Speech Prosody 2016
31 May - 3 Jun 2106, Boston, USA

Sentence Segmentation and Phrase Strength Estimation in Malay Continuous Speech

Haslizatul Mohamed Hanum, Zainab Abu Bakar

Universiti Teknologi MARA Shah Alam, Malaysia
haslizatul@salam.uitm.edu.my

Abstract

Continuous speech sentences are delivered in several shorter **phrasing segments** which can be considered as units of information. The paper proposes a technique to improve intonational speech (IP) segmentation into the normal-strong-normal structure. The segmentation process is carried out in two phases. First, each sentence is segmented into arbitrary segments by evaluating the pause duration. Then, phrase strength is estimated using repeated pitch and intensity patterns on each phrasing segments. Phrase strength defines how strong is the pitch or intensity at that particular phrasing segment compared to the adjoining segment. This technique equates the occurrence of local maximum on pitch and intensity contour with occurrences of phrases from Malay continuous speech sentences. The result of this study suggests that the intensity contour on Malay continuous speech vary systematically with the phrase structure. This finding is useful for identifying the phrase segments that a speaker emphasized in content-based classification and retrieval of speech recordings.

Index Terms: phrasing segment, phrase strength, speech segmentation, prosodic phrasing, Malay continuous speech

1. Introduction

Many speeches recordings are available on the Web and can be retrieved through Multimedia Retrieval Application. While the existing speech retrieval application is intelligible, the speech output still lacking appropriate structure indicating the information content. Current speech retrieval applications limit the output to speech recordings which have been time-word aligned or manually labeled with speaker name or topics. Hence, the users still need to listen to the whole recordings and decide if the whole recording is relevant or only small segment of the recording is what they want. Listening to a lengthy recording can be a stressful experience, particularly if the structure (and content) is not known beforehand, and there is no option for the user to skip any segment of the recording. In order to retrieve pertinent segments from speech recording, it is essential to study a technique to manipulate the speech segmentation into the phrasing structure. We propose to segment a speech sentence into a normal-strong-normal structure which can assists a user to choose recording segments that they want to listen to. Detection of correct phrasing structure has proven useful for speech technologies such as automatic speech recognition (ASR) [1, 2] that assists understanding of an utterance and a proper interpretation of speaker's speech intention.

Speech phrasing facilitates speakers to convey their speech intent through meaningful chunking or phrasing of the sentence. Words which 'belong together' from a meaning point of view, are grouped as *prosodic phrasing*. Prosodic phrasing is often studied as either the way it is perceived by the listener or the way it is produced by the speaker. When it relates to listener perception, detection of correct phrasing structure helps a listener to interpret speech content. On the other hand, when analyzing individual speaker's delivery style, prosodic phrasing suggests for speaker's intention; whether he or she is uttering a statement or a question, or identifying emotional states of a speaker; whether he is expressing anger, joy etc. [3].

In this work, we focus on detection of phrasing structure on a sentence to show which part of the sentence is emphasized by examining the speaker's voice. We identify speech contour directly from speech signal which is segmented into fixed-sized frames, and each speech frame is represented by the corresponding sequence of prosodic feature vectors. In order to perceive the phrasing contour, we integrate each pitch and intensity feature separately and approximate the prosody contour from sequences of three consecutive frames. The pitch and intensity contour is estimated by grouping sequences of the three-frames prosodic features that behave similarly, that consequently subdivide a sentence into smaller phrasing segments. We assume that resulting shorter length contours (the segment contour) may reflect the prosodic intonational phrases (IP). Then, the phrase strength [4] is estimated from the contours and segments are classified into a normal-strong-normal structure. In previous research, the task requires a linguistic expert to annotate the tone as well as the segment boundaries, on manual or ASR-based speech transcript. However, annotation by trained linguists for a huge collection of speech recording is costly for under-resourced language like the Malay language.

