

Heterogeneous Driver Behavior State Recognition Using Speech Signal

NORHASLINDA KAMARUDDIN
Faculty of Computer and Mathematical
Sciences
MARA University of Technology
40450 Shah Alam, Selangor
MALAYSIA
norhaslinda@tmsk.uitm.edu.my

ABDUL WAHAB
Kulliyah of Information and Communication
Technology
International Islamic University
50728 Gombak, Kuala Lumpur
MALAYSIA
abdulwahab@kict.iium.edu.my

Abstract: - Driver behavior is one of the major factors contributing to accident. Hence, if we are able to detect the abnormal driver behavior, we may be able to prevent such tragedy from happening. Based on the hypotheses that 1) driver behavior is influenced by emotion and 2) emotion can be measured using speech; we proposed an alternative way to recognize driver affective states. Emotion is assumed to be dynamic and changes gradually over time. Such assumption is collectively agreed by psychologist that emotion can be represented using the affective space model. In this paper, we derived the affective space model dynamically using the emotional speech data based on three different culture bases, namely; American, European and Asian. This is to show that such approach is well generalized and can be adapted to different cultures as well. Experimental results show potential of applying such approach to determine driver behavior states (DBS); namely: sleepy, talking through cell phone, laughing while driving as well as normal driving.

Key-Words: - Driver Behavioral State, Speech Emotion Recognition, Affective Space Model, Emotion Profiling, Neural Network, Fuzzy Neural Network

1 Introduction

Human judgment in reacting to changes and the environment are influenced by their emotion and perception. In addition the adverse effects of the aroused emotion can bias one's judgment and even change the priority of any action that were taken. Some psychologists proposed that emotion influences judgment directly by serving as experiential and bodily information regarding how one feels about the object that affect judgment [1]. Such experiential information sometimes were processed faster than rational thought resulting in an irrational driving situation. Emotions can affect drivers' concentration, invoking recklessness and disabling appropriate self-control.

Driving activity in general can be viewed as a complex cognitive task that constitutes attention, information processing and decision making task. These driving activities depend on the vehicle performance and environment itself. The current road condition, behavior of the other road user, the vehicle state and resources information such as petrol and duration of excursion are some examples of important inputs that are needed by the driver to make necessary choice. Bellet et al. [2] formulated driver mental representation as an interaction between long term memory (LTM) with working

memory (WM) and its environment. It is illustrated in Figure 1. The WM is a structure that hosts active knowledge instantiated through mental representations to serve the activity in progress that focus on the driver's current goal. The LTM stores driving knowledge permanently in latent state from learning process. These driving knowledge couples with information from the environment that are gathered through perceptive process construct the mental representation of the driving situation. Moreover, the construction of such mental representation is complemented with the aid of emotionally influenced reasoning process which includes decision making, anticipation and action planning. The output is then manifested through the realization of driver behavior. However, in some cases the cognitive resources may be influenced by some pre-cursor emotion, which in-turn will affect the output. From this understanding, we can hypothesize that even the smallest disturbance can potentially lead to severe repercussions [3]. Effect from illicit drug and alcohol consumption, fatigue, stress and excessive uncontrolled emotion exhibited formidable distraction level to the drivers.

A driver emotion indicator application based on speech can be incorporated into an intelligent motor vehicle system as a preventive measure from any

unwanted incidents. This is based on the hypothesis that emotion can influence the way the driver behave and emotional effect can be quantified using driver's speech. Any abnormalities in the driver behavior can be identified so that an appropriate warning mechanism can be developed. Such system can provide means to monitor the drivers' behaviour in complex traffic situation with practical implementation.

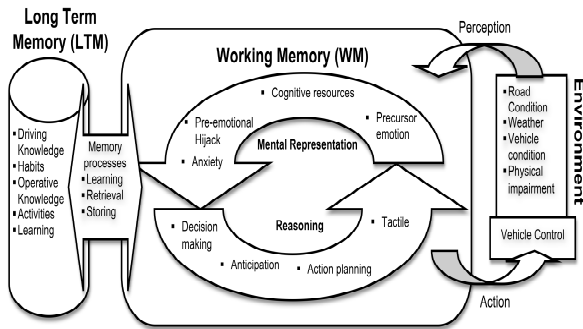


Fig. 1: Elementary Architecture of the Driver Cognitive System

1.1 Emotion and Perception

Relational model of emotion theory basically attempt to provide a description of the emotion generation process based on complex reaction that engages both the human mind and body [4]. Scherer provides a review of the various research paradigms into the vocal communication of emotions based on the modified Brunswik's functional model of perception [5]. The model represents briefly the concept of emotion perception through speech. The process begins by encoding or expressing the emotional states of the speaker using certain distinctive voice and speech characteristics which can be measured objectively.

