國立交通大學

電機資訊國際學位學程

碩士論文

陪伴機器人之基於情感辨識音樂播放器系統

Emotion-Based Music Player for Companion Robots

研究生：Carlos Cervantes,史威德

指導教授：宋開泰 教授

中華民國一百零一年六月

陪伴機器人之基於情感辨識音樂播放器系統

Emotion-Based Music Player for Companion Robots

研究生：史威德　　　　Student：Carlos Cervantes

指導教授：宋開泰　　　　Advisor：Kai-Tai Song

國立交通大學
電機資訊國際學位學程
碩士論文

A Thesis
Submitted to EECS International Graduate Program
National Chiao Tung University
in Partial Fulfillment of the Requirements
for the Degree of
Master

June 2012

Hsinchu, Taiwan, Republic of China

中華民國一百零一年六月

陪伴機器人之基於情感辨識音樂播放器系統

學生：史威德　　　　　　　　　　指導教授：宋開泰　博士

國立交通大學電機資訊國際學位學程碩士班

摘要

現今低資源需求與平價之嵌入式系統已普遍應用於各式產品中。透過使用者情感表達，可以增進人們與科技產品間之互動。本研究嘗試提出互動應用的人機介面設計，利用聲音訊號作為情感辨識基礎，進而創造一個可感知情緒的音樂播放器，以應用於智慧型手機與智慧機器人。由於人類具有複雜的情緒，例如快樂或高興，不能僅藉由某些特定的數據類別來加以區分，因此我們提出的系統可以將簡短的語音投射至二維 Valence – Arousal 情緒座標，使用者任何情感可以以相隊應之連續值表示。本論文之方法允許系統自動從歌曲資料庫中選取音樂，這些被選取歌曲也以二維 Valence – Arousal 數值呈現。此外，我們提出一個激勵策略，如果使用者聲音被偵測為負面情緒，即會自動選取並播放多樣化音樂類型，用於提升使用者情感進而達到更平靜及快樂的境界。此概念已被應用現於流行且平價嵌入式平台 Beagleboard XM，它提供使用者電子麥克風及觸控式螢幕作為輸入設備。本論文提出之系統可被使用於人機介面應用，舉凡陪伴式機器人、汽車音響及通訊設備。其中通訊設備，如手機，可以根據使用者情緒狀態播放合適音樂。

Emotion-Based Music Player for Companion Robots

Student：Carlos Cervantes Advisor：Kai-Tai Song

EECS International Graduate Program
National Chiao Tung University

Abstract

Currently resource constrained and inexpensive embedded systems have become virtually an ubiquitous technology. The way in that people interact with this technology can be improved by using emotional information from the user. In this work, a novel human-machine interaction system is proposed where emotional recognition from the speech signal is used to create an emotion aware music player that can be implemented on a standard smartphone-like embedded platform. Since a person's emotion cannot always be classified in a certain number of categories such as happy or angry, the proposed system maps an inputted short speech utterance to a two dimensional emotional plane of valence and arousal, where any given emotions can have continuous values. This strategy allows the system to automatically select a piece of music from a database of songs, in which emotions are also expressed using arousal and valences values. Furthermore, a cheer-up strategy is proposed where in case the detected emotional content is detected as negative, music songs with varying emotional content are played in order to cheer-up the user to a more neutral – happy state. The system has been implemented in a popular inexpensive embedded platform Beagleboard XM which uses an electret microphone and a touch screen panel as input from the user. The proposed system can be used in human-machine interface applications like companion robots, car sound systems and communication devices like cellphones, where music can be played according to user's emotional state.

Acknowledgments

I would like to thank my advisor Professor Kai-Tai Song for his insightful advices and support throughout the course of my work at the lab. I would like to thank also my oral defense committee members – Professor Shun-Feng Su, Professor Kuu-Young Young, and Professor Jwu-Sheng Hu for their valuable suggestions for improvement. I also thank to all members of the Intelligent Systems and Control integration Lab who were always very kind and helpfully whenever I needed. And last but no less I would like to thank my wife and son for their motivation and support during my studies at NCTU.

Carlos Cervantes

Hsinchu, Taiwan, Republic of China

June 2012

Table of Content

List of figures

List of tables

I. Introduction

1.1 Motivation

Current advance 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 kind of intelligence [2]. These pets' robots can perform some actions based on speech and face recognition capabilities. Besides pet robots, recently there has been a boom in portable hand-held devices like smartphones, of which processing capabilities are constantly increasing; some of them even have similar capabilities of those found in a desktop PC but with the advantage of consuming very low power and having a very small footprint. It would be good if these embedded systems can interact with us using the emotional information that is encoded in the human speech. Like that, 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.

1.2 Previous Related Works

The work to be proposed in this thesis focuses on the human-robot interaction (HRI) and emotional speech cognition (ESR) technology. ESR basically aims to automatically identify the emotional or physical state of a human being from his or her voice [3]. Emotion recognition has been a hot research topic and its main goal is to allow a robot or a machine to better behave or adapt to the user, thus making the interaction between human and machine more natural and friendly [4]. Although there are several methods in that emotion can be detected, for instance biosensors (EKG, ECG, blood pressure), or facial emotion recognition, in particular emotion recognition from speech requires less hardware and computational complexity. The following paragraphs will present some previous works that uses emotion recognition technology from speech applied to HRI applications.

In [5] the authors tackle the problem that uses combined audio-visual information in emotion recognition and works by employing very straightforward and simple rules. Authors then used speech emotion recognition as a part of a bimodal system comprising also face recognition, in which a novel probabilistic strategy is studied for a support vector machine (SVM) based classification design to assign statistically information-fusion weights for two feature modalities. Figure 1 shows their design. For visual recognition, the processing method has face detection and facial feature extraction algorithms. For speech processing, after endpoint detection and framing, features from audio signal used for emotion classification are statistics of short time energy and pitch. At last by setting proper weights to each modality based on theirs recognition reliability, a more accurate recognition decision is obtained. The complete system is implemented in a DSP-based embedded system which can recognize five facial expressions on-line in real time. Authors claim to achieve recognition rate of 86.9% which is 5% improvement if only image information is used.

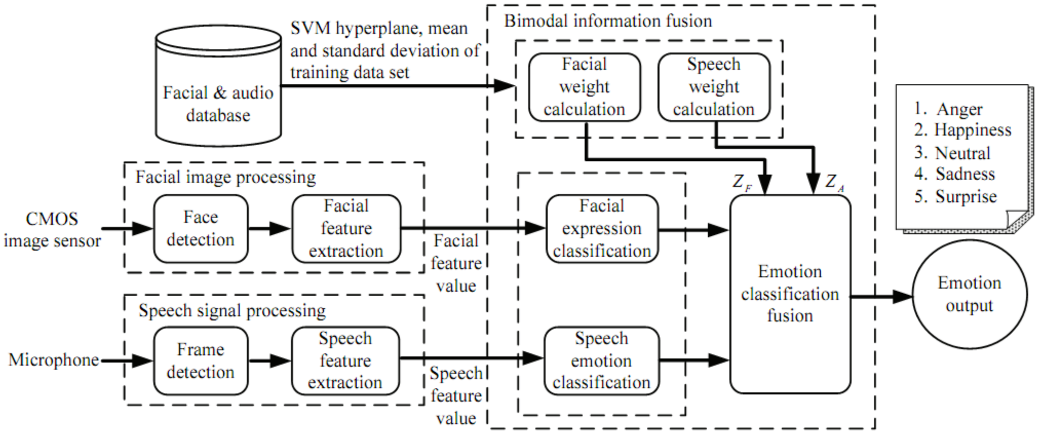

Facial & audio database
SVM hyperplane, mean and standard deviation of training data set
Bimodal information fusion
Facial weight calculation
Speech weight calculation
1. Anger
2. Happiness
3. Neutral
4. Sadness
5. Surprise
Z_F
Z_A
Facial image processing
CMOS image sensor
Face detection
Facial feature extraction
Facial feature value
Facial expression classification
Emotion classification fusion
Emotion output
Speech signal processing
Microphone
Frame detection
Speech feature extraction
Speech feature value
Speech emotion classification

Figure 1: Bimodal Emotion Recognition System in [5].

The work in [6] evaluates the performance of a speech emotion recognition method for affective human-robot interaction. Six emotions can be recognized (angry, bored, happy, neutral, sad and surprised). Basically the systems works by obtaining a feature vector from the sampled voice signal and using a Gaussian support vector machine as emotion classifier. The features obtained from speech utterances are log energy, shimmer, formant frequencies and Teager energy. The support vector machine maximizes the margin between two classes. Using a kernel function the SVM maps the space S ={x} consisting of the input samples into a high dimensional feature space so that the input samples become linearly separable in the high dimensional feature space. The kernel function K(x,x_i) of new input x and training input x_i is:

$$f(x) = \sum_{i=1}^{SV_s} \alpha_i\, y_i K(x_i, x) + \alpha_0 \qquad (1)$$

were SVs is the number of the training inputs and $y_i = \pm 1$ are the label of the training inputs x_i. The parameters $\alpha_i > 0$ are optimized during training phase. Figure 2 and Figure 3 show the support vector machine with a kernel function example and the system architecture used in [6].

