SPEECH EMOTION RECOGNITION

CAROL SHANTI G. ESTOLAS[1], MICAH MARIEL V. GARCIA[1], JOHN PAUL V. MAYO[1], JOHN MICHAEL NOLASCO[1] and ORLAND DELFINO TUBOLA[1,2]

[1]Department of Computer Engineering, College of Engineering, Polytechnic University of the Philippines
[2] Center for Engineering and Technology Research, Institute for Science and Technology Research, Polytechnic University of the Philippines

Abstract. This paper introduces a graphical user interface (GUI) that will classify the emotion of a recorded speech or utterance. Features embedded on the GUI include sound normalization extraction, wave formation and sound frequency detection. In this study, experiments were made to determine the effectivity and accuracy of the speech emotion recognition system through the GUI. This system deals with recognizing two databases of utterances expressing six emotions:joy, anger, sadness, fear, neutral and boredom state. The first database or local database comprises of 600 short recorded utterances which were portrayed by one hundred PUP Computer Engineering students. On the other hand, the second or foreign database composes of 240 wave file obtained from Medical Electronics Division of Technical University of Lodz, Institute of Electronics site. These sound files are then filtered and processed through WavePad Sound Editor to reduce the environment noise before feeding it to the GUI to be able to obtain a reliable collection of wave files. Researchers trained the neural network of the system through the speech emotion recognition tab for four to six hours to obtain a high accuracy rate. Experiments had demonstrated the following accuracy for the local database: neutral – 78.12 %, joy – 100 %, anger- 100%, sadness – 93.10%, fear – 81.48%, boredom -90.90%. On the other hand for the foreign database, results are as follows: neutral – 66.67 %, joy – 83.33 %, anger- 100%, sadness – 30.77%, fear – 76.92%, boredom -84.62%. The system can be potentially applied for the detection of emotion expressed by operators and clients during calls which are very important indicators for measuring the quality of service offered by call centers. Future researches are recommended to use a better recording device that will automatically remove or lessen the noise of the environment that might help to obtain a higher accuracy rate. Also, it will be better to create an additional feature that will detect more emotion such as disgust, surprise, sarcasm and anticipation.

Keywords: *neural network, graphical user interface, feature extraction, sound normalization, speech emotion recognition*

1. INTRODUCTION

Speech Emotion Recognition is a relatively new field of research, it has many potential applications however recognition of speeches is a complex task that is furthermore complicated by the fact that there is no unambiguous answer to what "correct" emotion is for a given speech sample (Scherer, 2003; Batliner *et al*., 2003). Currently, researchers are still debating what features influence the recognition of emotion in speech. There is also considerable uncertainty as to the best algorithm for classifying emotion, and which emotions to class together. In this study, we attempt to address the issue by utilizing the effectivity of a graphical user interface on recognizing speech emotion recognition. Researchers have used MATLAB GUI, a high-level technical computing language and interactive environment for algorithm development,

data visualization, data analysis to classify six emotions namely joy, anger, sadness, boredom, neutral and fear. Using the embedded features of the graphical user interface that includes sound normalization, feature extraction and frequency detection, the program compares the input sound file to the set of wave files in the database of each emotion to determine which type of emotional load the file is.

2. METHODOLOGY

Speech emotion recognition is one of the latest challenges in speech processing. Besides human facial expressions speech has been proven as one of the most promising modalities for the automatic recognition of human emotion. In this study, researchers utilized a MATLAB program to determine the emotion of a recorded speech or utterance through a graphical user interface.

2.1 Creating the Database

In creating a pattern recognition system to determine the emotion of a speech, a database is to be set to serve as a comparison or reference to the file that is to be input in the program. In this system, researchers have created two databases. The first one which is the local database comprises of 100 recorded utterances for each emotion (joy, sadness, anger, boredom, neutral and fear) which is portrayed by 100 PUP Computer Engineering students. They were given lines to utter. They had to imagine themselves uttering the given lines or speeches to an individual. Each utterance should correspond to one of the six emotional classes. Each example in the database was evaluated by human subjects who have to decide if they are appropriate or not. In this setup, a mobile phone was used as the recording device for the local database. On the other hand, the second or foreign database consists of 40 files for each emotional class. These are obtained from Medical Electronics Division of Technical University of Lodz, Institute of Electronics site on Polish Emotional Speech. With these two databases, researchers will be able to compare the effectivity and accuracy of the program for both collections of sound files.