1.1. Related Research

Speech processing approaches use supervised and unsupervised techniques to detect repeating acoustic patterns in speech or audio document. Methods for detecting repeating patterns are originally invented for music retrieval and summarization. The treatment of a time-series as a sample from a linear dynamical system is also known as a *dynamic texture (DT)* in the computer vision literature, where a video sequences is modeled as a sequence of vectorized image frames. The dynamic texture model has also been successfully applied to various computer vision problems, including video texture synthesis, video recognition, and motion segmentation [5]. The DT as a generic model can also be applied to any time-series data, such as sequences of feature vectors that represent fragments of musical audio [6]. For music, similar features are used to model simultaneously the instantaneous audio content (e.g., the instrumentation and timbre) and the melodic and rhythmic content (e.g., guitar riff, drum patterns, and tempo). DT was applied to the task of song segmentation (i.e., automatically dividing a song into coherent segments that human listeners would label as verse, chorus, bridge,

doi: 10.21437/SpeechProsody.2016-239

etc.), by modeling audio fragments from a song as samples from a DT model.

Recently, in spontaneous speech, automatic determination of prosodic phrasing assist understanding of an utterance as in [9, 10] and [11, 12]. Phrase detection using direct modeling of prosody features as described in [2] is less expensive than modeling using intermediate representations. The direct approach often requires minimum or no manual annotations of the speech. This is because prosodic features are extracted directly from the speech signal and using learning techniques, the algorithm is constructed to manipulate the features and predicting the target classes relevant to the speech technology tasks. Using prosodic features like pitch, intensity and durations, recurring intonation patterns are detected. This approach has produced a comparable result when compared to prosody modeling that requires time-consuming and laborious manual annotations.

Speech retrieval application requires annotation of speech excerpt that describes the speech content. Frequent words are often indicate importance in topic-based text or speech analysis. However, recent research which ”examined the lexical correlates of importance” where they find ”less frequent words tend to have higher average per-word importance” [7]. In addition, recent finding related to Malay speech suggests that frequent words do not convey significant prosodic features that indicate importance [8]. Thus, in this paper, would like to explore a technique to identify emphasized segment on speech recording. We identify strong phrasing segments, by evaluating prosodic features without any knowledge of the corresponding word. The result from this experiment may be useful for retrieving smaller speech segments from a lengthy collection of the Malay speech recording.

2. Method

2.1. Speech Corpus

The speech data is continuous speech recorded from Malaysian Parliamentary speech of the year 2008. There are 200 audio sentences manually extracted from the Parliamentary speech belong to two male and three female speakers. There are on average 16-28 words in each sentence with a total of 4142 words. We adapt the methodology described in [13] for phrase break detection task and extract sequences of pitch, intensity from speech frames, and pauses durations between words. We use the pitch and intensity features as separate feature to estimate the phrase contour and pause duration to segment the sentence into shorter phrase segments.

2.2. Segmentation

The segmentation is performed in several steps:

- select the audio .wav file.
- background noise is removed by setting up the noise time (in seconds) and filter frequency (in Hertz) ranges.
- set silent threshold for each speaker and segment the speech sentences into phrase segments.
- extract pitch and intensity sequences from each phrase segment.
- evaluate each phrase segments by:
  - compute phrase strength.
  - display speech segment contour.
- playback the audio segments.

2.3. Phrase Strength

We incorporate new phrase strength computed from sequences of three-framed features of pitch and intensity from the speech signal. We assume the local maxima in the pitch or intensity contour will show us the occurrences of phrasing segments on speech sentences. The phrase strength is used to define how strong is the pitch or intensity at that particular frame location compared to the adjoining frames. It is defined as the average intensity difference between the local maxima and its adjacent local minima. Particularly,

$$Phrase\_strength = \frac{1}{2}\left[(M_i - m_j) + (M_i - m_k)\right] \quad (1)$$

where M_i is the intensity value of the local maximum, while m_i and m_k are the intensity values at two adjacent local minima [4].

