ELSEVIER
Available online at www.sciencedirect.com
ScienceDirect
Procedia - Social and Behavioral Sciences 97 (2013) 46 – 53
Procedia
Social and Behavioral Sciences

The 9th International Conference on Cognitive Science

Investigation on dynamic speech emotion from the perspective of brain associative memory

Norhaslinda Kamaruddin[a]*, Abdul Wahab[b]

[a]*Faculty of Computer and Mathematical Sciences, MARA University of Technology, 40450 Shah Alam, Selangor, Malaysia*
[b]*Kulliyah of Information and Communication Tecnology, International Islamic University Malaysia, 53100 Kuala Lumpur, Malaysia*

Abstract

Many researchers have studied speech emotion for years from the perspective of psychology to engineering. To date, none has made the speech emotion recognition system intuitive enough in such a way that it can be embedded in automatic answering machines that can effectively detect the various affective states of human verbal communication. In most cases the underlying emotional information was misinterpreted thus resulting in wrong feedbacks and responses. The complexity of understanding and analyzing speech emotion is presented in the dynamics of the emotion itself. Emotion is dynamic and changeable over time. Hence, it is imperative to cater for this parameter to boost the performance of the speech emotion recognition system. In this paper, values of Valence (V) and Arousal(A) are used to generate a recalibrated affective space model. Such approach is adopted from psychologists' understanding that emotion can be represented using emotion primitives' values. The VA approach is then coupled with the brain associative memory concept that can provides a better means in understanding the dynamics of speech emotion. Results of such analysis tallies with the psychological findings and has its practical implementation.

Keywords: Speech emotion recognition; Affective Space Model; Mel Frequency Cepstral Coefficient; Emotion dynamic.

1. Introduction

Human are able to express their affective states through verbal and non-verbal communication in everyday life. Variation of emotional states can be used as essential evaluation criteria to access the efficacy of cognition processing particularly for action tendency [1] and decision making [2]. Facial expression, speech, gestures and body posture are the common channels to project an emotional reaction towards stimuli. Such combination responses are used by the audience to perceive the underlying messages. However, in most cases, emotion can still be propagated using single channel although it may compromise the recognition performance. For instance, an infant can detect the mother's affective state by just listening to her voice even when the child's vision capability is not fully developed [3]. Such observation is reported by Miles and Meluish who stated that very young children tend to respond to their mother's voices with greater sucking rate compared to other female stranger's voices [4]. Saito et al. [5] found that exaggerated parents voice (infant-directed speech) while communicating with their children can increased blood flow to the frontal area of the child's brain. In much recent work, Naoi et al [6] observed that

---

* Corresponding author. Tel.:+03-55211130
E-mail address: norhaslinda@tmsk.uitm.edu.my

doi:10.1016/j.sbspro.2013.10.203

significant activation occurs when the infant listened to the infant-directed speech produced by their own mother as opposed to unfamiliar mothers. Such findings provide neurophysiological evidence that support the fact that infant can recognize emotion although they process visual and auditory information at a slower rate. In addition, one can also recognize the affection portrayed by pantomime actors who employs facial expressions, gestures and body postures in their performance. Hence, it can be summarized that affective state can be disseminated by various channels towards the convergence and optimization of the interactive process. However, the focus for this paper is on the dynamic of speech emotion.

Currently, recognizing emotion from speech starts gaining interest from the researchers. It can be seen from the steady increase in the amount of quantitative researches. Verviridis and Kotropoulus [7] and El Ayadi et al. [8] provide comprehensive reviews on speech emotion recognition focusing on various data corpora, feature extraction methods and classifiers. To date, no researchers can boast highest performance recorded due to the complexity of speech emotion recognition task. It is very much dependent on many factors, such as: the data itself, number of emotions identified and cultural influence [9]. Hence, physiological changes can be used as a benchmark to gauge the perceived emotional feedbacks. These researches basically follow the James-Lange theory of emotion [10] that indicates physiological arousal instigates the experience of a specific emotion. The notion of emotion can be empirically discriminated boost the researchers' effort to measure emotion. For comprehensive discussion, Friedman [11] extensively explained the James-Lange theory and relates the theory to the contemporary researches.