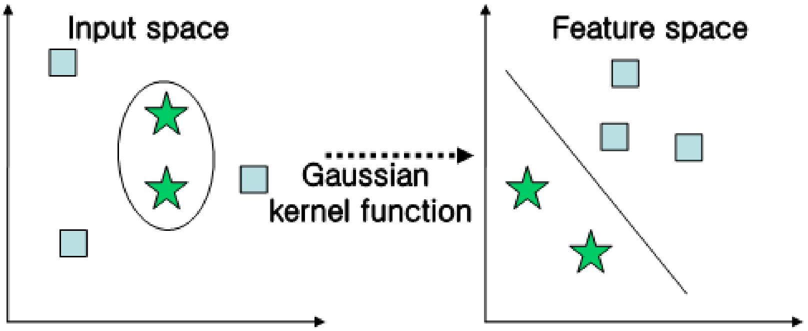


Figure 2: SVM with a kernel function in [6].

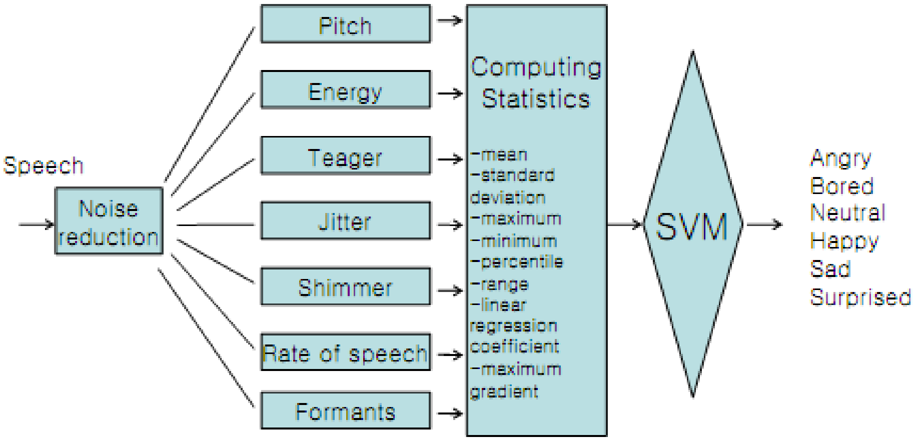


Figure 3: Emotion recognition system in [6].

The speech data based was recorded using Korean language according to a human-robot interaction scenario enrolling sentences, greeting dialogs, some conversations regarding weather and time and speech commands to control moving direction of robots. Authors claim to obtain a classification accuracy of 58.6% compared to manual classification obtained from human listeners of 60.4%.

Work in [7], authors proposes a robot driven camera that can be controlled manually by a joystick or using speech commands used in laparoscopic surgery (kind of minimal invasive surgery), but to overcome the problems arising from high stress and partially fatigue, the motional factors are also integrated in the human robot interaction. In the case of a confused or angry surgeon the interface initializes a security callback dialog to certify that understood camera direction is correct. Authors employ a high dimensional acoustic feature space and subset optimization for recognition of positive versus negative emotion for interaction adaptation, surgeon self-monitoring. As database authors discuss the recording of a 3,035 turns of spontaneous emotional speech in real life surgical operations. Vocabulary consists of highly limited ten terms (camera, quit, move, stop, up, down, left, right, forward, and backward). The speech recognition system is based on a support vector machine classifier Model. A strictly systematic generation of features was chosen for the construction of large feature space as basis for subsequent selection of relevant features. Features used include: duration, energy, pitch, formants, cepstral, and voice quality. Speech is sampled at 16 kHz, 16 bit. Detected emotion is mapped into 3 categories: positive negative and neutral. The authors say that the recordings showed the need of handling of emotional speech since only 53% of the surgeon-robot interaction turns were labeled neutral. After brute-force generation and subsequent space optimization 75.5% accuracy could be reached for the discrimination of positive and negative emotion by the SVM. Work in [08], introduced a human activity and emotion aware mobile music player Xpod, which is based on the idea of automating much of the interaction between the music player and its user. It learns user's preferences, emotions and activity and reproduces music selections accordingly.

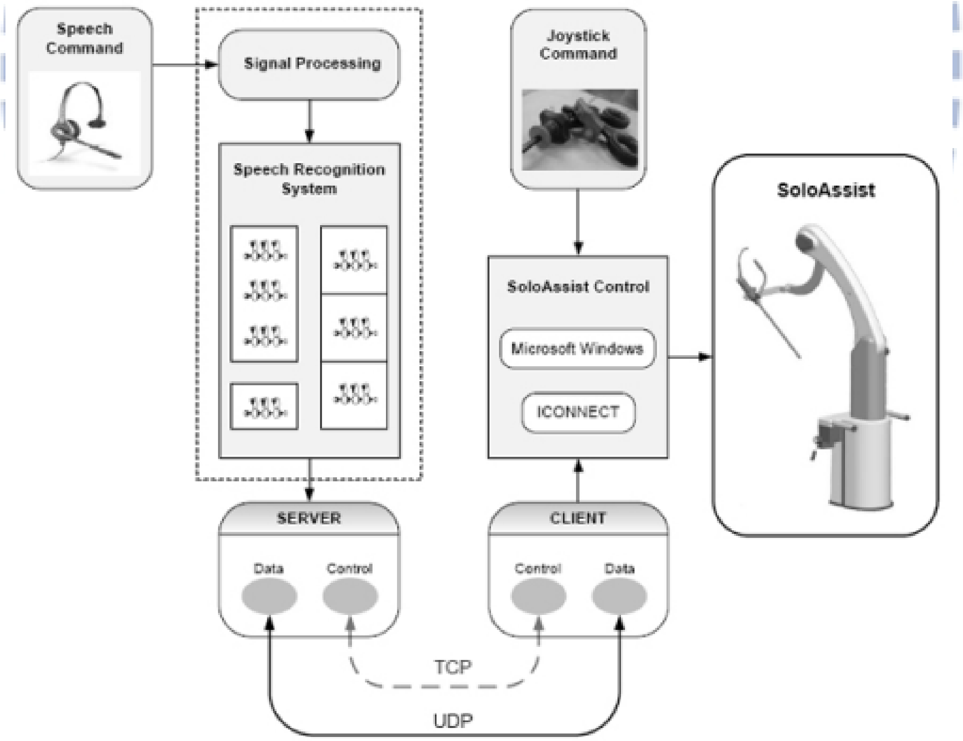


Figure 4: Speech control interface for surgeon assistant robot in [7].

The device in [8] 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. The Xpod relies on a BodyMedia Sense Wear device which is a real-time device that uses a series of sensors to measure the rate of body movement, acceleration, and body heat from user. This data is used by a series of algorithms that determine average, and standard deviation and compare the results to a pre-determined range that would correspond to "active", "passive" and resting. Once the user state is determined, information is passed to a neural network, which compares the user's current state, time and activity levels to past user song preferences matching the existing set of conditions and makes a musical selection. The XPOD neural network architecture is shown in Figure 5. The Xpod system is composed of a server and a client scheme. The server is a laptop PC in which processing is done. The client is a personal digital assistant (PDA) where the music player interfaces its running. Client and server communicate through Wi-Fi network. The Xpod concept was tested with a simple ranking rule according to the song Beats-per-minute (BPM).

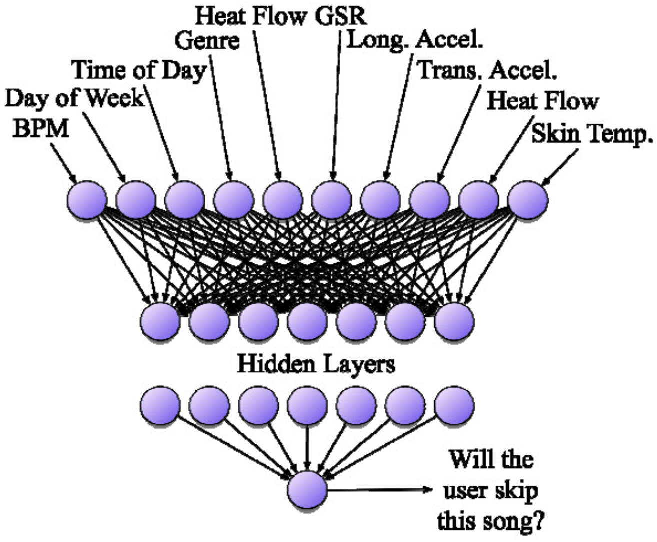


Figure 5: XPOD neural network format in [8].

