Indonesian Journal of Electrical Engineering and Computer Science
Vol. 15, No. 3, September 2019, pp. 1374~1381
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v15.i3.pp1374-1381

Emotional speech feature selection using end-part segmented energy feature

Noor Aina Zaidan, Md Sah Hj Salam
School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia

Article Info

Article history:

Received Dec 21, 2018
Revised Mar 11, 2019
Accepted Apr 20, 2019

Keywords:

Atir segmentation
Emotion recognition
Energy feature
Gatir segmentation
Segment-based
Utterance-based

ABSTRACT

The accuracy of human emotional detection is crucial in the industry to ensure effective conversations and messages delivery. The process involved in identifying emotions must be carried out properly and using a method that guarantees high level of emotional recognition. Energy feature is said to be a prosodic information encoder and there are still studies on energy use in speech prosody and it motivate us to run an experiment on energy features. We have conducted two sets of studies: 1) whether local or global features that contribute most to emotional recognition and 2) the effect of the end-part segment length towards emotion recognition accuracy using 2 types of segmentation approach. This paper discussed about Absolute Time Intervals at Relative Positions (ATIR) segmentation approach and global ATIR (GATIR) using end-part segmented global energy feature extracted from Berlin Emotional Speech Database (EMO-DB). We observed that global feature contribute more to the emotional recognition and global features that are derived from longer segments give higher recognition accuracy than global feature derived from short segments. The addition of utterance-based feature (GTI) to ATIR segmentation somewhat contributes to increase the accuracy by 5% up to 8% and conclude that GATIR outperformed ATIR segmentation approached in term of its higher recognition rate. The results of this study where almost all the sub-tests provide an increased result proving that global feature derived from longer segment lengths acquire more emotional information and enhance the system performance.

Corresponding Author:

Md Sah Hj Salam,
School of Computing,
Faculty of Engineering,
Universiti Teknologi Malaysia,
81310 Johor Bahru, Johor, Malaysia.
Email: sah@utm.my

1. INTRODUCTION

In natural human conversation, the information conveyed through speech along with speaker's emotion. It is significant to concentrate on how the information is relayed to acquire natural conversation between humans [1]. Emotional intelligence (EI) is the ability to recognize our own and a person's emotional states that we communicate with to guide our though and actions [2]. Human-computer interaction requires the presence of emotional intelligence to establish effective interaction between the human and machine.

The growth of today's high-tech system involves big data focusing on dimensionality and sample size and managing large data in an effective and efficient manner is a major challenge in today's industry. To ensure the efficiency of this high-tech system, feature selection is one of the most important phases that contribute most to the performance of the emotional recognition system. Human speech consists of a combination of sentences, words syllables and phonemes. Emotions can be presented by human in various ways and sometimes they display more than one emotion at a time [3]. Since there is more than one

perceived emotion in one speech, it is difficult to find the dominant emotions because it is not a simple task to determine the boundaries of emotional changes in an utterance. Segmentation approach can contribute in solving this issue where speech can be segmented according to its time interval to determine the dominant emotion.

The search for emotional feature representative set in optimum emotional extraction time scale is an issue that still needs to cater in the Speech Emotion Recognition (SER) field. The appropriate time-scale selection is crucial to produce a high performance SER system. Emotional features can be categorized into two types of time scale: 1) Low Level Descriptor (LLDs) known as local features and 2) Statistical function, known as global feature [4]. Local features define the temporal dynamics in the prosody while statistic value such as minimum, maximum, mean, standard deviation, and slope of the contours highlights the global features [5]. The High-level Statistical Functions (HSF) is used to provide a brief description of temporal variations and contours of the different LLDs during speech [6]. This statistic aggregation function (global) is applied to each of the LLDs (local) for the whole utterance, resulting in a long feature vector.

Until now, the researcher has been still unclear about the feature that brings the biggest emotional information. A commonly used approach is to extract a statistical feature at the utterance level resulting in a series of feature vector, dimension reduction method will be performed on a large dimension of feature vector to compress large data and finally classification algorithm will be performed [7-8].