In this paper, emotion is defined as a mental state that arises spontaneously from external stimuli that often accompanied with dynamic physical expressions which is heavily influenced by culture to serve a purpose of a particular situation. Hence, the emotion that we discussed in this paper refers to biologically induced emotion but its condition of elicitation and its content is conditioned by learning.

1.2 Archetypal of Emotion

In our judgment, it is best to accept that some emotions are more universal than others which can give us a direction in developing the speech emotion recognition system. Thus, in this study we focus on three basic emotions; namely: anger, sadness and happiness with neutral acting as reference to the

emotionless state. These emotions were selected based on the psychologists' view [6] and can be assigned in three different quadrant of the affective space model [7] as shown in Figure 2. The x axis represents the valence primitives value, the y axis represents the arousal primitives values and the z axis represents the dominance primitives values. Furthermore, based on Ververidis and Kotropoulos review of 32 emotional speech databases [8], anger, sadness and happiness emotion are the top most emotion recorded and analyzed by the researchers in the field of speech emotion.

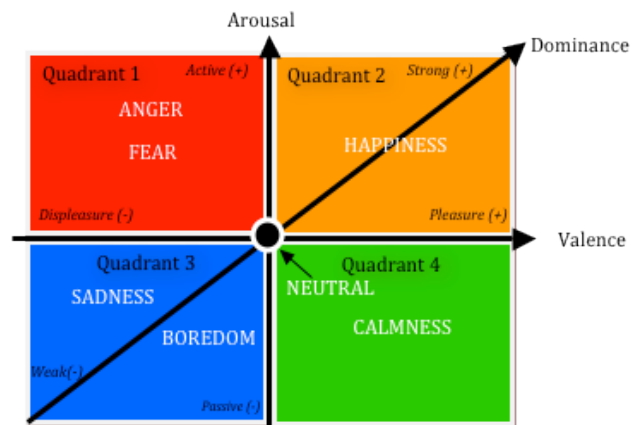


Fig. 2: The Affective Space with Emotion Labeling.

1.3 Organization of the paper

This paper is organized in the following manner; Section 2 presents the dataset used for the experiments. The Real-time Speech Driving dataset (RtSD) and the three culturally-influenced speech emotion datasets are described. The Valence-Arousal Analysis System (VAAS) are proposed in Section 3. Such system is constructed based on adoption of Mel Frequency Cepstral Coefficient (MFCC) as features extraction method coupled with Multi-layer Perceptron (MLP) and Adaptive Network-based Fuzzy Inference System (ANFIS). Experimental results of DBS and speech emotion identification using the VAAS approach are reported in Section 4. Finally, Section 5 concludes the paper with summary and discussion of the DBS and speech emotion understanding.

2 Data Corpus

The driver behavior state data are derived from the Real-time Speech Driving Dataset (RtSD) [9]. It was collected from 11 Singaporean drivers with age ranging from 20 to 54 years. Four DBS were selected for analysis; namely: talking through mobile, laughing, sleepy and normal driving

condition. The detail description of the dataset can be obtained from [10].

Subsequently, although the focus of this paper is not on culture influence on driver behavior analysis, we adopted three different culturally-influenced speech emotion dataset; namely: Berlin dataset [11], NTU_American dataset [12] and NTU_Asian dataset as a solid basis for training data to illustrate the well-generalized VAAS capability. The dataset selected represents European, American and Asian culture respectively for three basic emotions; namely; anger, sadness and happiness with neutral acting as reference for emotionless. These dataset form a foundation for supervised learning as the emotions had already known prior to the testing. The result of the speech emotion recognition system using these dataset had been reported in [9,12].

