International Journal of Engineering & Technology, 7 (3.15) (2018) 151-156

International Journal of Engineering & Technology

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

Research paper



Emolah: A Malay Language Spontaneous Speech Emotion Recognition on iOS Platform

Izzad Ramli[1]*, Nursuriati Jamil[1]*, Norizah Ardi[2], Raseeda Hamzah[1]

[1]Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Selangor, Malaysia
[2]Aacademy of Language Studies, Universiti Teknologi MARA, Selangor, Malaysia
**Corresponding author E-mail: lizajamil@computer.org*

Abstract

This paper presented the implementation of spontaneous speech emotion recognition (SER) using smartphone on iOS platform. The novelty of this work is at the time of writing, no similar work has been done using Malay language spontaneous speech. The development of SER using a mobile device is important for ease of use anytime and anywhere. The main factors to be considered is the computational complexity of classifying the emotions in real-time. Therefore, we introduced *EmoLah*, a Malay language spontaneous SER that is able to recognize emotions on the go with satisfactory accuracy rate. Pitch and energy prosody features are used to represent the emotions in the spontaneous speech and Naïve Bayes learning model is selected as the classifier. *EmoLah* is trained and tested using Malay language spontaneous speech acquired from television talk shows, live interviews from news broadcast and mini-parliament sessions conducted by children. Four types of speech emotions are collected that are happy, sad, angry and neutral. The total duration of all the speech emotion is four hours. The speech emotion training is using MATLAB scripts and the weights are implemented in XCODE as the iOS software for application development. *Emolah* accuracy is evaluated using cross-validation test and the result showed that it can discriminate angry, sad and happy. However, most emotions are misclassified as neutral emotion.

Keywords: *Speech emotion recognition, spontaneous speech, Internet of Things, iOS platform, Malay language.*

1. Introduction

Automated speech emotion recognition (SER) is conventionally designed mainly to work offline and for desktop applications due to the intensive nature of processing the speech data. However, with the advancements of processing speed and storage capabilities of smartphones, a substantial amount of work to implement automated SER on mobile devices has been done. Provisioning emotional care can greatly enhance a user's experience to improve user interaction, thus the quality of life [1]. Point of care applications using smartphones are seen as a necessity as field-portability, cost-effectiveness, rapid and easy-to-use interfaces have already been demonstrated [2]. However, computational limitation of smartphones still remains a challenge in real-time processing of intensive speech data. Therefore, this paper describes and demonstrates the development of a real-time SER using a smartphone catering to the computational limitations.

Emotion recognition on mobile platforms has been implemented using different modalities such as image, video, text or human gestures. As an example, text from the Short Message Service (SMS) can be analyzed to detect emotions of the users. Once detected, the system can respond by placing an 'emoji' at the end of the sentence. In image and video data, the system analyzes the facial expression of the user and responded by changing the wallpaper or playing a song that closely corresponds to the expressed emotions. Other than that, emotion can also be recognized from mobile devices using gestures such as user typing (i.e. backspaces mistakes, the number of special symbols used) and phone shaking. Speech emotion detection has been implemented in many domains as early as 1999 to detect the emotional state in call centre conversations to provide feedback to an operator and as a monitoring tool for supervisor [3]. A more recent application of SER is used in a simulated online learning in which student's response to a course is analyzed to enable customization based on the student's learning abilities [4]. Emotion based on speech is still popularly studied in many different platforms, with mobile devices as the current state of implementation.

To realize the implementation of the emotion recognition in a smart phone, the developed emotion recognition engine must be done in real-time, lightweight and can achieve acceptable recognition accuracy. Therefore, in this paper, we reviewed several emotion recognition methods which are most suitable to be implemented for the mobile platform. Then, we collected the Malay language spontaneous speeches and trained four types of basic emotions that are happy, angry, sad and neutral. The contributions of this paper are twofold: 1) Collection of annotated spontaneous emotional speech datasets for Malay language digital speech and 2) Prototype of a spontaneous speech emotion recognition system on iOS platform.

