Detection of Malay Phrase Breaks using Energy and Duration

Haslizatul Mohamed Hanum, Zainab Abu Bakar
Universiti Teknologi MARA, Shah Alam
Selangor, Malaysia

Abstract — A simpler approach to identify and classify phrase breaks in prosodic phrasing using energy patterns and duration is useful in speech segmentation. Prosodic phrasing is useful to segment lengthy spontaneous speech into smaller meaningful utterance without analysis of linguistic information. We propose a listening test that allow trained listener to classify the boundaries as minor or major breaks. This cheaper and faster approach is proven useful for under-resource language such as Malay which do not have comprehensive prosodic-annotated corpus. Word-related energy and duration features are extracted from the targeted phrase breaks. The training feature set is developed from evaluation of targeted phrase break from 100 sentences evaluated in the listening test. Evaluation of the features with RBF, MLP and logistics models reveal best detection accuracy of 80.6% which is comparable to existing context-based algorithm. Instead of labeling the phrase break using linguistic and phonetic meaning, the proposed listening test allows labeling of phrase break as perceived by listener. In addition, the results can be use as preliminary information for evaluation of boundary salience at the targeted boundary locations.

Keywords — phrase breaks; speech energy; phrase strength; phrase boundary detection

I. INTRODUCTION

Phrasing and prominence which belongs to intonation class, have improve efficiency of speech understanding applications such as speech retrieval and speech recognition. Phrasing organizes prosodic constituents into a tree of prosodic structure, by viewing constituents; words and phrases, which 'belong together' either from production or perception point of view. The most common practice is to code the hierarchical structures within a sentence as Intonational Phrase (IP) and Phonological or Intermediate Phrase (PPh) respectively following the initial suggestion by [1].

Researcher has focused on the cues on intonation to discover phrasing in speech. Factors, namely the syntax, semantics, and phonological length of sentences, contribute to determination of phrasing [2]. In addition, individual speaking rate affects the number of phrasing invoke on sentences in which rapid speech tend to have less number of phrasing structures. Prosodic phrase often studied as either the way it is perceived by listener, or the way it is produced by speaker. Specifically, when relates to how speech are produced, appropriate use of phrasing suggest for speaker's intention and help to discover prominence in speech [3-5].

In this work, we focused on detection of correct phrasing structure, with a goal of achieving proper interpretation of speaker's meaning, which was proven useful for speech technologies such as automatic speech recognition (ASR) [6-8] and speech synthesis [9, 10]. Using prosodic features like pitch, intensity and durations recurring intonation patterns are detected. In spoken language, especially in spontaneous speech [11], determination of prosodic boundaries assist understanding of an utterance. Wagner [9] described how speakers usage phrasing to convey few types of information such as identifying the purpose of the utterance; whether he is uttering a statement or a question, or identifying emotional states of a speaker; whether he is angry, joy etc.

Recently, phrase detection and labeling task are improved using direct modeling of prosody features as first reviewed in [7], which is less expensive than modeling using intermediate representations. The latter approach often requires availability of corpus with manual annotations of the prosodic events. Instead, prosodic events are automatically labeled using features which are extracted directly from speech signal. Using machine learning techniques, classifier algorithm is constructed to manipulate the features and predict the target classes relevant to the speech technology tasks. Direct modeling approach has produced comparable result when compared to prosody modeling that requires time-consuming and laborious manual annotations, but provide substantial time savings [12].

Research in phrasing strategies can be divided into detection of phrase breaks on using different prosodic units [3, 13]. Phrase break detection help to generate pauses and lengthening at phrase boundaries in text-to-speech (synthesis) systems [14, 15]. Intonational phrase (IP) break detection has also improved speech summarization [16] and recognition systems.

Prosodic strength evaluation (which observed on frequent words in [17]) is also important to discover acoustic and prosodic cues that best signaling existence of phrase boundaries [18, 19]. Although it is relative clear in English that phrase breaks coincides with pause, final lengthening and pitch reset, how the parameters utilized in Malay long utterance is still an open question. Limited literature on perceptual observation of Malay prosody as described in [20] indicated that duration and intensity may have some roles in differentiating stressed and unstressed syllables. The earliest instrumental study by Zuraidah [21] and recent syllable analysis by Wan Aslynn [20, 22] also
suggested that there is no stress in Malay. However, the conclusion was observed true on read speech. Hence, there is a need to validate this claim (in the desertion of pitch features that indicate stress) on more spontaneous style of speech. Other language like English has the accented syllable to indicate stress. Thus, there must be a method for speakers to arrange their phrasing strategies to emphasize speech intention.