Automatic segmentation has advantages over manual segmentation where it is possible to obtain high repeatability segments [9]. This occurs when the same method is applied to identify the boundary between two standard segments in different locations for the entire signal. There is several segmentation approaches used since the past decade. Segmentation approach refers to the method of decomposing the speech signal into the fraction of basic phonetic units [10]. Speech signal can be segmented on different level: phonemic, sub phonemic, syllabic, word level, syntagmatic level depending on the segmentation algorithm used. Speech feature can be extracted in both local and global from the utterance-based and segment-based approach.

2. RELATED WORK

Global features have advantages over local features as the feature vector series is smaller. The use of global features in the application of cross-validation and feature selection algorithms makes execution of the process faster and efficient [11]. But then is the global feature extracted from the utterance level the right choice for modelling emotion? This issue became the focus of many current researchers because of the difficulties to use this utterance-wise statistic without being affected by the content of the speech. Zhang et al. [12], state that most current studies often consider utterances (short sentences) as fundamental units and are recognized based on global utterance-wise of the derived local segment, resulting in a single feature vector for each utterance. A lot of information will be neglected if using an utterance-wise feature solely compared to the segment-wise feature that captures information from each segmented utterance. Motivated by those findings, they cultivate a novel approach that used purely segment-level features and completely abandoned the utterance level feature based on Absolute Time Intervals at Relative Positions (ATIR) segmentation approach with the advantages of smaller and fixed number of segments. The result shows significant improvements of more than 20% in the average level of accuracy.

Tzinis and Potamianos [4] run a study on both local and global features and evaluate the performance at various time-scales (frame, phoneme, word or utterance). The result shows that, global statistical feature extracted from speech segment that correspond to the duration of few words yield optimal accuracy using Recurrent Neural Networks (RNNs). On the other hand, Rao et al [5], they conduct a study involving the initial, middle and final positions of each word and syllable to determine whether the speech position affects emotional identification. They claimed that the final position of word in sentences and the final position of syllables in words carried more emotional information than other positions by using local prosodic features.

In this paper, we examine the use of local and global features and the effect of the segments size from the end-part of the speech utterance towards emotion recognition rate using ATIR and GATIR segmentation approach. We have selected the ATIR approach, but with a slight adjustment: we take more segments at the end-part of the speech utterance, inspired from the finding by Rao et al [5]. We observed that global feature contribute more to the emotional recognition as it slightly increase the result by the average 1.7% in ATIR approach and 4.0% in GATIR approach. Apart from that, we found that global features that are derived from longer segments give higher recognition accuracy than global feature derived from short segments as it slightly increase the result by the average 8.5% in ATIR approach and 3.5% in GATIR approach. The results of this study where almost all the sub-tests provide an increased result proving that global feature derived from longer segment lengths acquire more emotional information and enhance the

system performance. The rest of this paper is organized as follows: Section II contains the explanation about emotional speech features and its categories. Segmentation approach is discussed in Section III. Section IV describes about features selection. Section V discusses about the experimental result and analysis followed by the conclusion in the section VI.

3. EMOTIONAL SPEECH FEATURES

Emotion expresses the psychological state of the human. They experience different types of emotional state when dealing with the surrounding events and environment. Cultural differences or human personality cause different emotion [13] and speech production is substantially affected by the different emotional states of the human [14]. The presence of different emotions in human speech can be determined by several parameters in speech signal features such as pitch (fundamental frequency), energy, formant, duration, zero crossing rates, and spectral features like Mel Frequency Coefficient, wavelet and voice quality. The prosody is associated with a speech feature that covers the entire sentence and is used to describe the intonation, rhythm, loudness and stress in the speech structure [15].

Feature extraction is another important issue in speech emotion recognition. Some of the researchers believe that prosodic features carry the emotion information and describe the emotion effectively. This is supported by the review from Koolagudi and Rao [16] where they stated that pitch (also known as fundamental frequency), energy, duration and their derivatives are mainly used as the acoustic correlates of prosodic features. Vocal aspect of speech such as pitch, energy, speaking rate, fundamental frequency has been used in existing emotion recognition systems [17]. In an utterance, pitch represents tonal and rhythmic while energy (intensity) define the pause and accent of the speech signal [18]. Energy is said to be a prosodic information encoder and there is still a study on energy use in speech prosody and it motivate us to run an experiment on energy features.

