Modeling Spontaneous Affect Using Continuous Data

Katrina Ysabel Solomon
Center for Empathic Human-Computer Interactions
De La Salle University
Manila, Philippines
katrina.solomon@delasalle.ph

Avelino Alejandro Latorre
Center for Empathic Human-Computer Interactions
De La Salle University
Manila, Philippines
avelino.latorre@delasalle.ph

Juan Paolo Tensuan
Center for Empathic Human-Computer Interactions
De La Salle University
Manila, Philippines
paolo90@gmail.com

Jocelynn Cu
Center for Empathic Human-Computer Interactions
De La Salle University
Manila, Philippines
jiji.cu@delasalle.ph

Merlin Suarez
Center for Empathic Human-Computer Interactions
De La Salle University
Manila, Philippines
merlin.suarez@delasalle.ph

***Abstract*—Human affect is spontaneous. Thus, it is more appropriate to use spontaneous data rather than acted data when dealing with the identification of human affect. Moreover, emotions can be interpreted under various intensities, which show that categorical labeling of affect is not sufficient in representing them. This paper aims to present the dimensional annotation procedure of the Filipino Multimodal Emotion Database (FilMED2) database as well as to show results of the prediction of dimensional affect through built models from spontaneous data. The labels used were valence and arousal. Audio features namely pitch, intensity, formants, energy, and Mel-frequency Cepstral Coefficients (MFCC) were extracted from the clips. Facial point distances were extracted through the use of the Active Appearance Model (AAM). The models were built from the "best" annotations as the result of the inter-coder agreement. Accuracy was tested by means of computing for the error values using regression algorithms namely Multilayer Perceptron and Support Vector Machine for Regression. Since the classifiers were able to predict the valence and arousal values with minimal errors, the dimensional labels accurately represented the affect as much as the previous categorical labels did.**

Keywords- Emotion detection, dimensional labeling, continuous data, multimodal affect model, face recognition, voice recognition.

I. INTRODUCTION

Recent studies in empathic computing encourage the use of dimensional description of affect (e.g., [9]). Using categorical classes cannot fully represent the extensive continuum of human emotion [19]. The dimensional description of affect do not have direct equivalent in categorical description; Accordingly, a single affect, represented by a point in a multi-dimensional coordinate space, can provide information on affective attributes such as gradation (i.e., the ebb and flow of affect), richness and variations over time. This representation is free from linguistic constraint and semantic constraint. This concept is also supported by a number of researchers in the field of psychology (e.g., [12], [13]). Thus, aside from categorical labels, well-known spontaneous multimodal emotion databases like the Sensitive Artificial Listener Database (SAL-DB) [5] and the SEMAINE database (SEMAINE-DB) [17] are annotated with dimensional labels, i.e., using the arousal (how excited or apathetic the emotion is) and valence (how positive or negative the emotion is) space.

Using continuous data with dimensional labels for human affect analysis is not without its challenges. Several commonly used machine learning techniques like the Hidden Markov Model (HMM) and Support Vector Machines (SVM) cannot handle continuous affect recognition because these techniques mainly classify nominal classes. Many automatic affect recognition systems using audiovisual data and the aforementioned approaches still perform manual segmentation of their input data (e.g., [18]). Continuous affect recognition requires that the spontaneous data has to be segmented and pre-processed automatically prior to affect modeling. Issues like inter-coder agreement and baseline detection need to be dealt with.

The use of spontaneous data was suggested by [16]. They specified four guiding principles on building an ecologically valid emotion database. These principles require that emotions captures should show (1) genuine emotion, (2) emotion in interaction, (3) gradation, and (4) richness. According to their research, it is important that the emotion database should include materials generated by people experiencing genuine emotion, which is effectively captured when one is interacting with another person. This also entails that emotions typically expressed in everyday life could be mixed and may change over time. According to [10], certain acted expressions differ in appearance and timing compared to spontaneously occurring ones. Also, there are expressions that people cannot perform when they are acted; however, these actions can be done spontaneously [10]. Thus by modeling spontaneous affect, various human expressions may be represented more accurately that just relying on acted instances. Moreover,

spontaneous expressions better reflect human behavior in real-life situations. Expressing pure and intense emotions is a rare occurrence [16], which is why it is more advisable to spontaneous data; however, it is difficult to work with spontaneous data specifically in its data gathering procedure because for humans to express spontaneous behavior, they must be within a real working environment. The issue of privacy is a factor and data acquisition is challenging as the subjects get conscious and act in control when noticing that their actions are being captured, thus defeating the purpose of having spontaneous data.

