International Journal of Engineering & Technology, 7 (3.15) (2018) 137-140

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Instruction Task for Malay Phrase Boundary Annotation

Haslizatul Mohamed Hanum[1]*, Nur Atiqah Sia Abdullah[1], Zainab Abu Bakar[2]

[1]Faculty of Computer and Mathematical Sciences,
Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia
[2]Al-Madinah International University, Shah Alam, Selangor, Malaysia
**Corresponding author E-mail: haslizatul@salam.uitm.edu.my*

Abstract

The paper presents a refined instruction task to assist evaluation of prosodic phrase (PPh) boundaries by naive listeners. The results from the perceptual experiments were compared to the boundaries produced by online automatic tagger. The Kappa evaluation shows the average of 85% on inter-rater agreement. More than 60% of the boundaries which are detected by the automatic tagger matched the reference boundaries, showing that the refined instruction task can be used to evaluate perception on phrase boundaries on continuous speech.

Keywords: *prosody; prosodic phrase; agreement; speech boundaries.*

1. Introduction

Speech utterances are subdivided into prosodic segments that assist discourse comprehension. However, the definition of prosodic phrase (PPh) boundaries and how they are identified depends on the language construct. “Malay has major and minor prosodic structures corresponding to those of English, and while there are significant differences in the detail, these structures are broadly similar at the phonetic level” [1]. Patterns of falling pitch was observed on major prosodic boundaries of Malay, while a sustained contour with perhaps a slight rise in pitch was observed for minor structures [1]. However, those observation was done with a goal to detect turn-taking, thus the major boundaries have role in signaling the completion of turn-taking unit of a speaker only at the end segment of the sentence. Similar approaches for audio sentence presented in [2, 3] have been proposed to improve the automatic detection.

However, PPh segment is usually defined as a group of words that carries the intended speech meaning. In English, the boundaries are defined in a hierarchy of major, intonational (IP) which mark a phrase that carry meaning and minor, intermediate (ip) boundaries to mark smaller size phrase within an IP. PPh boundary marks a disjuncture on speech and it is often detected with occurrence of silence, change of pitch rate, or final lengthening. Thus, it is remained unclear how perception of prosody on the PPh boundaries is defined on Malay utterances, other than those other research that focuses at the end of a sentence or word [4].

Perceptual judgement of hierarchy of PPh boundaries are commonly observed from the sets of instructions that guide the listeners on the judging on what they hear. Previous research have shown some degree of consistency among the listeners in their perception of prosody on English [5] and French [6]. Rapid Prosody Transcription (RPT) approach was introduced [7] to allow perceptual experiments conducted by naïve listeners.

Other research uses voiced segments and valleys to predict PPh boundaries. Automatic prosody tagger is flexible enough to support language independent speech signal analysis and detection of prominence and boundaries at the PPh level using a combination of acoustic features, rather than merely F0 contours, as previous empirical and theoretical studies claimed (refer to [8] for review on approaches to automatic audio boundary detection). Acoustic information is derived from voiced/silence segments, which the valleys and peaks are the candidates for a PPh boundary. The automatic tagger [9] is used to identify phrasing boundaries creating a comparative result to this research.

2. The Proposed Method

Prosodic phrase (PPh) boundaries are perceptual judgement, which evaluate on the silence, lengthening, and change of amplitude heard by an individual listener. This approach focuses on how PPh boundaries are perceived on-line by the listeners with language comprehension in mind, as opposed to subtle listening by the experts for annotation purposes. To differentiate on the boundary type observed by the listeners, an instruction task for detecting the PPh boundaries is refined into two-steps listening task that reflects the speech content rather than mere detecting ‘gaps’ between sequence of words. This refinement is proposed in the second step of the boundary marking instruction to guide the listener in differentiating the boundary type. By using this improved method, listener should be able to not just observe the ‘phrase-related gaps’ with a high degree of consistency, but with higher agreement rates on major boundary perception than on minor boundary perception in accordance with previous studies [1].

In the first Instruction Task (IT_1), as the volunteer repeat playing the audio, he or she needs to identify the sentence boundary by tagging the sentence break segment when he or she heard a break, discontinuity or disconnection in the utterance [7]. In the second Instruction Task (IT_2), the volunteers weighted the strength of breaks by labelling the breaks as either minor phrase break, or major phrase break, as they listen through each sentence [10]. Each volunteer marked the identified boundaries as either major (label 1) or minor (label 0) phrase based on the following indicators.

i. When the speaker pause at this position, he/she has conveyed the message intended.
ii. When the speaker pause at this position/word, the current message is just additional to the previous one.
iii. When the speaker pauses at this position/word, he/she still has more to say.