Emotion itself can be classified as a discrete class or multi-dimensional basis functions. For simplification, selections of few basic emotions are listed for discrete class emotion recognition approach. Due to its subjectivity and dependancy on culture and environment-related factors, there have been many criteria to identify emotion resulting in a several number of basic emotion lists. The disparity between these emotion lists occur because each researcher claims that his/her list of emotion subset covers the most comprehensive description of basic emotion. Nonetheless, researchers in various disciplines agree that there are some emotions that can be universally accepted and is fairly consistent across cultures [12, 13]. Hence, it is sound to consider some of these emotions as basic emotions for analysis purposes. On the contrary, discrete class emotion approach is challenged with the arguments that emotions cannot be thoroughly explained with a limited set of class and it does not allow subjects to access an adequate level of discrimination. In addition, few weaknesses of discrete class emotion approach are encountered; namely:

- The discrete-class approach is only suitable for supervised learning where the testing instance must have mutual exclusive membership of the training class. For instance, if the training class consists of basic emotions; namely: anger, happiness, sadness and neutral classes, the testing instance must belong to either one of these respective classes.
- It is also found that if the correct training classes are not properly accounted for, valuable information is discarded resulting in inconsistent estimation and biased outcome.
- Furthermore, the dynamic characteristic of emotion is ignored although it is known that emotion is not a discrete event as proposed by the functionalist model of emotion.
- The rigidity of such approach restricts any other variation of class that is not trained to recognize the emotion, hence limiting the detailed study of emotion primitives and dynamics.

A multi dimensional basis functions emotion approach is thus selected for this work to account the dynamic nature of the speech emotion. This may reflect the subtlety and complexity of the affective states conveyed by such rich sources of information experience and physiological reactions.

The multi-dimensional basis functions emotion model is constructed from the understanding that emotion constituted from three emotion primitives values of valence (V), arousal (A) and dominance (D) [14]. These primitives are basis functions that describe the generic attributes of an emotion that act as a fully complementary description of the emotion and form the Affective Space Model (ASM). This model can be divided into quadrants where each quadrant represents positive or negative value of the emotion primitives; such as scale of pleasure (+) to displeasure (-) for valence, scale of active (+) to passive (-) for arousal as well as scale of strong (+) to weak (-) for dominance. Such analysis facilitates the understanding of the core attributes of emotion rather than limiting the focus to the customary discrete emotion classes. In this paper we only consider valence and arousal values to generate our recalibrated ASM. This is because dominance factor has accounted for the least variance in affective judgment [15]. Such notion is consistent with Russell's earlier finding [16] that valence (V) and arousal (A) are accounted for by far the major proportion of variance in the affect scales while dominance only accounted for a very small scale.

The idea of recalibrated ASM is then extended with brain associative memory concept. Such correlation is proposed due to the fact human brain can change its structure and function based on input from the environment and on the potential associations or consequences of that input. Brain adaptive ability links the component parts using direct or spatial, temporal or other kinds of relationships. These components are represented by neural activity in different part of the neocortical region of the brain that project to the medial temporal lobes (MTL) where they are integrated to create associative memory. During retrieval, one component can trigger recall of other components, which reactivates part or all of the memory. Therefore, recall is inherently associative. Further reading is provided by Mayes et al. review [17]. The concept of brain associative memory can be instantiated by the following scenario: if a person puts his hand in a fire and get hurt, he learns to avoid it. Later, the sight of fire has gained a predictive value, which in this case, is dangerous. Thus, combination of ASM and brain associative memory seems to be a natural process.