The paper is structured as follows. Section 2 describes the related background of emotional speech datasets and machine learning methods for emotion recognition. In Section 3, the overall methodology of the spontaneous SER prototype is presented and each process is described in detail. Section 4 elaborates on the implementation of *EmoLah* prototype by showing the screen snapshots. Section 5 presented the evaluation of *EmoLah*, while Section 6 draws a conclusion and recommends future work.

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

2. Research Background

This section begins with the overview of emotion speech corpus to understand the significance and limitations of our research. A review of learning classifiers is also presented to determine the classifier used in this paper.

2.1. Emotion Speech Corpus

Emotion speech corpus is basically categorized into three types that are actor-based emotion speech corpus, elicited-based emotion speech corpus and natural-based emotion speech corpus.

2.1.1. Actor-Based Emotion Corpus

As stated by [5], actor-based emotion speech corpus is most commonly used in emotion speech research. The corpus is collected from the professional and theatre actor or actress, and radio artist because it is the most convenient and reliable means of acquiring a large range of emotions. This type of corpus is available in many languages as the recording can be done repeatedly to produce the variations of expressiveness needed. However, the acted emotions tend to be exaggerated compared to the actual situation [6] and are somewhat unnatural.

2.1.2. Elicited-Based Emotion Speech Corpus

The elicited emotion speech is collected by inducing emotions from the speakers by pretending artificial emotion circumstances, without the knowledge of the speaker [5]. The induced speech is produced by involving the speaker with an anchor that created a different contextual situation to excite the emotion of the speaker. Sometimes, the collection of the elicited corpus is recorded by asking the speakers to get involved in verbal interaction with a computer whose speech responses are in turn controlled by the human being without the knowledge of the speakers [7].

2.1.3. Natural Emotion Speech Corpus

Natural emotion speech is acquired from spontaneous conversations between two persons in a public place or conducted interviews from televisions or radio talk shows. Examples of natural speeches are call-centre conversations, cockpit recording, dialogue between mother and child and office meetings. Emotions from natural or spontaneous speeches are commonly mildly expressed [5], thus annotations of these emotions are subjective and debatable. Another drawback of natural emotion speech acquisition is the difficulties of getting a wide range of emotions in a single session [8]. Since speech acquisition is done through live sessions, the dataset usually contains overlapping utterances and background noises. Another challenging matter is the legal issues such as privacy and copyright of using the public natural speech [9]. In this paper, we used a natural emotion speech dataset as our prototype is designed to recognize emotion in real-time.

2.2. Machine Learning for Speech Emotion Recognition

Various types of classifiers have been used to classify speech emotions with the most common being Support Vector Machines (SVMs), Hidden Markov Model (HMM), k-nearest neighbours (k-NN), Gaussian Mixture Models (GMMs) and Naïve Bayes (NB) [10]. HMM and GMM favourable characteristics are the ease of implementation and a strong mathematical basis. However, the need for a proper initialization for the model parameters before training and the long training time often associated with them [10] are the main drawbacks of HMM and GMM. Artificial Neural Network (ANN) is another popular classifier for pattern recognition. ANN is best for non-linear mapping and its performance is highly dependent on its design parameters. Therefore, in some speech emotion recognition systems (SER), more than one ANN is used [11]. The classification accuracy of ANN for SER is also normally lower compared to other classifiers. Support Vector Machine (SVM) is a classifier that is known to outperform other classifiers in pattern recognition applications. SVM also has good data-dependent generalization bounds [12] and achieved global optimal better than HMM and GMM [13]. However, SVM requires large training data and average training time is longer compared to HMM and GMM [14]. Another classifier is Naïve Bayes based on the Bayesian theorem with the Naïve assumption of independence between every pair of features. Naïve Bayes classifier does not require long processing time, has a good performance and works quite well in many real-world situations [10]. The main advantage of Naïve Bayes is it produced the conditional independence assumption that can obtain quick classification with the results of probabilities belonging to each class.