3 Valence-Arousal Analysis System (VAAS)

Assuming that each of the basic emotion can be represented by angry emotion, $f(a)$, sad emotion, $f(s)$ and happy emotion $f(h)$ in the emotion space \mathfrak{R} and each emotion is linear then each of the driver behavior state can be represented as Equation

$$f_n(d) = k_d + a_n^d f(a) + s_n^d f(s) + h_n^d f(h) \quad (1)$$

where d is the driver behavior state, in this case it is denoted as normal, sleepy, talking and laughing, n is the instant time sampled of the speech signal s , k_d is the constant for a particular driver behavior state, a_n^d is the coefficient for the basic emotion anger function, s_n^d is the coefficient for the basic emotion sadness function and h_n^d is the coefficient for the basic emotion happiness function. From Equation (3.1), each of the three driver behavior state can be represented as basis function to the basic emotion which in turn can be defined as part of the valence and arousal of the emotion space. Figure 6.15 shows a simplified block diagram of the proposed driver behavior analysis system where the learning process uses the basic emotion including the neutral as emotionless state while the testing process uses the rules generated from the learning system as part of the mechanism to determine the value of the basic emotion responsible for the driver behavior state.

The basic emotion approach of Equation (1) provides generated rules by training speech emotion data using fuzzy neural network from supervised

learning. Different driver behavior state speech data were tested with this set of rules to determine the influential emotion that have effect on the driver while he/she was in that particular state. Input features from drivers speech signal were extracted using MFCC features extraction method. These information were then used as testing data to be classified by the previously trained classifier. From the result, weightage of the each basic emotion were calculated. Since neutral is the common class for both basic emotion recognition and normal DBS in driver behavior analysis experiments dataset, it is excluded from being an accuracy performance indicator.

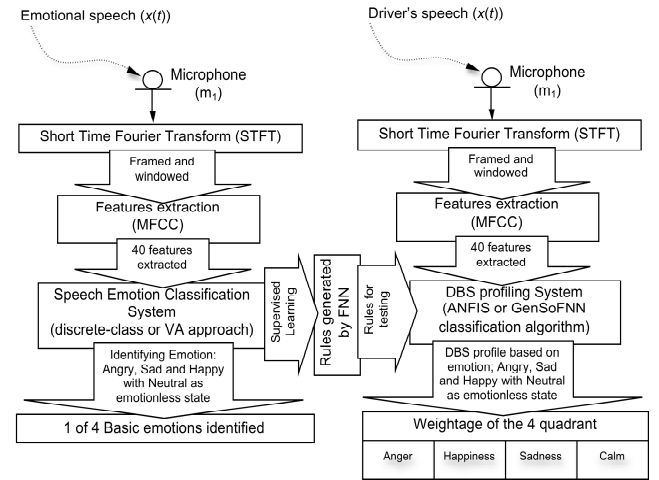


Fig. 3: Block Diagram of the Extension of Basic Speech Emotion Recognition System for Driver Behavior Analysis

Based on the notion that each basic emotion can be represented in the VA axes of the affective space model, Equation (1) can be rewritten as Equation (2).

$$D = N + \alpha E \quad (2)$$

where D is the driver behavior state defined in vector space of the valence and arousal axes, N is the neutral centroid vector in the valence and arousal space that shifted the origin due to cultural influenced, E is the valence and arousal space vector of the classifier output that determine the behavior state and α is a scalar constant due to the translation of signals from the basic emotion to driver behavior state (DBS). Clearly Equation (2) is simpler and able to facilitate the analysis of the DBS based on the affective space model understanding. The resulting effort can easily be translated for dynamic analysis of the driver behavior profile. Subsequently, the valence and arousal values are then transformed into a data-driven affective space

model. The origin and the quadrants of the affective space model need to be formed to facilitate emotion segregation.

In order to find the individual emotion centroid, particularly neutral, k-means clustering method [13] was adopted. K-means clustering treats each instance as a point with its own coordinate. The point coordinate in this case refers to its valence (x-axis) and arousal (y-axis) value. K-means finds the centroid by computing the distance between points and the cluster center iteratively so as to minimize the within-cluster sum of squares. Thus, it can be said that the centroid of the cluster is the point to which the sum of distances from all points in that cluster is minimized. However, instead of using Euclidean distance as a tool to compute the sum of absolute differences, Manhattan distance technique or also known as the L_1 distance was employed. Each centroid is the component-wise median of the points in the neutral cluster.