This approach to affect modeling enables models to become one step closer towards growth-centric affect modeling development. Growth-centric models are self-improving model which adapt and automatically learn and discover emotion without compromising accuracy [21]. Such applications of these models are e-Health systems, intelligent tutoring systems, and smart homes.

This research aims to build multimodal affect models of spontaneous data with the Filipino Multimodal Emotion Database (FiLMED2) [4] database. FiLMED2 is composed of spontaneous emotional speech and facial expression. It is an extension of the original Filipino Multimodal Emotion Database developed at the Center for Empathic Human-Computer Interactions (CEHCI) Laboratory of the College of Computer Studies, De La Salle University [4], which was created to serve as training data for the design and development of multimodal human affect analyzers.

II. RELATED WORKS

There are already several existing researches regarding spontaneous affect modeling using continuous data such as [20] and [19]. Each work has a different approach because there has not been an agreed methodology when it comes to this type of affect modeling.

The work of [20] made use of the SAL database. Three (3) modalities were used which are the face, voice, and shoulder gestures. Twenty (20) facial points were extracted from the face, five (5) points from the shoulders, and fifteen (15) features from the voice. Valence and arousal values were used in labeling.

For [19], the database used was also SAL. In this work, only one modality was utilized which is the voice modality. All in all, 4,843 features were extracted. Valence and arousal were also the dimensions considered for labeling.

III. METHODOLOGY

A. Data Preparation and Pre-processing

900 clips containing single and blended emotion labels from FilMED2 were included in the dimensional annotation process. Six coders, three male and three female, were invited to annotate the FilMED2 data using the FEELTRACE [15] tool, which allows labeling of video only, audio only, audio-video clips, and images with valence and arousal dimensions. The coders were students from De La Salle University and they were trained to use FEELTRACE before annotating the clips. The criteria in choosing the audio-video clips included in the database were: subject is in a realistic situation, emotion is clearly expressed by facial expression and/or voice, frontal facial regions are clearly visible (i.e., no occlusions of the left/right eyebrows, left/right eyes, nose, and mouth), and the subject is speaking in Filipino. The spontaneous data were collected from the following reality television programs: *Pinoy Big Brother Season One Collection* and the *Philippines' Scariest Challenge*.

Before the coders proceeded with the annotation process, the proponents made sure that they were empathic enough to give dimensional representations of emotion to the FilMED2 clips. This was determined through Baron-Cohen's Empathy Quotient Test [1] which measures empathic ability through self-report. All the coders scored average to above average results.

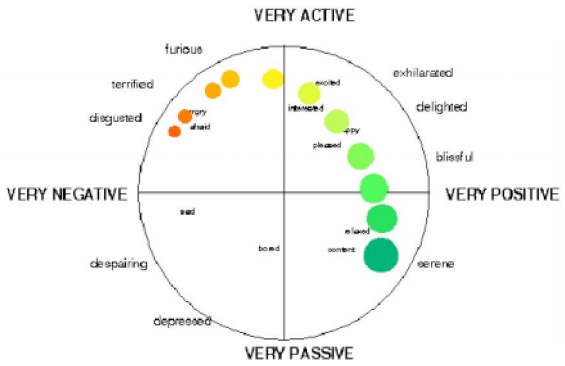

Figure 1. FEELTRACE's activation-evaluation space [15]

The coders were allowed to watch the clips before engaging in the annotation process, which involves plotting the emotion by holding the left mouse button down on FEELTRACE's activation-evaluation space as the clip is playing. Emotion dimensions of valence (measure of how positive or negative) and arousal (measure of how active or how passive) with respect to the time were the output of FEELTRACE. Categorical labels (such as joy, anger, fear, disgust, surprise, and sadness) served as guides for the coders' dimensional labeling as they are placed on the annotation space. The activation-evaluation space of FEELTRACE is shown in Figure 1.

