FILIPINO EMOTION CLASSIFICATION IN SPEECH SIGNALS BASED ON AUDIO FEATURES AND TRANSCRIBED TEXT

Ramon L. Rodriguez
Joel P. Ilao

Introduction

The Emotions of persons are manifested in their facial expressions, through the voices and words they use in conversations. Nevertheless, facial expression, or voice expression alone are not sufficient indicators to detect emotions. Emotions can be controlled and even concealed by persons depending on whom they are associated with. In the study of Elfenbein and Ambady [1] they documented evidence of an in-group advantage, which means that people are generally more accurate at judging emotions when the emotions are expressed by members of their own cultural group rather than by member of a different cultural group. In this case, culture affects the judgment of emotion. While the annotators for this study belong to the same group of students, we hope that they are better judges of what emotions are present in a conversation or in the sentences. On the one hand, Filipinos are not so expressive with their negative emotions because of the concept of "Pakikipagkapwa" and "Pakikisama." For example, even if they are angry; they still maintain the good way of saying words or maintaining the pitch of their voice. We tend to control our negative emotions. On the other hand, Filipinos are very expressive when they are happy. Filipino emotional displays tend to be influenced by our culture [2]. Detecting and modelling of Filipino emotions is a challenge for researchers since determining these emotions require a multimodal approach.

A number of affective computing researches were done in the past. Most of these researches were done using single modal approach, which most of the time focused on facial expression or voice signals [3]. The success of the affective system relies on effectiveness of the corresponding affective model. While different approaches are used in modelling and detecting emotions, mostly single modal, there are some researches that used multimodal approach. In this study we used text mode (transcribed text from the speech signal), audio signal, and the audio-video signal. We want to know which approach is better in classifying emotion given the cultural background of the raters. We investigated if Filipino words alone are good indicators in classifying emotion, since we assumed that our annotators can understand well the context of what the speaker in the audio file are saying primarily because they are on the same age bracket with the

annotators. Maybe they are using the same words when they want to convey emotion to others. We hypothesized that the combination of different modes is better in classifying and detecting Filipino emotion.

In this paper we used different approaches in detecting and modelling Filipino Emotion by relying only on audio file and detecting emotion solely on words. Section 2 discusses the related works in the field of affective computing and emotion recognition, section 3 discusses the tools, section 4 the methodology and sections 5 and 6 include results and discussion, conclusion and future works.

Related Works

This section discusses the related works on emotion recognition, Emotion annotation and emotion classification that use single, bimodal and multimodal approaches. EmoTV1 is a corpus of video clips recorded from French V channels containing interviews. EmoTV1 studies the influence of the modalities on the perceptions of emotions. The Anvil tool [17] was used to annotate this corpus for three conditions: audio without video, video without audio, and video with audio. The result of the annotations identified 14 labels of emotions [4].

Emotion recognition can also be done using speech signals and physiological measure (biosignals). The study of Jongwa [5] modeled the emotion using different speech features such as Mel-frequency cepstral coefficients (MFCC) and biosignals. Prosodic features such as intensity, pitch, and duration of utterance are commonly explored speech signal features in developing emotion recognition system. A rule-based method for emotion recognition was proposed by Chen, 2000 as cited in the study [5]. The data used in the work of Chen contained two foreign languages (Spanish and Sinhala). Some annotators did not understand the language, so their judgment was based on vocal expression without considering the linguistic/semantic content of the speaker. In our study, we asked the student to listen to the audio file for them to determine the appropriate emotion for the particular audio file. Aside from listening, we also asked the students to read the words transcribed from the audio file for them to determine the emotions. In this case the students take into account the linguistics and semantic content both for audio and text based mode.

Another study dealt with recognizing spontaneous emotion in a natural dialogue [6]. In this study, the researchers used pitch and energy for acoustic feature and images feature for modelling the emotion. The study classified the emotion as positive and negative, and they concluded that multi-modal approach enhances the model for emotion recognition. In this paper, five types of emotion were used: anger, fear, happiness, neutral and sadness. The basis for using these emotions was the FILMED2 database

annotation done in the previous study [2]. We used multi-modal approach via text and audio features to identify the five types of emotion.