When doing activity the user will give a high rank on high BPM songs being played and low rank to songs with low BPM, when resting the user will rank song in the opposite way. The system proves to learn the pattern of listening behavior exhibited by the test user. Author claims that although the pattern of listening was trivial and easily testable, much more complicated patterns would be learned by the proposed algorithms.

Work in [9] presents an emotion-based music retrieval platform, called Mr. Emo, for organizing and browsing music collections shown in Figure 6. Unlike common approaches to classification of emotion into classes, the music retrieval platform defines emotions in the two dimension valence-arousal emotional plane. Since music is associated with the valence arousal values, each piece of music is easily retrieved as a point in the 2D arousal-valence emotional model making easy for the user to retrieve music with certain emotions. The proposed system architecture is composed of two main parts, the prediction of arousal-valence values using regression models, and the emotion-based visualization and retrieval of music samples. The emotion prediction is done by using a support vector regression method for training.

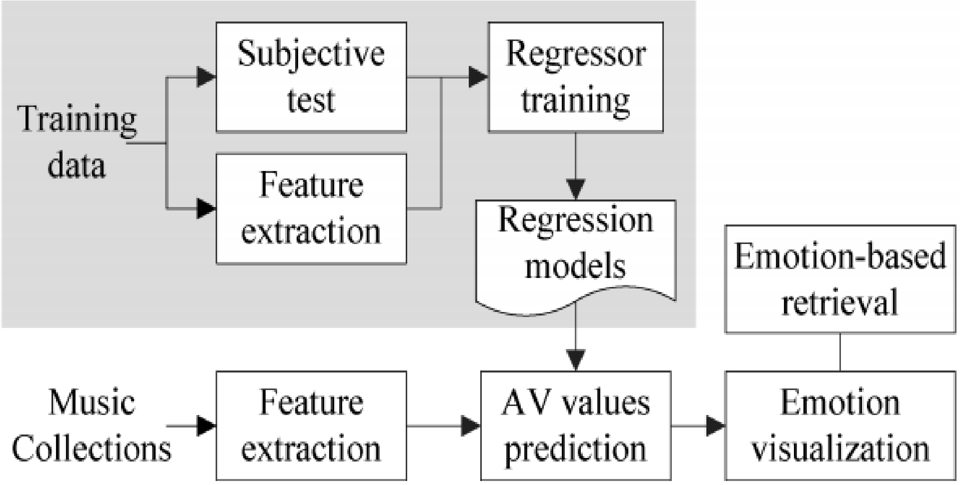


Figure 6: System architecture of Music retrieval platform in [9].

The training set is composed of 60 English pop songs, whose arousal-valence values are annotated by 40 participants. For feature extraction, authors uses 52 timbral texture features (spectral centroid, spectral roll off, spectral flux and MFCC) and 192 MPEG-7 features (spectral flatness measure and spectral crest factor). Prediction accuracy reaches 0.793 for arousal and 0.334 for valence which authors claims satisfactory in the light of valence modeling since previous studies show that even human subjects can easily perceive opposite valence for the same song. Emotion-based visualization and retrieval system predicts arousal-valence values for a new given song automatically using the regression model. Each piece of music is visualized as a point in the two dimensional emotional plane and similarity between music samples is measured by Euclidean distance. Figure 7 shows the samples distribution of some Beatles songs obtained by the music retrieval platform. Authors' system demonstration has three retrieval methods: the user can retrieve music of a certain emotion by specifying a point in the emotional plane, user can create a playlist by drawing a free trajectory representing a sequence of emotions in the emotion plane and the user can search by specify an author and a desired emotion.

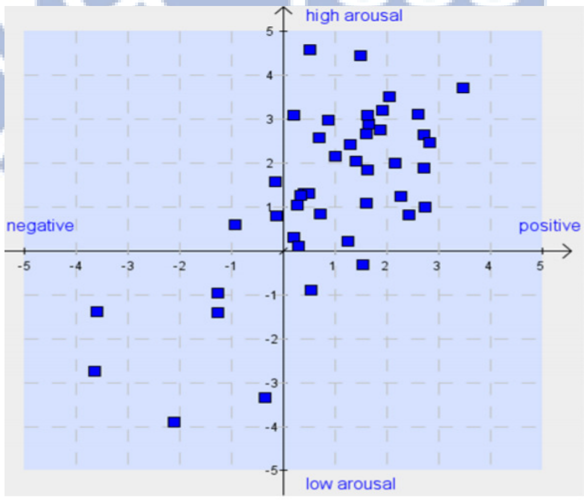


Figure 7: Music samples distribution of The Beatles according to music retrieval platform in [9].

1.3 Problem Definition

Research in emotion recognition technology and emotion in music has been a hot topic due to its important role in improving human-robot interfaces for robots to better understand and serve humans [10]-[12]. 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.

The idea of the current work is to design a music player that is expected to recognize emotional content in the speech using some basic features and will be able by some means to map them in to songs that will reflect the user emotional state. In a real scenario, this kind of system will be designed so it can be implemented in an inexpensive resource constraint platform as the ones used in small personal robots that are able to "understand" the user emotions by playing music accordingly and furthermore that in case of a "negative" emotion is detected, the system would try to cheer up the user just like a user good friend will try to do.

II. System Architecture and Design

2.1 System Architecture

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 purpose is 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 allow mapping of the obtained features from speech processing in 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. Figure 8 shows the system architecture. The three main blocks are identified as preprocessing (A), feature extraction (B) and arousal and valence mapping (C). In the Following paragraphs and sections a detail explanation of the overall system architecture will be presented.

Figure 8 also shows that the three main blocks in the system architecture can be further subdivided into several blocks named: preprocessing, feature extraction, post-processing, arousal and valence mapping (AV mapping), arousal and valence matching and emotion cheer-up. The preprocessing block does end-point and silence removal detection, signal framing and windowing. The output from this module contains the voiced data divided into 30 ms windowed frames (where speech is mainly stationary). The feature extraction block calculates pitch, energy and the first three formants on every frame. This module follows a post-processing block where filtering processing for signal smoothing get rid of some noise in the features waveforms. Then statistical values of these features (mean, standard deviation, maximum and minimum) are calculated to form a feature vector. The arousal and valence mapping maps the feature vector into arousal and valence values in the two-dimensional emotional plane by using a feed-forward back-propagation neural network architecture. The arousal valence matching module takes the emotion mapped in the two-dimensional emotional plane and will find a suitable song from the database that best matches the given emotion. The emotion cheer-up not only reflects your current emotional state by playing a song accordingly, but also tries to continuously change songs in order to keep user's emotional state in a neutral-positive state if a negative emotion has been detected.

The overall system is mapped and realized onto the Beagleboard XM embedded platform [13] which among several characteristics has an 1GHz ARM processor, and audio codec that is used for speech signal acquisition and audio reproduction, 512 MB of RAM memory, a HDMI connector for a touchpanel and a Micro-SD slot. The input to the system is small duration utterance captured with an electret microphone and a touchscreen panel for control interface. The inputted speech is processed and a song will be played after few seconds depending on the emotional content in the input utterance. The system contains a referenced music database composed of 60 songs where arousal and valence values have been manually annotated by several invited people. Every block on the system architecture will be explained in detailed in the next sections.

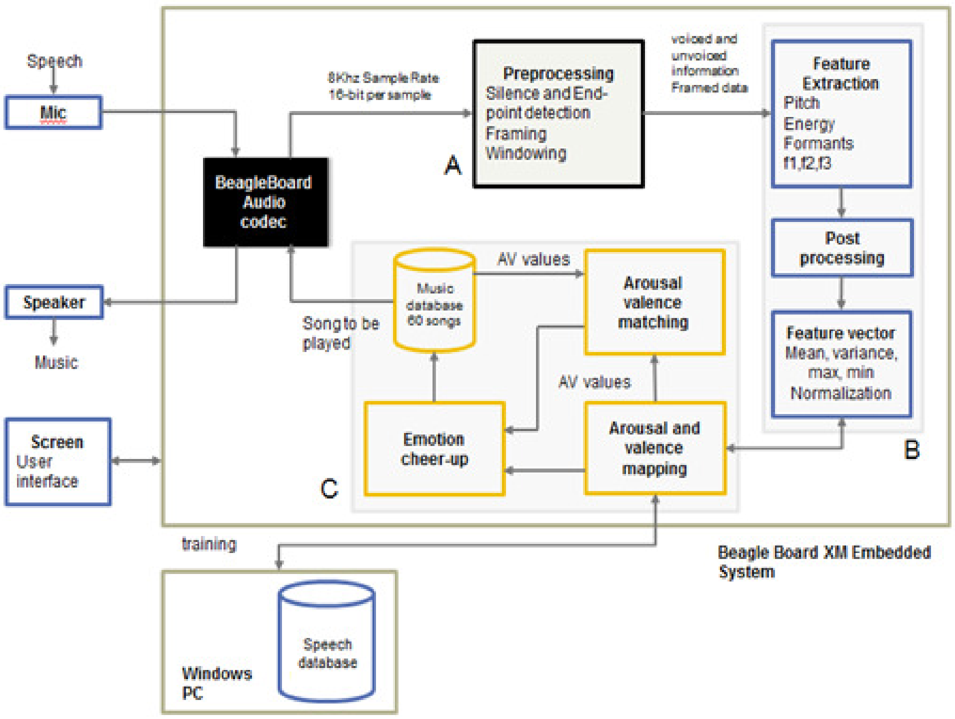


Figure 8: Emotion based music player system block diagram.