Database of the network plays the most important role since this serves a pattern tool in comparing the input file to the classes set for the system. Before feeding the system with set of sound files for its database, researchers first filtered the recorded files using WavePad Sound Editor which is a program used to remove all environment noise obtained from recording speeches using the mobile phone. Once most of the noise has been removed, samples are then fed to the program's database.

1.1 Graphical User Interface

A graphical user interface or GUI, is a type of interface that allows users to interact with electronic devices through graphical icons and visual indicators such as secondary notation, as opposed to text-based interfaces, typed command labels or text navigation. Figure 1 shows the GUI used for the system.

Figure 1. Graphical User Interface (GUI) of the Neural Network

1.2 Training the Network

The key for obtaining a good accuracy rate for a neural network or any pattern recognition system is a proper training of the network with a definite and concise sound files for the database. In this study, series of steps were implemented.

Using the Add Selected sound button of the GUI, the user can feed the program with a larger database. Once the button is clicked, the system will request the user to load a recorded sound file and ask the user to which emotion the input file is to be classified. Figure 2 shows the results of loading a sound file from the computer. It will also display the emotion where it was classified.

After feeding the program with a database for each emotional class, researchers made an initial training by using the speech emotion recognition tab of the program which has a sound feature extraction, sound normalization trait and frequency detection to determine the emotion of the input file. During the initial test four input files were loaded to the program (Figure 3).

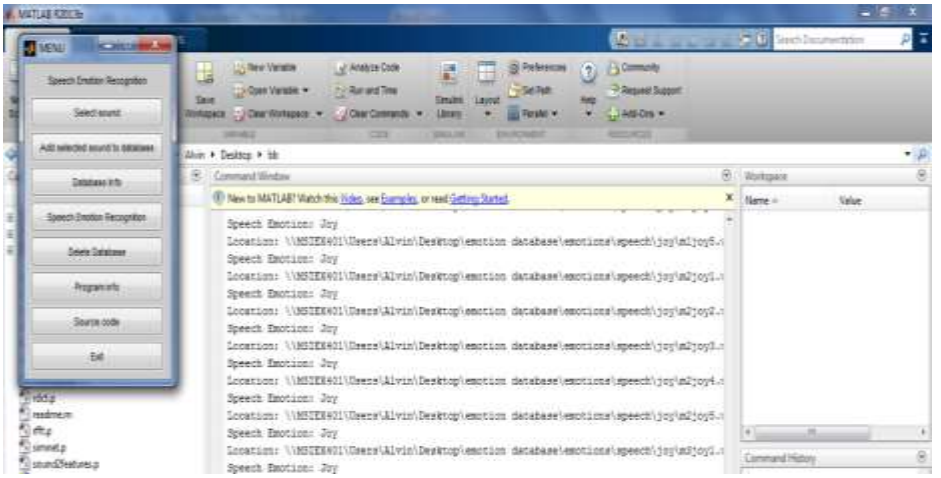

Figure 2. Feeding the Network's Database.

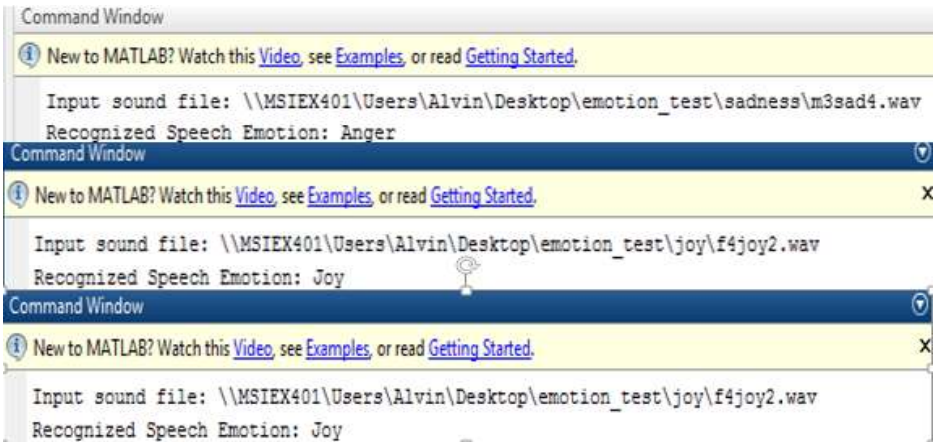

Figure 3. Network Training.