Another study was conducted to classify emotion, i.e., anger, happiness, neutrality, excitation and sadness. In this study the bimodal approach using facial expression and the voice was utilized. Thirty (30) facial features and acoustic feature such as prosodic and spectral were used in classifying emotion [7].

Tools Used

This paper used several tools to clean the audio file and to extract the features. Audacity software package was used to remove the unwanted signal of an audio file. jAudio software was used in feature extraction while WEKA software package was used in classifying and building a Filipino emotion model. The following are the short description of the tools used.

Audacity is a free, easy-to-use, multi-track audio editor and recorder for Windows, Mac OS X, GNU/LINUX and other operating systems [8]. In this paper Audacity was used to remove the unwanted signal or noise of the audio file.

jAudio is a new framework for feature extraction designed to eliminate the duplication of effort in calculating feature from an audio signal. It also provides a unique method of handling multidimensional features and new mechanism for dependency handling to prevent duplicated calculations [9]. In this research we used jAudio for feature extraction of the audio signal. This software is commonly used for music information retrieval but in this paper we used it as feature extractor for audio signal.

The Waikato Environment for Knowledge Analysis (Weka) is a comprehensive suite of Java class libraries that implement many state-of-the-art machine learning and data mining algorithms. We used Weka classification algorithm to develop a model of Filipino emotions, particularly decision tree, neural network, and support vector machine algorithm implementation [10].

Methodology

Experimental Set-Up

The audio files that we used in this paper were already labelled with the corresponding emotion in the prior study [2]. The aim of the study is to be able to discover other modes of emotion modelling. The other mode can be through the spoken words and through listening to speech. We conducted a quasi-experiment, to 50 selected computer science and information technology students from Mapua Institute of Technology. Out of 50

students, 20 are females and 30 are males with ages ranging from 18 to 19 years old. Two sets of experiment were conducted. In the first set, the students were asked to read the transcribed words and to put the appropriate label of emotions according to their context. In the second set, the students were asked to listen to the audio file and to label the emotion according to their perception. Each student was given the paper where the transcribed words were written on the left side. On the right side of the paper were boxes with the corresponding emotions. The students labelled the appropriate emotion when they read the words. In the same manner, the students who listened to the audio files checked the appropriate boxes that corresponded to the emotion. There was only one set of student annotators for both words and audio files.

The result was collated based on the words and the audio file. The pre-annotated dataset was re-labelled depending on the result of the annotation. To determine the emotion that should be labelled for each file the score of the students was used as the basis in determining what emotion should be labelled for each file.

For instance in audio file 1, 35 out of 50 students label them as angry, 10 labelled them as sad, 2 labelled them as neutral, 1 labelled them as fear, and 3 labelled them as sad. Given the score, audio file 1 has an emotion label of angry. This approach was used both for the audio labelling and word labelling. We used the Fleiss Generalized Kappa to compute for the student agreement. The kappa of the student who listened to the audio file is 0.48 while the kappa of the student who labelled the emotion based on words is 0.40. Looking closely at the agreement per emotion, it is noted that the agreement of the students for emotion labelled as angry for both words and audio file is higher compared to other emotions. The reasons behind this are the words used and for the audio file the speaker is talked with a high pitch.

Dataset

The study used the FILMED2 database [2] which is composed of clips taken from television series Pinoy Big Brother (PBB) Season one and the Philippine's Scariest Challenge (PSC). Ninety-percent (90%) of the clips in this database came from the Philippine Show Pinoy Big Brother (PBB) season one from August 21 to December 10, 2005. PBB created a social environment where the participants were forced to live with a group of strangers and to interact with one another. The audio files used in this study ranging from one (1) to 17 seconds consist of 53 audio files; 10 labelled as angry, 10 labelled as fear, 11 labelled as happy, 10 labelled as neutral and 12 labelled as sad. The labels were done in the previous study with 20 annotators who listened and watched the audio-video. The kappa

statistics of the previous study is 0.6, which means there is a higher inter-agreement rate among annotators.

