2013 IEEE International Conference on Systems, Man, and Cybernetics

Embedded Design of an Emotion-Aware Music Player

Carlos A. Cervantes
Institute of Electrical Control Engineering
National Chiao Tung University
Hsinchu, Taiwan R.O.C.
cervantes.ing@gmail.com

Kai-Tai Song, *Member, IEEE*
Institute of Electrical Control Engineering
National Chiao Tung University
Hsinchu, Taiwan R.O.C.
ktsong@mail.nctu.edu.tw

***Abstract*—In this paper, a novel human-robot interaction(HRI) design is proposed where emotional recognition from the speech signal is used to create an emotion-aware music player that can be implemented in an embedded platform. The proposed system maps an inputted short-speech utterance to a two dimensional emotional plane of valence and arousal. This strategy allows the system to automatically select a piece of music from a database of songs, of which emotions are also expressed using arousal and valence values. Furthermore, a cheer-up strategy is proposed such that music songs with varying emotional content are played in order to cheer up the user to a more neutral/happy state. The proposed system has been implemented in a Beagleboard. The online test verified the feasibility of the system. A questionnaire survey shows that 80% of subjects agree with the songs selected by the proposed cheer-up strategy based on the emotional model.**

Keywords-emotion recognition, human-robot interaction, emotional model.

I. INTRODUCTION

Current advances in technology make possible to build more complex intelligent systems in hardware constraint embedded platforms [1]. One result of this is the so called pet robots or companion robots that are typically small size robots with some human-robot interaction(HRI) functions [2, 3]. Besides pet robots, recently there has been a boom in portable hand-held devices like tablet and smartphones. It would be good if these embedded systems can interact with a user using the emotional information, such as encoded in the human speech. These systems could have some sort of better intelligence that allows them to adapt and even "better" behave instead of just acting according to some given instructions. Some previous works that relates emotion and music applied to HRI applications[3-4]. The device in [3] monitors a number of external variables to determine its user's levels of activity, motion and physical states to make a model of the task its user is undertaking at the moment and predict the genre of music that would be appropriate. Work in [4] presents an emotion-based music retrieval platform. Unlike common approaches to classification of emotion into classes, the music retrieval platform defines emotions in the two dimensional valence-arousal emotional plane. Since music is associated with the valence arousal values, each piece of music is retrieved as a point in the 2D arousal-valence emotional model making easy for the user to retrieve music with certain emotions.

This work was partly supported by the National Science Council, Taiwan, R.O.C., under grant NSC 100-2221-E-009-033-MY2.

Research in emotion recognition technology and emotion in music has been an interesting topic due to its important role in improving human-robot interfaces for robots to better understand and serve humans [6-7]. People like to listen to music and usually it is chosen based on feelings at the moment. Depending on the situation, people would like to hear music but maybe it would be inconvenient if, for example, emotions with negative feelings are being experienced. In that case it could be good if an automated system helps the user to select music that could reflect is emotions or even better that can be aware of his emotional feelings and can try to help him feel better by cheering up trough music of certain emotional content just like a real friend would do. This work tries to find a solution that can make possible to relate music and speech emotional content by instead of using discrete emotional categories, a continuous emotional model is used that allows emotional content to be expressed in a continuous way. This will allow to best differentiating songs that has similar emotional content but that are still different between each other.

II. FEATURE EXTRACTION OF SPEECH SIGNAL

The system proposed in this work is based on three main blocks: signal preprocessing, feature extraction and arousal and valence mapping. Signal preprocessing and feature extraction blocks works to get useful information from the speech signal. This information is used to get some features that are related to the emotional content in the speech. The focus in this part is to use the most useful features that has been reported in previous work on emotional speech recognition trying to avoid complex implementation since the target platform is a low cost hardware constrained embedded system. In the arousal and valence mapping block, the arousal and valence two-dimensional model is adopted to propose a method that allows mapping of obtained features from speech processing to a continuous emotional space. This kind of mapping is desired in order to relate emotional speech and emotional content in music. This block also includes an emotion cheer-up strategy that depending on detected emotional content, will select adequate tunes in order to cheer up the user to a more neutral-happy state. Fig. 1 shows the system architecture. The overall system has been realized in a Beagleboard XM embedded platform [8]. The input to the system is small duration utterance captured with a microphone and a touchscreen panel for control interface. The inputted speech is processed and a song will be