Consider Manhattan distance, d_1 between 2 vectors of p and q in 2-dimensional affective space model with fixed Cartesian coordinate system. Centroid location can be calculated using Equation (3) where the result is the sum of the lengths of the projections of the line segment between the points onto the coordinate axes.

$$d_1(p, q) = \|p - q\|_1 = \sum_{i=1}^2 |p_i - q_i| \quad (3)$$

where $p = (p_1, p_2, \dots, p_n)$ and $q = (q_1, q_2, \dots, q_n)$ are vectors. For instance, the L1 distance between coordinate (p_1, p_2) and (q_1, q_2) can be calculated as $|p_1 - q_1| + |p_2 - q_2|$. Figure 4 depicts the emotion and its centroid for NTU_American dataset using ANFIS classifier. It is noted that neutral distribution is concentrated in the center compared to anger that bias towards Quadrant 1, happiness bias towards Quadrant 2 and sadness bias towards Quadrant 3. This finding was consistent to the affective space model description [Error! Bookmark not defined.].

Once the origin and the quadrant of the affective space model has been established, the boundary between emotion and neutral need to be determined. The neutral-emotion boundary refers to the soft boundaries segregating neutral and emotion in its respective quadrant. For instance, anger and neutral need to be discriminate for quadrant 1. However, this situation posed as a challenge since neutral ratio in emotion is unknown due to the fact that neutral is embedded in all emotion. To date, no one has

concretely established the neutral region of the affective space model. However, in this paper, we are interested in segregating emotion and neutral rather than establishing neutral region per se. Hence, we developed the soft-boundary of neutral-emotion based on the assumption that the testing instant belongs to the class with nearest distance of either emotion centroid or neutral centroid in its respective emotion quadrant.

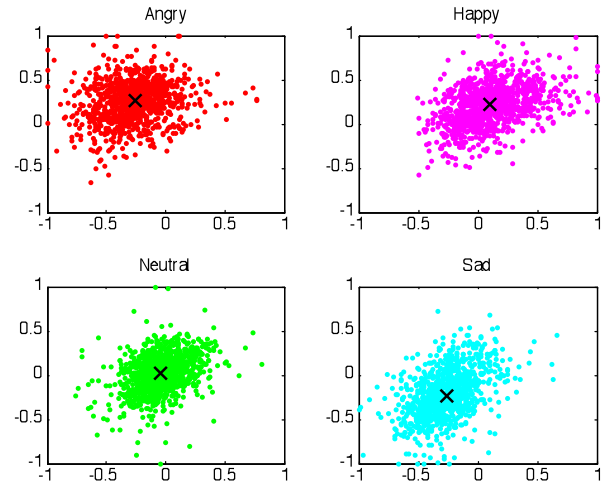


Fig. 4: NTU_American Dataset Emotions and its Centroid using ANFIS Classifier.

Consider each instance of the speech emotion the classifier produces output (x_i, y_i) where x_i represents the valence and y_i represents the arousal value respectively. The emotion centroid (x_e, y_e) for $e = (anger | happiness | sadness | neutral)$ emotion can be derived using the K-means approach discussed earlier. The centroid of neutral $(x_{neutral}, y_{neutral})$ serves as a hard boundary segmenting positive and negative values of emotion primitives.

Assuming instance i with distance to emotion centroid denotes as $d_{i,e}$ and neutral centroid $d_{i,neutral}$. Both distances can be computed by Equation (3.4).

$$d_{i,e} = \sqrt{(x_i - x_e)^2 + (y_i - y_e)^2} \quad (4)$$

$$d_{i,neutral} = \sqrt{(x_i - x_{neutral})^2 + (y_i - y_{neutral})^2}$$

where x_i and y_i are the valence and arousal values derived from the classifier outputs (testing instances), $x_{neutral}$ and $y_{neutral}$ are the valence and

arousal values of the neutral centroid and x_e and y_e are the valence and arousal values of the respective emotion centroid. Thus the instance i is more likely to be a member of a particular class if it has the nearest distance between the neutral and respective emotion centroid as depicted by Equation (5), where the boundary between neutral and the respective emotion can be derived.