In this study, the term local feature refers to the energy value for each energy contour while the term global feature referring to the statistical value derived from the local energy feature. The energy of a signal x in a certain window of N samples is given by:

$$En = \sum_{n=1}^{N} x(n) \cdot x^{\star}(n)$$

We focus only on the energy feature for this preliminary study to identify the contribution of position and segment length on the emotional recognition accuracy. In future research, we will integrate other potential prosodic features such as pitch and zero crossing rates.

4. SEGMENTATION APPROACH

Automatic speech segmentation is the process of dividing continuous speech into non-overlapping discrete units [19]. Segmentation of the non-stationary signal require an automatic signal boundaries detection, where the boundaries depends on the statistical characteristic like amplitude and frequency [20]. Basically, the segmentation algorithm can be categorized into those who use phonetic content information (supervised) and those who ignore the phonetic contents information (unsupervised) [21]. Speech signal can be segmented on different level: phonemic, sub phonemic, syllabic, word level, syntagmatic level [9] depending on the segmentation algorithm used. We are motivated by the existing automatic segmentation approach by Schuller and Rigoll [22] for running this study, where their automatic segmentation concept is as follows:

Numbers shown refer to segment-index. GTI: global time intervals (utterance-based feature); **ATI**: absolute time-intervals (segment-based feature); **RTI** relative time intervals (segment-based feature); **GRTI**: combination of GTI and RTI; **ATIR**: absolute time intervals at relative positions (segment-based feature); **GATIR**: combination of GTI and ATIR as shown in Figure 1.

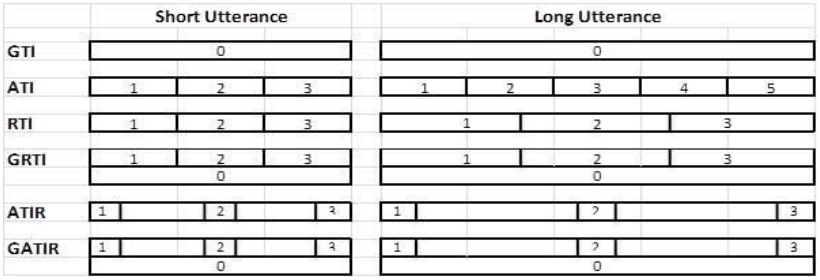

Figure 1. Numbers shown refer to segment-index. GTI: global time intervals (utterance-based feature); **ATI**: absolute time-intervals (segment-based feature); **RTI** relative time intervals (segment-based feature); **GRTI**: combination of GTI and RTI; **ATIR**: absolute time intervals at relative positions (segment-based feature); **GATIR**: combination of GTI and ATIR

1. GTI - Speech feature extracted from the entire speech utterance length
2. ATI - Speech utterances are segmented at the same fixed time interval
3. RTI - Speech utterances are segmented at the fixed relative positions.
4. ATIR - Combination of ATI and RTI. Fixed-length segments are constructed at fixed relative positions, and this overcomes the drawback of different segment lengths and numbers obtained from different utterance lengths.

This study is conducted according to the selected framework for design and procedure phase. Input is obtained from speech signal source, then converted to digital value by A-to D converter, then feature analysis module converts the digital value to a set of data series composed as feature vector and used for various DSP applications to get the recognition results. The information obtained from features extraction technique is data in the form of discrete values representing sound waves, non-sound and noise. This value is normalized and used as input data for the process of recognition. In producing feature vectors that is minimal dimensional yet effective in recognition emotion, we are motivated to run an experiment using ATIR and GATIR segmentation approach since it produces smaller and fixed number of segments. Both local and global features can be extracted using any segmentation approach as illustrated in Figure 2. To avoid data loss and the uses of zero padded for normalization; we choose only global features from utterance-based, and global/local features from segment-based to produce fixed number of feature vector.

