Recognizing Affect in Spontaneous Filipino Laughter

Jocelynn Cu
jiji.cu@delasalle.ph

David Manangan
davidmanangan@gmail.com

Jason Wong
wong.jason.trinidad@gmail.com

Christopher Galvan
chrgalvan@gmail.com

Michael Sanchez
mike.ponce.s@gmail.com

Merlin Suarez
merlin.suarez@delasalle.ph

ABSTRACT
Laughter is an important social signal that conveys other affect aside from happiness. In this study, we attempt to identify what types of affect are present in spontaneous Filipino laughter. Experiment results confirmed that Filipino affects such as *natutuwa*, *nasasabik*, *kinikilig*, *nahihiya*, and *mapanakit* are indeed present in laugher. Automatic feature extraction techniques are applied on the face and laughter audio signals to extract relevant features for modelling the affect. Machine learning techniques are used to classify the different affect. Tests revealed that the face is more important in determining the valence in laughter, and audio is more important in determine the arousal level in laughter. An attempt to plot these affect in Russel's circumplex model of affect resulted to these emotions converging in the positive valence and high arousal quadrant, which suggests that laughter is generally a positive affect, even if negative affect is embedded within it.

Keywords
Affective computing, laughter analysis, audio and video processing

1.0 INTRODUCTION

Affect is present in our day-to-day communication [10]. With this fact, researchers had been studying how to apply human affect analysis on human-computer interaction (HCI) systems. The goal of HCI systems is to become human-centered, which means that deliberate tasks are performed according to a user's affective state. Thus various research efforts are focused on building and developing automatic affect recognizer.

Researchers who study affect from audio signals usually focus on the speech or linguistic aspect of the signal. However, research in cognitive science states that the non-linguistic or paralinguistic sounds are also significant in identifying non-basic emotions like distress, anxiety, and boredom [12]. Murmurs, yawns, and laughter are just some examples of paralinguistic sounds that can also supply significant affect information. Laughter, which is considered one of the most noteworthy paralinguistic sound [5], can occur either by itself or interspersed with speech. It is both spontaneous and conversational [11].

Although paralinguistic sounds are important in affect recognition, only a handful of research is focused in this area. Devillers and Vidrascu [4] studied affect in laughter that is part of spoken dialogue. Their study found that there are three types of affect present in laughter: positive, negative, and ambiguous laughter. The study also found that negative laughs have more unvoiced sounds compared to positive laughs. Pantic et. al., [9] distinguishes laughter from speech using audio-visual inputs. Based on their study, visual information is more important than auditory information in distinguishing laughter from speech. However, combining both modalities will produce more reliable results than using a single modality. Escalera et. al. [5] also used a multimodal approach in recognizing laughter. Locally, a similar study was made by [7], from which a tool was developed to automatically separate laughter segments from speech segments. Another example is the study of Alonzo et. al. [1] which attempted to discover underlying affect present in Filipino laughter. The motivation for the study [1] is the fact that Filipinos are fond of laughing. Apart from using laughter as an expression of happiness and a form of defense mechanism, a whole lot of other emotions, both positive and negative, can be perceived from laughter. Their findings revealed that the commonly occurring emotions in Filipino laughter include *natutuwa*, *nasasabik*, *kinikilig*, *nahihiya*, and *mapanakit*, loosely translated in English as happiness, excitement, ticklish, shy, and hurtful, respectively. Nevertheless, Alonzo et. al. [1] used a database of posed female and male laughter in their study. Since posed affect differ from spontaneous affect in several aspects, it was argued that the results may not accurately reflect the actual affect present in the signal. The same argument also applies to the appropriateness of the label used to describe the affect; for example, excitement and ticklish do not fully express the affect described by *nasasabik* and *nahihiya*. These are the motivation of this research.

This study intends to discover affect in Filipino laughter that is expressed spontaneously and naturally. Primarily, a database of spontaneous Filipino laughter needs to be collected and annotated. Then relevant audio-visual features need to be extracted to build the affect model. A reliable affect model can classify emotions, even if these are embedded in laughter.

This paper is organized as follows. Section 2 explains the approach used to build the laughter database and the affect model. Section 3 presents the findings of the experiments and analysis of results. Section 4 concludes the paper.

2.0 METHODOLOGY

The objective of this research is to identify emotions that are present in spontaneous Filipino laughter. This task requires a useful database and a reliable affect model.