$$B = \frac{d_{i,e}}{d_{i,neutral}} = \begin{cases} \geq 1 & i = neutral \\ < 1 & i = emotion \end{cases} \quad (5)$$

From Equation (3.5), for $B \geq 1$ means that the VA value of the instance is nearer towards neutral centroid, thus the instance is classified as neutral. On the other hand, for $B < 1$ refers to instance that is in favor of respective emotion class as it is nearer to the emotion centroid. It is important to bear in mind that this neutral-emotion boundary (B) is only applicable for the respective emotion of a particular quadrant where i instance is located. This is to ensure the balance of the x and y axis of the affective space model is not disrupted. Figure 4 shows NTU_American dataset with neutral-emotion boundary for each quadrant with exception of Quadrant 4. This is because calm centroid is not available since calm is not included in the analysis. The soft boundary approach allows different size of neutral centroid to be used for segregating different emotion in the different quadrant. However, for simplification, neutral region in quadrant 4 is assumed to have similar value as valence of Quadrant 2 and arousal of Quadrant 3. Red dashed line in Figure 4 shows the projection of neutral for calm region in Quadrant 4.

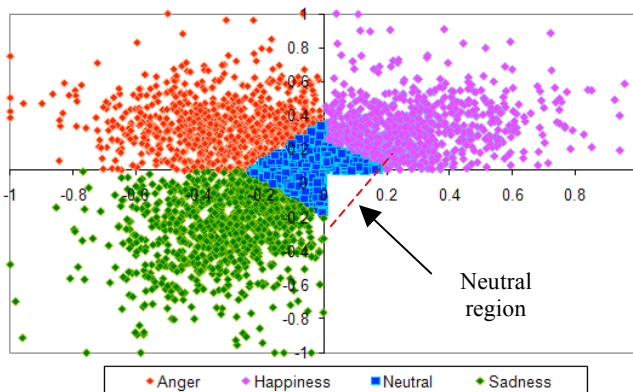


Fig. 3: NTU_American Dataset with Neutral-Emotion Boundary Using ANFIS

5 Results and Discussion

DBS-emotion correlation experiment was

conducted by training a single culture-influenced dataset to test the RtSD dataset in determining the driver behavior. Experimental results shown in Table 1-3 using both ANFIS and MLP classifier illustrate variation results recorded due to the cultural influenced that bias towards certain emotion. NTU_American emotion correlation result in Table 1 shows the inclination towards quadrant 4 with mean accuracy performance of 56% and 34% using ANFIS and MLP respectively. The Berlin correlation result in Table 2 is more evenly distributed for all four quadrants with performance ratio of 14.5% to 36.7% for ANFIS and 18.5% to 36.9% for MLP. In addition, the NTU_Aasian correlation result in Table 3 is more biased towards quadrant 2 with 45.1% accuracy using ANFIS. However, the bias for the same dataset is divided into quadrant 2 and quadrant 4 with mean accuracy of 42.2% and 52.5% respectively using MLP classifier. High distribution in quadrant 4 and quadrant 2 are expected for the correlation experimental result as the drivers were feeling calm and excited during the driving exercise and can be observed especially during normal DBS with exception of Berlin dataset trained with ANFIS classifier.

Table 1: NTU_American Emotion Correlation Result

	ANFIS				MLP			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Talk	7	21	18	54	11	30	35	24
Laug	10	28	14	48	17	36	23	24
Nrml	4	35	4	57	3	45	5	47
Slpy	4	16	15	65	9	41	8	42

Table 2: Berlin Emotion Correlation Result

	ANFIS				MLP			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Talk	22	48	11	19	23	37	23	17
Laug	18	41	13	28	21	43	22	14
Nrml	46	27	13	14	19	44	21	16
Slpy	18	30	21	31	11	23	39	27

Table 3: NTU_Aasian Emotion Correlation Result

	ANFIS				MLP			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Talk	27	44	13	16	1	39	5	55
Laug	21	50	12	17	2	41	4	53
Nrml	7	62	8	23	1	60	4	35
Slpy	7	24	30	39	1	29	4	66