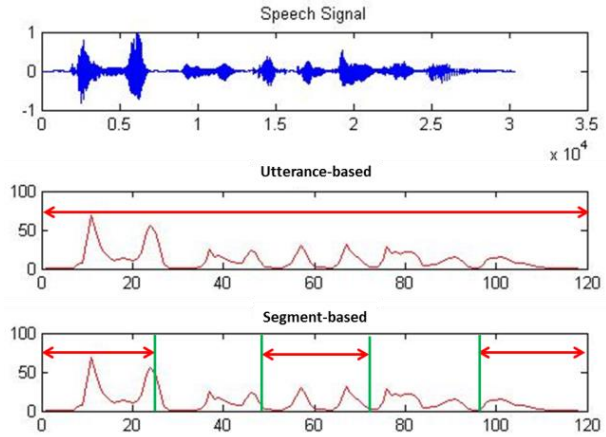

Figure 2. The features extracted using utterance-based and segment-based on energy feature

5. FEATURES SELECTION

Feature selection is an approach in selecting the relevant subset features, and the optimal feature selection methods that enhance the emotion recognition performance is still being studied [23]. Features that are extracted from a speech signal were transformed to a feature vector using the feature selection methods to reduce large sets of features before performing classification. In this paper, we focused on local energy feature and sets of global statistical features derived from it: min, max, mode, median, mean, standard

deviation, variation, skewness, and kurtosis from each segment using ATIR segmentation and compare the result with GATIR. To conduct this study, we use both local and global features to represent the feature vector using ATIR and GATIR segmentation approach. We are investigating the emotion recognition rate based on the three methods:

a) **Method 1**: 10 numbers of local energy feature (per segment)
b) **Method 2**: 9 numbers of global energy feature derived from 10 local features (per segment)
c) **Method 3**: 9 numbers of global energy feature derived from 20 local features (per segment).

All of these methods use one segment at the initial position of utterance, one segment in the middle position of utterance, and the increasing segment number at the final position of utterance as illustrated in Figure 3 below:

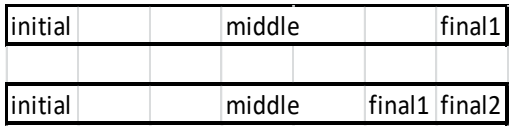

Figure 3. Increasing segment number in end-part of utterance

We have run two sets of tests from three methods above. As for both tests a) and b), we make comparisons between the local and global features, whichever one contributes most to the emotional recognition accuracy. As for both tests b) and c), we investigate about the end-part segment length whether it improves emotional recognition.

6. RESULT AND ANALYSIS

SVM is shown to overtake other classifier technique because of the excellent data-dependent generalization bounds and the global optimal of training algorithm [11]. In situations where the training data should be limited, support vector machine (SVM) can be used because of its good performance of classification compared to other classifier [24]. SVM is one of the most robust and effective as it provide less processing time. SVM provides an easy machine learning algorithm that is well-organized and has been widely used in classification tasks [24-25]. We have chosen Sequential Minimal Optimization SMO algorithm for training a support vector classifier with 10-fold cross-validation, built in Waikato Environment for Knowledge Analysis (WEKA) to run and evaluate this study. We have selected 10-fold cross-validation as a benchmark for comparison with previous researchers as it is widely used for validating general models.

The software we are using for this study is MATLAB and WEKA for the features extraction and classification accordingly. We have collected the empirical data, and the results are analyzed to identify the selected features which most contributed to the enhanced performance of emotional speech recognition system. We are using EMO-DB database as it contains seven emotions: happiness, neutral, sadness, boredom, anger, fear and disgust. Confusion matrix was generated and the summary of the recognition rate presented in Table 1 below:

Table 1. Comparison of Emotion Recognition Accuracy between ATIR and GATIR Approach using Different End-Part Local and Global Segmented Speech Feature.

Approach	Segment	Feature #	Method 1	Method 2	Method 3
ATIR	2+1	27	27.8	29.1	40.3
	2+3	45	34.1	35.5	39.7
	2+5	63	33.4	34.4	44.1
	2+7	81	29.7	32.8	41.9
GATIR	1+2+1	36	33.1	34.1	41.9
	1+2+3	54	38.1	47.5	43.8
	1+2+5	72	37.8	43.8	47.2
	1+2+7	90	38.1	37.8	44.4

Table 1 classifies the results between three methods mention earlier applied to existing segmentation approaches ATIR and GATIR. Segment number 2+1 represents the two segments from initial and middle position + 1 segment from final position. The increasing numbers of segments represent the increasing segment in final position.