2.1 Building the spontaneous Filipino laughter database

Both male and female subjects were invited to participate in the data collection. Pairs of volunteers are placed in separate rooms and are asked to chat with each other through Skype. While chatting, their facial expressions and voice are recorded. The subjects are given video clips and comic strips to help sustain

their conversation. The clips also serve to induce affective states found by [1] to be present in laughter. Figure 1 shows the physical setup of the data gathering process used in this study.

Fig 1. Physical setup of the data gathering process.

Data collection resulted to 139 useful laughter clips acquired from three (3) test subjects – one male and two females. Table 1 shows the number of useful clips taken from each subject.

Table 1. Number of laughter clips per subject.

Test Subject	Number of Clips
Subject 1	57
Subject 2	52
Subject 3	30

Table 2. EQT scores of coders.

Coder	EQT Score
Coder 1	67%
Coder 2	71%
Coder 3	67%

These clips were segmented automatically using LASER [7]. Clips range from 1 second to 5 seconds. Each laughter clip was annotated with Filipino categorical labels and dimensional labels by three coders, who scored at least 53 points[1] in the Baron-Cohen EQ Test [2]. Test scores of each coder are presented in Table 2. All coders used the FeelTrace [3] annotation tool to labels the clips in the valence-arousal dimensions. Since each coder may give different labels to a single clip, sign agreement [8] was used to arrive at a single label for each clip. Sign agreement measures the degree of agreement between the coders. Figure 2 shows an example of annotation done by three coders on a single clip, plotted on Russel's Circumplex Model of Affect. The *x*-axis represents valence (i.e., positive or negative affect) and the *y*-axis represents arousal (i.e., active or passive affect). Only those clips with a sign agreement value of 75% for both arousal and valence are selected to be part of the laughter database.

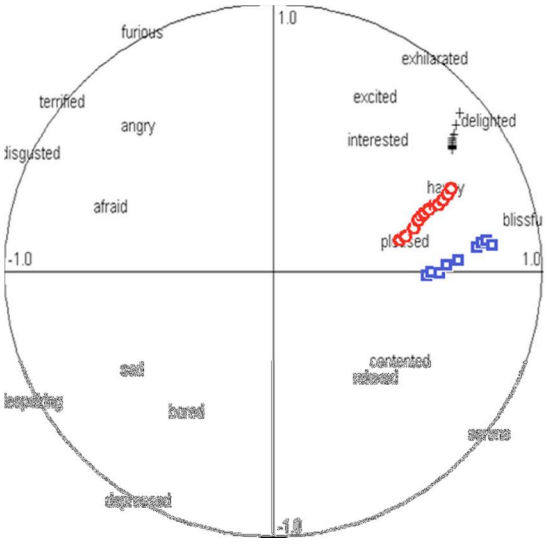

Fig 2. Annotations of three coders on a single laughter clip (Legend: "+" refers to annotations of Coder 1, "○" refers to annotations of Coder 2, and "□" refers to annotations of Coder 3)

2.2 Creating the affective laughter model

AV clips are separated into laughter and non-laughter segments using LASER [7]. Non-laughter segments are discarded. Laughter segments undergo feature extraction process, where significant features from the video and audio frames are computed. The resulting features vectors are then used to build the laughter model.

Features are extracted from both the image and the audio signal. Facial features are extracted with the help of the Active Appearance Model, where facial points are automatically placed on several regions of the face. Distances between these facial points are tracked to determine facial movement (refer to Figure 4). The collection of these distances, which totals to 170, makes up the facial features vector. Prosodic and spectral features are extracted from the audio using PRAAT. PRAAT is a software tool that implements speech processing techniques to compute the pitch, energy, intensity, formants, and Mel-Frequency Spectral Coefficients (MFCC) of the audio signal. A total of 30 prosodic and spectral features make up the audio features vector.

Since the data is made up of valence-arousal pairs, regression techniques were used to classify the features. For this study, the following techniques are used: linear regression, multilayer perceptron, and support vector machine-regression. Several models are built: face model using valence data, face model using arousal data, audio model using valence data, and audio model using arousal data. Output from each models are combined using decision-level fusion approach.

[1] The Baron-Cohen EQ Test was utilized by the research team to identify coders who can annotate clips consistently, meaning, their perception and interpretation of an emotional display is more coherent.