978-1-4799-0652-9/13 $31.00 © 2013 IEEE
DOI 10.1109/SMC.2013.431

IEEE computer society

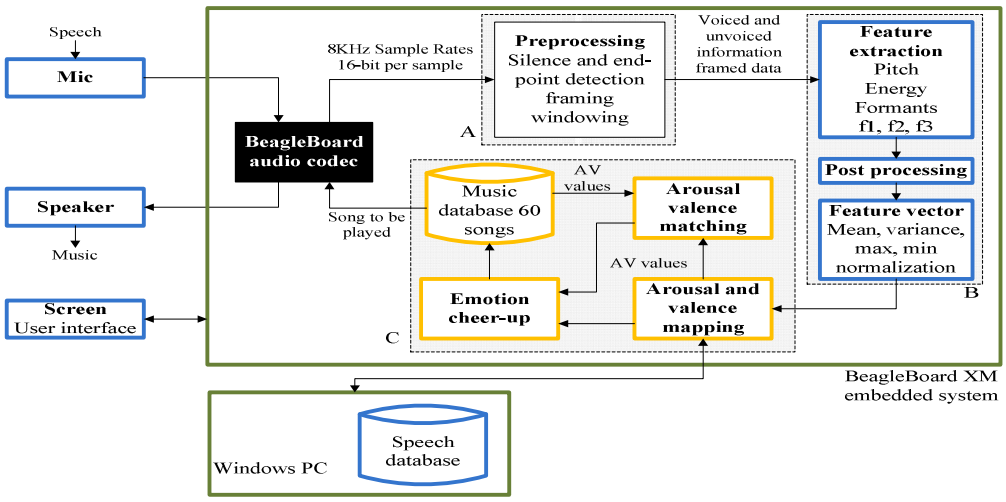

Figure 1. System block diagram of emotion-aware music player.

played depending on the emotional content in the input utterance.

A. *Emotion Recognition from Speech*

Emotional information is encoded in all aspects of language. Murray and Arnott [9] refer to a number of studies which give evidence of how emotion can be inferred from acoustics relate to speech signal. For instance, the most investigated vocal parameters are related to prosody of speech like pitch, intensity, speaking rate and voice quality. There is evidence that emotions are correlated to how these prosodic parameters change in the speech signal. Emotion can be inferred by processing some characteristics in the digital representation of the speech signal like the fundamental frequency, energy and frequency components. The emotion recognition processing is better explained in the following paragraphs.

B. *Preprocessing*

Before the selected features for emotion classification can be correctly extracted from the speech signal, it has to be quntized. Audio signal is relatively low bandwidth, usually between 100 Hz to 8 KHz and most of the energy is concentrated between the 100Hz to 4 KHz band [10].

1) Silence Removal and End Point Detection: Once having the digitized signal, silence removal and end point detection signal processing must follow in order to get rid of unnecessary information that otherwise will take time and resources to be processed. For silence and endpoint detection the algorithm proposed in [11] is used. The algorithm is based on the probability density function of the signal background noise and also the physiological aspect of speech production processes.

2) Framing and windowing: The speech signal is a slowly time varying signal [12] in the sense, that, when examined over sufficiently short period of time (between 5 and 100 milliseconds), its characteristics are fairly stationary: however, over a long period of time (on the order of 1/5 second or more) the signal characteristics change to reflect the different speech sounds being spoken. The speech signal is divided in frames of 20 to 30 milliseconds where the signal is supposed to be time-invariant with 2-3 signal periods [12]. Usually the frames are overlapped, meaning that the next frame does not start after the last sample on the last frame but instead starts in a certain sample inside the last frame. This overlap can be up to 50%.

C. *Feature Extraction*

Features for emotion recognition are mainly derived from speech recognition technology. But up to date is still unclear which set of features give the best emotional content [13]. Classical features used are based on prosody of speech, and are sometimes classified as short time features. Common prosodic features are based on pitch and energy of segments of an utterance [14]. In this work, following features are selected to be relevant to emotion speech recognition and at the same time targeting and embedded platform as a test bench.

1) Energy: The short-term energy is useful for speech emotion recognition since it is related to the arousal level of emotions.

2) Pitch: The pitch signal contains information about emotion, because it depends on the tension of the vocal folds and the sub glottal air pressure [15]. The autocorrelation method is used to calculate the pitch signal.