ATIR segmentation approach uses the minimum segment by relative position where a lot of emotion information may be omitted. Global statistical features that have been extracted from the whole utterance are considered as one segment. The addition of utterance-based feature (GTI) to ATIR segmentation somewhat contributes to increase the accuracy of recognition rate by 5% up to 8% as it retrieves more potential emotional information. Refer Figure 4 and 5 for the comparison of the result.

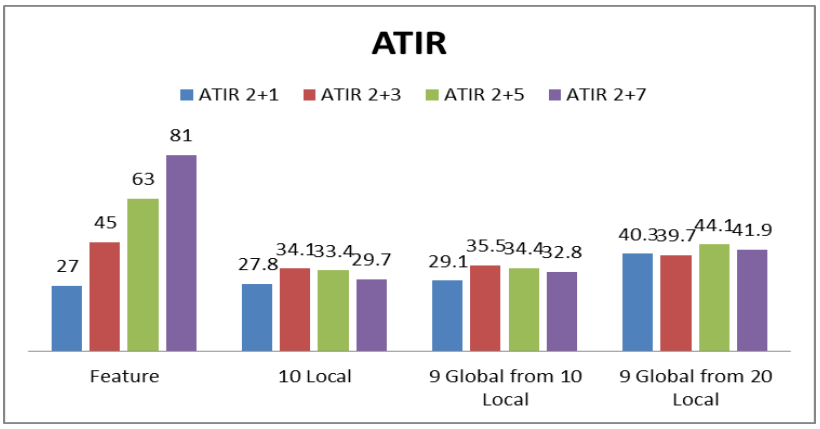

Figure 4. Comparison of the result from three methods for ATIR segmentation approach

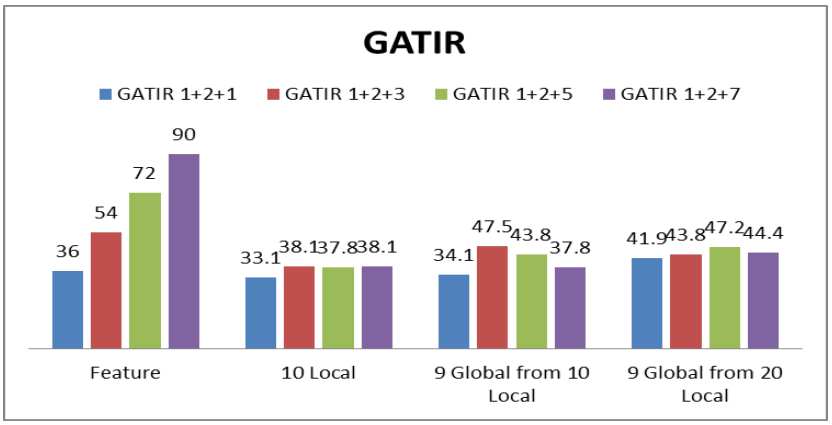

Figure 5. Comparison of the result from three tests for GATIR segmentation approach

We have run two tests to investigate the result for the contribution of local, global features and segment length towards emotional recognition accuracy. From the first tests, we observed that global feature contribute more to the emotional recognition as it slightly increase the result by the average 1.7% in ATIR approach and 4.0% in GATIR approach. The highest result is 47.5% for GATIR with 2 initial and middle segments + 3 final segments, using the global feature. The result shows that the global feature in end-part features did contribute to the performance of the system. This outcome indicates that, the recognition accuracy using global energy features is performed better compared to the accuracy of local energy features.