In this paper, four emotions from NTU_American dataset [9, 18] are studied, namely; anger, happiness, sadness and neutral acting as emotionless state. It would be briefly explained in Section 2. The proposed combination of VA approach and brain associative memory concept is provided in Section 3. The Mel Frequency Cepstral Coefficient (MFCC) [19] feature extraction method coupled with Adaptive Network Fuzzy Inference System (ANFIS) [20] classifier is employed to train the speech signal into valence and arousal emotion primitives' values and map the generated value to the recalibrated affective space model (ASM) [21, 22]. Subsequently, nearest neighbor technique is applied to simulate the brain associative ability onto the recalibrated ASM. Experimental results are analyzed and discussed in Section 4. This paper concludes with summary and future work in Section 5.

2. Data Corpus : NTU_American Dataset

A validated data is imperative to ensure the correctness of the proposed approach. In this work, NTU_American [9, 18] dataset is used. It is collected at the Center for Computational Intelligence (C^2iLab), Nanyang Technological University, Singapore. The data is specifically designed by collecting speech emotion data samples from movies and television sitcom video clips downloaded from the Internet. The database includes short utterances contributed by four to six speakers covering the basic discrete emotions, namely: anger, happiness, sadness, disgust, surprised and neutral serving as the emotionless state. All speakers are using the American English language as a medium of interaction.

The emotions portrayed by the speakers were manually identified based on the speech semantic, interlocutor facial expression and the basic understanding of the video clip's plot. These video clips were then converted to audio wave file at 8,000 samples per second, mono-streamed and its amplitude value is normalized to [-1:+1] V. To ensure 'pure' speech emotion data, the artifacts such as background music, long pauses and other non-verbal sounds (sigh, sobbing and filled pause utterances) were removed. In this paper, we only consider anger, happiness, sadness and neutral emotion with 235 utterances. Such selection is made due to the fact that these emotions are highly recognized [9,18] and extensively reported in the literatures.

3. VA Approach on Recalibrated Affective Space Model using Brain Associative Memory Concept

Raw speech emotion signals are collected in the data collection module and pre-processing steps need to be taken to ensure the data is clean and without artifacts. Then, the Mel Frequency Cepstral Coefficient (MFCC) feature extraction method [19] is employed to extract relevant features to be processed by the Adaptive Network-based Fuzzy Inference System (ANFIS) classifier [20]. Slaney's 40 features MFCC approach is selected based on Ganchev et al. observation that it can yield reasonable accuracy and able to provide the lowest decision cost [23]. Once the raw signals are transformed into features, ANFIS will classify the instances based on the k-fold training and testing approach. 80% of the data are used as the training dataset whereas the remaining 20% data will be used as the testing dataset. Such training and testing process are iteratively conducted to ensure all data are completely used for testing. This is to get generalization value where the classifier does not see the testing instances prior to the testing process. Table 1 illustrates the training target set for the different emotions. These values are conformed to the psychologists' understanding of affective space model [14]. Neutral is set as VA value of (0,0) based on hypothesis that neutral is emotionless hence the value for the emotion primitives must be 0.

