Emotion Recognition using Audio Speech Modality

Enoch T. Puno
College of Computer Science
De La Salle University
Manila, Philippines
enoch_puno@dlsu.edu.ph

Miguel Alfonso M. Quiambao
College of Computer Science
De La Salle University
Manila, Philippines
miguel_quiambao@dlsu.edu.ph

Carl James M. Raymundo
College of Computer Science
De La Salle University
Manila, Philippines
carl_raymundo@dlsu.edu.ph

***Abstract*—According to recent trends, development in systems that make use of emotion detection/recognition have started to rise. In the current emotion recognition market, analysts have stated that it would be set to grow at a CAGR of 32.56% around period of 2019 to 2027. With this in mind, the importance of expanding the field of emotion recognition has never been greater. In addition to this, there has also been an increase in interest regarding emotional recognition through audio speech signal.**

Keywords—emotion recognition, audio modality, mel-frequency cepstral coefficients.

I. Introduction

Emotion plays a significant role in daily human interactions. This is essential to our rational as well as intelligent decisions. It helps one to understand one. Through previous works and continuing research it has been revealed that emotion is a powerful role that plays in shaping human social interaction. This also opened up new fields in emotion recognition, having basic goals to retrieve and recognize emotions.

II. Related Works

A. *Detecting Emotions in Speech*

According to them, one must make the following distinctions when studying emotions in speech: the emotional attitude of the speaker towards the hearer, the emotional attitude of the speaker towards the message, or the emotional state of the speaker. Finally, Psolzin and Waibel concluded: Acoustic and Prosodic information can be combined and added into an HMM speech recognition system, that is augmented by suprasegmental states. Consequently, their study focused on the determination of the emotional state of a speaker through acoustic and prosodic features.

B. *An Efficient Approach for Emotion Detection from Speech Using Neural Networks*

For the dataset of the research, the dataset is composed of 354 instances wherein 121 instances are labeled under happiness, 117 instances that are labeled under sadness and the rest of the instances are labeled under neutral. For the feature extraction of the audio files in the dataset, the researchers considered the pitch, speaking rate and accent of the user. For the limitations of the research, the limit of the research is that it only identified emotions such as happiness, sadness and neutral. In order for the researchers to perform the research, the researchers used MATLAB.

C. *Automatic Speech Emotion Recognition Using Machine Learning*

After extracting the features in the dataset, the researchers used a technique called recursive feature elimination with linear regression (LR-RFE) which was presented in the research and was proven effective in selecting the features in the dataset. For the Spanish Dataset, the dataset consists of 2 actors, 1 female and 1 male. The German dataset consists of 10 actors, 5 actors were female while the other 5 actors were male. For the dataset of the research, the researchers used the German and Spanish datasets in order to perform feature extraction and emotion detection.

D. *Emotion Detection from Speech*

If the speech is unvoiced the corresponding marker in the pitch vector was set to zero" The researchers used these to identify the characteristics of the emotions used by the statistics calculated from the pitch. " The speakers uttered numbers and dates in different emotions, but it would have been better if they used more words because word choice can indicate emotion. The data used for this project comes from the Linguistic Data Consortium's study on Emotional Prosody and Speech Transcripts The subjects were professional actors uttering common phrases with fourteen different emotions used. It also states the limitations of the field which includes what features influence the recognition of emotion in speech, and which algorithm for classifying emotions is best for the research. Their results also showed that agitated emotions such as happiness, elation, and interest have similar properties as with their subdued counterparts despair and sadness, and because of this 'agitated' and 'subdued' could be an emotion class. Speaking rate (inverse of the average length of the voiced part of the utterance) The researchers eventually found that the methods they have used are extremely promising. "The performance of these methods should be evaluated for multi-class classification (using multi-class SVMs and K-Means. There were five female speakers and three male speakers. The researchers used Mel Frequency Cepstral Coefficients (MFCCs) and Formants of speech as inputs to classification algorithms so they could see which features carry the most information and why. They worked on improving previous studies such as separating male and female speakers, and it did well for their study.

E. *Speech Emotion Recognition Methods: A Literature Review*

Chen's goal was to enhance speech emotion recognition in a speaker-independent 3 level speech emotion recognition method. Method is used to classify the emotion into six different categories then uses the private dataset to train and test the new system. In this research paper by Babak Basharirad and Mohammadreza Moradhaseli [3], they provided a speech database which is utilized to validate the proposed methods in speech emotion recognition. This classifies different emotions from coarse to fine then pick a fit feature using Fisher rate. Most of the current research concentrate on investigating different features and their correlation with emotional state in spoken speech. The majority of the current datasets are not capable for evaluation of speech emotion recognition. Thus, results expose that LFPC is a better option as feature for emotion classification than the standard features.