On the other hand, from the second test, we make comparisons between the end-part segment lengths to determine the highest emotional result. We find that global features that are derived from longer segments give higher recognition accuracy than global feature derived from short segments as it slightly increase the result by the average 8.5% in ATIR approach and 3.5% in GATIR approach. As stated in the result in Table 1, GATIR segmentation approach provides a relatively low average result compared to ATIR approach in the second test. This happens because there is a slight decrease in accuracy in segment 1+2+3

(47.5% to 43.8%). The reason may be that there are some overlapping data as the segment used in method three is longer. The highest result obtained from second test is 47.2% for GATIR approach with 2 initial and middle segments and 5 final segments, using global feature derived from 20 local energy features. The results of this study where almost all the sub-tests provide an increased result proving that global feature derived from longer segment lengths acquire more emotional information and enhance the system performance.

7. CONCLUSION

From the result above, we observed that global feature contribute more to the emotional recognition as it slightly increase the result by the average 1.7% in ATIR approach and 4.0% in GATIR approach. Apart from that, we found that global features that are derived from longer segments give higher recognition accuracy than global feature derived from short segments hence, GATIR was outperformed ATIR segmentation approached in term of its higher accuracy. We believe that no speech information should be truncated as the whole speech data hold the specific emotional characteristic including the unvoiced, silent and pause's signal. Since GATIR used the global feature from the whole utterance, more emotional information has been obtained. Additionally, our idea in getting more features at the end-part features did contribute to the performance of the system. Although large number of segments in the end-part also can enhance system performance, it does not guarantee high accuracy as it possibly will contain redundancy of data. In this work, we only use energy features to see the effect of having more features at the end-part of the speech to see the emotion classification improvement. For future work, additional feature will be taken into account as it is expected to provide improvements in emotional recognition accuracy. We will investigate the integration among prosodic feature categories and study the efficiency of the system performance.

ACKNOWLEDGEMENTS

This research is supported by UTM VicubeLab Research Group at School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia under Research University Grant Q.J130000.2528.14H51.

REFERENCES

[1] R. B. Lanjewar and D. S. Chaudhari, "Speech Emotion Recognition : A Review," *Int. J. Innov. Technol. Explor. Eng.*, vol. 2, no. 4, pp. 68-71, 2013.
[2] A. M. Colman, "A Dictionary of Psychology (3rd ed.)," *Oxford University Press*, 2009. [Online]. Available: http://www.oxfordreference.com/view/10.1093/acref/9780199534067.001.0001/acref-9780199534067. [Accessed: 29-Jan-2015].
[3] M. Argyle, "Bodily Communication," *Routledge; 2 edition*, 1988. .
[4] E. Tzinis and A. Potamianos, "Segment-based Speech Emotion Recognition using Recurrent Neural Networks," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2017, no. September, pp. 190-195.
[5] K. S. Rao, S. G. Koolagudi, and R. R. Vempada, "Emotion Recognition from Speech using Global and Local Prosodic Features," *Int. J. Speech Technol.*, vol. 16, no. 2, pp. 143-160, 2013.
[6] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic Speech Emotion Recognition using Recurrent Neural Networks with Local Attention," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2017, pp. 2227-2231.
[7] B. Schuller, D. Arsic, F. Wallhoff, and G. Rigoll, "Emotion Recognition in the Noise Applying Large Acoustic Feature Sets," *Speech Prosody, Dresden*, pp. 276-289, 2006.
[8] A. Álvarez *et al.*, "Feature Subset Selection based on Evolutionary Algorithms for Automatic Emotion Recognition in Spoken Spanish and Standard Basque Language," vol. 3206, no. September, 2006, pp. 565-572.
[9] Y. Amirgaliyev, M. Hahn, T. Mussabayev, and O. Access, "The Speech Signal Segmentation Algorithm using Pitch Synchronous Analysis," pp. 1-8, 2017.
[10] M. Kalamani, S. Valarmathy, S. Anitha, and R. Mohan, "Review of Speech Segmentation Algorithms for Speech Recognition," vol. 3, no. 11, pp. 1572-1574, 2014.
[11] M. El Ayadi, M. Kamel, and F. Karray, "Survey on Speech Emotion Recognition: Features, Classification Schemes and Databases," *Pattern Recognit.*, vol. 44, no. 3, pp. 572-587, Mar. 2011.
[12] H. Zhang, S. Warisawa, and I. Yamada, "An Approach for Emotion Recognition using Purely Segment-Level Acoustic Features," *Keer2014, Int. Conf. Kansei Eng. Emot. Res.*, 2014.
[13] C. N. Scollon, E. Diener, S. Oishi, and R. Biswas-Diener, "Emotions Across Cultures and Methods," *J. Cross. Cult. Psychol.*, vol. 35, no. 3, pp. 304-326, May 2004.
[14] A. Iliev, "Emotion Recognition using Glottal and Prosodic Features," University of Miami, 2009.
[15] S. A. Ali, S. Zehra, M. Khan, and F. Wahab, "Development and Analysis of Speech Emotion Corpus using Prosodic Features for Cross Linguistics," *Int. J. Sci. Eng. Res.*, vol. 4, no. 1, pp. 1-8, 2013.
[16] S. G. Koolagudi and K. S. Rao, "Emotion Recognition From Speech : A Review," no. July 2011, pp. 99-117, 2012.