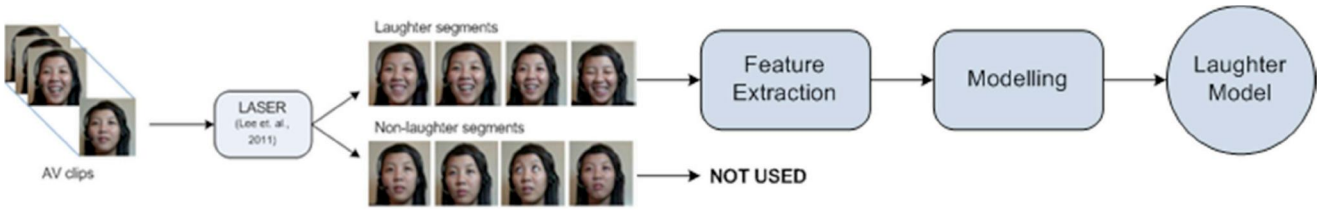

Fig 3. Affect model building process

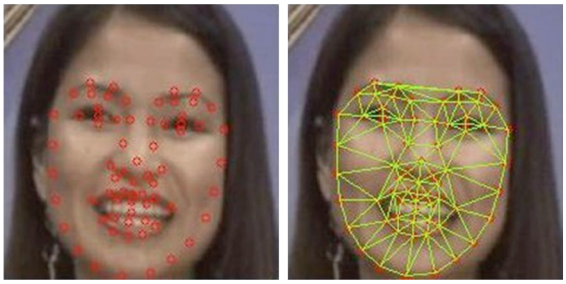

Fig 4. Image of a face with facial points plotted (left) and distances between facial points computed (right).

3.0 EXPERIMENT RESULTS AND ANALYSIS

The dataset used in the experiments is summarized in Table 3. There are 300 instances of audio and 2,507 instances of video. Each audio instance is made up of 30 prosodic and spectral features. Each video instance is made up of 170 facial feature points. WEKA, a machine learning tool, was used to build and train the model for classification. The 10-fold cross validation is used to measure the accuracy of the regression approach. The mean absolute error (MAE), which measures the distance between the predicted values from the original value, is used to assess the error rate of the classification process.

Table 3. Dataset used in the experiments.

	Audio	Video
Instances	300	2,507
Features	30	170

The first experiment is to determine which classifier is best for audio data and determine the minimum set of audio features needed for classification. Table 4 showed the results of running the valence and arousal values over several regression algorithms.

Table 4. Results of classifying audio signals only, where LR – Linear Regression, SVM – Support Vector Machine, MLP – Multilayer Perceptron, and FS – feature selection.

Classifier	Valence		Arousal	
	w/o FS	w/ FS	w/o FS	w/ FS
LR	0.1564	0.1520	0.1589	0.1564
SVM	0.1553	0.1657	0.1617	0.1665
MLP	0.2323	0.1530	0.2637	0.1903

Based on the results of classification using all features, the following points can be observed:

- SVM is better at classifying valence data compared to LR, with a difference of 0.0011 in their MAE values.
- LR is better in classifying arousal data compared to SVM, with a difference of 0.0039 in their MAE values.
- MLP yielded the largest error when classifying both valence and arousal values.

Hoping to improve the results, automatic feature selection in WEKA was used to derive the minimum feature set that will produce minimum error. These were presented in Table 4 as well. As expected, a minimum feature set improved the performance of all classifiers as evidenced by the decrease in their MAEs, except SVM. Interestingly, only SVM errors increased after the feature selection process. A look at the resulting minimum feature set revealed that only the mean pitch, intensity, energy and selected MFCCs (i.e., #2, 3, 4, 5, 6, 8, 11, 12, 13) are used in the minimum set. It seems that SVM relies on all features to make better classification. However, as to which specific feature, further experiments and analysis has to be conducted. With feature selection, MLP performs better than SVM, but LR still achieved the least error, regardless if it is valence or arousal data.

The second experiment aims to determine which classifier is best for facial features and determine which facial points are relevant for classification. This time, only LR and MLP are tested. The results of this experiment are presented in Table 5.

Table 5. Results of classifying facial points only.

Classifier	Valence		Arousal	
	w/o FS	w/ FS	w/o FS	w/ FS
LR	0.1481	0.1542	0.1498	0.1604
SVM	0.0512	0.0512	0.0506	0.1261

When classifying facial points using all 170 facial features, experiment results suggest that SVM is approximately 1.5 times better at classifying facial points compared to LR. Even with feature selection, which resulted to only 15 facial features, SVM is consistently better albeit the difference in error rate is not that significant when classifying arousal data. The result of feature selection on significant facial points is shown in Figure 5 and summarized in Table 6.