The annotations done by the coders were not comparably similar to each other due to the differences of human perception. Moreover, the use of dimensional representations of valence and arousal does not explicitly show the agreement between the coders, thus inter-coder agreement measures were carried out. Pre-processing of the files was done before computing for the agreement, which mainly involves binning, a process that groups the annotations based on the number of frames per second of the clip. There were some instances where there are missing values from the annotations due to distractions or possible technical problems – these were

handled through the use of Not a Number (NaN) notations; however, these values are not be included in the computation of agreement.

The sign-agreement method (SAGR) and the correlation formula (COR) were used to compute for the agreement between the coders, which determined the set of annotations that was used for building the model. For each coder file, it was paired with the other coder files and the COR and SAGR were computed on the valence and arousal values, which were averaged with the other pairs with the same file to get the average COR and the average SAGR. The averaged results served as the basis of comparison to which file has the highest agreement between the other coder files. The annotation file with the highest agreement was used as part of the training set as it served as the best representation of annotations for a given clip.

The sign-agreement formula calculates the agreement between a pair of coders based on how much they agree on a per-frame basis. Its value ranges from zero (0.0), denoting no agreement, up to one (1.0), denoting full agreement. SAGR is formally defined as:

$$SAGR = \frac{\sum_{f=0}^{|frames|} e\left(C_i(f).val, C_j(f).val\right)}{|frames|} \quad (1)$$

Ci and *Cj* correspond to the pair of coders, *Ci(f).val* and *Cj(f).val* are the valence values for coder *Ci* and *Cj* at frame *f*. The function *e* checks whether the values between two coders on the same frame agree. This is defined as:

$$e(i,j) = \begin{cases} 1 \; if \; (sign(i) == sign(j)) \\ 0 \; else \end{cases} \quad (2)$$

The SAGR is computed for the valence values since the valence deals with how positive or how negative the coders perceive the clips to fall under. The values of the annotations for valence are disregarded; only the sign of the annotation is considered for this agreement computation. The SAGR can also be expressed in percent form which would denote the percentage of agreement between entities.

The COR describes the relationship between the annotations and the value can be positive or negative. Best case scenario for the COR is to have a positive value which depicts that one coder's annotations follow the same or nearly the same pattern as the other coder being compared to, thus drawing the visible line of agreement. It is also possible to get a negative value which denotes an inverse relationship between the annotations. The correlation is computed as:

$$COR = \frac{N\sum xy - \left(\sum x\right)\left(\sum y\right)}{\left[N\sum x^2 - \left(\sum x\right)^2\right]\left[N\sum y^2 - \left(\sum y\right)^2\right]} \quad (3)$$

Where *N* is the number of pairs of valence or arousal annotations, *x* is the valence or arousal annotation of one coder and *y* is the valence or arousal annotation of the pair coder. While SAGR is computed for the valence values, the COR is computed for the arousal values since it represents the perception of how intense or how passive the expressions are. With positive COR results, it denotes that there is a similar trend between the coders, thus having more agreement between each other.

Figure 3 is a representation of the "best" annotations on the activation-evaluation space as well as their categorical equivalent. As a simple verification for consistency of the pure emotions, the coder's annotations were compared with the categorical labels in a way that the annotations are near the quadrant of the label. Since the activation-evaluation space of FEELTRACE consists of categorical labels as guides for the annotation process, this could also be the basis for comparison of the categorically labeled clip.

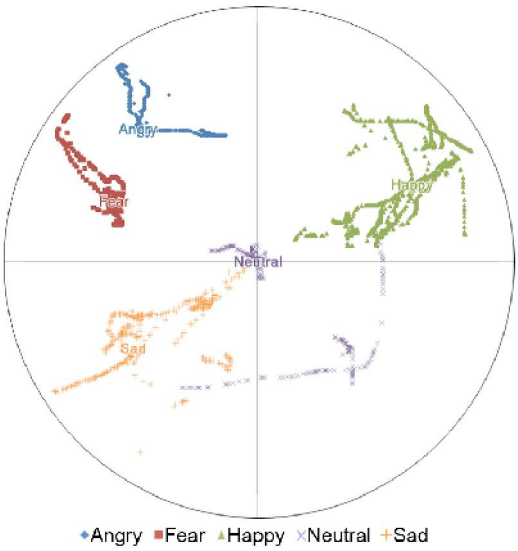

Figure 2. The annotations as plotted in the activation-evaluation space along with their categorical equivalent

The annotations for clips that were previously labeled as angry, fear, happy, and sad were located in the same quadrant as their categorical equivalent. In terms of distance, the dimensional labels are close to their counterparts. The annotations for the neutral state, however, are scattered along the four quadrants. This may be attributed to the different perception of each coder.