When the listener answers 'Yes' for question (i) the phrase break is classified as major phrase break, then the break is marked as major phrase (indicated by label (1)). When the listener answers 'Yes' to question (ii) or (iii), the boundary is considered a minor break, and then the boundary is marked with a label 0.

3. Research Method

The audio are extracted from the Malay Parliament speech sessions [4], which is a spontaneous Malay speech collection. A graduate assistant was appointed to process and extract audio from the selected video recordings. A graduate assistant was trained to listen to each of the recordings and identify speaker's speech paragraph. The assistant listened to each of the three selected videos, identify the start and end of a speaker's session, and extracted audio from each individual speaker.

Listening and extraction tasks were done using Audacity tool. The audio files are subdivided into speaker-paragraph collections (as individual speaker-audio .wav file with the utterance id, start time, end time and duration as well as word location in the speech) referred as speaker's Audio-Paragraph (named as AP_DATA) collection, is recorded to ease the later tracking process. Each AP_DATA dataset contains speech paragraphs which then, manually transcribed using Praat linguistic tool and its corresponding transcribed text are referred as speaker's Text-Paragraph (named as TP_DATA) collection.

The first 20 paragraphs in the speakers' spontaneous recordings are initially chosen for sampling. Only those audio paragraphs without other speaker's interruptions (by evaluating the overlapping voices) or environment noises (such as laughter, clapping hands etc.) are investigated. Disfluent utterances were excluded since they may contain acoustic disfluencies associated with another domain than the prosodic phrase (see reference that disfluencies). A total of seventy three (73) passages are extracted from the original Parliament 2008's recordings with a total of (more than) 3369 words transcription. The audio are extracted and converted to .wav format and saved as individual .wav files.

Boundaries are marked through three approaches, first the manual annotation using Listening experiment, in which the results are defined as reference boundaries. In addition, the boundaries are detected using automatic boundary detection algorithms. First approach is using acoustic features evaluation, while the second approach is using a silence detector.

3.1. Listening Experiment

Reference boundaries are marked through perceptual evaluation by four native Malay volunteers aged between 22-26 years old. There are 2 male and 2 female volunteers, whom are randomly selected. The volunteers had no training in phonetics and were not shown any visual display of the speech waveform, spectrogram or pitch track.

In order to assist the volunteers, a pre-requisite (training) listening session was conducted for each volunteer. During the training, each volunteer listened to a speech recording and taught on how to differentiate between speech pause, and correlate the speech pause according to the speech content. All the listening experiments were conducted in an air-conditioned private room to minimize outdoor noise that may interrupt the listening session.

Listening experiments were conducted using a self-developed Listening Tool. The tool has two modules that displays the names of the speakers, and play an audio when a speaker is selected. The first module allows the user to select the speaker that they want to listen to. The next module allow listener to select speech paragraphs. Once the paragraph is selected, the corresponding audio file is played. There is a button for playing the audio file, and a button to progress to the next phase. Printouts of the speech transcripts were also presented to each volunteer.

3.2. Speech Corpus

This paper reports the results from single recording as shown in Table 1.

Table 1: The speech collection

Speaker ID	No of passages	No of words	Total duration (minutes)
M01	7	618	6.5
M02	7	400	4.2
F01	6	390	3.4
M04	6	484	3.8

The audio recording from 4 different parliament speakers with multiple topics were extracted, with a total duration of 16.2 minutes are reported in this paper. The recordings were manually subdivided and transcribed into 26 passages, each discussing a topic, containing a total of 1892 words.

3.3. Corpus Annotation

The speech material was manually annotated for perceived prosodic boundaries by two experienced transcribers. Four native speakers of Malay with age between 20 and 26 were chosen as the volunteers and presented with a Listening tool and printouts of the speech transcripts. Printed transcripts of the word content from each passage is provided for each volunteer (later identified as listener), with passages ordered to match the ordering of the sound files they will hear. Words are separated by a space with no punctuation. Each word was either marked with a slash (/) when followed by a boundary, or not marked if it was followed by no boundary. In the second phase, each marked boundary was classified as being followed by a strong or weak boundary. Each strong boundary is labelled with value (/1).