Results from the test above show that the system has recognized all except for the first input where the sound file which should be recognized as sadness is classified as anger. As more input file is loaded in the system, the more error it might produce. To prevent these errors, the network /system should be retrained and more sound files is to be loaded in the database. The figure below shows the re –training of the network. Researchers have added set of sound files to each emotional class.

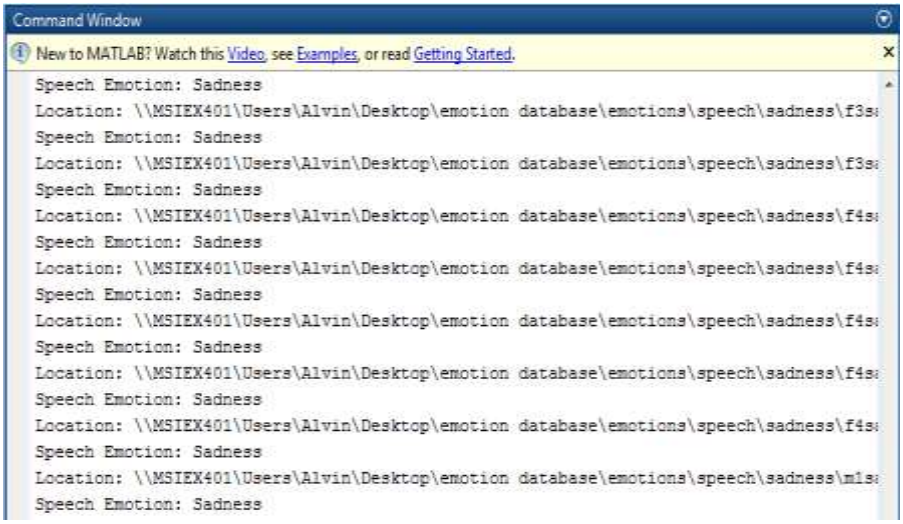

Figure 4. Retraining the Network.

The accuracy of the network depends on the training and database. The system comprises of seven trainings. The first training is to compare the input wave file to the sound files that is stored on anger database. Second training is for boredom. The third up to sixth trainings are set for fear, joy, neutral and sadness respectively. Lastly, the seventh training is for the network to decide in which emotion the input file is to be classified.

3. EXPERIMENTAL RESULTS

The table below shows the number of audio files that was recognized for each emotional load. Boxes were colored orange for each emotion that was properly recognized. For example for "Joy", the researchers input 12 audio files. 10 out of 12 was correctly recognized and 2 audio files were classified as anger.

Table 1. Tabulation of recognized and unrecognized Emotion for Foreign Database

Emotions	JOY	ANGER	FEAR	BOREDOM	SAD	NEUTRAL	TOTAL
JOY	10	2	0	0	0	0	12
ANGER	0	11	0	0	0	0	11
FEAR	0	1	10	0	1	1	13
BOREDOM	0	0	1	11	0	1	13
SAD	0	0	3	3	4	3	13
NEUTRAL	2	0	2	0	0	8	12

Table 2. Tabulation of recognized and unrecognized Emotion for Local Database

Emotions	JOY	ANGER	FEAR	BOREDOM	SAD	NEUTRAL	TOTAL
JOY	31	0	0	0	0	0	31
ANGER	0	30	0	0	0	0	30
FEAR	0	2	21	0	3	1	27
BOREDOM	0	0	0	30	2	1	33
SAD	0	0	0	1	27	1	29
NEUTRAL	3	0	0	4	0	25	32

For the local database, another table was created to determine the number of recognized audio files for the six types of emotion. Results are shown in table 2.

For the foreign database, the voice recordings requested from the University of Lodz were saved to the neural network. These were sorted base on the emotion of the wav file. Details are shown in Table 3.