[17] M. Sezgin, B. Gunsel, and G. Kurt, "Perceptual Audio Features for Emotion Detection," *EURASIP J. Audio, Speech, Music Process.*, vol. 2012, no. 1, p. 16, 2012.
[18] K. Wang, N. An, B. N. Li, Y. Zhang, and L. Li, "Speech Emotion Recognition using Fourier Parameters," *IEEE Trans. Affect. Comput.*, vol. 6, no. 1, pp. 69-75, 2015.
[19] O. Scharenborg, V. Wan, and M. Ernestus, "Unsupervised Speech Segmentation: An Analysis of the Hypothesized Phone Boundaries," *J. Acoust. Soc. Am.*, vol. 127, no. 2, pp. 1084-1095, 2010.
[20] H. Azami, K. Mohammadi, and H. Hassanpour, "An Improved Signal Segmentation Method using Genetic Algorithm," *Int. J. Comput. Appl.*, vol. 29, no. 8, pp. 5-9, 2011.
[21] R. Makowski and R. Hossa, "Automatic Speech Signal Segmentation based on the Innovation Adaptive Filter," *Int. J. Appl. Math. Comput. Sci.*, vol. 24, no. 2, pp. 259-270, 2014.
[22] B. Schuller and G. Rigoll, "Timing Levels in Segment-based Speech Emotion Recognition," in *Proc. INTERSPEECH 2006, Proc. Int. Conf. on Spoken Language Processing ICSLP*, 2006, pp. 1818-1821.
[23] M. K. Safdarkhani, S. P. Mojaver, S. Atieghechi, and M. S. Riahi, "Emotion Recognition of Speech using ANN and GMM," *Aust. J. Basic Appl. Sci.*, vol. 6, no. 9, pp. 45-57, 2012.
[24] P. Shen, Z. Changjun, and X. Chen, "Automatic Speech Emotion Recognition using Support Vector Machine," *Int. Conf. Electron. Mech. Eng. Inf. Technol.*, pp. 621-625, Aug. 2011.
[25] B. A. Ingale and D. . Chaudhari, "Speech Emotion Recognition using Hidden Markov Model and Support Vector Machine," *Int. J. Adv. Eng. Res. Stud.*, vol. 1, no. 3, pp. 316-318, 2012.

BIOGRAPHIES OF AUTHORS

Noor Aina binti Zaidan obtained her bachelor degree in Computer Science (Graphic Multimedia) from Universiti Teknologi Malaysia, Malaysia, Johor. Currently, she is a Ph.D student at the Faculty of Computing, Universiti Teknologi Malaysia. Her research interests include speech segmentation, emotion recognition, and computer vision.

Dr Md Sah Hj Salam is a lecturer at Computer Science Department, Faculty of Computing, Universiti Teknologi Malaysia. He obtained his bachelor degree in Computer Science (Software Engineering) from University of Pittsburgh, PA, USA (1997), MSc (2001) and PhD(2010) in Speech processing and AI from Universiti Teknologi Malaysia. He is a member of UTM ViCubeLab Research group under Faculty of Computing, Universiti Teknologi Malaysia, Skudai, Johor. His research interests include speech segmentation and recognition, artificial intelligent and computer vision.