3) Formants: The speech production system is usually modelled as a source filter system where the source is a train of impulses representing the glottis and vocal cords, the filter is a representation of the vocal tract [16]. The resonant frequencies are called as formants and they give information about the vocal track shape. Formants are important feature for emotion recognition since the vocal track is affected by emotional states. Formants can be obtained using linear predictive coding (LPC) analysis [16], [17].

D. *Feature Vector*

After feature extraction processing, some of the feature waveforms have some noise that can affect the statistics values that will be measured in the feature vector extraction. In order to reduce the noise, a smooth of the extracted features is done using the moving average filter.

After the feature extraction and filtering processing, a feature vector is built by computing some statistical values (mean, standard deviation, maximum and minimum) on the obtained features. The feature vector has the form $F_V = (F_1, F_2, \ldots, F_n)$ where n is the feature vector length which is 20 for this work. The obtained feature vector will have values of significant difference in magnitude between each component, thus to avoid a component being dominant. The feature vector is normalized using linear normalization [18].

III. Proposed Emotion Model and Cheer-up Scheme

As people show their emotional expressions with various degrees individually, it is not an easy task to judge or to model human emotions [19]. In the literature, there are mainly two methods to model emotions. One method is to map emotions into discrete categories, e.g. joy, sadness, surprise, anger, love, fear, etc. This method has the drawback that real emotions range vary significantly and cannot be adequately expressed in discrete word labels. The second method to classify emotion is by using one or multiple dimensions or scales. Two commons scales are valence and arousal, where valence represents the pleasantness of stimuli and arousal represent the energy or activation level of stimuli. This kind of model was proposed by Thayer [20] and is termed *AV mapping*.

A. AV Mapping

In this work, the Thayer dimensional model is adopted. We divided the *AV plane* into the most common used emotion categories in order to assign arousal and valence values to the input utterances used in the neural network. Based on Thayer's work the emotional plane has been subdivided as shown in Fig. 2. For convenience, the maximum and minimum ranges of the arousal and valence values in the plane have been chosen to fall between -1 and 1. The arousal and valence value assigned to an utterance in the training phase is taken as the central value of the emotion category quadrant in the plane. So if the emotional category quadrant's origin is set as the most left and lower point of the quadrant in the emotional plane, the arousal and valence values assign to each emotion category E(A,V) used for training are calculated using (1).

$$E(A,V) = (O_A + 0.25\,, O_V + 0.25\,) \qquad (1)$$

where O_A and O_V are the quadrants origin in the emotional plane. In this work, a feed-forward back-propagation (FFBP) neural network is used to map the obtained feature vector to the 2D emotional plane [21].

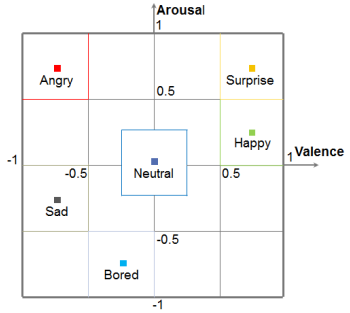

Figure 2. Thayer's emotional model and the emotional categories used for training of the neural network.

Music songs will be the output of the music player system. To adequately compare arousal and valence values from the AV mapping block and the individual songs, a fair assignment must be done since a given song emotional content must try to represent the same emotional content inferred from the user. For this purpose, the work of [22] is referenced, where a database of 60 English famous popular songs has been annotated by 40 individuals in a range between -1 and 1.

After obtaining arousal and valence values from the AV mapping module, the obtained arousal and valence values are matched to the songs ones by calculating the maximum Euclidean distance between the speech utterances and all songs on the dimensional plane using the expression (2).

$$min\left(D_n = \sqrt{(V_{n2} - V_{n1})^2 + (A_{n2} - A_{n1})^2}\right) \qquad (2)$$

where D is the Euclidean between two points (V_2, A_2) and (V_1, A_1) in the arousal and valence emotional plane. Therefore the song to be played is the one which has the minimum Euclidean distance to the imputed speech arousal and valence values. Using the Euclidean distance, the selected song for this case would be the one having the distance d1 which is the one which better represent the emotional content of the inputted speech utterance.