3.4. Automatic Boundary Detection

The speech audio from each speaker is pre-process using open source tool, PRAAT functions. The speech .wav files are converted to mono and re-sampled to 16 kbits (with 50 samples precision), for easier processing. With the window length set to 10 ms and the dynamic range to 50 dB and noise is removed with spectral subtraction using filter range of 200-3000 Hz with 40Hz smoothing technique [4].

To automatically detect the boundaries, the audio passages are run through automatic prosody tagger (known as Praat Web [11]) system, and the resulted boundaries, called PraatWeb boundaries are evaluated, any boundary that marked by any one of the human listener and the system is considered as the true positive boundary (refer to for the confusion metric in Table 2).

Boundaries are also detected using Praat's function. The unvoiced and voiced segments are extracted using the *To TextGrid (silences)* function set with 200Hz for minimum pitch and minimum silence interval duration is set to 0.2s to avoid plosive being considered. The results are called the Praat-SIL boundaries. From the initial observation, minimum pitch of 100Hz gives the best VAD results, in which the value bigger than 100Hz leads to over-segmentation (a number of in-between word-syllable silence region are identified as silence regions). The standard silence threshold of -25 dB is retained.

3.5. Boundary Analysis

For each word on the transcript, the auditory impression from each volunteer is identified and assigned with a prominent score (P-score) and a boundary scores (B-score). First, the labels assigned by multiple volunteers are aggregated to assign each word in the transcript a Boundary Score (B-Score) (draw a figure). The B-

Score assigns the boundary label following each word with one of the boundary types; no boundary (0), minor boundary (1), or major boundary (2).

The distribution of b-score are analysed and discussed, however, the analysis of p-score is beyond the scope of this paper. The first analysis is to evaluate the agreement between different listeners (as in Table 2).

Table 2: Confusion matrix for comparing the results [12]

	system true	system false
reference true	true +	false +
reference false	true -	false -

The aim is to score high on true-positive and compare the ratio to true-negative boundaries. Setting up the results from human perceptual marking as the reference boundary, the other words marked by the system are the automatic boundaries. True means the tool output a mark for the boundary for the word, while false means there is no boundary marked by the system

4. Results and Analysis

Fig.1 shows the reference (tier PPhword) and automatic boundaries (tier PPh) on Praat tool. The blue line at highlighted segment PPh_1 is an example of false-positive boundary, while the line at end of PPh_2 and PPh_3 are the true-positive boundaries. From the automatic tagger, a word that has a boundary mark is given a B_score of 1 while the other words are given B_score of 0. Only marked boundary at the end of word is considered, as the automatic tagger is working on syllable unit, while the naïve human evaluation is limited to the smallest word unit. Thus, to make a feasible comparison, all B_score of 2 in the reference boundaries are re-assigned to 1, indicating a boundary is present.

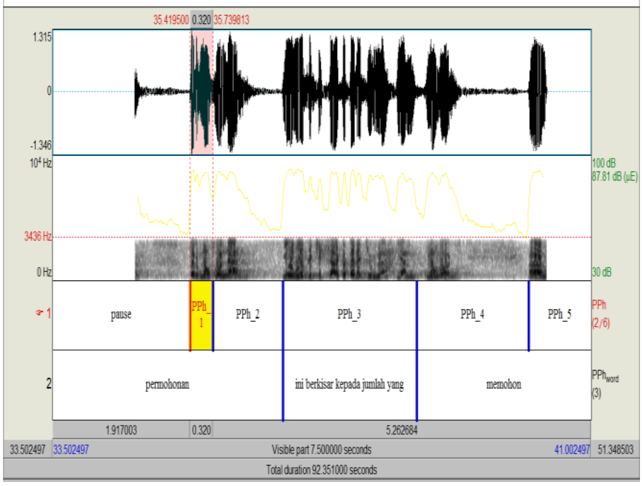

Fig. 1: Example of boundaries marked at M01006 audio fragment

The agreement among the listeners was assessed using a modified form of Cohen's Kappa [13]. From the observations, we found that naïve listeners' pairwise agreement scores on the perception of boundary range from 0.78 (adequate) to 0.82 (very good) with a mean agreement score of 0.815. The result suggests that most of the pairs of the listeners have similarly labels in their judgment of boundary, i.e., there is no significant variation in perception across the listeners.

When the results from automatic tagger is compared to the reference boundaries, the Average Pairwise Percent Agreement using Krippendorff's Alpha metric, the agreement between each listener and the automatic tagger ranges between 83.981% to 92.233%, that gives the mean agreement of 88.35% (0.553). Out of 153 boundaries in the reference collection, the tagger identified 60.8% true-positive boundaries. However, 23.5% of the reference boundaries are true-negative boundaries, and another 32% are missed by the tagger.