We transcribed the audio files to determine whether the words used are good indicators in detecting emotion since most of the words used are in Filipino. This transcribed data was used in the quasi experiment. Out of 53 voice clips, 47 audio clips were transcribed. Five audio clips were not transcribed because of so much noise even after the application of noise removal function of Audacity. These five audio files were not comprehensible so they were excluded on the dataset. Below is a sample transcribed audio file.

"Kailangan tawagin niya ko sa pangalan ko. Hindi ako hoy, hindi ako aso para tawaging hoy" (Call me by my name. Not "hoy." I'm not a dog.) [He/She should call me by name. I'm not hoy, I'm not a dog that can be called hoy.]

To balance the dataset, we decided to use the transcribed 47 audio clips as our final dataset. The audio and words both contains 47 files. The original label was retained for the 47 audio files. Out of 47 audio files, 10 were labelled as angry, eight (8) labelled as fear, 11 labelled as happy, nine (9) labelled as neutral and nine (9) labelled as sad. We disregarded the three (3) audio files originally labelled as sad, one (1) audio file labelled as neutral and two (2) audio files labelled as fear since the content of the audio file is not comprehensible, hence difficult to transcribe. Only those audio clips that were clearly transcribed were included in our final dataset. After the experiment, two datasets from the same audio file were created. This is because we asked the students to annotate the data based on the transcription data and through listening to the audio file. The same dataset was used in building the emotion model. This was done by extracting audio features for the 47 audio files. In our experiment the output model for the segmented audio file is better than those without applying the segmentation. The 47 audio file were segmented using rectangular window function of jAudio. The window size is 512 with window overlap of 0.01 and a sample rate of 16 kHz. After the feature extraction, and segmentation process, 5101 instances were produced and 36 features were extracted.

Data Cleaning and Feature Extraction

The previous dataset which consist of 53 audio file contained unwanted sounds. The hissing and humming sounds and clippings in the speech waveform are factors that might affect in the annotation of the students and might yield to the low classification accuracy result. Because of these unwanted sounds, the 53 audio files were analyzed using Audacity [8] in order to determine which part of the audio file contains noise and to remove the unwanted sounds. In removing the noise, we first selected the audio file that contained unwanted sounds and we got the noise profile (a

short section of audio containing only the noise to be removed) of the selected audio clip. After getting the noise profile, we selected the audio file that we wanted to clean and then applied the noise removal function of Audacity. The noise removal algorithm uses Fourier analysis: it finds the spectrum of pure tones that make up the background noise in the quiet sound segment that are selected as discussed in the Audacity wiki [18]. For each windowed sample of the sound, a Fast Fourier Transform (FFT) was taken and then the statistical data were tabulated for each frequency band. The noise removal phase started by setting a gain control for each frequency band such that if the sound has exceeded the previously determined threshold, the gain is set to 0; otherwise the gain is set lower, to suppress the noise. Then frequency-smoothing was applied so that a single frequency is never suppressed or boosted in isolation, followed by time-smoothing so that the gain for each frequency band moves slowly. The gain controls were applied to the complex FFT of the signal, and then the inverse FFT was applied, followed by a Hanning window; the output signal was then pieced together using overlap/add of half the window size [18]. The size of FFT is 2048, which result in 1024 frequency bands. We used the default value of Audacity in noise removal which is 24dB noise reduction, 0dB sensitivity, 150Hz Frequency smoothing and 0.15 sec attack/decay time. Figure 1a represents an audio signal that contains noise while Figure 1b represents the cleaned audio signal. For the cleaned signal the silent signal was remove. We applied the noise removal function of audacity to the three audio files but they could not be used as dataset because the content was not comprehensible. After cleaning the audio file, it was decided that only 47 audio files were subjected to feature extraction since only 47 audio files were transcribed and the other five audio files were not transcribed because of so much noise even after noise removal process.