Laughing is consistently distributed highest in quadrant 2 that representing happiness. This result is

in-line with psychologists' hypothesis that laughter is often cue for happiness especially in verbal communication without face-to-face interaction. The highest performance is recorded in correlation experiment using NTU_Asian dataset coupled with ANFIS classifier with performance of 50.2%. The most interesting finding is that sleepy DBS is highly distributed in calm region with 45.0% and 44.8% using ANFIS and MLP respectively. This result again empirically verified the psychologists' hypothesis that sleepy should fall in quadrant 4. The result in Figure 5 shows that although talking through mobile (denotes as talking) is assumed to pose as a distraction thus eliciting angry emotion, it is not necessarily perceived as an unwanted interruption by the driver. This perception is conveyed in the drivers' speech respond which resulting different emotion rather than angry to be elicited. In fact the result in the experiment shows that quadrant 2 distribution was the highest for the talking DBS which can be interpreted as the drivers were welcoming the call although in the same time the drivers needs to delegate his/her attention between responding appropriately to the caller conversation and to drive safely.

6 Conclusion and Future Work

The VAAS has shown potential in recognizing driver behaviour state (DBS) using the three basic emotions and neutral as the basis function. Results from table 1-3 show the potential of detecting sleepy driver from just the speech signals. This will enable car manufacturer to used such concept in recognizing fatigue driver or driver under stress in future. Tremendous amount of work can be carried out to complement the work of Psychologist to an engineering field to benefit the community at large.

References:

- [1] G. L. Clore, & J.R. Huntsinger, How Emotions Inform Judgment and Regulate Thought, *Trends in Cognitive Science*, Vol. 11, 2007, pp. 393-399.
- [2] T. Bellet, B. Bailly-Asuni, P. Mayenobe & A. Banet, A Theoretical and Methodological Framework for Studying and Modelling Drivers' Mental Representations, *Safety Science*, Vol. 47, 2009, pp. 1205 – 1221.
- [3] C. Nass, I-M. Jonsson, H. Harris, B. Reaves, J. Endo, S. Brave & L. Takayama, Improving Automotive Safety By Pairing Driver Emotion and Car Voice Emotion. *In: Proceeding of Conference on Human Factors in Computing Systems (CHI'05)*, 2005, pp. 1973-1976, Portland, OR, USA.
- [4] R. S. Lazarus, & B. N. Lazarus, *Passion and Reason: Making Sense of Our Emotions*, USA: Oxford University Press, 1996.
- [5] K. R. Scherer, Vocal Communication of Emotion: A Review of Research Paradigm, *Speech Communication*, Vol. 40, 2003, pp. 227-256.
- [6] R. Cowie & R. Cornelius. Describing the Emotional States That Are Expressed in Speech, *Speech Communication*, Vol. 40, 2003, pp. 5-32.
- [7] H. Schlosberg. Three Dimension of Emotion, *Psychological Review*, Vol. 61, Issue 2, 1954, pp. 81 -88.
- [8] D. Ververidis & C. Kotropoulos, A State of the Art Review on Emotional Speech Databases, *In: Proceeding of 1st Richmedia Conference*, 2003, pp. 109–119, Lausanne, Switzerland.
- [9] M. Khalid, A. Wahab & N. Kamaruddin, Real Time Driving Data Collection and Driver Verification Using CMAC-MFCC.” *In: Proceeding of the 2008 International Conference on Artificial Intelligence (ICAI '08)*, 2008, pp. 219-224.
- [10] N. Kamaruddin & A. Wahab, Driver Behavior Analysis Through Speech Emotion Understanding *In: Proceeding of the 2010 IEEE Intelligent Vehicle Symposium (IV 2010)*, pp. 238-243, 2010, San Diego, California, USA
- [11] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier & B. Weiss, A Database of German Emotional Speech”, *Proc. INTERSPEECH-ISCA*, 2005, pp. 1517-1520.
- [12] Kamaruddin N. & Wahab A. (2009). Features Extraction for Speech Emotion, *Journal of Computational Methods in Science and Engineering (JCMSE)*, Vol. 9, Issue Supplement 1, 2009, pp. S1 – S12.
- [13] Jain, A. K. Data Clustering: 50 Years beyond K-means. *Pattern Recognition Letters*, Vol. 31, Issue 8, 2010, pp. 651-666.