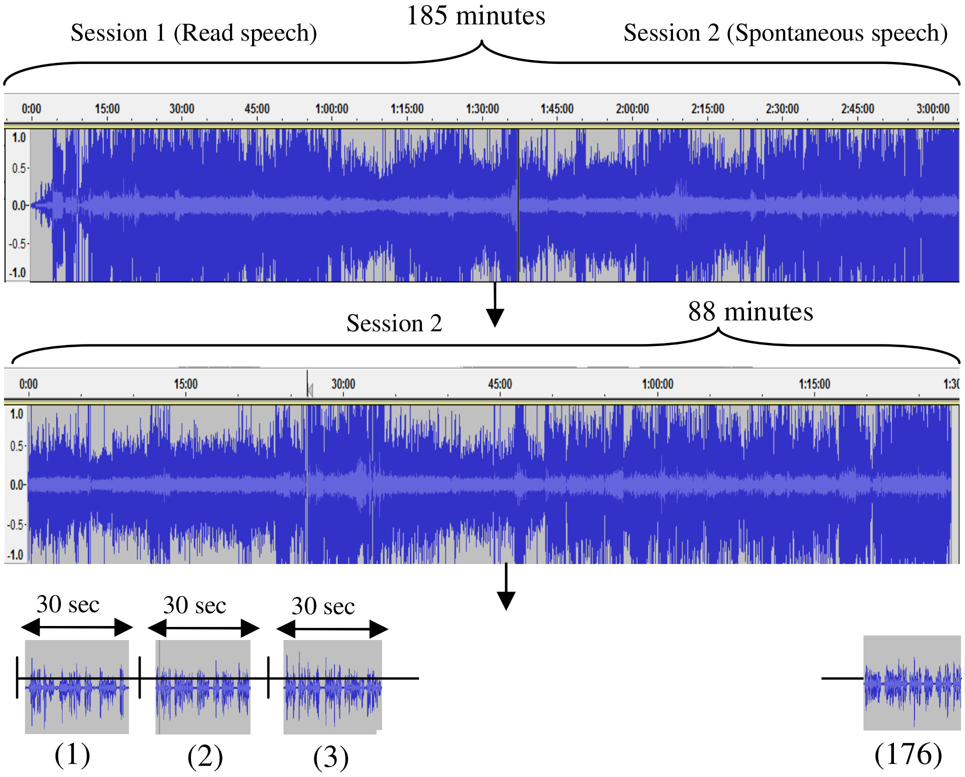


Figure 1: Speech data selection

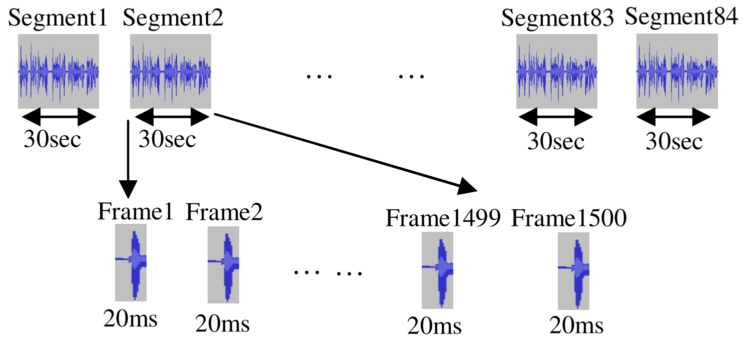
After analyzing the audio data of both sessions, the first session is omitted as it consists of formal speeches with read text prepared before the session. Only the second session of the debate is used as they are from the unplanned questions and answer (*Q/A*) session spontaneously answered during the 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 totalling 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 at least two continuous sentences with minimum total duration of speech of at least 30 seconds. This is important as each speaker has different style of speaking, rate of speech and fundamental frequency. If one speaker dominates a session over the other speaker, his/her speech features will be biased data during training. In the 42-minutes dataset, there are a total of 227 sentence boundaries. From this dataset, eighty percent is used for training of AdaBoost classifiers and twenty percent is used as testing dataset. Figure 1 illustrates the process of speech data selection for this paper.

4. Research methodology

There are four major stages involved in our experiments of sentence boundary detection: 1) audio segmentation 2) speech/non-speech classification 3) boundary candidate feature extraction 4) AdaBoost training 5) speech boundary detection.

3.1. 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 into a total of 126,000 frames. Figure 2 demonstrates 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.