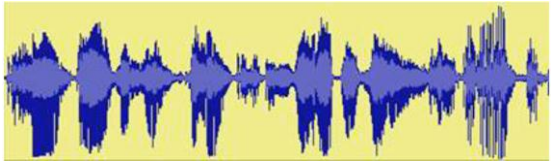

Figure 1a. Audio Signal that contains noise

Figure 1b. Cleaned Audio File

jAudio was used to extract all the audio features. This tool was commonly used for music feature extraction. jAudio is a digital signal processing project built to provide an easy-to-use program for audio feature extraction. Audio feature extraction has extracting properties, such as beat points, statistical summaries, along with many other less obviously useful properties [11]. The following features and its description were extracted using jAudio:

a) Spectral Centroid is calculated as a weight mean of the frequencies present in the signal. Thus, the lower the signal-to-noise ratio of a signal is, the less meaningful is the spectral centroid value [12].
b) Spectral Roll-off Point is the frequency below which 95% of the power in the spectrum resides. This is a measure of the right-skewedness of the power spectrum [13].
c) Spectral Flux is a good measure of the amount of spectral change of a signal. Spectral flux is computed by first calculating the difference between the current value of each magnitude spectrum bin in the current window from the corresponding value of the magnitude spectrum of the previous window. Each of these differences is then squared, and the result is the sum of the squares.
d) Compactness extract the Beat Sum from a signal. This is calculated by finding the sum of all values in the beat histogram.
e) Spectral Variability is the standard deviation of the magnitude spectrum. This is a measure of the variance of a signal's magnitude spectrum
f) Root Mean Square (RMS) is a good measure of the power of a signal. RMS is calculated by summing the squares of each sample, dividing this by the number of samples in the window, and finding the square root of the result.
g) Fraction of Low Energy Windows is a good measure of how much of a signal is quiet relative to the rest of a signal. This is calculated by taking the mean of the RMS of the last 100 windows and finding what fraction of these 100 windows are below the mean.
h) Zero-crossing rate is a measure of the number of time the signal value cross the zero axe. Periodic sounds tend to have a small value of it, while noisy sounds tend to have high value of it. It is computed at each time frame on the signal.
i) Linear Prediction Coeffecients are calculated using autocorrelation and Levinson-Durbin recursion
j) Method of Moments calculates the first five statistical method of moments.
k) Area Method of Moments of MFCCs is the statistical computation of Mel Frequency Cepstral Coefficients (MFCCs). The MFCC

represent the shape of the spectrum with very few coefficients. The cepstrum, is the Fourier Transform of the logarithm of the spectrum. The Mel-cepstrum is the cepstrum computed on the Mel-bands instead of the Fourier spectrum. The use of mel scale allows to take better into account the mid-frequencies part of the signal. The MFCC are the coefficients of the Mel cepstrum. The first coefficient being proportional to the energy is not stored; the next 12 coefficients are stored for each frame [12].

Machine Learning and Classification

In this paper, we used three different machine learning classification algorithms from WEKA namely: J48 decision tree, and two functional classifier, Multilayer Perceptron (MLP), and SMO support vector machine implementation in Weka. The following are the short description of the classifier used in the study.

J48 classifier is a simple C4.5 decision tree for classification. It creates a binary tree. The decision tree approach is most useful in classification problem. With this technique, a tree is constructed to model the classification process. Once the tree is built, it is applied to each tuple in the database and in the result of classification for the tuple [14, 15]. This J48 algorithm is generally used in emotion analysis.

Multilayer Perceptron is an implementation of Neural Networks algorithm in WEKA. This is the most prominent type of neural network which belongs to a class of networks called feedforward networks because they do not contain any cycles and the network's output depends only on the current input instance. This algorithm is usually trained by minimizing the squared error of the network's output essentially treating it as an estimate of the class probability. A serious disadvantage of MLP that contain hidden units is that they are essentially opaque [10].

SMO is a support vector machine (SVM) implementation in WEKA. This algorithm use linear model to implement nonlinear class boundaries. SVM makes use of a separating line called hyperplane (linear model) to classify two clusters. SVMs are based on an algorithm that finds a special kind of linear model: the maximum margin hyperlane. The maximum margin hyperlane is the one that gives the greatest separation between the classes – it comes no closer to either than it has to [10].