III. Methodology

This chapter describes the details of the group's emotion recognition system. Included here are the discussions on data collection, preprocessing data, feature extraction, and lastly feature selection.

A. Data Collection

The data retrieved by the group are raw .WAV (Waveform Audio File Format) sound files from the Berlin Database for Emotional Speech. This database contains at least 500 utterances performed by 9 different actors, 4 males and 5 females. Each sound file displays a certain emotion; these emotions are happy, angry, anxious or fearful, bored, disgusted, or neutral. Along with this, there are 10 different texts that each actor utters while depicting a different emotion. In some cases, two sound files may come from the same actor, with the same text, and with the same emotion yet with different styles of utterances. Furthermore, the data files collected were named in such a way that specific data can be inferred in each file. This naming convention follows:

- The first 2 characters refers to the speaker id
- The following 3 characters refers to the text spoken
- The following letter refers to the emotion displayed in the file
- The last character refers to the different versions the utterance

The speaker id, text code, and the letters referring to an emotion can be viewed in Tables. 1, 2, and 3 respectively.

Table 1. *Speaker Information.*

Id	Sex	Age
03	Male	31
08	Female	34
09	Female	21
10	Male	32
11	Male	26
12	Male	30
13	Female	32
14	Female	35
15	Male	25
16	Female	31

Table 2. *Code for the text spoken in the file.*

Code	German Text	English Translation
a01	Der Lappen liegt auf dem Eisschrank.	The tablecloth is lying on the fridge.
a02	Das will sie am Mittwoch abgeben.	She will hand it in on Wednesday.
a04	Heute abend könnte ich es ihm sagen.	Tonight I could tell him.
a05	Das schwarze Stück Papier befindet sich da oben neben dem Holzstück.	The black sheet of paper is located up there besides the piece of timber.
a07	In sieben Stunden wird es soweit sein.	In seven hours it will be.
b01	Was sind denn das für Tüten, die da unter dem Tisch stehen?	What about the bags standing there under the table?
b02	Sie haben es gerade hochgetragen und jetzt gehen sie wieder runter.	They just carried it upstairs and now they are going down again.
b03	An den Wochenenden bin ich jetzt immer nach Hause gefahren und habe Agnes besucht.	Currently at the weekends I always went home and saw Agnes.
b09	Ich will das eben wegbringen und dann mit Karl was trinken gehen.	I will just discard this and then go for a drink with Karl.
b10	Die wird auf dem Platz sein, wo wir sie immer hinlegen.	It will be in the place where we always store it.

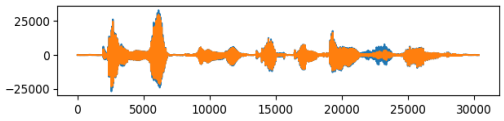
Table 3. *Code for the emotion show in the file*

Code	Emotion
N	Neutral
W	Anger
L	Boredom
E	Disgust
A	Anxiety/Fear
F	Happiness
T	Sadness

B. Preprocessing

In order to fully utilize the data retrieved from the database as a test set for an emotional recognition system, each sound file needed to be processed first. Firstly, the group separated each sound file depending on the gender to which the files belonged to. This was done because the authors goals were to create two different models, one for males and another for females, in order to create a more accurate system. Afterwards, each sound file was then turned into a set of numerical values, coming from the sine wave emitted by the files, using pyAudioAnalysis [2]. These signal values , an example can be viewed in Fig 1, were then used for the feature extraction process.

Figure 1. *The numerical values from the audio signal plotted into a graph.*



C. Feature Extraction

Using pyAudioAnalysis [2] again, the authors were able to extract various plausible features that the system may use. These are:

- Zero Cross Rating,
- Energy,
- Spectrum,
- Mel-frequency cepstral coefficients,
- and the Chroma Vectors