B. Emotion Cheer Up

The emotional mapping module not only serves to detect emotional content and reflect it as a music tune, but is also used in a strategy that aims to cheer-up the user based on the detected emotional content. Depending on the detected arousal and valence values, the system also tries to cheer up the user to a more neutral-happy state by gradually playing songs by increasing arousal and valence values towards a more neutral-happy area in the emotional plane. This target "emotion cheer-up value" can be set-up by the user as an input parameter to the system according to his personal taste since depending on the user's age or gender, arousal and valence values can be lower or higher and can be difficult to predict.

In order to accomplish emotion cheer up, once the inputted utterance location on the emotional plane has been obtained, a straight line between this location and a target point in the happy neutral zone (see Fig. 3) of the motional plane is traced using the equation of a straight line as shown in (3).

$$y = mx + b \qquad (3)$$

where y and x are valence and arousal values of a given point in the emotional plane and m and b are the slope and the intercept with the arousal axis respectively. m and b can be solved by having two points in the emotional plane which in this case are the coordinates of the detected emotion and target emotion desired for cheer up. Once having the equation of the straight line linking the desired points, it is divided evenly into a certain number of sections which is decided by the user according to how many songs he would like to hear until cheer-up target value has been reached. Then the initial and end points coordinates of each section of line in the emotional plane are calculated by knowing that the distance between each initial and end points of a section is the distance d of the

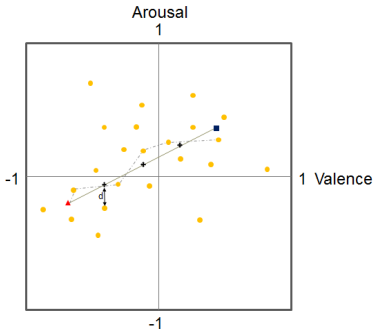

Figure 3. Songs (red dot) that are played on the cheer-up mode are selected based on their distances to the straight line between input (black dot) and target (blue dot) locations on the formed line.

straight line between the detected and target points divide by the number of sections used. Then using the equation of the straight line as (3) and the Euclidean distance as (2) yields the following expression (4).

$$(1+m^2)x_2^2 + (-2x_1 + 2mb - 2my_1) + (x_1^2 + b^2 - 2by_1 + y_1^2 - d^2) = 0 \quad (4)$$

where x_2 and x_1 are the arousal components for the beginning and end points of a line segment respectively, y_1 and x_1 the arousal and valence component of the beginning point of a line segment. Solving for x_2 using the solution for a quadratic equation yields:

$$x_2 = \frac{-B \pm \sqrt{B^2 - 4AC}}{2A} \quad (5)$$

where

$$A = (1+m^2)x_2^2 \quad (6)$$

$$B = (-2x_1 + 2mb - 2my_1) \quad (7)$$

and

$$C = (x_1^2 + b^2 - 2by_1 + y_1^2 - d^2) \quad (8)$$

After x_2 has been found, y_2 is also found by using the next equation:

$$y_2 = mx_2 + b \quad (9)$$

After all the coordinates for all the desired points along the straight line has been found, the shortest Euclidean distance between each one of these points and the songs locations in the dimensional plane are computed and for each one of the desired points in the straight line there will be a song that will be selected based on the shortest distance computed. This applies if a detected emotional content has negative arousal or negative valence values. If arousal and valences values for the inputted utterance are positive then the system just assumes that the user is in a good emotional state and a determinate number of songs with arousal and valence values close to the ones of the detected emotion calculated also using the shortest Euclidean distance to the target cheer-up emotion location are played. This process can be seen graphically in Fig. 3, where the straight line between the detected arousal and valence values (red triangle) and the target cheer-up emotion location (black square) is divided in 4 sections and then the locations of the asterisks are calculated using (8) and (9).

IV. Experimental Results

The system has been tested by a series of experiments where two kinds of databases have been used. One proprietary database was built in order to test the system in an on-line scenario, where a real person can interact with the system. For this case, the system audio codec and microphone was used to capture the speech audio signal, so it was expected that for testing purposes the recording conditions would not be altered too much. The other referenced database was used for comparison and proof of reliability of the system architecture. In the next paragraphs a description on the emotional databases used in this work will be presented, the experiments and the results made with them will be discussed.

A. Emotional speech database

In this work, emotional recognition from the speech is treated as a pattern recognition problem, where a classifier (neural network) is employed. A training data set is used in order to predict outputs from unknown input values. A database with enough utterances from different people is necessary to generate mapping of the input utterances to the two-dimensional emotional plane. A common method to build emotional speech databases is to use professional actors or normal people that can try to speak phrases emulating a determinate emotion. This method was adopted in the databases used in this work.