In predicting emotions, performance measures such as Precision, Recall and F-Measure were used. This strategy was adopted in the study of Azcarraga and Suarez [16]. They used this performance measure to predict academic emotion. The following are short description of the performance measure taken from this study [16].

Precision is the probability that a class A is true among all that have

been classified as class A. This is also referred to as the Positive Predictive Value (PPV).

Recall is the proportion of examples which were classified as class A among all instances of class A. This is also referred to as the True Positive Rate or Sensitivity.

F-Measure is the combined computation of precision and recall. This is the harmonic mean of Precision and Recall in which both are evenly weighted. The three performance measures were used to assess the performance classification of the models in the study of Azacarraga and Suarez [16] as adopted in this paper. Aside from this performance measures, we look at the correctly classified instances and the kappa statistics. The correctly classified instances and the kappa statistics is a determinant of the emotion model.

Results and Discussion

Based on the result of the annotation, empirical test has shown that listening to audio is better in terms of recognition of emotion compared to the printed word. The kappa statistic for listening to audio is 0.48 while the kappa statistic of the word is 0.40. The interrater reliability analysis shows that the measure of agreement among students annotators in listening to audio file is moderate while in reading the transcribed files is fair. As a rule of thumb, values of Kappa from 0.21 – 0.40 are considered fair, 0.41 – 0.60 moderate, 0.61 – 0.80 substantial, and 0.81 – 1.00 has almost perfect agreement as mentioned in Landis and Kock [19]. Comparing the result of the previous annotation, it is noted that the students classified 11 files as angry while listening to the audio files and five (5) as angry based on the transcribed words out of 10 classified as angry in the original annotation. Six (6) audio files classified as fear by listening to audio alone, and based on transcribed words five (5) audio file classified as fear, out of eight (8) audio files in the original annotation labelled as fear. It is noted that most of the audio files are classified as neutral for both transcribed words and listening to audio, 26 using words alone and 21 through listening to audio. In the original annotation the audio files with the neutral classification are only nine (9). For happy in the original annotation are 11 files, using the words it is only seven (7) and five (5) through listening from the audio file. For sad emotion the original annotation is nine (9), while both for words and listening to audio is four (4). The charts below represent the comparison of the original annotation and the two approaches we used: the words alone and listening to the audio file.

Given the transcribed data, it was noted that the student judgement was based on the thought and content of the words or sentences. Single words and short phrases were classified as neutral. Table 1 shows the example of

single words and phrases that are incomplete in thought or vague hence classified as neutral.