Table 1. Valence and Arousal Values Trained for Specific Emotion

Emotion	Valence	Arousal
Anger	-1	+1
Happiness	+1	+1
Neutral	0	0
Sadness	-1	-1

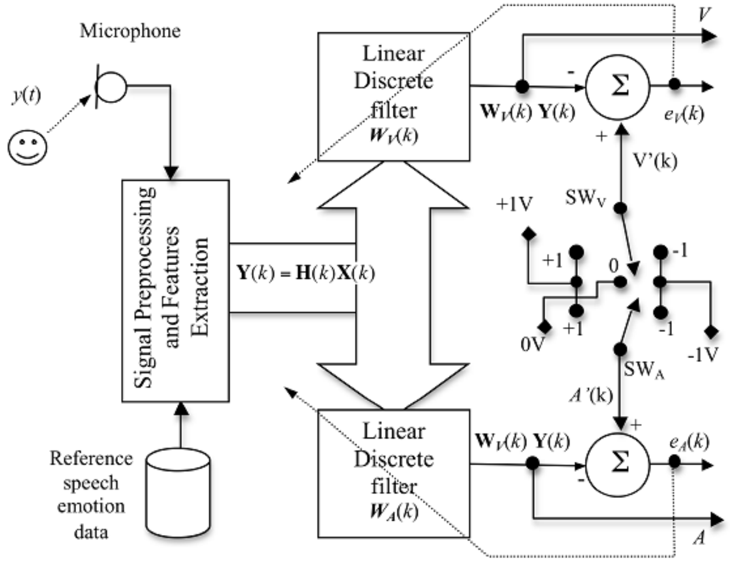
Microphone
$y(t)$
Linear Discrete filter $W_V(k)$
$\mathbf{W}_V(k)\ \mathbf{Y}(k)$
-
Σ
$e_V(k)$
V
+
V'(k)
+1V
SW_V
+1
0
-1
0V
-1V
SW_A
$A'(k)$
Signal Preprocessing and Features Extraction
$\mathbf{Y}(k) = \mathbf{H}(k)\mathbf{X}(k)$
Reference speech emotion data
Linear Discrete filter $W_A(k)$
$\mathbf{W}_V(k)\ \mathbf{Y}(k)$
+
-
Σ
$e_A(k)$
A

Fig. 1. Block diagram of the VA approach

The two-dimensional data-driven ASM will be generated based on these two outputs. A particular emotion can be discriminated based on its axes position in the newly generated ASM. Valence is denoted in the x-axis of the model and it refers to the impact of emotion to oneself ranging from positive to negative. In addition, arousal values range from excited to passive to describe emotion activation and it is represented by y-axis. Figure 1 illustrated the block diagram of VA training and testing. Once the ASM is generated, the weight of the points that is located at a specific coordinates of the ASM will be incrementally calculated. Such approach is implemented to measure the density of the specific point populated in the ASM. Although we can discriminate emotion directly using the generated ASM, the weight is important to provide us with the dynamic of the emotions based on the valence and arousal (VA) emotion primitives' values. In addition memory associations are also reflected by the weights of the ASM. The weights of each point are interconnected in such a way that any changes in one point may affect the neighboring points. Hence, such situation explains the gradual changes of emotion dynamic from one emotion to another over time.

4. Experimental Result and Discussion

VA values of each instance in the NTU_American dataset are then mapped onto the generated ASM. Figure 2 shows the scatter plot of the ASM for the three basic emotions of anger, happiness and sad and neutral as the emotionless state. Majority of the respective emotions lie within the specified quadrant of the emotion except for about 20% of the emotion instances seems to be outside the quadrant. This is due to the fact that emotion is dynamic and needs to be build-up to the respective emotions. The gradual changes of one emotion to another are natural because human does not have the ability to automatically switch on/off emotion. Anger populated heavily in Quadrant 1 that has positive value for arousal and negative value for valence. Happiness and sadness are distributed in almost diagonal mirror in such a way that happiness positioned in positive valence and arousal whereas sadness is in the contrary. Neutral yielded the most interesting result where it is tabulated in all four quadrants. Such observation is inline with the psychologists' hypothesis that neutral is the basis of other emotion that the emotion transition must undergo through the neutral region first prior to the emotion changes.