2.2 Emotion recognition from speech

Emotional information is encoded in all aspects of language. Murray and Arnott [14], 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 recognition from speech is usually done by processing extracted acoustic information from the speech signal. First, signal preprocessing is done and then SER is tackled as a pattern recognition task which implies the following steps: feature extraction, feature selection, classifier choice and testing [15].The process starts when human voice is produced: vocal cords open and close (glottis) very quickly generating vibration in air flow

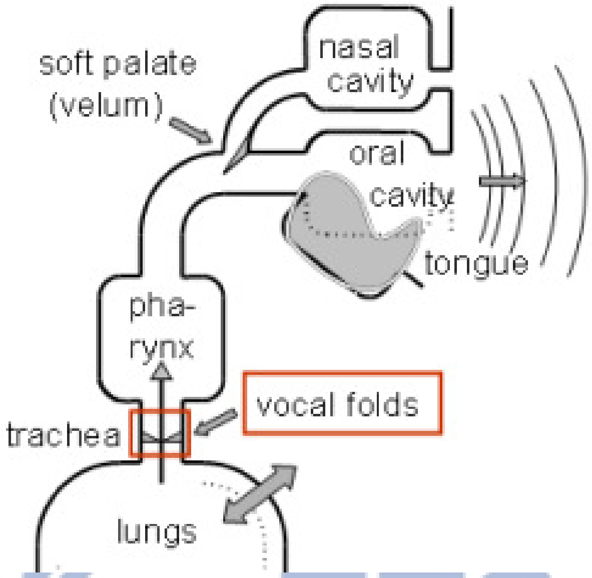


Figure 9: Human Voice production system [17].

coming from the lungs which fundamental frequency is called pitch. The air vibration is then modulated by the resonances of the pharyngeal/nasal/oral cavities, forming different timbre in the voice [16]. Figure 9 shows a graphical representation of the human voice production system.

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 human voice is then digitized and quantized to represent it as a digital waveform that can be then processed by a digital system. The emotion recognition processing is better explained in the following paragraphs.

2.3 Preprocessing

Before the selected features for emotion classification can be correctly extracted from the speech signal, it has to be preprocessed in the following way: sampling and quantization, silence and end-point detection, framing and windowing. These steps will be explained next.

The audio signal is inherently analog so in order to be processed by a computer it has to be represented as a digital wave form. Two important parameters when dealing with digital signals is the sample rate (number of samples took from the original analog signal per second) and the bits per sample. 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 [18], so for a good digital representation the sample rate has to be at least two times of the higher frequency to be analyzed according to the Nyquist theorem. If higher frequencies are needed, a higher sampling rate of usually 16 KHz can be used but this requires more computing power and memory. The higher bits per sample the better representation of the input signal is obtained, but this also increases computing power and memory needed, so usually 8 12 or 16 bits are used when dealing with speech signals. The digitization of the analog signal is made by the audio codec embedded in the Beagleboard XM. Once properly configured, it takes the audio signal coming from an electret condenser type microphone (line-in input) and outputs the raw sampled signal.

2.3.2 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 [19] 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. The normal density has the traditional "bell-shaped curve" and is completely determined by the numerical value of the two parameters the mean μ and the variance σ^2:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad (1)$$

The distribution showed in (1) is symmetrical about the mean, the peak occurring at $x = \mu$, and the width of the bell is proportional to the standard deviation σ, usually, distributed data point tend to cluster around the mean. A natural measure of distance from x to mean is the distance $|x - \mu|$ measured in units of standard deviation and which is analytically expressed as:

$$r = \frac{|x-\mu|}{\sigma} \tag{2}$$

which is also defined as the "Mahalanobis distance". A standardized normal random variable $r=(x- \mu)/\sigma$ has zero mean and unit standard deviation that is:

$$p(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{\mu^2}{2}} \tag{3}$$

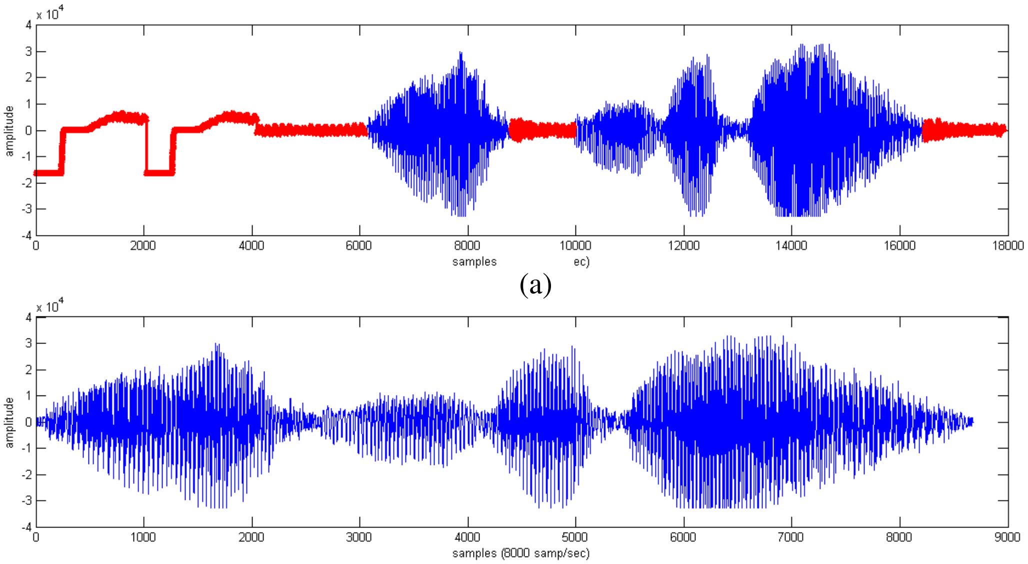
In an utterance (piece of speech) usually the first 200 milliseconds (which are 1600 samples using a sampling rate of 8 kHz) corresponds to silence or background noise because the speaker takes some time to start speaking when recording starts. Knowing this and using the above theory the algorithm for silence removal and endpoint detection goes as follows:

- Calculate the mean μ and standard deviation σ of the first 1600 samples of the given utterance. Then going from first to last sample of utterance in each sample x, check if $|(x- \mu)/\sigma| > 3$ or not. If is greater than three, the sample is treated as voiced otherwise it is silence/unvoiced. Noting here that the threshold reject the samples up to 99.7% since the probability P_r of data values clustering around the mean is given by the next relation:

$$P_r(|x-\mu| \leq 3\sigma) = 0.997 \tag{4}$$

- Mark the voiced sample as 1 and unvoiced sample as 0. Divide the whole signal into 10ms non-overlapping windows.
- In each window if the number of ones is greater than zeros, mark zeros as one and vice versa. This method keeps in mind that the speech production system cannot change abruptly in a short period of time window (taken 10 milliseconds if 8 kHz sampling rate adopted).
- Collect the voiced part only according to the labeled “1” samples and dump in a new array.

Figure 10 shows an input speech and the processed signal using the algorithm of silence removal and endpoint detection described above. In (a) the input speech signal is composed of both speech and silences marked as blue and red respectively. In (b) after processing using the endpoint detection and silence removal, just the speech signal is left marked as blue. As can be seen, the total number of samples has been reduced from 18000 to almost 9000 which will reduce processing time considerably.



(a)

(b)

Figure 10: (a) Input utterance and (b) processed signal using the silence removal and endpoint detection algorithm.

2.3.3 Framing and windowing

The speech signal is a slowly time varying signal [20] 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. Time invariant signals are usually assumed in signal processing techniques, therefore 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 [20]. Thus at 8 KHz sampling rate, a frame is composed of 160 to 240 samples. So instead of processing the whole utterance, every frame on the signal is processed separately and then a new representation of the signal is created where each sample is the result of processing the corresponding frame. 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%. Figure 11 shows a speech signal waveform and how it is divided into overlapping frames F1, F2, etc.