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

$$\text{Total frames} = 1{,}500\text{frames x } 84 \text{ segments} = 126{,}000 \text{ frames}$$

Figure 2: Audio speech segmentation

3.2. 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. The non-speech segments are further used as boundary candidates for sentence boundary detection. The speech segments are regarded as non-boundary candidates, thus is not used for sentence boundary detection. Before the experiment is conducted, a groundtruth dataset is constructed by manually labelling the speech/non-speech segments of the speech datasets using Audacity 1.3.12-beta. A total of 6,413 segments are annotated from 84 segments consisting of 3,206 speech segments and 3,207 non-speech segments. Due to hardware constraint, the 84 segments are further divided into 20 milliseconds, non-overlapping frames totalling 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. 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 [18]. However, we improved the rule by adding fundamental frequency feature and achieved 97.8% accuracy rate as described in our earlier work [19]. The improved classification rules are presented in Figure 3. 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 further used as boundary candidates in sentence boundary detection experiment.

```
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 3: Improved V/C/P classification rules

3.3. Feature extraction of boundary candidates

Table 1 lists 10 audio features consisting of 7 prosodic features, 2 rate-of-speeches (*ROS*) and a volume feature. These features are extracted from the 2,272 boundary candidates for sentence boundary detection.

Table. 1 Description of features used for SBD

No.	Feature type	Features	Description
1.	Prosodic features	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.	Rate-of-Speech (ROS)	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	Volume change rate	Volume change of rate preceding boundary candidate

For sentence boundary detection, we only consider non-speech segments as boundary candidates because 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 the shorter non-speech segments, we are left with 2,272 sentence boundary candidates. The extracted features are illustrated in Figure 4 and descriptions of each feature is depicted in Table 1.

Table. 1 Description of features used for SBD

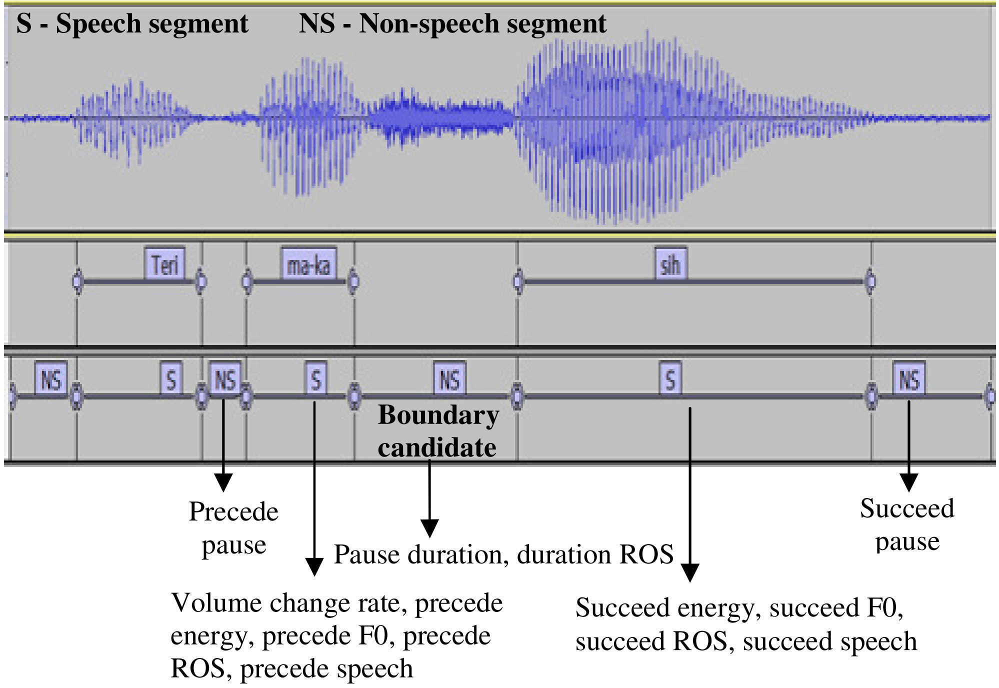

S - Speech segment
NS - Non-speech segment
Teri
ma-ka
sih
NS
S
NS
S
NS
S
NS
Boundary candidate
Precede pause
Pause duration, duration ROS
Succeed pause
Volume change rate, precede energy, precede F0, precede ROS, precede speech
Succeed energy, succeed F0, succeed ROS, succeed speech

Figure 4. Extracted features of a boundary candidate

3.4. Training of AdaBoost classifier

Training is important to build the classification function to detect sentence boundaries. Training dataset comprised 80% of total boundary candidates and the remaining 20% is used as testing dataset. Since there are 2,272 boundary candidates, 1,187 boundary candidates are used in training and 455 boundary candidates are reserved for sentence boundary detection testing data. Furthermore, 207 true sentence boundaries amounting to 80% of the total true sentence boundaries in the whole speech dataset are used for training; while the remaining 20 sentence boundaries (20%) is used for testing.