The annotations of the six coders have been binned and have been processed for their inter-coder agreement measurements to find out which the "best" annotations are between each of them. Table 3 shows the average computed sign-agreement between the valence values and the computed correlation between the arousal values across all the sets of clips previously labeled with categorical emotions.

TABLE I. SAGR AND COR RESULTS

Emotion	SAGR	COR
Angry	1.0	0.2496
Happy	1.0	0.1043
Neutral	0.5396	0.0431

Fear	1.0	0.1341
Sad	1.0	0.5386
AVERAGE	**0.8973**	**0.2140**

The annotations between the coders for the chosen clips have been in full agreement for almost all the clips categorized under their previous categorical labels. The average SAGR denotes that the annotations were 89.73% in agreement with each other based on how positive or negative the clips were. The average COR signifies that there is a similar pattern with how the six coders perceived the intensity of emotion expressed by the spontaneous clips due to its positive value. Annotations for the neutral clips garnered a smaller value for the SAGR and COR. This is because the coders have identified some of the neutral clips as either one of the other emotions based on how they perceived them.

B. Feature Extraction

Since the affect model will be multimodal, two modalities were considered, which are the face and the voice. According to [14] vocal cues are important in the expression of emotion. Vocal emotion expression has powerful effects on interpersonal interaction and social influence [11]. For building the model, both prosodic and spectral features were considered. Prosodic features involve the way how sounds are acoustically realized and disambiguate a text transcription such as a question or statement or add new information like the speaker's emotional state [7]. The extracted prosodic features include pitch (frequency of a sound), intensity (power of the voice in speech), energy (absolute values of a speech sample over a defined period), and formants F1, F2, F3. Twelve Mel-frequency Cepstral Coefficients (MFCC) were taken as the spectral features where in these represent the speech signals. The features were extracted from the FilMED2 clips using the software PRAAT [2].

Facial features were also considered, which will be through facial point distances. The features were extracted through the use of the Active Appearance Model (AAM). The Active Appearance Model (AAM) is an approach that matches a statistical model of the shape and appearance of the face to a new image. Although it fits a new face on the original inputs, it is disregarded and only the facial point distances are utilized. 68 points were taken from the face and from these points, 170 facial point distances were derived to serve as the features for the data set. Figure 3 shows a sample image fitted with the points and the distances by the Active Appearance Model.

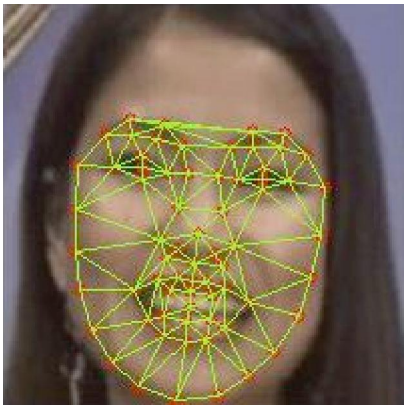

Figure 3. An image fitted with the points by AAM

The features were extracted per frame. The videos from the FilMED2 database have varied frame rates. To establish uniformity, videos were converted so that all have a frame rate of 25 frames per second. Feature extraction was then done every 0.04 seconds, representing a frame. Each frame with its corresponding features is treated as in instance. The face data set contains one hundred seventy (170) features per instance while the voice data set contains eighteen (18) features per instance.

C. Model Building and Validation

The number of instances for the face and voice data sets are listed in Table 4. These instances were balanced to ensure that all emotions, as determined by the previous categorical labels, are well-represented. Balance between the positive and negative emotions was also noted. Annotations for neutral clips were classified under positive emotions since most annotations selected through agreement were located near the calm and relaxed states.

The instances for the voice data set are greater than the face data set due to the fact that some frames cannot be tracked by AAM even though there is a face and therefore, not producing the facial point distances. While the face is not detected, the speech present in the clip can still be evaluated by PRAAT thus, producing a greater amount of instances for the voice data set.