From the foreign data given, the average accuracy rate obtained from the sets of input files loaded in the system is 73.72%. Only anger emotion had been recognized clearly by the system without any errors while, sad emotion had the least recognition rate of 30.77% where it was not clearly interpreted and recognized by the GUI interface, same as for those emotions who have unrecognized files such as voice recordings with lack of emotions or input recording files that has noise were not yet filtered producing a different output.

Table 3. Summarization of Results for Foreign Database

Emotion	No. of Input Files	Correctly Recognized	Unrecognized	Accuracy Rate
JOY	12	10	2	83.33%
ANGER	11	11	0	100%
FEAR	13	10	3	76.92%
BOREDOM	13	11	2	84.62%
SAD	13	4	9	30.77%
NEUTRAL	12	8	4	66.67%
			AVERAGE RATE:	73.72%

Table 4 shows the data of the speech recognition files sorted according to the emotion given. The researchers obtained the local data from students and movies. The data was trained and compared to the different emotions to decide where the input file is most similar with.

Table 4. Summarization of Results for Local Database

Emotion	No. of Input Files	Correctly Recognized	Unrecognized	Accuracy Rate
JOY	31	31	0	100.00%
ANGER	30	30	0	100%
FEAR	27	21	6	81.48%
BOREDOM	33	30	3	90.90%
SAD	29	27	2	93.10%
NEUTRAL	32	25	7	78.12%
			AVERAGE RATE:	90.6%

From the local data given, the joy and anger emotions were easily recognized by the system for they have a recognition rate of 100%. Fear, boredom and sadness on the other hand had medium or average recognition rate which are 81.48%, 90.90% and 93.10% respectively. Boredom, of all, had the lowest recognition rate of 78.12%,

4. CONCLUSIONS AND RECOMMENDATIONS

For each voice recording, samples were created to check the accuracy of the learning after the sets of training. The results include the recognition rate for each emotion. In the foreign voice recording database, the network performs a high rate of more than 75% for the emotions such as anger, boredom, fear and joy while for sadness and neutral emotions, it is recommended to feed the network of more input data file to be trained to recognize the respective emotions. On the other hand, local voice recording database having more than 600 samples of voice recordings obtained plausible rate greater than 75% for each emotion which shows a high rate of accuracy of recognizing and learning of the network for the local voice recording.

Future researchers are recommended to add another set of emotional load such as disgust, surprise, sarcasm and anticipation for the improvement of the neural network scope. For the better enhancement of the loaded voice recordings, it is recommended to use a better device that will automatically remove or lessen the noise of the environment to obtain a higher accuracy rate of learning and recognition of the system.

5. REFERENCES

(n.d.). Retrieved from http://www.cse.unr.edu/~bebis/MathMethods/NNs/lecture.pdf.

Aastha Joshi, Rajneet Kaur. 2013. A Study of Speech Emotion Recognition Methods. *International Journal of Computer Science and Mobile Computing*, *2(4)*, 28-31.

Ayaz Keerio, Bhargav Kumar Mitra, Philip Birch, Rupert Young, and Chris Chatwin. 2009. On Pre-processing of Speech Signals. *International Journal of Signal Processing*, 5;3.

B. Heuft, T. Portele, and M. Rauth. 1996. Emotions in time domain synthesis. *Proc. of International Conference on Spoken Language Processing, Philadelphia*, 1974-1977.

Christopher J.C. Burges. 2000. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery,* 2, 121-167

Douglas Reynolds. (n.d.). Gaussian Mixture Models. Retrived from advancedsourcecode.com

Gongde Guo, Hui Wang, David Bell, Yaxin B and Kieran Greer,. (n.d.). KNN Model-Based Approach in Classification. *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE*, 2888, 986-996

Kismat Maredia Fall. 2010. A Study in Decision Analysis using Decision Trees and Game Theory.

N. Amir , S. Ron. 1998. Towards an automatic classification of emotion in speech. *Proc. of International Conference on Spoken Language Processing, Sydney*, 555-558.

R. Cowie, and E. Douglas-Cowie . 1998. Automatic statistical analysis of the signal and prosodic signs of emotion in speech. *Proc. of International Conference on Spoken Language Processing, Philadelphia*, 1989-1992.

Yang, L. 2000. The expression of emotions through prosody. *International Conference on Spoken Language Processing 2001, Beijing*, 74-77.