According to [14], there is still no agreement on which classifier is the most appropriate for speech emotion classification. The literature showed that each classifier has its own strength and limitations. Several factors need to be considered when choosing a classifier for the dataset and the purpose of the classification. One factor is the size of the dataset and the density distribution of the dataset for each class [15]. The second factor is the appropriate choice of parameters that has considerably affected the accuracy of the classifier. In our work, speed and computational complexity is also an important factor as the development platform is a smart phone. In a work done by [2], the performance of five classifiers to detect malware using smart phones is compared in terms of True Positive Rate (TPR). The classifiers are such as Random Forest, Decision Tree J48, Multi-Layer Perceptron (MLP), Bayes network, and K-Nearest Neighbour (K-NN). Classification on MalGenome dataset showed that the highest TPR is achieved by Bayes network and Random Forest with TPR of 99.97%, followed by MLP at 93.03%. Since Bayes network is known for its fast and efficient computation with quick data training [2], we selected the Naïve Bayes network as the classifier for our spontaneous speech emotion recognition.

3. Methodology

The development of our proposed spontaneous speech emotion recognition called *Emolah* is divided into two phases: Offline and online (real-time) processing. Figure 1 illustrates the process flow of these two phases. In the offline processing phase, spontaneous speeches are collected from live television shows and interviews to be used as training datasets. Speech prosody features are then extracted from the datasets to be used in training the SER engine using a Bayes network on MATLAB. The training results produced weights for each type of emotion to be used in the learning model in online processing. The spontaneous speech emotion recognition is done in online processing phase running on an iOS mobile platform. The spontaneous speech is captured real-time using the smart phone's microphone and serves as the input to the SER engine. The predicted emotion is then displayed in the smart phone's display screen to the user.

In this section, details of the spontaneous speech datasets used in the offline processing are presented in section 3.1; followed by the speech pre-processing in section 3.2. After the speech datasets are selected for training and testing, they are pre-processed to attenuate the speech artefacts prior to feature extraction. In this paper, pitch and energy are extracted from the spontaneous speeches to represent the emotions. Finally, the features of the training dataset are used in the training of the Naïve Bayes model in classifying four types of emotions. The weights from the trained Naïve Bayes model are later used for the real-time *Emolah* spontaneous speech emotion recognition.

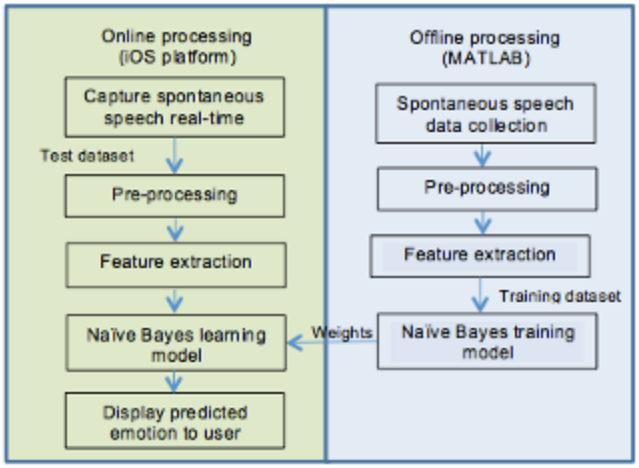


Fig. 1: System diagram of *Emolah* spontaneous speech emotion recognition.

3.1. Data Collection

There are many available databases created by the speech processing community which are widely used in speech-related work. Some of the commonly used databases are Danish Speech Database (DES), Berlin Emotional Speech Database (BES), German Emotional Speech Database (GES) and speech under Simulated and Actual Stress (SUSAS) Database as described in [16]. In this paper, we collected a new natural emotion speech dataset in the Malay language to add to the collections of existing speech emotion corpus and to enrich the digital speech corpus in the Malay language. Our speech dataset which is used to develop the training model is as shown in Table 1.

Table 1: Collection of spontaneous speech emotion

Emotion	Duration	Age range	Sources
Angry	1 Hour 4 seconds	6-55 years old	Dialogue on mini parliament and Malaysia parliament, interview on *Melodi* talk show
Sad	1 Hour 3 seconds	18–65 years old	Interview on *Melodi* and *Bersamamu* talk show
Happy	1 Hour 3 seconds	25-50 years old	Interview from *Melodi* and *Meletop* talk show
Neutral	1 Hour 3 seconds	18-50 years old	Interview from *Melodi* and local news reading from a radio.