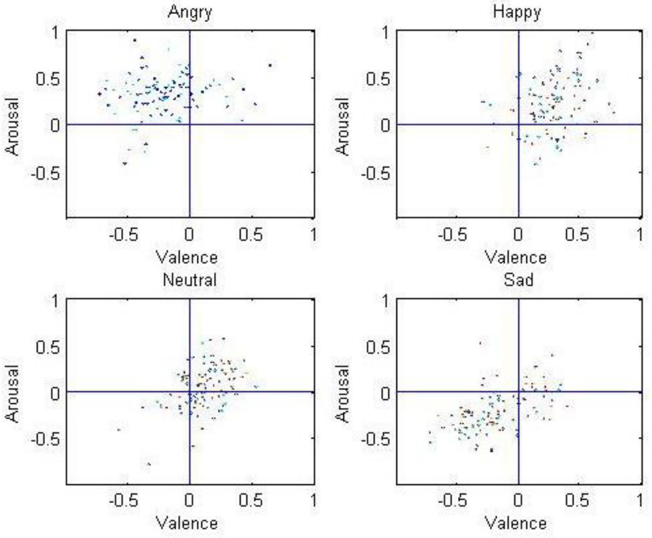

Fig. 2. NTU_American VA Weight Scatter Plot Without Brain Associative Concept.

The weights of the ASM are then measured to correlate with brain associative concept. Figure 3 and 4 depict the NTU_American VA weight mesh and contour plot respectively. In Figure 3, x-axis represents valence, y-axis represents arousal and z-axis represent weight. Figure 4 is provided to give a clearer view of the distribution of the weight. From the experimental results in Figure 3 and 4, it is interesting to note that neutral (emotionless state) seems to be distributed in all four quadrants which explain the four different peaks in the plot. This is because neutral is the base of all emotion and may exist in all emotions. However, neutral is observed to be slightly biased towards happy with positive valence and arousal value. Such result can be explained that the speaker are typically in a positive affective state and actively engaged when they associate themselves with neutral. Although neutral is denoted with (0,0) VA values, it is a challenge to obtain as speech emotion is influenced by cultural bias. In addition, all other three emotion plots illustrate that the points are populated correctly as depicted by their single peak in their respective quadrants. Subsequently, the mesh plot of Figure 3 and the contour plot of Figure 4 also show that neutral seems to be clustered towards the origin which is expected thus indicating the concept of plotting the valence and arousal as basis function for the emotions reflect most psychological findings and can be used in analyzing the dynamics of speech emotions.

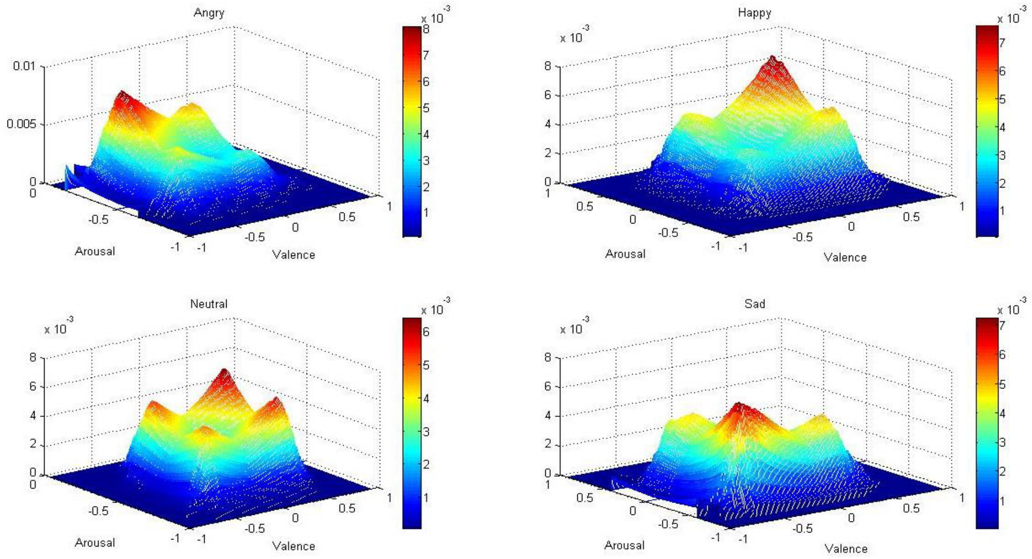

Fig. 3. NTU_American VA Weight Mesh Plot With Brain Associative Concept.