For this project, observations are done among the candidate phrase boundaries to identify quantitative distinctions in the acoustic patterns that indicate a difference in boundary strength; major versus minor boundaries. A listening test is constructed to assist identification of phrasing in development of the training set, and evaluation of the classifier performance. This approach better matched existing technique showing how human perceived speech as in [19] which useful for language that do not have lexical and syntactic annotations on the speech transcripts available.

The goal is to discover if there are systematic differences in boundary strength, and if exist, investigation is done to construct the representation of such differences. Roles of the feature vectors (from duration and energy) and their representations from words; positioned at before and after the potential phrase break, are manipulated in several conditions.

The observations and results contribute to identification of phrase break locations throughout speech utterance, and definition the Malay prosodic structure and its characteristics. Then, a classifier is developed for identifying phrasing boundaries used in Malay Parliamentary speech. Classifier is build using energy and duration. Also, a new constructed local maxima features are computed to detect the presence of phrase break. Evaluation is done on classifier performances using individual and combinations of features, by comparing the results with the labels identified from listening test.

II. Method

The speech and non-speech segments from continuous speech can be detected using Voice Activity Detection (VAD) algorithm. The sequence of speech and pause segments are extracted by evaluating the pause duration on each sequence at threshold value of 0.03sec. The lowest valley in the energy sequence is set as possible sentence boundary location. Each spoken sentence is stored in individual .wav files. At each targeted phrase boundary, the spectrograms of the extracted speech files are analyzed using open source tool, PRAAT with the window length set to 10 ms and the dynamic range to 50 dB.

The spectrograms are viewed using the Fourier method with Gaussian window shape. The audio are re-sampled to 16kbits for easier processing. Noise are removed with spectral subtraction using filter range of 200-3000 Hz with 40Hz smoothing technique.

A. Speech corpus

The speech under investigation consisted of continuous speech from a male speaker, delivering speech paragraphs containing 140 sentences. There are between 16-28 words in each sentence. Adapting the methodology described in [23] for phrase break detection task (using syllable and word as prosodic unit), prosodic features such as duration and energy from speech signal and pause duration between words are extracted. The energy features are extracted at a frame-by-frame level using autocorrelation algorithm in addition to the word and pause duration. The features are all normalized by the mean value across the whole sentence. The feature dimensions considered are as follows.

- the duration of a pre-word, post-word (2)
- the silence duration between two words (1)
- the energy min, max, mean values of the pre-word and post-word (6)
- the energy standard deviation values of pre-word and post-word and their ratio values (3)

B. Listening Test

Initial perceptual analysis of overall phrase characteristics on long continuous speech is carried out to detect and label perceived utterance, phrase and word boundaries. The cross-listener parsing of multiple-phrase speech sentences reveals how the speech paragraphs were heard and what the listener look for when listening to continuous speech. In order to test the validity of phrase detection method, we compare the hypothesized phrase with the phrase occurrence as annotated by listener by answering ‘Yes’ or ‘No’ to the following questions.

a) When the speaker pause at this position, he/she has convey the message intended.

b) When the speaker pause at this position/word, the current clause is just additional to the previous one.

c) When the speaker pause at this position/word, he still have more to say.

Two postgraduate students are trained to carry out the listening test. They both listen to each sentence twice and weighted the break strength [24]. Then, they marked the target breaks as either major (label 1) or minor (label 2) phrase. When a listener answer ‘Yes’ for question (a) the phrase break is classified as major phrase break, while answer ‘Yes’ to question (b) and (c) shows a minor break. Statistical analysis shows agreement on breaks labels between listeners is 88% for phrase breaks.
From the corpus collection, 100 spoken sentences were manually transcribed and used for feature vectors observation at targeted boundary segments and developed into train set. Speech spectrograms were annotated with two text grids: the first text grid was for the segmented words together with their orthographic transcription, and the second text grid was for the break label for the word at phrase breaks. We observed that as most of the sentences are long sentences, each sentence has at least one major break, and one or more minor breaks. The parameters are used to train algorithm for detecting phrase boundaries. Lastly, the detection algorithm are tested on remaining speech collection.