We collected a total of approximately 7 hours spontaneous speeches from Malaysia parliament sessions, entertainment talk shows and live interviews in local news broadcast from a radio channel. The speech utterances consisting of phrases with emotions are manually segmented using PRATT software. The duration of each utterance ranges from 3 to 5 seconds. The utterances are manually labelled by a language expert from a local university to categorize the speech emotions into four categories: happy, angry, sad and neutral. For each category of emotion, roughly about an hour of utterances is randomly selected as the training dataset. Only one hour of each emotion category is used to ensure the equal size of training dataset is used. The total number of utterances differs from each emotion's category because each speaker has different rate-of-speech and the speeches are spoken spontaneously. A total duration of 4 hours and 13 seconds with the size of 450,493 MB is finally collected and used as the natural emotion speeches in this work. Ninety percent of this dataset is used to train the Naïve Bayes model and ten percent is used as test data.

3.2. Speech Pre-Processing

Speech pre-processing is the process of transforming the speech signal from sound pressure wave to digital signal and deal with the identified artefacts in the speech signal to reduce any subsequent problems. The pre-processing steps applied are sampling, framing, windowing and filtering. The sampling frequency of collected speech is 16 kHz in the mono channel with 16-bit bit resolution. Framing is done in the second step by blocking the speech signal into frames of *N* samples. In our work, we used a frame length of 1000ms [17], with adjacent frames separated and shifted by 10ms [17]. The frame length is set to 1000ms because our proposed SER analysed the emotion at every alternate second of a speech's frame before concluding the final emotion for a speech utterance. Windowing is applied after framing to minimize the signal discontinuities at the beginning and end of each frame. Windowing pre-multiplies the signal with a window function that smoothly decreases to zero value at each start and end frame. In our work, the Hanning window is chosen because it produced a smoother and accurate signal [18] compared to Hamming or Blackman functions. The final step of pre-processing is filtering to suppress interfering signals and reduce environmental noise. High pass filtering is done using Audacity with a cut-off value of 18dB, frequency smoothing (band) of 3 and sensitivity of 6 is used in our work to attenuate frequencies above 18dB and reduces frequencies below 18dB.

3.3. Feature Extraction

In this paper, prosody features are used to represent the speech emotions as the computations of prosody features are simpler compared to spectral or qualitative features. Despite its simplicity, the prosody features give a clear view of understanding of emotion-specific knowledge [5]. Prosodic features include fundamental frequency (FO), pitch, energy, duration, and formants. Even though more features may improve the emotion classification results, bigger feature space suffers the curse dimension problem. Since the speed of execution is important for a mobile-based platform, we used only pitch and energy to describe the emotions in our speech datasets.

Pitch is the quality of a sound represented by the rate of vibrations producing it and a vital prosody parameter having the perceptual property that makes utilization of frequency-related scale to requesting of sounds [19]. Pitch of the utterances represents periodicity candidates as a function of time. It is sampled into a number of frames centred around equally spaced times. Meanwhile, fundamental frequency (*F0*) is defined as the lowest frequency of a periodic waveform. A period of the waveform is the shortest possible time after which the waveform repeats itself. This single period is the smallest repeating unit and it will describe the signal completely. The equation of *F0* is defined as in Equation (1), where *F0* is the fundamental frequency and *T* s the fundamental period.

$$F0 = \frac{1}{T} \tag{1}$$

Intensity is referring to the energy of the speech. A few estimations of the speech energy exist, from the ordinary short-term intensity measure to the long haul coordination over prosodic units [19]. Energy represents an intensity contour at linearly spaced time points with values in dB. Energy (E_k) for *k*-th segment is defined in Equation (2), where g(t) amplitude for t-the frame and *N* is the number of frames.

$$E_k = \sum_{n=1}^{N-1} |g(t)|^2 \tag{2}$$

The pitch and energy are extracted based on a 1000ms frame length from the emotion speech dataset and stored as a feature vector.

3.4. Naïve Bayes Classification Model

The mean and variance of pitch and energy for each emotion are calculated and used as weights in the Naïve Bayes classification model. They are shown in Table 2.