1) Proprietary database: This database was built in order to train the system and also to be used for on-line testing for real persons. For this database, ten people were invited to utter some short phrases using acted emotional content. The subjects were males between 20 and 30 years old. Six basic emotions were used: anger, boredom, happiness, neutral, sadness and surprise. Six invited persons uttered 36 different phrases, six phrases for each emotion category and 4 others uttered 10 phrases for each emotion category. That is a total of 456 utterances. Every utterance consists of a common used phrase in Chinese language of about 3 to 4 seconds duration. The recording were done in a quiet environment using the Beagleboard embedded system and a microphone. The setting sampling frequency was 8 KHz and 16 bit per sample.

2) Berlin Database: As a project funded by the DFG (German Research Community) a database of emotional speech (Emo-DB) [23] was built where ten actors (5 female an 5 male) uttered simulated emotional phrases. Actors uttered 10 sentences (5 short and 5 longer) which could be used in everyday communication. To achieve a high audio quality the recordings were taken in an anechoic chamber of the Technical University Berlin, using high-quality recording equipment. In total the where about 500 utterances. Recordings were taken with a sampling frequency of 48 KHz and later down sampled to 16 KHz and 16 bit per sample. This database has free access in the internet.

B. Off-Line Experiments

To test the performance of the proposed system, off-line experiments were performed where the database was divided in two parts, one part is for training the neural network and the second part is used only for validation of the architecture. The obtained data is plotted in the arousal and valence emotional plane for analysis purposes. The results are basically the arousal and valence values detected for every inputted utterance in the AV mapping module which are values in the range of -1 and 1. For every database used, after the training phase, a different FFBP neural network structure was obtained. The first test was performed using the proprietary database, where 75% of the phrases were taken for training and the other 25% were used for validation. The result is plotted in the two dimensional emotional plane in Fig. 4. As can be seen in Fig. 4, the distribution of arousal and valence pairs with boredom (b), surprise (f) and neutral (c) emotional content falls around the target emotion location. Anger distribution can be confused with neutral and some of the values have lower arousal level in the emotional plane. Values with sad, neutral and bored emotional content seems to be difficult to differentiate in the correct area since these emotional categories has very similar arousal and valence characteristics.

The second experiment was done using the Emo-DB emotional speech database. A total of 389 utterances are used as training set and 27 as the validation set using the leaving-one-speaker-out validation method, where part of uttered phrases of a subject is used as the validation set and the others as the training set. The results show that arousal and valence values has an acceptable overall classification except from values with sadness emotional content, since sadness is confused with boredom because their values are located very closed to each other.

Results obtained from the emotional databases show that an input utterance is adequately placed in the emotional plane since for a certain emotion category. This desirable characteristic is useful to later selecting songs based on the arousal and valence values detected. In the emotional planes where the songs to be proposed by the system are also plotted (red circle). In other words, the song with the closest emotional content is selected.

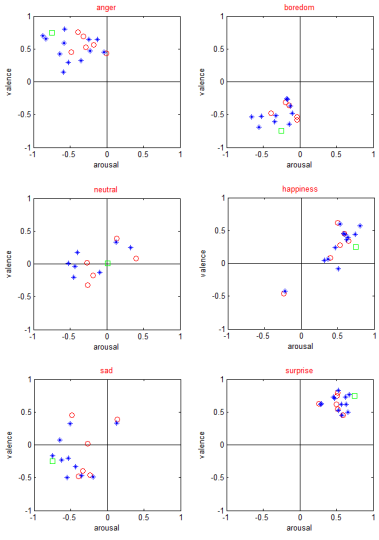

Figure 4. System output of using proprietary database. Green square indicates target emotional value, blue asterisk indicates the detected input utterance AV values and the red circle indicates the proposed song.

C. On-Line Experiment of Cheer-up Music Playing

An online experiment has been carried out in order to test system performance on a real scenario. The prototype system consists of a Beagleboard, a touch screen panel, speakers and a microphone. To interact with it, the user has to touch a record button whenever he wants to speak and then touch a stop button to stop the recording. Then the system will process the speech and the estimated valence and arousal values are generated. After few seconds a song or songs will be played depending if the cheer-up strategy has been activate. The screen also shows the emotional plane and the music song locations that will be played as green dots. If a negative emotion has been detected (negative or arousal values), the system will continuously play the songs resulting from the emotion cheer-up strategy. The user can also configure the cheer-up desired value by entering its arousal and valence value on the cheer-up configuration input boxes and can also choose the number of song he would like to hear during emotion cheer-up by entering the number in the number of songs input box.