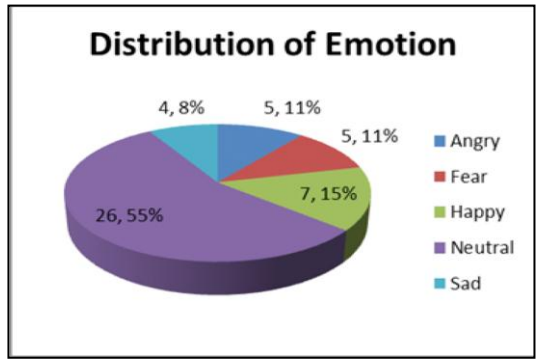

Figure 2. FILMED2 Annotation

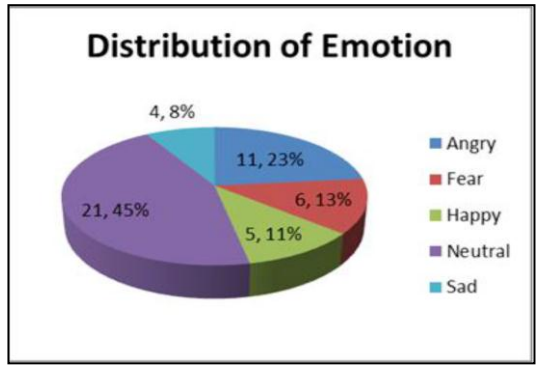

Figure 3. Annotation of Students based on words/text

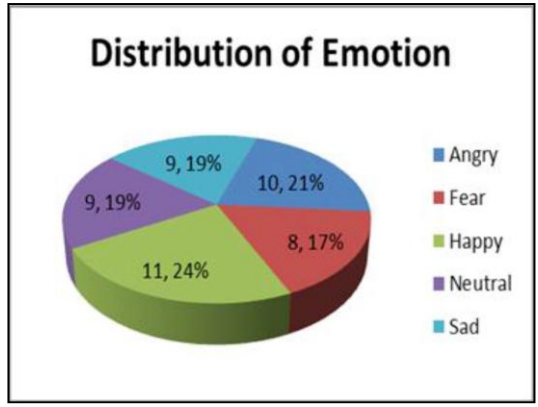

Figure 4. Annotation of Students through listening from Audio file

Table 1
Example of transcribed words/phrases

Single Words/ Phrases in Filipino	Translation
'ba yung reaksyon mo	What's your reaction
at kung ikaw, big brother	What about you big brother
hello po ate!	Hello ate
tama na	Enough; that's enough
ping pong pa	[more] ping pong
Kat	Kat
gaya ng sinabi ko	As what I've said
Images	Images
Mga figures	There are figures
Personality	Personality

Another measure that we considered is the kappa statistic per category. Based on the result presented in Table 2, the Angry category yielded a higher kappa statistic, which is 0.544 using the audio files, while for the words/text, the kappa is 0.494. This is clearly an indicator that student annotators have considered common features in determining the angry emotion. Using words alone, the student agreement is also high, which means that the angry category using words can be easily detected by the annotators. When we examined the transcribed data it was found out that those transcribed data that contain the word "galit" and the thought of the sentences implies negative most of the annotators classified them as angry. For the audio files, we found out that those that are high in pitch and shouting were classified by most of the student annotators as angry. This is an indicator that if the voice is high pitched it implies that the message is negative and for most Filipinos it means angry. Another good point to look at is the Fear category. Based on the actual listening to the audio, most annotators label them as fear if the speakers from the audio file were screaming and calling the name of Jesus. For the words alone, the interrater agreement is 0.322, which means that words alone are not good predictor for fear unless, the student, encounter the word "nakakatakot" or "takot," which means that the speaker is afraid. In the sad category using words, it

Table 2
Interrater Agreement per Emotion
Kappa Statistics-Interrater Agreement by Category

	Angry	Fear	Happy	Neutral	Sad	Overall Kappa
Audio	0.544	0.555	0.421	0.436	0.477	0.487
Word/Text	0.494	0.332	0.375	0.386	0.439	0.405

was found out that most annotators agree that it is sad, when they encounter the word "naawa" or "nalulungkot" or the message of the whole statement is sad, depending on the context of the speaker. Using words, the students easily determined the negative emotion, since they understand the content and context of the statements. For the audio, Angry and Fear have higher interrater agreement as shown in Table 2.

Three different classification algorithms were used in classifying Filipino emotion from acoustic information. J48 decision tree, MLP, and SMO were used as classifier. These classification algorithms are WEKA implementation. Table 3 presents the performance of each model using the 10-fold cross validation. Based on the result shown on Table 3, MLP performed generally better than J48 and SMO when FILMED2 annotations were used. For Audio and Word/Text mode, J48 perform generally better than MLP and SMO. It is noted that for Audio and Words/Text mode, MLP and SMO slightly decreased the performance maybe because of the not balanced distribution of data as shown in the pie chart in Figure 3 in the previous section. It also observed that among the five emotions, Angry was predicted most accurately using the three classifiers. Moreover, Happy was a difficult emotion to predict for word/text mode specifically using SMO. Looking at the result precision, recall and F-measure values are 0. This is interesting since we mentioned that the primary indicators of the annotators when they listened to audio is the high pitch or the shouting while talking or communicating to someone. For the words/text mode usually the content of the entire text or phrases/sentences and the use of words related to angry emotion were the bases. Looking at their annotations those five (5) angry annotations in the word mode is also Angry in the audio mode. This can be a contributing factor why Angry is predicted accurately. On the other hand, Happy is difficult to predict since in the word/text mode this is