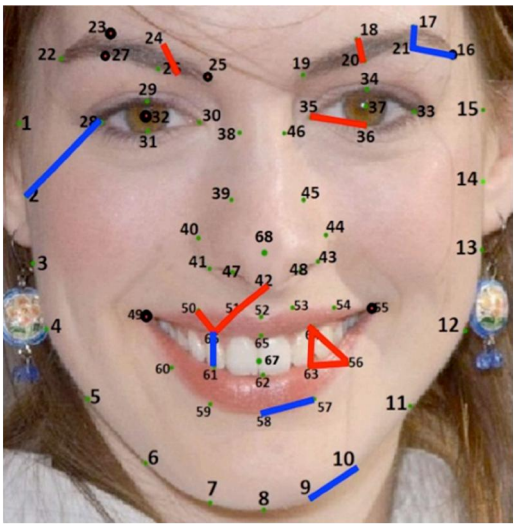

Fig 5. Facial features relevant to classifying affect are highlighted.

Table 6. Summary of valence and arousal facial features relevant to classifying affect in laughter.

Valence	Arousal
Dist (2,28) cheek-right eye	Dist (18,20) left eyebrow
Dist (9,10) chin	Dist (24,26) right eyebrow
Dist (16,21) left eyebrow	Dist (35,36) left eye
Dist (17,21) left eyebrow	Dist (42,51) nose to lip
Dist (57,58) lower lip	Dist (50,66) upper lip
Dist (61,66) upper-lower lip	Dist (51,66) upper lip
	Dist (56,63) lower lip
	Dist (56,64) upper-lower lip
	Dist (63,64) upper-lower lip

The third experiment was carried out to determine how much weight should be placed on which modality to achieve minimum error in classification. Summary of results are present in Table 7.

Table 7. Results of applying decision-level fusion on face and audio data.

Distribution of A/V weights	Valence	Arousal
100 / 0	0.1174	0.1619
90 / 10	0.1094	0.1496
80 / 20	0.1035	0.1384
70 / 30	0.0993	0.1317
60 / 40	0.0958	**0.1261**
50 / 50	0.0923	0.1267
40 / 60	0.0889	0.1303
30 / 70	0.0857	0.1339
20 / 80	0.0828	0.1380
10 / 90	0.0803	0.1425
0 / 100	**0.0788**	0.1478

Fusion results indicate that video is more important in determining the valence of laughter, i.e., how positive or negative a specific laughter is. This may be due to the fact that when a person expresses positive or negative laughter; there are noticeable changes in the upper facial region (above the nose). On the other hand, the combination of 60% audio and 40% video determines the arousal level of the laughter, i.e., how active or passive the laughter is. This may be attributed to the volume of expressed laughter that can be heard. The audio reveals the extent of how the person feels when he/she laughs. These findings are consistent with that of [6].

The final experiment attempts to plot all instances of each laughter type on Russel's Circumplex Model of Affect, in an effort to explain the actual affect conveyed by the person and find the most appropriate corresponding English label. Figure 6 shows the plot for each affect type.

Reviewing the annotated AV clips, it was observed that both *natutuwa* and *nasasabik* exhibit breathing dominant laughter with moderated pitch and evident rise-and-fall of intonation. *Kinikilig* is characterized by high pitch, close breathing intervals (similiar to panting), and loudness of the sound. *Mapanakit* is characterized by breathing dominant laughter with sudden change in moderation at the onset of laughter. And nahihiya exhibits long breaths and high pitch.

The subtle differences between these affects in laughter are also evident in the posed laughter database of Alonzo et al (2010), although it is more subtle in the spontaneous form. This may be attributed to the recognized laughter display rules in Filipino. Unless these different laughter affect are exaggerated, they are difficult to differentiate, especially if contextual information is removed. Such is the case in the posed laughter database, where the actors are explicitly asked to express the emotion in laughter, with no contextual information provided; while in the spontaneous database, annotation was done after the laughter clips had been segmented that also effectively removed any contextual information. Based on the results, it cannot be determine which English label best describes the affect displayed.