TABLE II. INSTANCES FOR EACH DATA SET WITH REGARDS TO THE PREVIOUS CATEGORICAL LABEL

Emotion	**Face**	**Voice**
Angry	385	592
Fear	391	555
Happy	1173	1800
Neutral	328	449
Sad	335	544
TOTAL	**2612**	**3940**

Models were built from data sets which consist of the "best" valence-arousal annotations from the computation of SAGR and COR, extracted facial point distances and the extracted audio features from the clips of the FilMED2 database. This was done to be able to observe the accuracy of the classification of dimensional emotions through valence

and arousal predictions. The WEKA [22] machine learning software was utilized to build the models. Since the aim is to predict dimensional labels, the use of algorithms with regression capabilities were used. Based on previous experiments, the Multilayer Perceptron as well as the Support Vector Machine for Regression performed the best among other tested regression algorithms such as kNN with Regression and Linear Regression, thus, they were chosen to build the models. 10-fold cross validation was utilized as well to assess the accuracies of the models.

- Support Vector Machine with Regression

The idea of SVM came from finding an optimal hyperplane that can separate classes with maximum margin. Computations are performed directly in the input space through the use of kernels; it does not involve computations within the high-dimensional space. The classification finds a boundary that can divide the input space into regions. Given that the data are known to be separable by a hyperplane, the minimal distance of a sample to the decision hyperplane is used to define the margin [3].

- Multilayer Perceptron

Multilayer Perceptron (MLP) is a feedforward artificial neural network which makes use of backpropagation as its learning technique for the network, which searches the space of possible hypotheses to reduce the error in the network fit to the training examples. Perceptrons take real-valued inputs and calculates combinations of these inputs under a specific threshold. MLP is an extension of the perceptron, which can also contain hidden layers of neurons. Neurons found in the MLP network use weighted sums to yield the neuron's activation value. MLP could distinguish data which are not linearly separable [8].

IV. RESULTS AND OBSERVATIONS

Two models were built from the data sets containing the annotations and features. The clips were hand-picked to balance the amount of positive and negative emotions as the elements within FilMED2 were previously labeled with categorical labels namely angry, fear, happy, neutral, and sad. The first model contained the "best" annotations from computing the agreement between the coders and audio features. The second model includes the facial features as well as the "best" annotations. Following the Circumplex of A ect [12], emotions can be identified according to the quadrant it is placed on. These quadrants are: low arousal positive, high arousal positive, low arousal negative and high arousal negative.

As the valence and arousal values were predicted, the mean absolute error, root mean squared error, and correlation coefficient are computed and used as a value for comparison and measure of accuracy. These measures are presented in Table 5 and Table 6.

- Mean Absolute Error (MAE)

MAE is the average of the differences between the predicted and actual annotations. It is defined as:

$$MAE = \frac{\sum |predicted - original|}{N} \quad (4)$$

- Root Mean Squared Error (RMSE)

RMSE is an error measure which is affected by deviations between actual and predicted values. The formula is defined as:

$$RMSE = \sqrt{\frac{\sum (|predicted - original|)^2}{2}} \quad (5)$$

- Correlation Coefficient (COR)

COR depicts the presence of patterns from the predictions and the actual data. The formula is defined in (3).

TABLE III. RESULTS FOR THE VOICE MODEL

Classifier	Measure	Valence	Arousal
SVM with Regression	**MAE**	**0.2911**	**0.1999**
	RMSE	**0.3923**	**0.2774**
	COR	**0.6573**	**0.6071**
MLP	MAE	0.403	0.2896
	RMSE	0.4801	0.364
	COR	0.429	0.2211

TABLE IV. RESULTS FOR THE FACE MODEL

Classifier	Measure	Valence	Arousal
SVM with Regression	**MAE**	**0.458**	**0.2416**
	RMSE	**0.5353**	**0.3261**
	COR	**0.3456**	**0.0942**
MLP	MAE	0.5015	0.2713
	RMSE	0.5572	0.3517
	COR	0.029	-0.0155

Based on the gathered errors and correlations for both the face and the voice models, the voice model yielded a better result than the face model. This may be attributed to the large dimensionality of features for the face, having one hundred seventy (170) features as opposed to the voice that only has eighteen (18) features. The algorithms predicted valence and arousal with considerably small error due to the inter-coder agreement measures that selected the "best" annotations or the most fit to use for prediction.