2.4. Evaluation

In order to test the validity of segmentation method, we compare the hypothesized phrasing segments with the phrase occurrence as annotated by trained listeners. Two female postgraduate students were trained to listen to phrase boundaries. Each listener was instructed to identify how the speech paragraphs were heard, and to identify her perceived understanding of the speech content. For each paragraph, listener first identifies the location of phrase boundary breaks and second classify each of the phrase boundaries as major or minor breaks. In other words, we try to extract some of the perceptual features listeners used to understand speech content.

3. Results

A program that automatically segment spoken sentences into phrase segments is written by integrating the Praat as a linguistic tool to extract prosodic features and Audacity for batch signal processing. The program is implemented in Matlab for the comfortable extraction and evaluation of speech features. The program then tested for estimation of phrase strength among each of the phrasing segments.

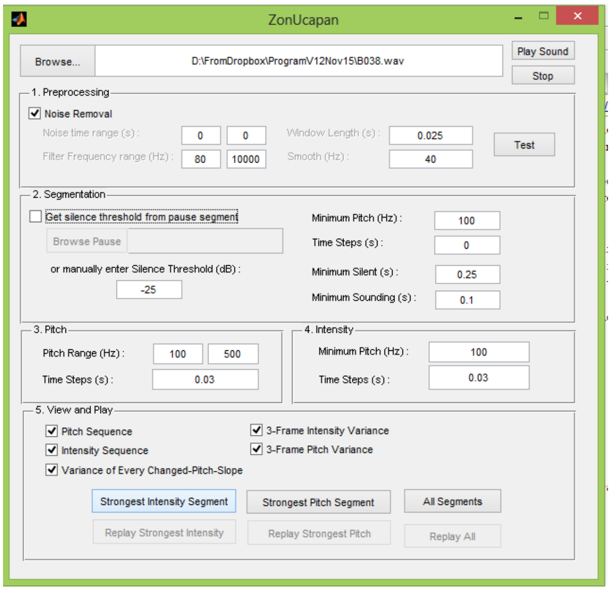
The tool *zonUcapan* (as shown in Figure 1) runs in four stages. It starts with the preprocessing and segmentation of speech audio. Then, feature extraction for both pitch and intensity, and finally, computation of phrase strength.

The tool incorporates following processes:

- select the audio .wav file
- option of computed or user-determined silent threshold
- assisted user interaction to:
  - set pitch range (Hz) and time steps (s) for pitch feature
  - set minimum pitch (Hz) and time steps (s) for intensity feature
  - choose option on types of frame size and sequences
- visualize and save phrase segments

It also assists user to view the pitch and intensity contours detected from each speech sentence.

There is also an option to either listen to or view the strongest phrase segment. An example of strong segment is the