The training of AdaBoost classifier happened in the steps enlisted below:

Step 1. The ten features extracted from the speech segments earlier and the correct class marker (i.e. 1: true sentence boundary; 0: non-sentence boundary) are fed into the boosting process.

Step 2. Divide data into two sets that is training and control set. Control set is used to evaluate the training set and is selected based on the performance of the previous weak learners.

Step 3. A decision tree is constructed for the weak learner and is used for boosting. Combination of two or more features become weak learner and multiple weak learners can be combined to generate a more accurate ensemble, known as strong learner.

Step 4. The weak learner is boosted using Gentle AdaBoost algorithm to produce *learner* and *weight*.

Step 5. Calculate classifier output based on *learner* and *weight* values.

Step 6. Calculate error by comparing with control set.

A diagram of the AdaBoost classifier is shown in Figure 5.

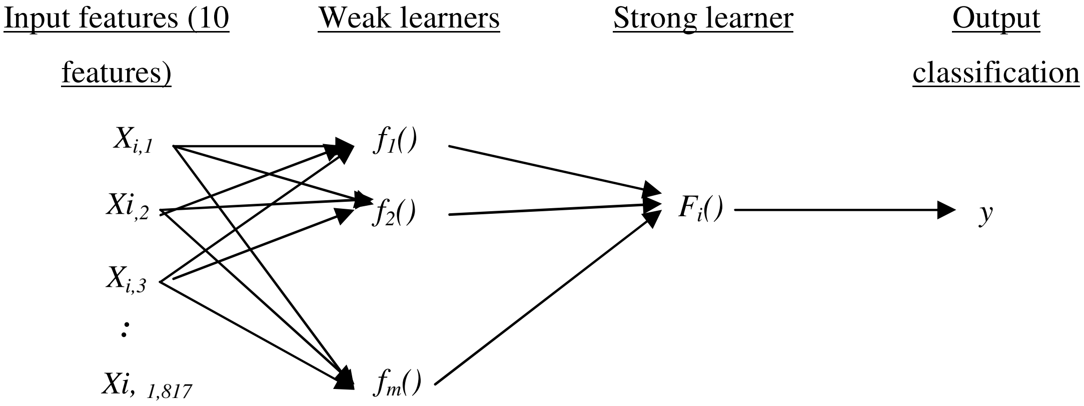


Figure 5. AdaBoost Classifier

In training experiments, true sentence boundaries percentage is increased comparable to non-sentence boundaries to achieve a commendable sentence boundary accuracy rates with an acceptable false error rate. This is done by reducing the number of non-sentence boundaries in the training dataset that has high potential to cause errors in learning. Six experiments are conducted to identify the number of training datasets that can built a classification function producing the highest sentence boundary detection with low false alert. Summary of the training experiments listing the number of true and non-sentence boundaries is shown in Table 2.

Table. 2 Data for training experiments

Experiment	No. of true boundaries	No. of non-boundaries	Total training data
Experiment 1	207	623	830
Experiment 2	207	525	732
Experiment 3	207	451	658
Experiment 4	207	345	552
Experiment 5	207	251	458
Experiment 6	207	233	440

Based on the results of the training experiments, AdaBoost classification model is constructed and further used for speech boundary detection.

3.5. Speech boundary detection

As indicated earlier, twenty percent of the total boundary candidates (i.e. 2,272) is used for speech boundary detection. 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 label [SB] based on the symbol period ' . ' and question mark, ' ? '. There are a total of 227 sentence boundaries in the speech dataset comprising 84 segments. The groundtruth is later used to evaluate the sentence boundary detection's performance.

4. Performance evaluation

Performance of sentence boundary detection is evaluated using accuracy rate as can be seen in equation (2), while total error rate (i.e. equation (3)) is calculated as a sum of false alert of equation (4) and missing alert of equation (5) [15].

$$Accuracy = \frac{Total\ correct\ sentence\ boundary}{Total\ sentence\ boundary} \tag{2}$$

$$False\ Alert = \frac{False\ detection\ \ of\ \mathrm{sentence\ boundary}}{Total\ sentence\ boundary\ candidates} \tag{3}$$

$$Missing\ Alert = \frac{Missing\ \mathrm{sentence\ boundary}}{Total\ sentence\ boundary\ candidates} \tag{4}$$