Decision-level fusion was then performed to combine the results from both models. This test involved assigning different weights to each model and finding out what is the best combination of weights to achieve the best result possible. Results are presented in Table 7 in terms of MAE and RMSE.

TABLE V. RESULTS FOR THE FUSION OF THE TWO MODELS

Weights		MAE		RMSE	
Face	**Voice**	**Valence**	**Arousal**	**Valence**	**Arousal**
0%	100%	0.4135	0.2776	0.4723	0.3215
10%	90%	0.4134	0.2735	0.4644	0.3129
20%	**80%**	**0.4163**	**0.2703**	0.4604	0.3052
30%	70%	0.4225	0.2683	0.4603	0.2983
40%	**60%**	0.4305	0.2666	**0.4638**	**0.2925**
50%	50%	0.4407	0.2656	0.4705	0.2876
60%	40%	0.4529	0.2655	0.4802	0.2840
70%	30%	0.4666	0.2658	0.4928	0.2817
80%	20%	0.4844	0.2666	0.5085	0.2804
90%	10%	0.5049	0.2682	0.5271	0.2809
100%	0%	0.5264	0.2699	0.5485	0.2827

In terms of the combined MAE, the assignment of 20% for the face and 80% for the voice yielded the best result. If the basis is RMSE, then the combination of 40% face and 60% voice produced the best result. The weight for the voice model for both measures is heavier compared to the weight of the face model mainly because the performance of the voice model is better than the face model as presented earlier in the separate model evaluation.

The next table presents the results of each work using the classification by SVM with regression as it is the common algorithm implemented by the researches. The accuracy of results is measured in terms of the Root Mean Squared Error (RMSE).

TABLE VI. RESULTS OF EXISTING WORK ON SPONTANEOUS AFFECT MODELING USING CONTINUOUS DATA

Author/s	Face (Valence, Arousal)	Voice (Valence, Arousal)
Nicolaou et al. [20]	0.21, 0.27	0.25, 0.26
Wöllmer et al. [19]	N/A	0.45, 0.35
Solomon et al. [This work]	0.54, 0.33	0.39, 0.28

The work of [20] gained smaller errors which can be attributed to the features that were considered. This work made use 170 facial point distances while [20] only made use of 20 facial points for the face modality. Not all of the 170 distances may have been relevant when the model was built. Some may have even be the ones that contributed to the higher error obtained. This behavior can be observed as well in the voice modality. [19] garnered the highest error rate which may be caused by the large number of features used. The work of [20] also gained the lowest error for this modality. They only used 12 features as opposed to this work and [19] which considered 18 and 4,843 features respectively.

V. CONCLUDING REMARKS AND FUTURE WORK

The results show that Support Vector Machine for Regression has better prediction accuracy than Multilayer Perceptron. The computed errors for SVM with regression show that dimensional labels were predicted with suitable correctness; however, the algorithms are yet to be tested with more instances which may improve or worsen the prediction errors.

Some prediction errors can be attributed to inconsistencies in annotations and features. Some frames may be detected by the AAM tracking algorithm but without corresponding audio features. There are also frames with audio features but no facial features because the face cannot be recognized by the algorithm. Because of this, it is expected that results will get better when the results for both models are combined.

Future work may involve the use of other fusion methods to merge the prediction from the voice model and the face model. Furthermore, additional clips will be included to represent spontaneous emotion better. Feature selection and consideration of more, or possibly less, features will also be performed to achieve the highest prediction accuracy. There can also be a consideration for additional modalities. As seen in [20], their use of three modalities produced lesser errors in prediction.

With FilMED2 having clips labeled as mixed emotions, a good venue for further research is the investigation of dimensional labeling in mixed emotion instances. Since there is a possibility that emotions can occur simultaneously, dimensional labeling should not be limited to dealing with pure emotions alone.

ACKNOWLEDGMENT

We would like to acknowledge the De La Salle University – University Research Coordination Office (DLSU-URCO) for the support funds; the College of Computer Studies (CCS), the Center for Empathic Human-Computer Interactions (CEHCI), the Filipino Department, and the Psychology Department for providing us the theoretical background and raising research issues concerning this project. This research is supported in part by the Department of Science and Technology – Philippine Council for Industry, Energy and Emerging Technology Research and Development (DOST-PCIEERD).