Table 2: Mean and variance of the prosody features for each emotion category

Emotion	Pitch		Energy	
	Mean (Hz)	Variance (Hz)	Mean (Hz)	Variance (Hz)
Angry	440.40	158360.62	59.49	79.81
Sad	372.99	146535.04	57.37	74.72
Happy	450.42	198652.74	63.21	136.56
Neutral	410.81	247025.11	64.72	60.91

In our work, the posterior probability of the Bayes theorem is used to calculate the probability for each class of emotion. The class with the highest posterior probability is the result of the emotion prediction. The calculation is as shown in Equation (3).

$$P(c|x) = \frac{P(c)P(x|c)}{P(x)} \quad (3)$$

where $P(c|x)$ is the posterior probability of *class* (c, target) given *predictor* (x, attributes)
$P(c)$ is the prior probability of *class*
$P(x|c)$ is the likelihood which is the probability of *predictor* given *class*
$P(x)$ is the prior probability of *predictor*

Suppose the mean pitch and energy of a speech segment (1000ms/frame) is 340Hz and 60Hz, respectively. In order to predict the speech segment emotion, the posterior probabilities of angry, sad, happy and neutral for the speech segment are calculated. An example of posterior probability's calculation of angry emotion is as follows:

The prior probability of the speech segment being angry is P(angry) = 0.25. Therefore, its likelihood is calculated as:

$$P(pitch|angry) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(\frac{-(\beta-\mu)^2}{2\sigma^2}\right) = 0.00097;$$

where β = 340, μ = 440.40 and σ^2 = 158360.62, are the parameters of the normal distribution which are previously determined (Table 2) from the training dataset.

β = Pitch of sample
μ = Mean pitch of angry
σ^2 = Variance pitch of angry

Next, we applied the same calculation of $P(x|c)$ for the likelihood of class energy, $p(energy|angry)$= 0.0436. The prior probability of the predictor, $P(x)$ is then calculated as shown below, resulting in $P(x)$ = 0.000179.

$P(x) = P(angry)\ P(pitch \mid angry)\ P(energy \mid angry) + P(sad)\ P(pitch \mid sad)\ P(energy \mid sad) + P(happy)\ P(pitch \mid happy) P(energy \mid happy) + P(normal)\ P(pitch \mid neutral) P(energy \mid neutral)$

Therefore, the posterior probability of the speech segment to be classified as angry emotion is

$$P(angry \mid predictor) = (P(angry)\ P(pitch \mid angry) P(energy \mid angry)) / P(x)$$
$$= (0.25)(0.00097)(0.0436)/0.000179$$
$$= 0.0589$$

The posterior probabilities of sad, happy and neutral emotions are further calculated and the results as $P(sad|predictor) = 0.0637$, $P(happy|predictor) = 0.8308$, and $P(neutral|predictor) = 0.0464$. Since the posterior probability of happy emotion is the highest, the speech segment is classified as happy. The classified emotion is finally displayed to the user by showing a happy emoticon with an audio response saying *"Are you happy?"*

4. Implementation of *Emolah*

In this section, the implementation of *EmoLah* spontaneous speech emotion recognition is described. The system requirement to develop *EmoLah* on iOS platform is as tabulated in Table 3.

Table 3: Hardware and software requirements

Computer	Macintosh computer
Operating system	Mac OS X 12.2 (Sierra)
Language	Swift 3.0
Software	Xcode 9.0
Toolkit	AudioKit
Reference Website	http://developer.apple.com/iphone

A Macintosh computer with MAC OS X 12.2 (Sierra) or above is needed to install Integrated Development Environment (IDE) or software called Xcode. The minimum version of Xcode 9,0 is required to use programming language Swift 3.0. Further details can be found at http://developer.apple.com.

4.1. The Main Character

The main character is created for *EmoLah* as illustrated in Figure 2 to express emotions in a graphical manner. The character appears in the Splash screen during start-up of *EmoLah*.



Fig. 2: The main character depicting angry and neutral emotions

Figure 3 illustrates the graphical expressions of the main character as indicators of the classified emotions. A Silent expression is added as *EmoLah* classifies emotions in real-time and if a non-speech segment is detected, a Silent icon will appear on the output screen. The other emotion expressions that are angry, sad, neutral and happy are displayed accordingly.