C. Phrase Break Detection

The boundary detection task is set as a binary classification, detecting presence or absence of phrase break at targeted locations. Logistic regression, Multilayer Perceptron (MLP) and RBF Network models from Weka version 3.6 are used to detect the presence of phrase breaks. Feature vectors (from duration and energy) and their representations from words; positioned at before and after the potential phrase break, are grouped into 5 tasks as shown in TABLE I. The first and second tasks are carried out to compare effects of syllable and word energy individually. The third task detects phrase break using pause duration and pre-word energy only, while the fourth task combines pause duration, with sequence of pre and post energy features. In the final task energy strength from pre and post words are compared.

### III. RESULTS AND DISCUSSION

TABLE II shows result of classifying 40 sentences as presented in [25] from the three classifiers; RBF, MLP and Logistic models, with best classifier performance of 84.6% achieved using before and after phrase boundary features with MLP. Overall performance of detection accuracy improved from 79.5% to 84.6% and F-score of 0.82 to 0.86 using MLP models. The results suggest that energy and duration features from both sides of (preceding and after) the targeted phrase break offer better characteristics for the phrase breaks.

In this paper, we report that observations of various aggregate energy features, word duration features and pause duration features on all 140 sentences from the same male speaker named ‘men27’, also show an outstanding characteristic of energy from pre-boundary word that shape major phrasing in Malay. MLP classifier achieves the highest performance of 80.6% using before and after phrase boundary features. Even though the result is not yet conclusive, it opens wider possibility to automatically detect phrasing in Malay even without existence of pitch reset. Pitch reset feature is widely used in English phrase boundary detection.

This result is comparable to auto-context algorithm in [23] with 89.0%, although lower by more than 8%, but that might be caused by the limited features (only energy and duration) used in the experiment. Further evaluation can be considered using more robust features. Next task is to conduct some lexical analysis on the sentences to find relation of this findings to the Malay grammar forms which also relevant to validate the listening test output.

### IV. CONCLUSION

This paper presents a technique to detect breaks in prosodic phrasing by equating the combinations of word energy and pause duration features with occurrences of phrase breaks. The training feature set consists of 100 continuous sentences is developed from evaluation of targeted phrase break in listening test. We proposed a listening test that help to classify the phrase breaks as minor or major breaks corresponding to Intonational Phrase (IP) and Phonological or Intermediate Phrase (PPh) respectively. In addition, the listening test allow labeling of phrase break as perceived by listener, instead of labeling the phrase break using linguistic and phonetic definitions.

Speech features are directly extracted from the speech signal, and this faster approach can benefit under-resourced language such as Malay which do not have comprehensive prosodic-annotated corpus. Then, word-related energy and duration features are extracted from the targeted phrase break. Evaluation of the features with RBF, MLP and logistics models reveal best detection accuracy of 80.6% which is comparable to existing context-based algorithm.
This simpler approach for detection of phrase break from acoustic signal can be used to segment lengthy spontaneous speech into smaller meaningful utterance without analysis of linguistic information. In addition, the results can be used as preliminary information for evaluation of boundary salience on targeted boundary locations.

As listener decides on the perceptual phrase breaks in the proposed listening test, the classifier built are able to model the energy variations from words preceding and after the phrase breaks. When combined with word and phrase durations, the classifiers perform better.

The results of this work suggest that the energy and duration of spontaneous speech vary systematically with Malay phrase structure and can be use appropriately for phrase break detection. Automatic detection and classification of intonational events can yield segmentation results which is more suitable for speech retrieval and understanding.

ACKNOWLEDGMENT

We would like to thank the reviewers for their valuable comments and suggestions to improve this paper.

REFERENCES


DOI 10.5013/IJSSST.a.17.32.26 26.4 ISSN: 1473-804x online, 1473-8031 print