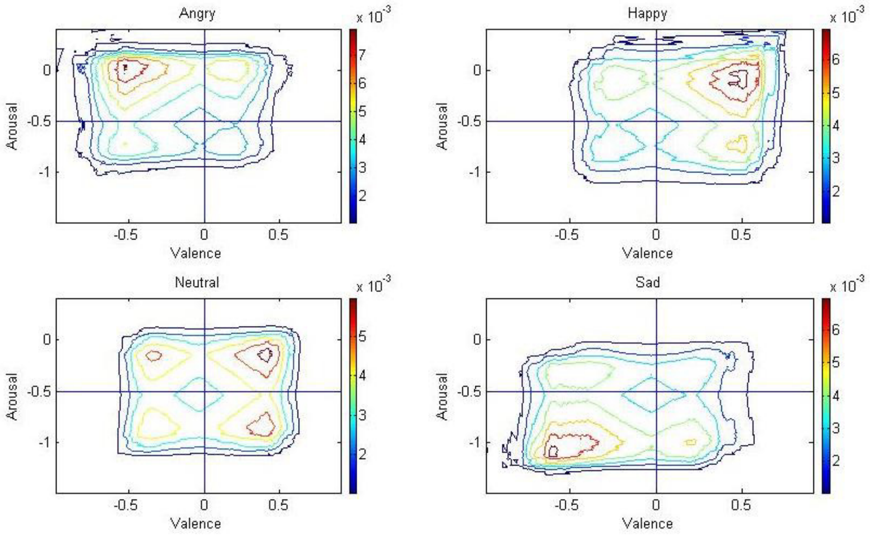

Fig. 4. NTU_American VA Weight Contour Plot With Brain Associative Concept

5. Conclusion and Future Work

Speech emotion is dynamic and can gradually change over time. In this paper, a VA approach using recalibrated affective space model is employed to visualize the emotion primitives' values of valence (V) and arousal (A). Then, correlation of recalibrated ASM with brain associative concept is introduced from the understanding that brain relates one memory to another to store the information. Such notion is translated in the experiment and experimental results show that it is feasible to use such approach to understand the dynamic of speech emotion. Anger, happiness and sadness are fairly distributed in their respective quadrant. Such result can be clearly observed in its single peak in Figure 3. Interestingly, using the proposed approach, neutral can be explained in more comprehensive manner other than it is only has 0 value for valence and arousal. Neutral is populated in all quadrant and such observation can be linked to the hypothesis of neutral is the base of all emotion.

More works can be extended from this paper. For instance, comprehensive analysis of the proposed approach can be implemented by employing other datasets. By doing so, cultural influence analysis can be conducted. Human behavior analysis particularly on driver behavior is another research ground that has lots of potential. This is because driving is an activity that requires full attention to ensure no accident occurs. Some preliminary results are presented in [24,25,26]. In addition, analysis on dynamic emotion data such as Electroencephalography (EEG) signal using this approach can also be applied. It is envisages that the proposed approach can be used as an alternative to the state-of-the-art approaches in complementing the effort to understand speech emotion.

Acknowledgements

The authors would like to thank MARA University of Technology, Malaysia for providing financial support through the excellence fund under the Research Intensive Faculty (RIF) grant (600 – RMI / DANA 5/3/RIF (769/2012)) to conduct the work published in this paper.

References

[1] Fridja NH. Emotion, Cognitive structure and action tendency. *Cognition and Emotion* 1987;**1**(2):115–143.

[2] Damasio AR. *Descartes' error: emotion,reason and the human brain.* New York: Grosset/Putnam; 1994.

[3] Mehler J, Bertoncini J, Barrière M, Jassik-Gerschenfeld D. Infant recognition of mother's voice. *Perception* 1978;**7**(5):491 – 497.

[4] Miles M, Meluish E. Recognition of mother's voice in early infancy. *Nature* 1974;**252**:123-124.