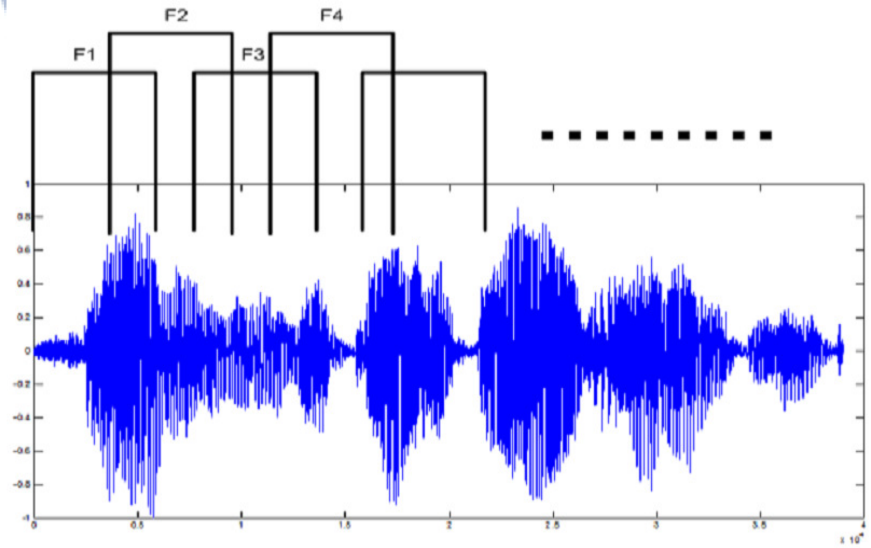
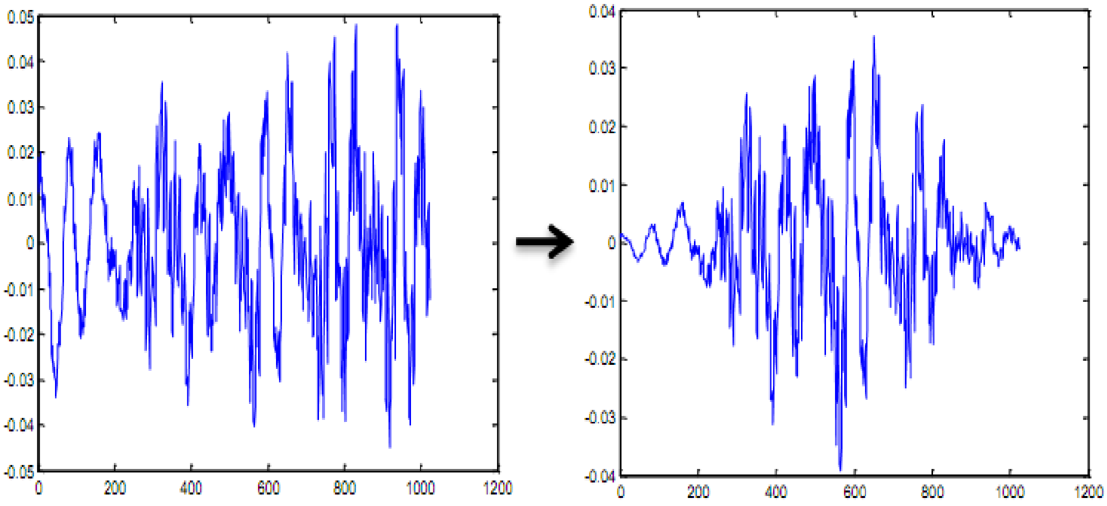


Figure 11: Example of framing of a signal [20].

The total number of frames in which a signal is divided is a function of the frame length, frame shift (number of samples that separates two overlapped frames. If overlapping is 50% and frame length is 160 samples, then frame shift will be 80) and signal length which is the total number of samples in a given speech signal as given by (5):

$$\#of\ frames = \frac{signal\ lenght - frame\ lenght}{frame\ shift} \tag{5}$$

For some speech processing, the abruptly cut off of the signal at the boundaries (Figure 12a) of frames can cause problems. Therefore a windowing operation is used where the signal in every frame is processed so samples at boundaries are altered. Figure 12 shows a frame signal (a) and the obtained windowed signal after it has been processed using a Hamming window. If samples in each frame of the signal remain unaltered it is assumed that a rectangular window has been used. On the contrary,



(a) (b)

Figure 12: Example of signal windowing. (a) original signal and (b)obtained signal after multiplication with a hamming window.

there is the hamming window which shrinks the value of the signal towards zero at the window boundaries (Figure 12b), avoiding discontinuities and is defined as follows:

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi(n-1)}{N-1}\right), 1 \le n \le N \qquad (6)$$

where *N* is the total number of samples per frame. The windowed signal *y(n)* is obtained by multiplying the signal *s(n)* composed of all the samples in a given frame and the hamming window *w(n)* as in (7):

$$y(n) = \mathrm{s(n)} * \mathrm{w(n)}, 1 \le n \le N \qquad (7)$$

2.4 Feature Extraction

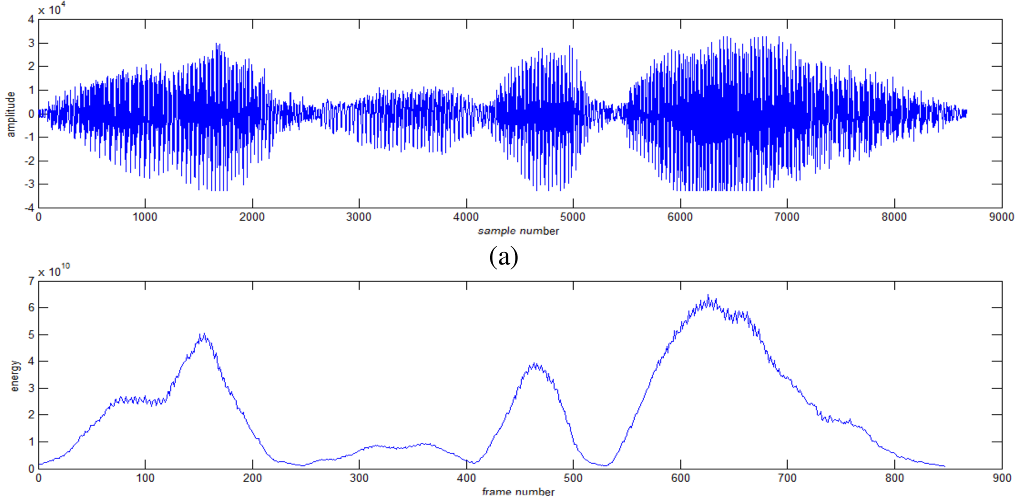
A critical issue in the realization of the emotion recognition system is the selection of feature set to use [21]. Features for emotion recognition are mainly derived from speech recognition technology. But up to date is still unclear which set of features give the best emotional content [22]. There is evidence, suggesting that only the use of few combined features can contribute for more accurate classification, and yet the excessive use of many elements degrades system performance [23]. 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 [24]. Short time features and its statistics provide important emotion-related information and are usually calculated on a frame-basis since fundamental frequency of speech is assumed to be stationary in these frames. For these work the next features are selected to be relevant to emotion speech recognition and at the same time targeting and embedded platform as a test bench.

2.4.1 Energy

The short-term energy is useful for speech emotion recognition since it is related to the arousal level of emotions [25]. The short term energy E_m of a speech signal *s(n)* that has been divided in *m* number of frames *f(n)* of a given is calculated as:

$$E_k = \sum_{n=1}^{N} f[n]^2, 1 \leq k \leq m \tag{8}$$

where *N* is the total number of samples in the framed signal *f(n)*. The result of the processing is a waveform in which every sample represents the energy contained in the respective frame used to processing. Figure 13 shows the result of short-term energy calculation over a speech signal. Figure 13 shows when the speech signal has a low amplitude value so does the energy so it can be inferred that for calm uttered speech like sad ore bored emotional speech, the result average energy will be lower if compared to that for exited speech where speech amplitude will be high.



(a)

(b)

Figure 13: Short-term energy (b) of a speech signal (a). Frame size is 320 and frame shift is 160 using rectangular window.