An automatic tagger, trained on English [11] is used to automatically identify PPh boundaries, but the results shows a high score on missed boundaries, and even over-segmentation. The problem may rely on evaluation of the acoustic features on each boundary. The aggregate features leads to over-segmentation on Malay utterances as shown in Fig. 2.

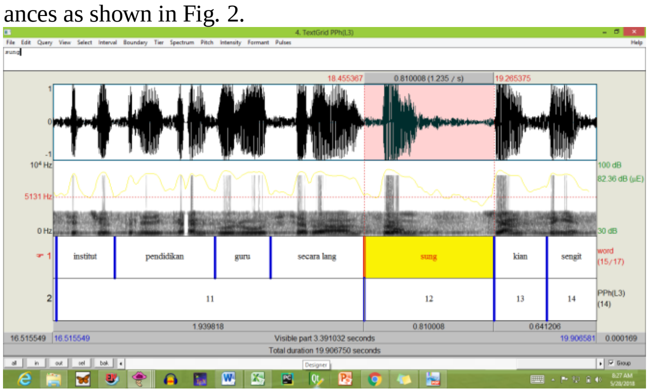

Fig. 2: Example of boundaries marked at M01006 audio fragment

Sample of a PPh(L3) number 12, the boundary is marked in the middle of a word '*lang-sung*' (direct), while at number 13, boundary is set at single word '*kian*' (as much).

5. Conclusion

The experiments have produced a reference to boundary perception from Malay speech with a high agreement score, thus relevant and useful for further language analysis. The result shows that the listeners have high agreement on identifying a word as boundary word.

As a conclusion, instruction tasks are suitable to assist listener identifying the phrase boundaries on spontaneous Malay speech. Results also show that listeners are able to effectively differentiate between major and minor boundaries with the refined instruction task.

Acknowledgement

The Research Is Funded By Universiti Teknologi MARA Under LESTARI Grant 600-IRMI/DANA 5/3/LESTARI (0111/2016).

References

[1] Mohd Don, Z. and Knowles, G., Prosody and turn-taking in malay broadcast interviews, Journal of Pragmatics, vol. 38, pp. 490-512, 2006.

[2] Jamil, N., Ramli, M. I., and Seman, N., Sentence boundary detection without speech recognition: A case of an under-resourced language, Journal of Electrical Systems, vol. 11, 2015.

[3] Izzad, M., Jamil, N., and Bakar, Z. A., Speech/non-speech detection in malay language spontaneous speech, in Computing, Management and Telecommunications (ComManTel), 2013 International Conference on, 2013, pp. 219-224.

[4] Seman, N., "Coalition of genetic algorithms and artificial neural network for isolated spoken malay speech recognition," PhD, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia, 2012.

[5] Cole, J., Mahrt, T., and Hualde, J. I., Listening for sound, listening for meaning: Task effects on prosodic transcription, in Proceedings of Speech Prosody, 2014, pp. 859-863.

[6] Simon, A. C. and Christodoulides, G., Perception of prosodic boundaries by naïve listeners in french, Proceedings of Speech Prosody 2016, 2016.

[7] Mo, Y., Cole, J., and Lee, E. K., Naïve listeners' prominence and boundary perception, in 4th International Conference on Speech Prosody 2008, SP 2008, 2008.

[8] Theodorou, T., Mporas, I., and Fakotakis, N., An overview of automatic audio segmentation, IJ Inf. Technol. Comput. Sci, vol. 1, pp. 1-9, 2014.

[9] Dominguez, M., Farrús, M., and Wanner, L., An automatic prosody tagger for spontaneous speech, in Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, 2016, pp. 377-386.

[10] Krivokapić, J. and Byrd, D., Prosodic boundary strength: An articulatory and perceptual study, Journal of phonetics, vol. 40, pp. 430-442, 2012.

[11] Domínguez Bajo, M., Latorre, I., Farrús, M., Codina-Filbà, J., and Wanner, L., Praat on the web: An upgrade of praat for semi-automatic speech annotation, in 26th International Conference on Computational Linguistics (COLING): System Demonstrations; 2016 Dec 11-17, Osaka, Japan, 2016, pp. 218-222.

[12] Liu, Y. and Shriberg, E., Comparing evaluation metrics for sentence boundary detection, in IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2007, 2007, pp. IV-185-IV-188.

[13] Randolph, J. J., Online kappa calculator [computer software], ed, 2008.