[5] Saito Y, Aoyama S, Kondo T, Fukumoto R, Konishi N, Nakamura K, Kobayashi M, Toshima T. Frontal cerebral blood flow change associated with infant-derected speech. *Archives of Diseasein Childhood - Fetal and Neonatal Edition* 2007;**92**:F113 – F116.

[6] Naoi N, Minagawa-Kawai Y, Kobayashi A, Takeuchi K, Nakamura K, Yamamoto J, Kojima S. Cerebral responses to infant-directed speech and the effect of talker familiarity. *NeuroImage* 2012;**59**:1735 – 1744.

[7] Ververidis D, Kotropoulus C. Emotional Speech Recognition: resources, features and methods *Speech Communication* 2006;**48**(9):1162 – 1181.

[8] El Ayadi M, Kamel MS, Karray F. Survey on speech emotion recognition: features, classification schemes and databases. *Pattern Recognition* 2011;**44**:572-587.

[9] Kamaruddin N, Wahab A, Chai Q. Cultural Dependency Analysis for Understanding Speech Emotion. *Expert System with Application* 2012; **39**(5):5115-5133.

[10] James W. What is an emotion. *Mind* 1884;**9**:188-205.

[11] Friedman BH. Feelings and the body: the Jamesian perspective on autonomic specificity of emotion. *Biological Psychology* 2010;**84**:383 -393.

[12] Buck R. The biological affects; a typology. *Psychological Review* 1999;**106**:301-336.

[13] Ekman P. Heider KG. The universality of a contempt expression - a replication. *Motivation and Emotion* 1988;**12**(4):303-308.

[14] Schlosberg H. Three dimensions of emotion. *Psychological Review* 1954;**61**(2):81-88.

[15] Bradley MM, Lang PJ. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry* 1994;**25**(1):49-59.

[16] Russell JA. Affective space is bipolar. *Journal of Personality and Social Psychology* 1979;**37**(3):345-356.

[17] Mayes A, Montaldi D, Migo E. Associative memory and the medial temporal lobes. *Trends in Cognitive Sciences* 2007;**11**(3):126-135

[18] Kamaruddin N, Wahab A. Features Extraction for Speech Emotion. *Journal of Computational Methods in Science and Engineering* 2009;**9**(Suppement 1):S1-S12.

[19] Slaney M. Auditory Toolbox (Version 2). *Technical Report #1998-010*, Interval Research Corporation 2007 [online] http://cobweb.ecn.purdue.edu/~malcolm/interval/1998-010

[20] Jang J-SR. ANFIS: Adaptive-Network-Based Fuzzy Inference System. *IEEE Transaction on Systems, Man and Cybernetics* 1993;**23**(3):665-685.

[21] Kamaruddin N, Wahab A. Human behavior state profile mapping based on recalibrated speech affective space model. *34th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* 2012;San Diego, USA:2021-2024.

[22] Kamaruddin N, Wahab A. Heterogeneous driver behavior state recognition using speech signal. *The 10th WSEAS International Conference on System Science and Simulation in Engineering* 2011;Penang, Malaysia:207-212.

[23] Ganchev T, Fakotakis N, Kokkinakis G. Comparative evaluation of various MFCC implementations on the speaker verification task. *Proceeding of the SPECOM* 2005; Patras, Greece **1**:191-194.

[24] Wahab, A, Wen TG, Kamaruddin, N. Understanding Driver Behavior Using Multi-dimensional CMAC. *The 6th International Conference on Information, Communication & Signal Processing* 2007; Singapore:1-5.

[25] Khalid, M, Wahab, A, Kamaruddin, N. Real Time Driving Data Collection and Driver Verification Using CMAC-MFCC. *The 2008 International Conference on Artificial Intelligence* 2008; Las Vegas, Nevada, USA:219-224.

[26] Kamaruddin, N, Wahab, A. Driver Behavior Analysis Through Speech Emotion Understanding. *The 2010 IEEE Intelligent Vehicle Symposium* 2010; San Diego, California, USA: 238-243.