2.4.2 Pitch

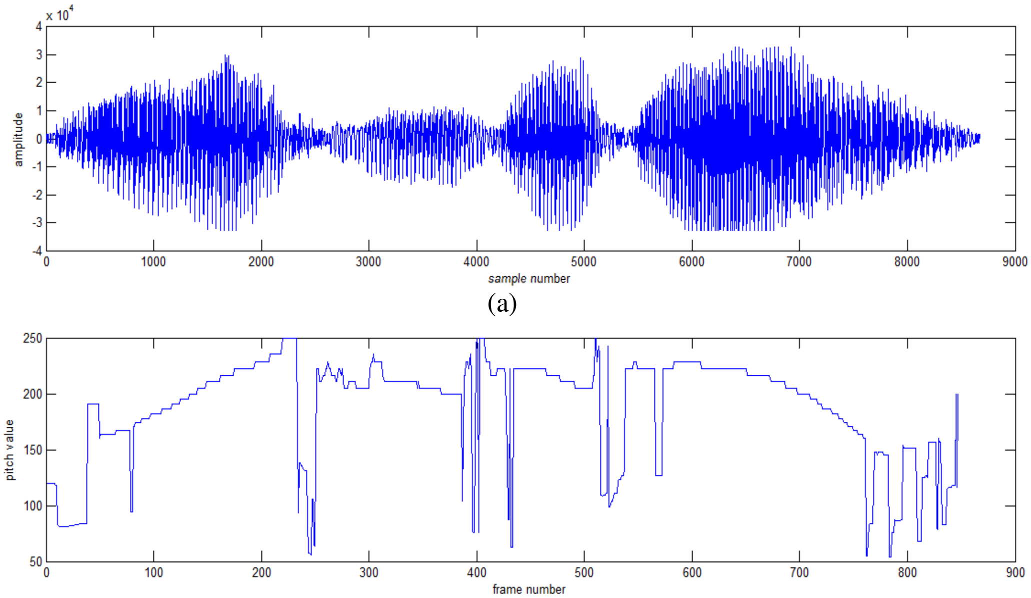
The pitch signal contains information about emotion, because it depends on the tension of the vocal folds and the sub glottal air pressure. The pitch signal is produced from the vibration of the vocal folds [26]. Pitch signal represents the vibration rate of the audio signal which can be represented by the fundamental frequency or the fundamental period. The glottal volume velocity denotes the air velocity through glottis during the vocal fold vibration. High velocity indicates music like speech like joy or surprise. Low velocity is in harsher styles such as anger or disgust [27]. There are many methods for calculating pitch signal. Among them is the autocorrelation method which is time domain based and is very suitable for embedded systems since required computing power is less compared to other more elaborated methods using frequency domain. The autocorrelation function estimates the similarity between a frame *s(i), i = 0 ~ n-1*, and its delayed version:

$$ACF(\tau) = \sum_{i=0}^{n-1-\tau} s(i)s(i+\tau) \tag{9}$$

where τ is the lag tab in terms of sample points. The value of τ that maximizes the autocorrelation function *ACF(τ)* over a specified range is selected as the pitch period in sample points. The pitch is then obtained using the expression:

$$pitch = \frac{Fs}{max\,|autocorrelation\ function|} \tag{10}$$

where *Fs* is the sampling frequency which for this work is 8000 samples per second. Figure 14 shows the obtained pitch waveform (b) for an inputted utterance (a) using the Beagleboard system. It's seen that it has a range between 60 and 250 Hz. But there is some noise that will be removed in the post-processing module.



(a)

(b)

Figure 14: (a) Speech signal and (b) pitch signal calculated using autocorrelation function (b). Frame size is 320 and frame shift is 160 using rectangular window.

2.4.3 Formants

The speech production system as shown in Figure 9 is usually modeled 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 [28]. The vocal track can be represented as a concatenation of lossless tubes; these tubes shaped the incoming wave as if the sources signal where being filtered by the resonances frequencies associated to the tubes. 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. Figure 15 shows a simplified diagram of the speech production model [28].

Formants can be obtained using linear predictive coding (LPC) analysis. Linear prediction is an adequate all-pole model to voiced speech signals. Parameters of all-pole model are representative of formant positions [28].

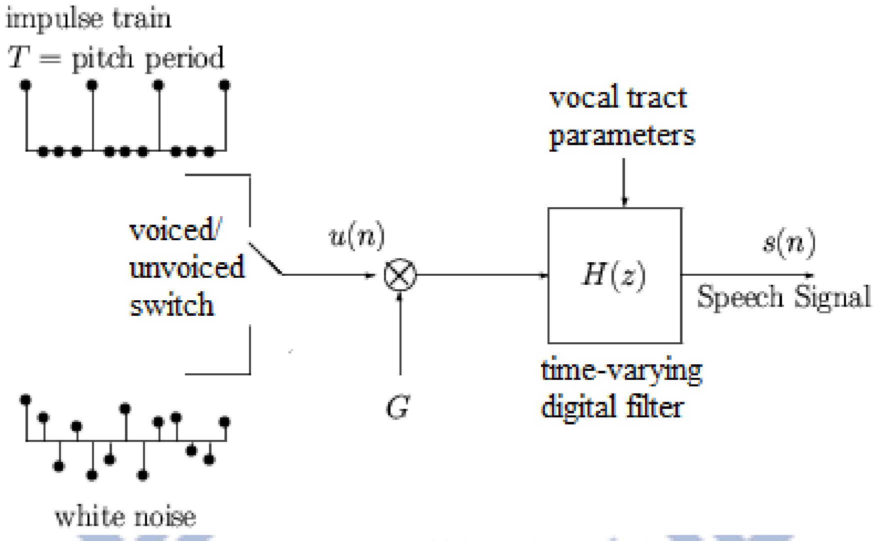


Figure 15: Block diagram of simplified model for speech production [28].

Figure 16 shows the frequency spectrum of the “ae” sound using the Fourier Transform (a) and suing linear predictive coding (b). And the first three formant frequencies named as F1, F2 and F3.

The basic idea of linear prediction is that a current speech sample can be closely approximated as a linear combination of past samples:

$$s(n) = \sum_{k=1}^{p} \alpha_k s(n-k) + Gu(n) \tag{11}$$

Applying the z transform:

$$H(z) = \frac{S(z)}{U(z)} = \frac{G}{1 - \sum_{k=1}^{p} \alpha_k z^{-k}} \tag{12}$$

Formants are found relaying on the α_k coefficients on Equation (12) which are also known as the LPC coefficients which are the solution to the autocorrelation equation in (13) [29]:

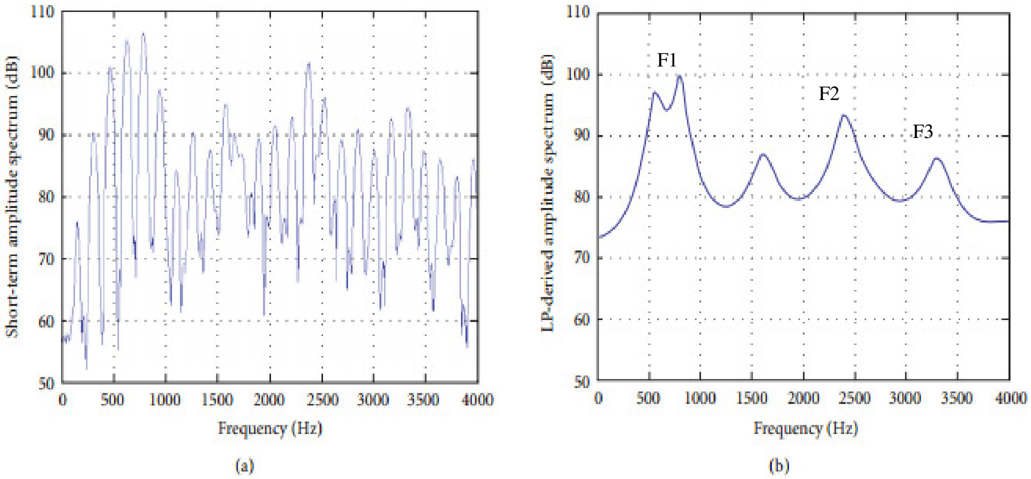


Figure 16: (a) Signal spectrum for the "ae" sound obtained using Fourier Transform and (b) the smooth spectrum obtained using linear prediction coding. First three formants frequencies are marked as F1, F2 and F3 [28].

$$\begin{bmatrix} R_0 & R_1 & . & . & R_{p-1} \\ R_1 & R_0 & . & . & R_{p-2} \\ . & . & . & . & . \\ . & . & . & . & . \\ R_{p-1} & R_{p-2} & . & . & R_0 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ . \\ . \\ \alpha_p \end{bmatrix} = - \begin{bmatrix} R_1 \\ R_2 \\ . \\ . \\ R_p \end{bmatrix} \tag{13}$$

A computational effective solution for the LPC coefficients is using the Levinson-Durbin algorithm [29]:

```
Given $\alpha_n(0) = 1$ and $E_0 = R_0$

For n=1 to p
    $k_n = -\frac{1}{E_{n-1}} \sum_{i=0}^{n-1} \alpha_{n-1}(i) R_{n-i}$
    $\alpha_n(n) = K_n$

        For i=1 to n-1
        $\alpha_n(i) = \alpha_{n-1}(i) + k_n \alpha_{n-1}(n-i)$
        End For

    $E_n = E_{n-1}(1 - k_n^2)$

End For
```

where $\alpha_p(1)$, $\alpha_p(2)$, …, $\alpha_p(p)$ are the p LPC coefficients. The formant frequencies are then estimated by finding the first three peaks on the spectrum signal obtained by using the discrete Fourier transform on the 256 samples signal formed from the LPC coefficient and some zero padding. This process is done on a frame-basis and results in formants frequencies waveforms as the one shown in Figure 17. In this case, an input speech utterance is processed and then on a frame basis the LPC coefficient for a P order of 11 (3 formants are needed) are founded using Levinson-Durbin algorithm giving 11 LPC coefficients. To form a 256 points signal, the 11 founded LPC coefficients are zero padding which means that 256 minus 11 zeros are added and the discrete Fourier transform is used to obtain the frequency spectrum. This obtained spectrum has the advantage that is very smooth so formants frequencies can be easily founded. A quick inspection on Figure 17 shows that formant frequency F1 has a value changing from 1000 Hz to 2000 Hz. Some noise appearing on the signal waveform can be reduced in the post-processing block using some filtering like median filter as explained next.