In the experiment, the user set up the cheer-up target values as 0.61 for valence and 0.53 for arousal, and the number of songs to be listened until cheer-up target AV-value has been reached has been set as 10. The FFBP neural network with the proprietary database was used. In this case the input speech had bored emotional content and it has been placed by the system nearby the boredom area, this corresponds to a negative arousal and valence values on the emotional plane therefore the emotion cheer-up strategy has started and the songs that will be played are plotted as green dots and are connected with a line for visualization purposes. The output screen of the system after one of the inputted speech utterances has been processed is shown in Fig. 5. It can be seen that the last song to be played has a positive emotional content. System suggested songs are also shown on the screen. Fig. 6 shows the results of emotion cheer-up for the example of Fig. 5. The red dot is the detected arousal and valence

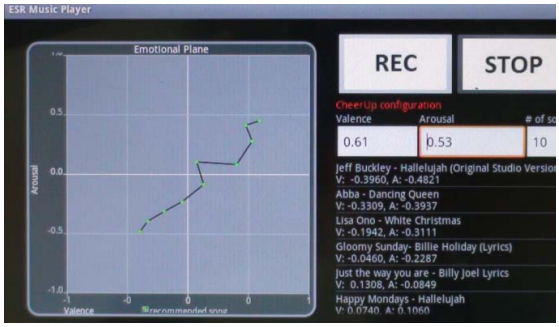

Figure 5. A screen capture of the music player.

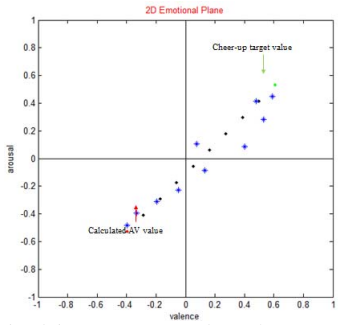

Figure 6. Example of the system output for an input utterance with bored emotional content. Songs played are blue numbered asterisks.

values and the green dot is the cheer-up target value entered by the user. The other dots colored in black are the ones calculated in the emotion cheer-up strategy according to the number of songs given by the user. The blue asterisks are the suggested songs that will be played.

The cheer-up strategy has also been evaluated by a questionnaire survey of 10 people after tested on it. In general, people agree to the system proposed songs during cheer-up mode and 80% subjects agree with the songs selected; the highest value for degree of agreement is for "somewhat agree" with 70%. From the online test is also seen that if negative arousal or valence values are detected, the cheer-up strategy successfully selects songs that will be played until a neutral happy emotional content is reached.

V. Conclusion and Future Work

An emotional speech based music player has been proposed and implemented using an embedded system platform. In order to allow the system to automatically select a song based on the user emotional state, a method to map an input speech utterance into a two dimensional emotional plane of valence and arousal has been developed. The system can automatically find a song that best matches the detected location on the emotional plane. Furthermore, a cheer-up strategy has also been proposed that according to the detected emotion will recommend songs whose emotional content will continuously change to a more neutral or happy emotional state in order to cheer up the user if a negative emotion (arousal and valence values) has been detected. Two off-line tests and an online test proved the feasibility of the proposed system. A questionnaire survey further shows that the 80% subjects agree with the songs selected by proposed cheer-up strategy based on the emotional model.

Some aspects of the system can be further improved. Adding more emotional related features can improve mapping in the emotional plane and robustness in speaker independent mode. In the future, more songs can be added to the actual music database to improve music recommendation.

References

[1] C. Wan and L. Liu, "Research of Speech Emotion Recognition Based on Embedded System," in *Proc. IEEE Int. conf. on Computer Science and Education,* Cape Town, South Africa, 2010, pp. 1129-1133.

[2] T. Shibata and K. Tanie, "Physical and Affective Interaction between Human and Mental Commit Robot," in *Proc. of 2001 IEEE International Conference of Robotics and Automation*, Seoul, Korea, 2001, pp. 2572-2577.

[3] S. C. Wang, M. J. Han and K. T. Song, "Human Emotion Recognition of a Pet Robot Using Natural Speech Information," in *Proc. of 2008 National Symposium on System Science and Engineering*, Ilan, Taiwan, 2008, P450S.