In extracting the features for the system, the sampling rate of each audio signal were discovered to be Fs=16000Hz, the audio signal was then segmented into multiple frames/windows consisting of frame_size=50 milliseconds of data from the audio signal, while moving at a frame step of 25 milliseconds and an overlap of 12.5 milliseconds per step until the end of the audio signal, thus creating a total of 20 (Fs/frame_size) frames/windows because each file contained a 1 second audio recording. By doing this, the resulting value of the feature extraction is an MxNx33 numpy [5] ndarray, where each row contains an Nx33 matrix; where M is the total number of frames recorded from the audio signal, N is a feature-specified length (features such as zero-cross rating, energy, etc. return different lengths of values), and 33 columns that relate to: the zero cross rating; the energy and its entropy; the spectrum and its centroid, spread, entropy, flux, and rolloff; the 12 MFCCs (Mel-frequency cepstral coefficients); and the 12 element chroma vector. Additionally, the pitches of the extracted signal were estimated using the crepe python package [4] and was added as another feature. Upon extracting these values, each value was put through the L2-norm (Euclidean normalization) [1] process to regularize the data in order to prevent overfitting for the machine learning models—the equation for this process can be seen in Eq. 2—afterwards the features were statistically described by getting the mean, standard deviation, and the skewness of each feature; thus creating the final and actual features that the authors plan to use in their system.

Equation 1. *Euclidean normalization or L2-norm*

$$w^{*} = arg\,min\ {}_{w}\sum_{j}(t(x_j) - \sum_{i} w_i h_i(x_j))^{2} + \lambda \sum_{i=1}^{k} w_i^{2}$$

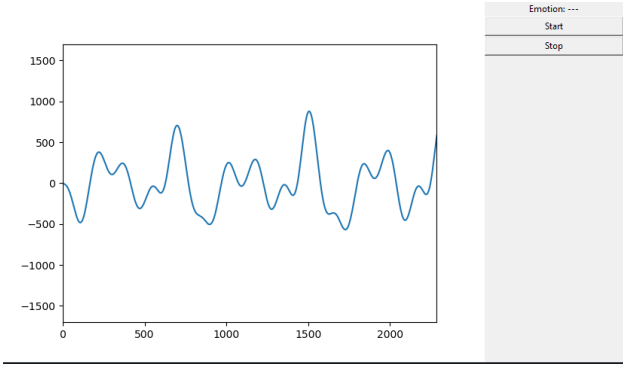
D. Feature Selection

After getting the statistical descriptions of all the initial features, they were then reduced to the ones that relate to the MFCC and pitch. This was done because the results of the previous works regarding the topic at hand, made use of only the MFCCs and the pitches of their dataset in accurately detecting emotion.

E. Realtime Feature Extraction

The process for the feature extraction of both non-realtime (NRFE) and realtime (RFE) are somewhat similar. The only difference is the way the audio signal is extracted and how the audio frames/windows are sent to the extraction process. During the feature extraction of the recorded audio signals, a fixed number of frames were created because of the finite length each audio recording had, this is in contrast to the realtime feature extraction. In realtime, there is no definite length of recording time, and as such, the authors developed a way to work around this limitation. The default sampling rate (Fs) of the microphone used in the RFE was Fs=44100Hz, in order to simulate the feature extraction from NRFE, each time a chunk (each chunk consisted of Fs/20 bytes of data) in the data stream was recorded—using pyAudio [6]—the values recorded were concatenated into a numpy [5] ndarray, once the total rows of the ndarray were reached 20, the signal values in the said ndarray were passed on to the feature extraction function, thus simulating an NRFE in an RFE environment (The program screen can be viewed in Fig 2). However, this process contains some latency, due to both the requirements of the crepe python package [4]—which has to create a machine learning model, of its own, to estimate the pitch—and the fact that a certain number of chunks must first be recorded before a feature can be extracted.

Figure 2. *The system implemented for realtime feature extraction and emotion recognition*



IV. Experimentation and Results

The author's experimentation utilized RapidMiner to assess which classification model is best suited for the male and female datasets. The F1 score of these models in detecting each emotion can be seen in Tables 4-7 for the male dataset, and Tables 8-11 for the female dataset.

For the male dataset, the results that applied forward selection performed better compared to other feature selection techniques that used the male dataset (Table 5). For the female dataset, the results that applied backward selection performed better compared to other feature selection techniques that used the female dataset (Table 10).

Table 4. *Results of the male dataset with complete features*

	k-NN (k = 3)	k-NN (k = 5)	k-NN (k = 10)	SVM	MLP	Decision Tree
Happiness	0.00%	40.00%	18.18%	0.00%	46.15%	0.00%
Neutral	20.00%	27.27%	18.18%	0.00%	57.14%	50.00%
Anger	52.38%	66.67%	55.32%	0.00%	91.89%	62.74%
Sadness	0.00%	15.39%	30.77%	0.00%	66.67%	0.00%
Fear	17.39%	31.58%	22.22%	0.00%	38.09%	0.00%
Boredom	40.00%	32.00%	41.66%	0.00%	35.30%	0.00%
Disgust	0.00%	0.00%	0.00%	8.34%	50.00%	0.00%