D:\FromDropbox\ProgramV12Nov15\B038.wav

1. Preprocessing

☑ Noise Removal

Noise time range (s) : 0 0

Filter Frequency range (Hz) : 80 10000

Window Length (s) : 0.025

Smooth (Hz) : 40

2. Segmentation

□ Get silence threshold from pause segment

or manually enter Silence Threshold (dB) : -25

Minimum Pitch (Hz) : 100

Time Steps (s) : 0

Minimum Silent (s) : 0.25

Minimum Sounding (s) : 0.1

3. Pitch

Pitch Range (Hz) : 100 500

Time Steps (s) : 0.03

4. Intensity

Minimum Pitch (Hz) : 100

Time Steps (s) : 0.03

5. View and Play

☑ Pitch Sequence

☑ Intensity Sequence

☑ Variance of Every Changed-Pitch-Slope

☑ 3-Frame Intensity Variance

☑ 3-Frame Pitch Variance

Figure 1: *The* zonUcapan *speech segment tool*

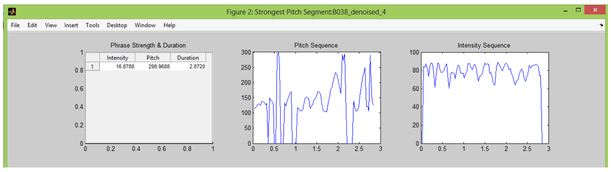


Figure 2: Strongest Pitch Segment:B038_denoised_4

	Intensity	Pitch	Duration
1	16.8788	298.9688	2.8720

Figure 2: *Screenshot of visualized pitch sequences extracted from audio file*

third segment indicated in dark red as in Figure 4 which was estimated from the variances of the the intensity sequences. The segments can be visualized using D3 components and Java programming language.

An example of several strong phrase segment estimated from the sample speaker, 'men27' is shown in Table 1. As an example, file audio A006, both pitch and intensity contours identified segment 8 as the strongest segment. However, for file audio A007, segment 3 is identified as the strongest using intensity contour, but segment 1 is identified as strongest using pitch contour. Overall analysis of 75 sentences from this speaker shows that 75% of the strong segments identified using intensity contour matched the strong segments identified using pitch contour. 67% of the strong segments identified using intensity contour matched the strong segments that our linguist annotated.

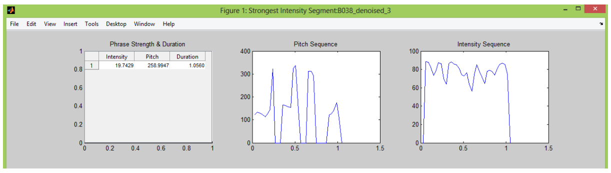
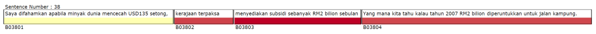


Figure 1: Strongest Intensity Segment:B038_denoised_3

	Intensity	Pitch	Duration
1	19.7429	258.9947	1.0560

Figure 3: *Screenshot of visualized intensity sequences extracted from audio file*



Sentence Number : 38

Saya difahamkan apabila minyak dunia mencecah USD135 setong, | kerajaan terpaksa | menyediakan subsidi sebanyak RM2 bilion sebulan | Yang mana kita tahu kalau tahun 2007 RM2 bilion diperuntukkan untuk jalan kampung.

B03801 | B03802 | B03803 | B03804

Figure 4: *Visualization of phrase strength*

Table 1: *Example of identified strong segments using pitch and intensity contour.*

audio index	intensity	pitch	match
A006	8	8	Yes
A007	3	1	No
A019	1	1	Yes
A037	7	2	No
A060	9	9	Yes
A066	1	1	Yes
A080	2	4	No
A097	7	4	No
A098	1	1	Yes

4. Discussion and Conclusions

We found out that evaluation of intensity feature is more useful for sentence segmentation into phrasing segments compared to using pitch feature. Result from intensity contour estimation discovers the language-specific patterns of intensity distribution which show the prosody organization of Malay language as suggested in [14]. Recent studies also suggested that for Malay speech, duration and intensity may have some roles in the determination of phrase boundaries [8, 15] and Malay speech rhythm in [16]. Previous work also shows that intensity distribution is 14% more efficient in the task of indicating prosodic phrase segments than prosodic word [14].

The uncalibrated contours on strong phrase segments are showing strong variances due to pitch or intensity changes throughout the speech segments. A preliminary test shows that 67% of the strong segments estimated by the process matched the major segments classified manually in listening test, suggests for further inquiry on the segmentation process. This result will also help to discover why the number of segments produced by the automatic process varies from the segments that identified manually. We need to further analyze the difference (if any) the phrase strength estimation on between the different speaker to confirm the pattern. The segmentation also produces very short segments (less than 0.8 seconds) which based on our observations, can be combined with the neighboring segments.