[4] Dornbush, K. Fisher, K. McKay, A. Prikhodko and Z. Segall, " XPOD - A Human Activity and Emotion Aware Mobile Music Player," in *Proc. Of International Conference on Mobile Technology, Appications and Systems,* Guangzhou, China, 2009, pp. 1-6.

[5] Y. H. Yang, Y. C. Lin, H. T. Cheng and H. Chen, "Mr. Emo: Music Retrieval in the Emotion Plane," in *Proc. ACM Int. Conf. Multimedia,* Vancouver, Canada, 2008, pp. 1003-1004.

[6] C. M. Thibeault, O. Sessions, P. H. Goodman and F. C. Harris Jr., "Real-Time Emotional Speech Processing for Neurobotics Applications," in *Proc. Int. Conf. Computer Applications in Industry and Engineering,* Las Vegas, NV, USA, 2010, pp. 239-244.

[7] N. Sebe, I. Cohen and T. S. Huang, "Multimodal Emotion Recognition," *Handbook of Pattern Recognition and Computer Vision,* World Scientific, 2005, pp. 1-23.

[8] (2011, May 5). *BeagleBoard-XM* [Online]. Available: http://www.beagleboard.org

[9] I. R. Murray and J. L. Arnott, "Toward the Simulation of Emotion in Synthetic Speech: a Review of the Literature on Human Vocal Emotion," *Journal of the Acoustic Society of America*, vol. 93, no. 2, pp. 1097-1108, 1993.

[10] R. Jang. (2011, June 10). *Audio Signal Processing and Recognition* [Online]. Available: http://neural.cs.nthu.edu.tw/jang/books/audioSignalProcessing/.

[11] G. Saha, "A New Silence Removal and Endpoint Detection Algorithm for Speech and speaker recognition applications," in *Proc. National Conf. Communications,* India, 2005, pp. 291-295.

[12] X. Huang, A. Acero and H. Hon, *Spoken Language Processing*. New Jersey: Prentice Hall, 2001.

[13] T. Iliou and C. N. Anagnostopoulos, "Classification on Speech Emotion Recognition – a Comparative Study, " *Journal on Advances in Life Sciences* , vol. 2, no. 1-2, 2010, pp. 18-28.

[14] D. Gharavian, M. Sheikhan and M. Jainipour, "Pitch in Emotional Speech and Emotional Speech Recognition Using Pitch Frequency," *Majlesi Journal of Electrical Engineering,* vol. 4, no. 1, pp. 18-28, 2010.

[15] D. Morrison, R. Wang and L. C. De Silva, "Spoken Affect Classification Using Neural Networks," in *Proc IEEE Int. Conf. Granular Computing* , Beijing, China, 2005, pp. 583-586.

[16] C. Kim, K. D. Seo and W. Sung, "A Robust Formant Extraction Algorithm Combining Spectral Peak Picking and Root Polishing," *EURASIP Journal on Applied Signal Processing*, vol. 2006, pp. 1-16.

[17] E. Bozkurt, E. Erzin, Ç. E. Erdem and A. T. Erdem, "Formant Position Based Weighted Spectral Features for Emotion Recognition," *Speech Communication,* vol. 53, pp. 1186-1197, 2011.

[18] R. Jang. (2011, November 10). *Preprocessing Data for Neural Networks* [Online]. Available: http://www.tradertech.com/preprocessing_data.asp.

[19] Y. Yoshitomi, "Effect of Sensor Fusion for Recognition of Emotional States Using Voice, Face Image and Thermal Image of Face," in *Proc. Int. Workshop on Robot and Human Interactive Communication*, Osaka, Japan, 2000, pp. 178-183.

[20] R. E. Thayer, *The Biopsychology of Mood and Arousal,* New York: Oxford University Press, 1989.

[21] J. Heaton, *Introduction to Neural Networks with JAVA*. Missouri: Heaton Research, 2008.

[22] Y. H. Yang, Y. F. Su, Y. Ch. Lin and H. Chen, "Music emotion recognition: the role of individuality," in *Proc. Int. Workshop on Human-Centered Multimedia,* Augsburg, Germany, 2007, pp. 13-22.

[23] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier and B. Weiss, "A Database of German Emotional Speech," in *Proc. Int. Speech Communication Association*, Lisboa, Italy, 2005, pp. 1517-1520.