Table 3

Performance measure per dataset

Classifier	Emotion	FILMED2			AUDIO			WORD/TEXT		
		PR	RE	FM	PR	RE	FM	PR	RE	FM
J48	Angry	0.941	0.962	0.952	0.949	0.948	0.948	0.984	0.917	0.949
	Fear	0.902	0.994	0.946	0.984	0.965	0.974	0.998	0.019	0.957
	Happy	0.988	0.881	0.931	0.975	0.877	0.923	0.994	0.861	0.923
	Neutral	0.986	0.828	0.900	0.891	0.959	0.923	0.915	0.993	0.953
	Sad	0.959	0.961	0.960	0.992	0.958	0.974	0.993	0.956	0.974
MLP	Angry	0.962	0.965	0.963	0.951	0.949	0.950	0.891	0.862	0.876
	Fear	0.978	0.979	0.978	0.919	0.980	0.949	0.843	0.833	0.838
	Happy	0.901	0.948	0.924	0.803	0.771	0.787	0.838	0.707	0.767
	Neutral	0.906	0.886	0.896	0.877	0.851	0.864	0.865	0.911	0.888
	Sad	0.981	0.954	0.968	0.986	0.962	0.974	0.942	0.928	0.935
SMO	Angry	0.941	0.917	0.923	0.942	0.886	0.913	0.869	0.754	0.807
	Fear	0.962	0.946	0.964	0.884	0.927	0.905	0.588	0.015	0.029
	Happy	0.747	0.938	0.832	0.821	0.210	0.334	0.000	0.000	0.000
	Neutral	0.929	0.733	0.820	0.559	0.772	0.649	0.594	0.864	0.704
	Sad	0.974	0.938	0.956	0.766	0.626	0.689	0.651	0.803	0.719

labelled as Happy while the audio mode is labeled as Neutral. The classification of Happy for all algorithms was misclassified as Neutral.

We also considered as our performance measure for the emotion model the correctly classified instances (CCI), and the Kappa Statistics. As shown in Table 4, for the FILMED2 dataset (used the original annotation of the previous study [2]), Multilayer Perceptron classification algorithm has 95.53% correctly classified instance with kappa statistics of 0.9427, though all classification algorithm yielded a good value ranging from 91% - 94.45%, with kappa statistics ranging from 0.88 to 0.93. This means that Multilayer Perceptron performed better compared with J48 and SMO for this particular dataset. The other two datasets (voice and word/text, J48 classification algorithm) have 95% correctly classified instances with kappa statistics of 0.94 using the audio file as a mode for the student to classify the emotion and 95.35% correctly classified instances with kappa statistics of 0.93 for word/text as the basis for the student to classify the emotion. Overall, the two classification algorithms (J48 and MLP) performed better in all the given datasets, while SMO did not perform well in the word/text mode. We can generally say that all modes can be used to build an emotion model based on the result given on Table 4. The results further demonstrate that all of these approaches that we used are good indicators to detect emotion in a conversational set-up.

Table 4

Correctly Classified Instances and Kappa Statistics

Classifier	Dataset	CCI	ICI	Kappa Statistics
J48	FILMED2	94.45%	5.55%	0.9284
MultiLayer Perceptron		95.53%	4.47%	0.9427
SMO		91.37%	8.63%	0.8894
J48	VOICE	95.00%	5.00%	0.9358
MultiLayer Perceptron		91.57%	8.43%	0.8919
SMO		75.59%	24.41%	0.6809
J48	WORD/TEXT	95.35%	4.65%	0.9332
MultiLayer Perceptron		87.59%	12.41%	0.8229
SMO		63.50%	36.50%	0.4244

Future Works

We did not consider the balancing of the dataset in this paper. As we presented in the previous section of this paper, the length of the audio file is not the same and the number of audio file is not equal to each category of emotion, especially those datasets that the students annotated (voice and word/text). Balance dataset is necessary in creating a model. It is good to explore a balance dataset, which means that the length of the audio file should be equal and the category should have the same number of audio

file. The result of this study can be further analyzed to produce good emotion model that can be used in emotion recognition system.