Angry Sad Neutral Happy Silent

Fig. 3: Graphical emoticons used in EmoLah

One of the main strength of *EmoLah* is its ability to recognize emotions from Malay language spontaneous speeches. It continuously recognizes emotions based on a 1000ms speech segment once the Record button is pressed. If the same emotion is recognized in three consecutive speech segments, *EmoLah* concludes the emotion as is. An example of Sad emotion recognized in 3000ms is demonstrated in Figure 4. The recognized emotion is displayed on top of the screen with a caption "Are you feeling sad?"



Fig. 4: Example of Sad emotion recognized by *EmoLah*

3.4. The Help

Help in *EmoLah* can be accessed by clicking an Information button, at the bottom of the main screen. Instructions on the Record and Play buttons are explained in the Help screen as illustrated in Figure 5(a). *EmoLah* also provides some texts to the user to familiarize with *EmoLah* by providing sample texts for the user to read. Figure 5(b) shows the active voice signal located at the bottom of the display screen as the user speaks. If the user wishes to perform the voice recording again, press the 'try again' button. When new recording begins, the previously recorded speech is deleted to allow space saving. Tips on how to produce the best recording are displayed to the user before recording begins is shown in Figure 5(c).



(a) (b) (c)

Fig. 5: Help session in *EmoLah*

5. Results and Discussions

As stated in Section 3.1, 10% (24 minutes) of the total spontaneous speech is allocated as a testing dataset. Based on the manual emotion labelling, there are a total 39 utterances of angry emotion, 30 utterances of sad, 35 utterances of happy and 48 utterances of neutral emotion in the testing dataset. The testing dataset is playback as input to *EmoLah* to simulate spontaneous speech, A ten-fold cross validation test is done using *EmoLah* on the testing dataset. The results of the evaluation are shown in the confusion matrix in Table 4.

Table 4: Confusion matrix of the test dataset

	Angry	Sad	Happy	Neutral	Accuracy (%)
Angry	18	0	7	14	46%
Sad	0	18	3	9	60%
Happy	5	0	20	10	57%
Neutral	2	3	12	31	65%

The results of the evaluation are shown in the confusion matrix in Table 4. It can be seen that neutral emotion achieved the highest accuracy rate of 65%, followed by 60% for sad emotion, 57% for happy and 46% for angry. Accuracy rate is computed as the number of correctly classified emotion over the number of actual labelled emotions. Overall, the accuracy rate is considerably on average due to the use of minimal features to describe the emotions. Upon closer inspection of the results, the main error is caused by misclassifying all emotions as neutral. As supported by the literature, emotions are mildly expressed in spontaneous or natural emotion compared to acted- or elicited-based speech emotions. This is the main problem in recognizing emotions in natural speeches. Angry and sad emotions are easily discriminated because the results showed that angry is never misclassified as sad, and vice-versa. However, happy and angry have similar pitch and energy, thus some speech fragments with happy emotions are classified as angry, and vice versa.

6. Conclusion

Mobile emotion sensing has gained interest for future consumer electronic devices and services. This paper introduced a spontaneous speech emotion recognition mobile application called Emolah and presents the details of the design, development and implementation of iOS platform which can be executed on iPhone, iPod and iPad. The classification accuracy achieved a reasonably satisfaction result ranging from 46% to 65% of accuracy rate. This is due to the limited prosodic features used to represent the speech dataset. As this is the first effort of implementing a mobile application on iOS platform, we only used the available functions provided by the operating system. We, however, developed the Naive Bayes model from scratch and further improvement on the model is needed to achieve better classification accuracy.

After our first implementation of *Emolah*, we are certain of further upgrading the prototype. Firstly, the selection of features used should consider spectral features which are known to represent emotions in speech better. It is hoped that the use of spectral features should discriminate angry and happy better. However, the computational complexity of spectral features is to be considered for real-time execution. Deep learning models should also be investigated to classify the emotions provided enough training data can be generated to effectively train the deep models. Finally, robustness is an important factor of a good SER that we have not yet comprehended. Background noise and noise generated during real-time speech acquisition degrade the emotion recognition rate. Robustness evaluation with various background levels also needs to be done in the future. Our spontaneous speech dataset also does not consider age, ethnic and ethnic factors. Previous work has shown that these factors also have a considerably profound effect on the accuracy of an SER.