REFERENCES

[1] S. Baron-Cohen and S. Wheelwright, "The empathy quotient (EQ): an investigation of adults with asperger syndrome or high functioning autism, and normal sex differences," *in Journal of Autism and Development Disorders*, 2004.

[2] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (version 5.2.09)," computer program retrieved from http://www.praat.org/, 2005.

[3] W. Cao, L. Li, and X. Lv, "Kernel function characteristic analysis based on support vector machine," *in 2007 International Conference on Machine Learning and Cybernetics*, 2007, pp. 2869-2873.

[4] J. Cu, M. Suarez,, and M. Sta. Maria, "A filipino multimodal emotion database," *in Proceedings of the Int'l Workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality, in assoc. with the 7th Int'l Conf. on Language Resources and Evaluation*, 2010, pp. 37-42.

[5] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. McRorie, et al., "The HUMAINE Database: Addressing the

Collection and Annotation of Naturalistic and Induced Emotional Data," *in ACII '07: Proceedings of the 2nd International Conference on Affective Computing and Intelligent Interaction*, Berlin, Heidelberg: Springer-Verlag, 2007, pp. 488–500.

[6] M. A. Hearst, "Support vector machines," *IEEE Intelligent Systems*, 13 (4), 1998, 18–28.

[7] K. Koumpis, & S Renals, "Automatic summarization of voicemail messages using lexical and prosodic features," *in ACM Trans. Speech Lang. Process*., 2 (1), 1, 2005.

[8] T. M. Mitchell, Machine Learning, McGraw-Hill Science/Engineering/Math, 1997.

[9] M. A. Nicolau, H. Gunes, and M. Pantic, "Automatic segmentation of spontaneous data using dimensional labels from multiple coders," *in Proc. of Int'l Workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality, in assoc. with the 7th Int'l Conf. on Language Resources and Evaluation*, Valletta, Malta, 2010, p. 43-48.

[10] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," *in Proc. IEEE Int'l Conf. Multimedia and Expo*, 2005, pp. 317–321.

[11] J. Rong, Y.-P. Chen, M. Chowdhury, and G. Li, "Acoustic features extraction for emotion recognition," *in 6th IEEE/ACIS International Conference on Computer and Information Science*, 2007, p. 419 -424.

[12] J. A. Russell, "A circumplex model of affect," *in Journal of Personality and Social Psychology*, 39 (6), pp. 1161-1178, 1980.

[13] K. R. Scherer, "Psychological models of emotion," *in J. Borod (Ed.), The Neuropsychology of Emotion*, Oxford/New York: Oxford University Press, 2000, pp. 137-166.

[14] K. R. Scherer, "Vocal communication of emotion" in *Speech and Communication*, 2003, 40(1–2), pp. 227–256.

[15] M. Schröder, R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, and M. Sawe, "'FEELTRACE': An Iinstrument for recording perceived emotion in real time," *in Proceedings of the ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research*, Belfast: Textflow, 2000, pp. 19-24.

[16] M. Schröder, E. Douglas-Cowie, R. Cowie, "A new emotion database: considerations, sources and scope," *in Proceedings of the ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research,* Belfast: Textflow, 2000, pp. 39-44.

[17] Semaine Database. Retrieved from http://semaine-db.eu/

[18] M. Song, M. You, N. Li, and C. Chen, "A Robust Multimodal Approach for Emotion Recognition" *in Neurocomputing*, 71 (10-12), pp. 1913–1920, 2008.

[19] M. Wöllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, et al., "Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies," *in Proc. of 9th Interspeech Conf*, 2008, pp. 597-600).

[20] M. A. Nicolaou, H. Gunes, M. Pantic, "Continuous Prediction of Spontaneous Affect from Multiple Cues and Modalities in Valence-Arousal Space," *in IEEE Transactions on Affective Computing*, vol. 2, pp. 92-105, 2011.

[21] R. Legaspi, S. Kurihara, K. Fukui, K. Moriyama, M. Numao, "Self-improving empathy learning," *in Proc. of ICITA 2008*, 2008.

[22] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, "The WEKA Data Mining Software: An Update," *in SIGKDD Explorations*, 11(1), 2009.

AFFILLIATIONS

The authors are members of the Center for Empathic Human-Computer Interactions of the College of Computer Studies from De La Salle University.