The consideration of the equal number of male and female annotators can be further explored if there is a change in the result of the annotation, since this study did not consider balancing the number of male and female annotators. The experiment set-up can be further improved, wherein all annotators can listen individually to the audio file to avoid destruction and being influenced by their peers.

Hence, the machine learning algorithms used in this study demonstrate a higher accuracy result for all dataset given. It is also good to re-examine the audio file dataset which is more than four (4) seconds by segmenting it to two (2) seconds window as cited in the study of Azcarraga & Suarez [16], because emotions persist from 0.5 to four (4) seconds. By using this approach, it is guaranteed that all transition of emotion can be captured. This may result to a better emotion model. Another research can be undertaken which bayes rule should be considered in the preparation of the dataset.

References

[1] Elfenbein, HA, and Ambady, N. Universals and cultural differences in recognizing emotions. Haas School of Business, University of California at Berkeley, California & Department of Psychology, Harvard University, Cambridge, Massachusetts, Current Direction in Psychological Science, 12 (5), (October 2003), 159 – 163.

[2] Cu, J, Solomon, KY, Suarez, MT. & Sta. Maria, M. A multimodal emotion corpus for Filipino and its uses. J Multimodal User Interfaces. (2012), DOI 10.1007/s12193- 012-0114-8.

[3] Zeng, Z., Pantic, M., Roisman, G., Huang, T. A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. (2009), 31 (1).

[4] Abrilian, S, Devillers, L, Buisine, S, and Martin, J.C., EmoTV1: Annotation of Real-life Emotion for the Specification of Multimodal Affective Interfaces. LIMSICNRS, BP 133, 91403 Orsay Cedex, France

[5] Jonghwa K. Bimodal Emotion Recognition using Speech and Physiological Changes. Robust Speech and Understanding, Michael Grimm and Kristian Kroschel (Ed.), ISBN 987-3-90213-08-0.

InTech, (2007) Available from: http://www.intechopen.com/books /robust_speech_recognition_and_understanding/bimodal_emotion_ recogniti.

[6] Zeng, Z., Hu, Y, Roisman, G., Wen, Z., Fu, Y., and Huang, T. Audio-Visual Spontaneous Emotion Recognition. ICM/IJCAI Workshop (2007).

[7] Metallinou, A., Lee, S., Narayanan, S. DecisionLevel Combination of Multiple Modalities for Recognition and Analysis. n.p., n.p., 2010.

[8] http://audacity.sourceforge.net/about/

[9] McEnnis, D., McKay, C., Fujinaga I, and Depalle, P. jAudio: A feature extraction library. *Proceeding of the International Conference on Music Information Retrieval*, (n.p., 2005), 600-3.

[10] Witten, I. H., Frank, E., Trigg, L., Hall, M., Holmes, W. Practical Machine Learning Tools and Techniques with Java Implementations. n.p., n.p., n.d.

[11] http://jaudio.sourceforge.net/

[12] G. Peeters. A Large Set of Audio Features for Sound Description. Technical report published by IRCAM. (2004).

[13] jAudio 1.0 Feature Appendix. http://jaudio.sourceforge.net/jaudio10 /features/.

[14] Dunham, M. H. and. Sridhar, S. *Data mining, Introductory and Advanced Topics*, (1st ed.) Person Education, n.p., 2006.

[15] Aman, K. S. and Suruchi S. A Comparative Study of Classification Algorithms for Spam Email Data Analysis. IJCSE, 3 (5), 2011, 1890-1895.

[16] Azcaraga, J. & Suarez, M.T. 2013. Recognizing Student Emotions using Brainwaves and Mouse Behavior Data. *International Journal of Distance Education Technologies*, 11 (2). (2013)

[17] http://www.anvil-software.org/

[18] How Noise Removal Works. Audacity Wiki. Retrieved from: Table 3:

Performance measure per dataset

[19] Landis, J.R., and Koch, G. G. The measurement of observer agreement for categorical data. *Biometrics* (1977), 33: 159-174.