$$Total\ Error = \frac{False\ detection + Missing\ detection}{Total\ sentence\ boundary\ candidates} \tag{5}$$

5. Results and discussions

Results are presented in two sections: 1) AdaBoost classification model and 2) speech boundary detection.

5.1. AdaBoost classification model

The main purpose of conducting six experiments during the training of AdaBoost classifier is to construct an AdaBoost model that is able to detect sentence boundary at a low error rate. In Table 3, the accuracy rates of the six training experiments are tabulated. The highest accuracy rate of sentence boundary detection at 86.34% is achieved by Experiment 6 using 440 training dataset comprising 207 true sentence boundaries and 233 non-sentence boundaries. However, it produced the highest total error rate of 22.95%. The lowest accuracy rate at 41.85% is attained by Experiment 1 with a high missing alert of 58.15%. In this study, Experiment 3 using 658 training data is considered as the best AdaBoost classification model because of its lowest total error rate of 18.24%. Furthermore, the accuracy rate of 62.11% is an acceptable sentence boundary detection of a learning model. Therefore, AdaBoost classification model using 207 true sentence boundaries and 451 non-sentence boundaries is further tested for sentence boundary detection.

Table. 3 AdaBoost classifier training results

Experiment No.	No. of training dataset	Total error (%)	False alert (%)	Missing alert (%)	Accuracy rate (%)
1	830	19	3.13	58.15	41.85
2	732	20.08	4.51	44.5	55.5
3	658	18.24	5.16	37.88	62.11
4	552	20.29	7.60	30.83	69.16
5	458	18.34	6.98	22.90	77.09
6	440	22.95	15.90	13.65	86.34

5.2. Sentence boundary detection using AdaBoost classifier

The sentence boundary detection test is conducted on 455 boundary candidates and an example of the test on one boundary candidate is shown in Figure 6. The upper plot shows the 30-seconds speech signal and the lower plot illustrates the sentence boundary labelling.

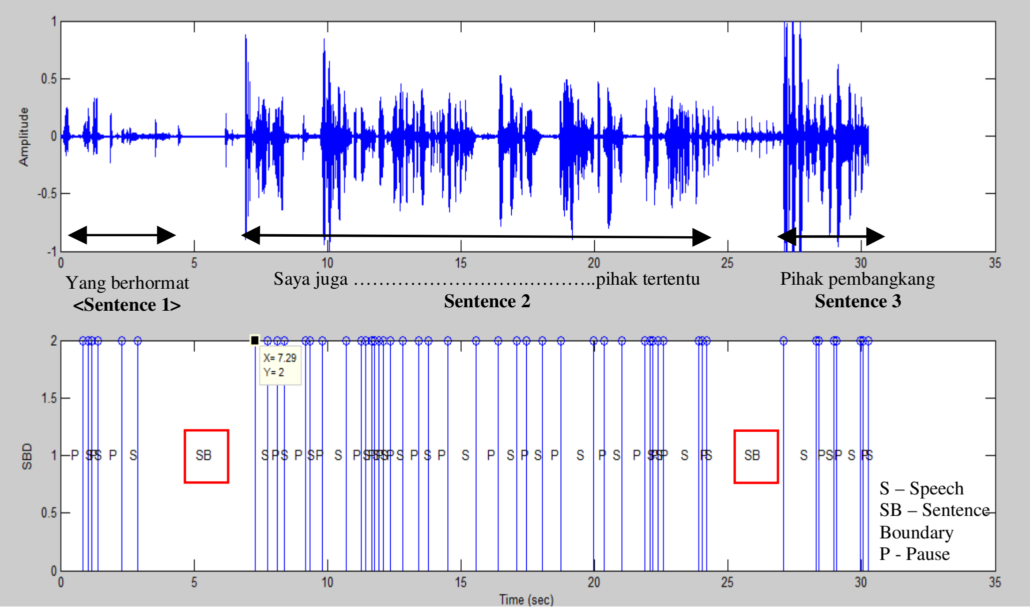


Figure 6. Labelled sentence boundary

Overall results of sentence boundary detection are tabulated in Table 4. The fusion of different features are done to investigate the effects on boundary detection's performance. Prosodic features alone is able to achieve 90.3% accuracy rate of sentence boundary detection with missing alert of 9.25% and false alert of 19.09%. When prosodic features are combined with other features such as volume and rate-of-speech, a 100% accuracy rate is attained and the total error is reduced by 2.23%. However, in both feature fusions whenever the accuracy of sentence boundary detection increases, the false alert also increases. Missing alert decreases if there is high rate of sentence boundary detection. It is because when the detection of sentence boundary is high, the number of missing sentence boundary becomes low. The accuracy of sentence boundary detection depends on the total error of sentence boundary detection. In Table 4, the total error rate is shown to be further reduced while maintaining an accuracy rate of 100% when all prosodic features, rate-of-speech and volume are combined.

Table. 4 Sentence boundary detection using AdaBoost Classifier

Features	Total error (%)	False alert (%)	Missing alert (%)	Accuracy rate (%)
Prosodic	23.86%	19.09%	9.25%	90.3%
Prosodic + ROS	21.63%	21.63.%	0%	100%