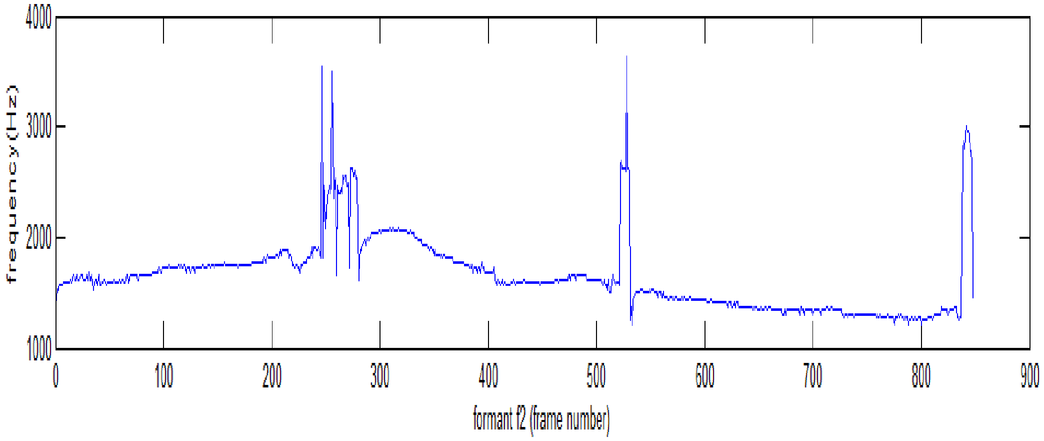


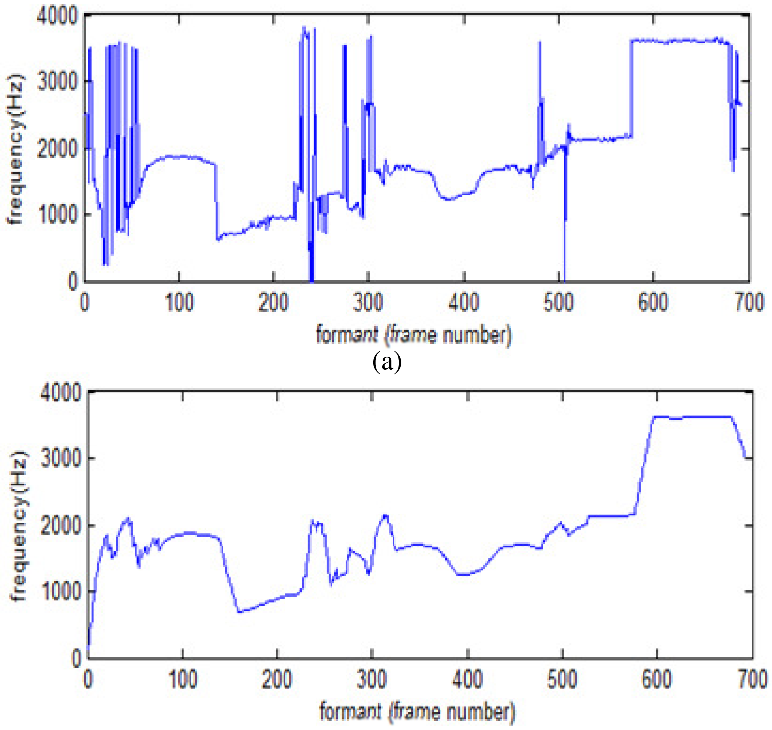
Figure 17: Formant frequency waveform obtained from a speech utterance using linear predictive coding. Frame size is 320 and frame shift is 160.

2.5 Post processing

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. This filter works by taking the average sum values around the value to be filtered in order to compensate from any abruptly change due to some noise. The moving average filter has the following expression:

$$(Y_k)_S = \sum_{i=-n}^{i=n} \frac{Y_{k+i}}{2n+1} \tag{14}$$

Where $(Y_k)_S$ is the filtered signal, Y_k is the signal to be filtered and n is a the filter order. Figure 18 shows the result of applying the filter over a formant waveform.



(a)

(b)

Figure 18: (b) Effect of the moving average filter on (a) original formant waveform.

2.6 Feature Vector

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. For these work, the feature vector length is 20:

F_V = *(mean Energy,*

Standard deviation of Energy,

Max of Energy,

Min of Energy,

Mean of Pitch,

Standard deviation of Pitch,

Max of Pitch,

Min of Pitch,

Mean Formant1,

Standard deviation of Formant1,

Max of Formant1,

Min of Formant1,

Mean Formant2,

Standard deviation of Formant2,

Max of Formant2,

Min of Formant2,

Mean Formant3,

Standard deviation of Formant3,

Max of Formant3,

Min of Formant3)

The obtained feature vector will have values having significant difference in magnitude between each component, thus to avoid a component being dominant, the feature vector is normalized using the next Equation 15 [30]:

$$I = I_{min} + (I_{max} - I_{min}) * (D - D_{min})/(D_{max} - D_{min}) \qquad (15)$$

where D_{max} and D_{min} are maximum and minimum values of the input data, I_{max} and I_{min} are the maximum and minimum range for normalization and I and D, are the normalized data and data to be normalized respectively. By using the normalization all components of the feature vector will fall between -1 and 1 range which is also the input range desired for the neural network classifier.

2.7 AV Mapping

As people show their emotional expressions to their various degrees individually, it is not an easy task to judge or to model human emotions [31]. In the literature, there are usually two methods to model emotions. One method is to map emotions in to discrete categories; this means that a given emotion has to be classified within a prescribed list of word labels, e.g. joy, sadness, surprise, anger, love, fear, etc. This method has the drawback that real emotions range vary significatively and although classify them in to discrete states could be helpful for simplification; still emotions cannot be adequately expressed in discrete word labels. The second method to classify emotion is by using one or multiple dimensions or scales. In this way a stimulus can be categorized in continuous scales like for example, pleasant-unpleasant, attention-rejection, simple-complicated, etc. two commons scales are valence and arousal where valence represent the pleasantness of stimuli and arousal represent the energy or activation level of stimuli. This kind of model was proposed by Thayer [32] and is shown in Figure 19. For example, valence has a positive valence or

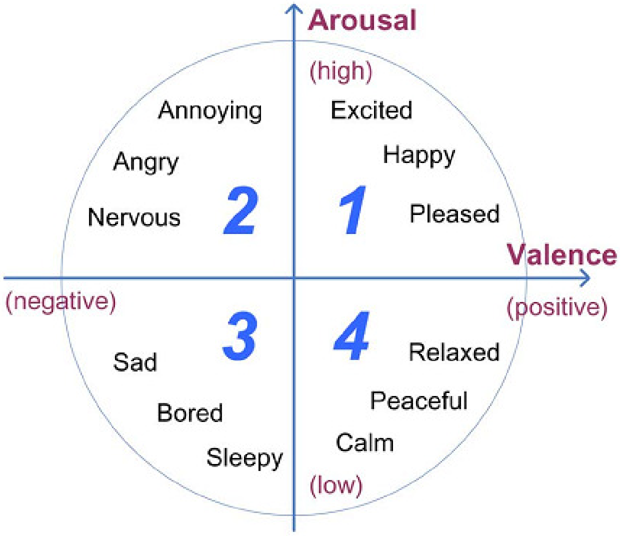


Figure 19: Thayer’s two-dimensional emotional plane of emotion [32].

pleasant, while disgust has a negative valence (or unpleasant). In the same way sadness has low arousal whereas surprise has a high arousal level. Using these two scales, a two axis plane results and a two-dimensional model can be constructed where different emotional levels can be plotted.

The two dimensional model is very useful because it can map emotions in a continuous way but at the same time it can also be divided in order to account for some categorization of emotions. Taking advantage of this property, the Thayer dimensional model is divided in the most common used emotion categories in order to assign arousal and valence values to the input utterances used in the neural network. This division will make possible also to later categorize the obtained arousal and valence values in order to make a fair comparison of the proposed system architecture to other works using emotional categorization as output. Based on Thayer’s work the emotional plane has been subdivided as shown in Figure 20. Just the emotional categories used for training are shown.

For convenience, the maximum and minimum ranges of the arousal and valence values in the dimensional 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 (16):

$$E(A,V) = (O_A + 0.25\ , O_V + 0.25\ ) \qquad (16)$$

where O_A and O_V are the quadrants origin in the emotional plane. 0.25 is half the side of a quadrant of 0.5 by 0.5 so by doing the calculations, the arousal and valences values assigned to training utterances can be written as given in Table 1.