Table 5. *Results of the male dataset applying forward selection*

	k-NN (k = 3)	k-NN (k = 5)	k-NN (k = 10)	SVM	MLP	Decision Tree
Happiness	15.38%	33.33%	0.00%	42.42%	53.33%	0.00%
Neutral	31.58%	45.46%	66.67%	50.00%	61.54%	41.67%
Anger	86.49%	85.71%	60.87%	10.01%	91.89%	62.07%
Sadness	50.00%	57.14%	50.00%	0.00%	66.66%	33.33%
Fear	34.78%	46.67%	28.57%	30.77%	33.33%	14.28%
Boredom	34.78%	40.00%	26.67%	46.15%	52.63%	42.10%
Disgust	0.00%	0.00%	50.00%	30.77%	0.00%	0.00%

Table 6. *Results of the male dataset applying backward elimination*

	k-NN (k = 3)	k-NN (k = 5)	k-NN (k = 10)	SVM	MLP	Decision Tree
Happiness	16.67%	42.86%	18.18%	0.00%	61.54%	61.54%
Neutral	41.67%	40.00%	31.58%	25.00%	50.00%	50.00%
Anger	54.54%	55.56%	60.72%	40.00%	94.44%	94.44%
Sadness	23.53%	18.19%	15.39%	0.00%	61.54%	61.54%
Fear	42.10%	22.22%	13.33%	15.78%	47.62%	47.62%
Boredom	23.53%	44.44%	38.09%	0.00%	38.46%	38.46%
Disgust	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

Table 7. *Results of the male dataset applying correlation*

	k-NN (k = 3)	k-NN (k = 5)	k-NN (k = 10)	SVM	MLP	Decision Tree
Happiness	21.43%	22.22%	33.33%	0.00%	87.50%	40.00%
Neutral	33.33%	21.43%	28.00%	0.00%	69.57%	53.33%
Anger	33.33%	46.67%	43.75%	0.00%	81.25%	60.00%
Sadness	45.45%	42.86%	25.00%	11.00%	64.29%	50.00%
Fear	20%	16.67%	0.00%	0.00%	87.50%	100.0%
Boredom	22.22%	12.50%	29.41%	0.00%	70.00%	35.71%
Disgust	42.86%	50.00%	0.00%	0.00%	100.0%	90.91%

Table 8. *Results of the female dataset with complete features*

	k-NN (k = 3)	k-NN (k = 5)	k-NN (k = 10)	SVM	MLP	Decision Tree
Fear	20.00%	18.18%	0.00%	0.00%	58.82%	0.00%
Happiness	18.18%	9.52%	33.33%	0.00%	60.00%	0.00%
Boredom	34.48%	48.28%	46.67%	0.00%	69.56%	13.33%
Neutral	23.08%	26.66%	37.04%	0.00%	66.67%	0.00%
Anger	28.57%	25.00%	31.11%	0.00%	61.54%	54.84%
Sadness	41.38%	15.38%	0.00%	0.00%	91.67%	57.15%
Disgust	47.62%	35.30%	22.22%	20.00%	70.00%	34.78%

Table 9. *Results of the female dataset applying forward selection*

	k-NN (k = 3)	k-NN (k = 5)	k-NN (k = 10)	SVM	MLP	Decision Tree
Neutral	30.00%	16.67%	16.67%	30.18%	25.00%	18.18%
Anger	44.44%	55.17%	45.45%	32.00%	18.18%	0.00%
Boredom	66.66%	42.86%	38.09%	38.71%	34.78%	40.82%
Disgust	54.55%	34.78%	40.00%	0.00%	53.84%	0.00%
Fear	47.62%	60.00%	69.39%	0.00%	57.70%	58.07%
Happiness	88.00%	78.26%	88.00%	58.82%	86.96%	60.87%
Sadness	60.00%	56.00%	61.54%	57.15%	55.56%	0.00%

Table 10. *Results of the female dataset applying backward selection*

	k-NN (k = 3)	k-NN (k = 5)	k-NN (k = 10)	SVM	MLP	Decision Tree
Neutral	62.50%	27.27%	55.17%	28.57%	52.18%	75.00%
Anger	41.90%	50.09%	53.33%	46.67%	25.53%	60.74%
Boredom	48.45%	45.36%	45.45%	35.10%	66.66%	33.62%
Disgust	47.36%	36.16%	77.50%	29.41%	74.07%	47.36%
Fear	0.00%	30.84%	15.38%	18.18%	62.50%	45.65%
Happiness	31.00%	41.39%	33.33%	33.16%	37.03%	24.84%
Sadness	84.39%	88.00%	95.65%	70.97%	91.70%	81.82%