Prosodic + Volume	21.63%	21.63%	0%	100%
Prosodic + ROS + Volume	19.44%	19.44%	0%	100%

6. Conclusion

This paper presents sentence boundary detection without the need of speech recognition linguistic component. Even though advanced method of speech boundary detection incorporate linguistic component, we managed to recall all sentence boundaries. However, precision of the detection still need to be improved. The introduction of volume change rate as one of the prosody feature seemed appropriate as Malay language has some unique properties of unvoiced segment. Our research direction is to reduce the error rate by looking into possibilities of end-point detection.

Acknowledgment

This work was financially supported by Research Entitive Iniatitive (REI) Grant, Universiti Teknologi MARA, Shah Alam, Malaysia, no. 600-RMI/DANA 5/3/REI (3/2014).

References

[1] Dictionary.Com http://dictionary.reference.com/

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

[3] 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, January 2014.

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

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

[6] S. Krauwer, The basic language resource kit (BLARK) as the first milestone for the language resources roadmap, 2003 International Workshop Speech and Computer SPECOM-2003, Moscow, Russia, pp. 8–15.

[7] N. Seman, N. Jamil and R. Hamzah, Dynamic Connection Strategies (DyConS) for spoken Malay speech recognition, 2013 IEEE International Symposium Signal Processing and Information Technology, pp. 40-45.

[8] 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.

[9] F. D. Wong, L. S. Chao and X. Zeng, iSentenizer-: Multilingual sentence boundary detection model, The Scientific World Journal, vol. 2014, pp. 1-10, 2014.

[10] N. Jamil, M. Izzad, Z.A. Bakar and N. Seman, Prosody-based sentence boundary detection of spontaneous speech, 2014 Fifth International Conference on Intelligent Systems, Modelling and Simulation, 27-29 January 2014, pp. 311-317.

[11] A. Mikheev, Periods, capitalized words, etc., Computational Linguistics, vol. 28, no. 3, pp. 289–318, 2002.

[12] D.D. Palmer and M.A. Hearst, Adaptive multilingual sentence boundary disambiguation, Computational Linguistics, vol. 23, no. 2, pp. 240–267, 1997.

[13] N. Seman, Z.A. Bakar and N.Jamil, Improving speech recognizer using neuro-genetic weights connection strategy for spoken query information retrieval, LNCS 8281 Information Retrieval Technology, 2013, pp. 528-539.

[14] S. Yildirim and S. Narayanan, Automatic detection of disfluency boundaries in spontaneous speech of children using audio–visual information, IEEE Transactions on Audio, Speech, and Language Processing,vol.17, no.1, pp.2-12, Jan. 2009.

[15] S. Cuendet, D. Hakkani-Tur, G. Tur, Model adaptation for sentence segmentation from speech, 2006 IEEE Spoken Language Technology Workshop, pp.102-105, 10-13 Dec. 2006.

[16] R.E. Schapire, Explaining Adaboost, in Empirical Inference, B. Schölkopf, Z. Luo and V. Vovk, Eds. Spriger Berlin Heidelberg, 2013, pp. 37-52.

[17] 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), pp. 1628-1635, Jun 2010.

[18] 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.

[19] 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), pp. 219-224, Jan 2013.