5. Acknowledgements

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

6. References

[1] K. Chen, M. Hasegawa-Johnson, A. Cohen, S. Borys, K. Sung-Suk, J. Cole, et al., *Prosody dependent speech recognition on radio news corpus of American English*, IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, pp. 232-245, 2006.

[2] E. Shriberg and A. Stolcke, *Prosody Modeling for Automatic Speech Recognition and Understanding*, in Mathematical Foundations of Speech and Language Processing. vol. 138, M. Johnson, S. Khudanpur, M. Ostendorf, and R. Rosenfeld, Eds., ed: Springer New York, 2004, pp. 105-114.

[3] A. Wagner, *A comprehensive model of intonation for application in speech synthesis*, PHD, Adam Mickiewicz University, 2008.

[4] E. Cheng and E. Chew, *A Local Maximum Phrase Detection Method for Analyzing Phrasing Strategies in Expressive Performances*, in Mathematics and Computation in Music. vol. 37, T. Klouche and T. Noll, Eds., ed: Springer Berlin Heidelberg, 2009, pp. 347-353.

[5] A. B. Chan and N. Vasconcelos, *Modeling, Clustering, and Segmenting Video with Mixtures of Dynamic Textures*, IEEE Transactions on Speech and Audio Processing, vol. 30, pp. 909-926, 2008.

[6] L. Barrington, A. B. Chan, and G. Lanckriet, *Modeling music as a dynamic texture*, Trans. Audio, Speech and Lang. Proc., vol. 18, pp. 602-612, 2010.

[7] N. G. Ward and K. A. Richart-Ruiz, "Patterns of importance variation in spoken dialog," 14th SigDial, 2013.

[8] N. H. I. Mohd Paudzi, H. Mohamed Hanum, and Z. Abu Bakar, *Evaluation of prosody-related features and word frequency for Malay speeches*, in 2014 International Conference on Computer, Communications, and Control Technology (I4CT), pp. 15-20, 2014.

[9] A. Beke and G. Szaszk, *Unsupervised Clustering of Prosodic Patterns inSpontaneousSpeech*, in Text, Speech and Dialogue. Proc., 15th Int. Conf., TSD 2012, Brno, Czech Republic, September 3-7, ser. LNAI. vol. 7499, P. Sojka, A. Hork, I. Kopeek, and K. Pala, Eds., ed: Berlin, Heidelberg: Springer 2012, pp. 648-655.

[10] A. Beke, G. Szaszak, and V. Varadi, *Automatic phrase segmentation and clustering in spontaneous speech*, in Cognitive Infocommunications (CogInfoCom), 2013 IEEE 4th International Conference on, 2013, pp. 459-462.

[11] P. Bell, T. Burrows, and P. Taylor, *Adaptation of prosodic phrasing models*, Speech Prosody, Dresden, Germany, 2006.

[12] A. Wagner, *Acoustic cues for automatic determination of phrasing*, in Speech Prosody, Chicago, IL, USA, 2010.

[13] J. Zhao, W.-Q. Zhang, H. Yuan, M. Johnson, J. Liu, and S. Xia, *Exploiting contextual information for prosodic event detection using auto-context*, EURASIP Journal on Audio, Speech, and Music Processing, vol. 2013, pp. 1-14, 2013/12/28 2013.

[14] C.-Y. Tseng, S.-H. Pin, Y. Lee, H.-M. Wang, and Y.-C. Chen, *Fluent speech prosody: Framework and modeling*, Speech Communication, vol. 46, pp. 284-309, 2005.

[15] Mohamed Hanum, Haslizatul and Abu Bakar, Zainab, *Evaluation of energy and duration on Malay phrase breaks*, in Asia Modelling Symposium (AMS) 2015, Ninth Asia International Conference on Mathematical Modelling and Computer Simulation, 2015.

[16] Wan Ahmad, Wan Aslynn, *Intrumental Phonetic Study of the Rhythm of Malay*, PhD Thesis, Newcastle University, 2012.