Table 11. *Results of the female dataset applying correlation*

	k-NN (k = 3)	k-NN (k = 5)	k-NN (k = 10)	SVM	MLP	Decision Tree
Neutral	33.33%	11.11%	0.00%	0.00%	70.59%	0.00%
Anger	26.09%	17.39%	20.00%	0.00%	64.00%	0.00%
Boredom	32.26%	51.61%	46.67%	0.00%	76.92%	13.33%
Disgust	24.00%	22.22%	28.57%	25.53%	59.26%	0.00%
Fear	36.36%	37.50%	29.79%	50.00%	80.95%	54.84%
Happiness	42.10%	16.67%	0.00%	19.05%	91.67%	57.15%
Sadness	50.00%	38.09%	33.33%	0.00%	84.21%	34.78%

In most of the tables presented, the SVM classification model presented the lowest f1 scores in classifying emotions both in the male and female datasets. The lowest overall accuracy that the SVM

classification model performed is at 4.35% for the male dataset and 11.11% for the female dataset.

For the most effective classification model, the classification model that performed better was Multi-Layered Perceptron (MLP). The MLP classification model presented high f1 scores in classifying emotions in both male and female datasets. For the overall accuracy performed by the MLP classification model, the highest overall accuracy performed by MLP model gained 60.87% for the male dataset and 77.78% for the female dataset.

For both male and female datasets, the emotion anger gained high f1 scores in all of the tables presented. For the lowest f1 scores gained, the emotion disgust evidently gained low scores for the male dataset and the emotion fear and happiness gained low scores for the female dataset.

V. Conclusion

In conclusion, the results of the experimentation shows that the most accurate and precise machine learning model that was used for recognizing emotions through MFCC and Pitch data is a Multi-Layered Perceptron classification model for both male and female datasets. While the least effective is a Support Vector Machine, however, this result somewhat contrasts the findings of most of the related literature. This may be due to the inclusion of additional Prosodic/suprasegmental phonology features in previous works. Furthermore, the authors believe that, in terms of the real-time program, the denoising filter that was implemented may not be sufficient and may not remove some noise in the extracted features. Additionally, due to how sklearn's MLP classifier is built differently from RapidMiner, a different accuracy rate may be produced each time the program is run.

References

[1] Differences between L1 and L2 as Loss Function and Regularization. [Online]. Available: http://www.chioka.in/differences-between-l1-and-l2-as-loss-function-and-regularization/.

[2] T. Giannakopoulos, "pyAudioAnalysis: An Open-Source Python Library for Audio Signal Analysis," Plos One, vol. 10, no. 12, 2015.

[3] A. M. Insights, "Global Emotion Recognition Market is set to Grow at a CAGR of Around 32.56% During the Forecast Period (2019-2027), Owing to Demand for Improvements in Defense Technology: Says Absolute Markets Insights," PR Newswire: press release distribution, targeting, monitoring and marketing, 19-Mar-2019. [Online]. Available: https://www.prnewswire.com/news-releases/global-emotion-recognition-market-is-set-to-grow-at-a-cagr-of-around-32-56-during-the-forecast-period-2019-2027-owing-to-demand-for-improvements-in-defense-technology-says-absolute-markets-insights-300814756.html.

[4] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, "Crepe: A Convolutional Representation for Pitch Estimation," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018.

[5] F. Nelli, "The NumPy Library," Python Data Analytics, pp. 49–85, 2018.

[6] M. Wickert, "Real-Time Digital Signal Processing Using pyaudio_helper and the ipywidgets," Proceedings of the 17th Python in Science Conference, 2018.

[7] L. Kerkeni, M. Mbarki, K. Raoof, M. A. Majoub, and C. Cleder, "Automatic Speech Emotion Recognition Using Machine Learning". 31-Jan.-2019.

[8] T. Psolzin and A. Waibel, "Detecting Emotions in Speech". 1998.

[9] B. Basharirad and M. Moradhaseli, "Speech Emotion Recognition Methods: A Literature Review"

[10] Gonvil and Caius, "Emotion Detection From Speech"

[11] S.Kim, P.Georgiou, S.Lee, S.Narayanan. "Real-time emotion detection system using speech: Multi-modal fusion of different timescale features", Proceedings of IEEE Multimedia Signal Processing Workshop, Chania, Greece, 2007

[12] K. Rajvanshi and A. Khunteta, "An Efficient Approach for Emotion Detection from Speech Using Neural Networks"