References

[1] Hossain MS, Muhammad G. An Emotion Recognition System for Mobile Applications. *IEEE Access*. 2017;5:2281-7.
[2] Narudin FA, Feizollah A, Anuar NB, Gani A. Evaluation of Machine Learning Classifiers for Mobile Malware Detection. *Soft Computing*. 2016 Jan 1; 20(1):343-57.
[3] Petrushin V. *Emotion in Speech: Recognition and Application to Call Centers*. Proceedings of Artificial Neural Networks in Engineering. 1999:710.
[4] Cen L, Wu F, Yu ZL, Hu F. *A Real-Time Speech Emotion Recognition System and its Application in Online Learning*. In Emotions, Technology, Design, and Learning. 2016; 27-46.
[5] Koolagudi SG, Rao KS. Emotion Recognition from Speech: A Review. *Int J Speech Technology*. 2012; 15:99-117.
[6] Williams C, Stevens K. Emotions and Speech: Some Acoustical Correlates. *J. Acoust. Soc. Am.1972;* 52 (4 Pt 2):1238–1250.
[7] Wolfgang W. Mobile Speech-to-Speech Translation of Spontaneous Dialogs: An Overview of the Final Verbmobil System. In: Wahlster, W. *Editor*. Verbmobil: Foundations of Speech-to-Speech Translation. Springer; 2000: 3–21.
[8] Apandi N, Jamil, N. *Emotional Speech Corpus Development for Emotion Recognition in Malay Language*. Proceedings of the In-

dustrial Electronics and Applications Conference (IEACon). Kota Kinabalu, Sabah. 2016; 225–231.
[9] Mustafa MB, Ainon RN. Emotional Speech Acoustic Model for Malay: Interative versus Isolated Unit Training. *J. Acoust. Soc. Am.* 2013 Oct; 134(4):3057-66.
[10] Urbano Romeu Á. Emotion Recognition Based on the Speech, using a Naive Bayes Classifier. Bachelor Thesis, Universitat Politècnica de Catalunya: 2016.
[11] Nicholson J, Takahashi K, Nakatsu R. Emotion Recognition in Speech using Neural Networks. *Neural Comput. Appl*. 2000; 9:290–296.
[12] Cristianini N, Shawe-Taylor J. An Introduction to Support Vector Machines. Cambridge University Press; 2000.
[13] Burges CJC. A Tutorial on Support Vector Machines for Pattern Recognition, *Data Mining Knowl. Discovery*. 1998; 2 (2): 121–167.
[14] El Ayadi M, Kamel MS, Karray F. Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases. *Pattern Recognition*. 2011 Mar 1; 44(3):572-87.
[15] Seman N. Coalition of Artificial Intelligence (AI) Algorithm for Isolated Spoken Malay Speech Recognition. PhD Thesis. Universiti Teknologi MARA: 2011.
[16] Wu CH, Liang WB. Emotion Recognition of Affective Speech Based on Multiple Classifiers using Acoustic-Prosodic Information and Semantic lLabels. I*EEE Transactions on Affective Computing*. 2011 Jan;2(1):10-21.
[17] Paliwal KK, Lyons JG, Wójcicki KK. *Preference for 20-40 ms window duration in speech analysis*. Proceedings of 2010 4th International Conference on Signal Processing and Communication Systems (ICSPCS), 2010 Dec 13; 1-4.
[18] Podder P, Khan TZ, Khan MH, Rahman MM. Comparative Performance Analysis of Hamming, Hanning and Blackman Window. *International Journal of Computer Applications.* 2014 Jan 1; 96(18):1-7.
[19] Ali SA, Zehra S, Khan M, Wahab F. Development and Analysis of Speech Emotion Corpus using Prosodic Features for Cross Linguistics. *International Journal of Scientific & Engineering Research*. 2013 Jan;4(1):1-8.