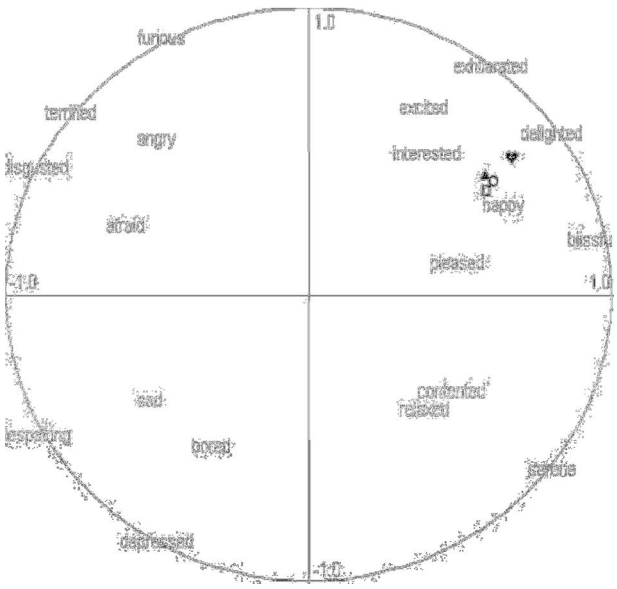

Fig 6. Mapping of different affect found in Filipino laughter, where "♥" represents *nahihiya*, "▲" represents both *natutuwa* and *nasasabik* (overlap of two affect), "○" represents *mapanakit*, and "□" represents *kinikilig*.

4.0 CONCLUSION

This study is the first attempt at differentiating various affect found in Filipino laughter. The study was able to confirm the following:

- that the face is more reliable in distinguishing the valence of different affects in laughter, where a minimum facial feature set can be derived;
- that the audible intensity of laughter is more useful in determining the level of arousal of the laughter, where a minimum prosodic and spectral feature set can also be derived;
- that generally, laughter is characterized with a positive valence, even if there is negative affect embedded in it.

Although the study was able to accomplish some of its objective, i.e., to analyse affect in spontaneous Filipino laughter and map Filipino laughter affect in the circumplex model; it was not able to conclusively determine the appropriate English label for the Filipino affect. Further studies are needed to properly characterize each affect in laughter, additional experiments are needed to determine which specific quantitative features that can be attributed to specific affect. Of course, a more comprehensive spontaneous database is also needed to accomplish these tasks.

5.0 ACKNOWLEDGEMENT

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 Center for Language and Translation (CeLT), the Department of English and Applied Linguistics (DEAL), 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).

6.0 REFERENCES

[1] Alonzo, J., Campita, J., Lucila, S. and Miranda, M. (2010) Discovering affect in Filipino laughter using audio features. In 3rd Int'l Conf HumanCom 2010, 1 – 6.

[2] Baron-Cohen, S. and Wheelwright, S. (2004) The Empathy quotient: an investigation of adults with asperger syndrome or high functioning autism, and normal sex difference. Journal of autism and developmental disorders. 34(2), 163 – 175.

[3] Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M. and Schroder, M. (2000) FeelTrace: an instrument for recording perceived emotion in real-time. In Proc. of ISCA workshop on speech and emotion, 19 – 24.

[4] Devillers, L. and Vidrascu, L. (2007) Positive and negative affectal states behind the laughs in spontaneous spoken dialogs. In interdisciplinary workshop on the phonetics of laughter, 37 – 40.

[5] Escalera, S., Puertas, E., Oriol, P. and Radeva, P. (2009) Multimodal laughter recognition in video conversations. In computer vision and pattern recognition workshop CVPR 2009.

[6] Kanluan I., Grimm M., & Kroschel K. (2008). Audio-visual emotion recognition using an emotion space concept. In 16th

[7] Lee, J., Miranda, M., Cu, J. and Suarez, M. (2011) Automatic laughter segmentation in meetings. In Proc. of 11th PCSC.

[8] Nicolaou, M., Gunes, H. and Pantic, M. (2010) Continuous prediction of spontaneous affect from multiple cues and modalities in valence and arousal space.

[9] Pantic, M. and Petridis, S. (2008). Audiovisual discrimination between laughter and speech. In acoustics, speech and signal processing, 5117 – 5120.

[10] Riekki, J., Yu, C. and Zhou, J. (2009) Expression and analysis of affect: survey and experiment. In symposia and workshops on ubiquitous, autonomic and trusted computing.

[11] Truong, K. and Van Leeuwen, D. (2005) Automatic detection of laughter. In Interspeech 2005.

[12] Zeng, Z., Pantic, M., Roisman, G. and Huang, T. (2009) A survey of affect recognition methods: audio, visual and spontaneous expressions. In IEEE Trans on pattern analysis and machine intelligence, 31(1).

7.0 AUTHOR'S INSTITUTIONAL AFFILICATIONS

Center for Empathic Human Computer Interactions, De La Salle University, 2401 Taft Avenue, Manila, Philippines, 1004.