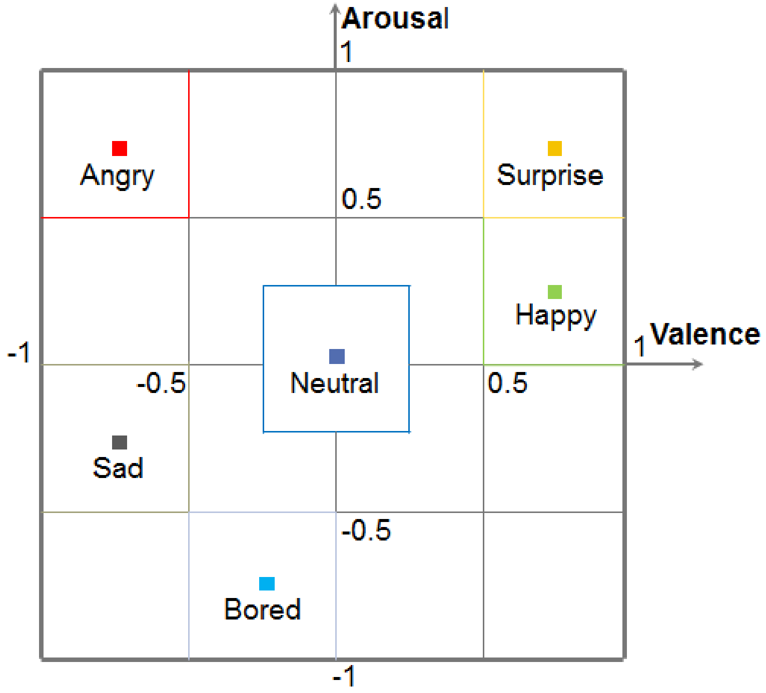


Figure 20: Thayer's emotional model and the emotional categories used for training of the neural network

In this work a Feed-forward Back-propagation (FFBP) neural network is used in order to map the obtained feature vector to the 2D emotional plane. In a Feed-forward Back-propagation neural network layers of nodes are interconnected in a feed-forward manner, this is data flows in a forward direction [33]. Usually this kind of network has three layers, an input layer, one or more hidden layers and an output layer. Figure 21 shows a typical structure of a Feed-forward Back-propagation Neural Network where L is the number of nodes in each layer and can be different for each layer.

Table 1: Training utterances' assigned arousal and valence values according to emotion category used.

	Valence	**Arousal**
Anger	-0.75	0.75
Boredom	-025	-0.75
Happiness	0.75	0.25
Neutral	0.0	0.0
Sadness	-0.75	-0.25
Surprise	0.75	0.75

What a FFBP neural network does is that when an input pattern is presented in the input layer, an output pattern will result at the output layer according to a training phase that has previously run through the network. Every input neuron should represent an independent variable that has an influence over the output of the neural network [33]. For this work there are three feed-forward back-propagation neural networks structures used since there were 3 emotional speech databases used that will be introduced later. To get the network structures, there were an iterative process were

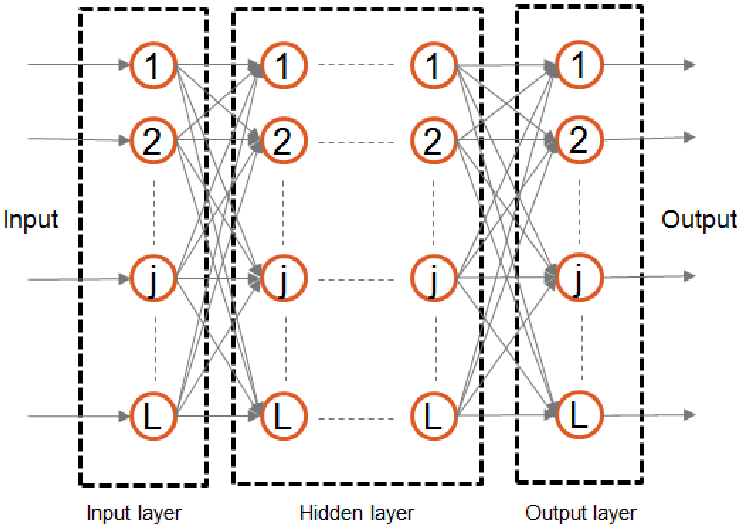


Figure 21: Feed-Forward Back-Propagation neural network architecture.

after each training phase targeting a certain error rate, the structure were used on the validation set to check the error again and the process finished after the "best" possible error rate was founded.

2.8 Music Database

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 [34] is referenced, where a database of 60 English famous popular songs has been annotated by 40 individuals. The music samples were trimmed to 25 seconds by manually trimming the chorus part. Figure 22 shows the arousal and valence values of each song in the music database plotted on the emotional plane. Each of the 40 persons manually annotated arousal and valence vales for each of the 60 trimmed song in a range between [-1 1]. At the end the mean value of the overall annotation has been assigned as arousal and valence values for a given song.

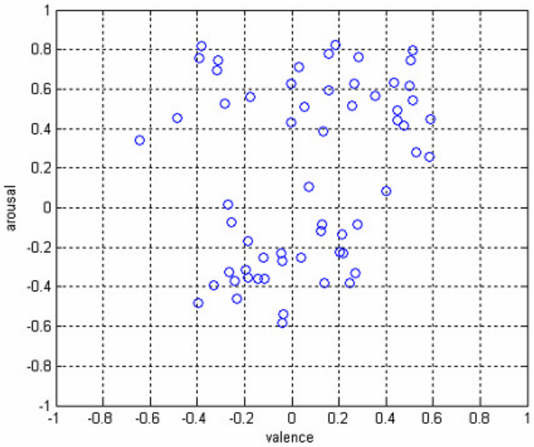


Figure 22: Music samples distribution of the song database [34].

2.9 Music and speech emotional matching

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 next expression:

$$min\left(D_n = \sqrt{(V_{n2} - V_{n1})^2 + (A_{n2} - A_{n1})^2}\right) \qquad (17)$$

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. A graphical description can be seen in Figure 23. Where the red triangle represents a detected arousal and valence value and the black dots represents arousal and valences values for each song. 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.

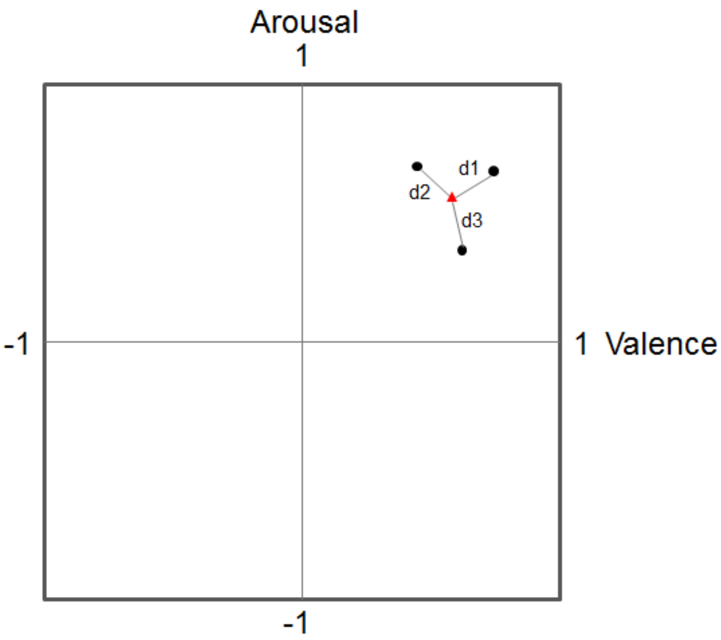


Figure 23: Distances between an utterance (red triangle) and song (black dot) in the two dimensional plane.

2.10 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 Figure 18) of the motional plane is traced using the equation of a straight line:

$$y = mx + b \tag{18}$$

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, this it is divided 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 straight line between the detected and target points divide by the number of sections used, then using the equation of the straight line as in Equation (18) and the Euclidean distance as in Equation (17) yields the following expression:

$$(1 + m^2)x_2^2 + (-2x_1 + 2mb - 2my_1) + (x_1^2 + b^2 - 2by_1 + y_1^2 - d^2) = 0 \tag{19}$$

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:

$$x_2 = \frac{-B \pm \sqrt{B^2 - 4AC}}{2A} \tag{20}$$

where:

$$A = (1 + m^2)x_2^2 \tag{21}$$

$$B = (-2x_1 + 2mb - 2my_1) \tag{22}$$

and

$$C = (x_1^2 + b^2 - 2by_1 + y_1^2 - d^2) \tag{23}$$

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

$$y_2 = mx_2 + b \tag{24}$$

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 Figure 24 where in this case 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 (23) and (24).

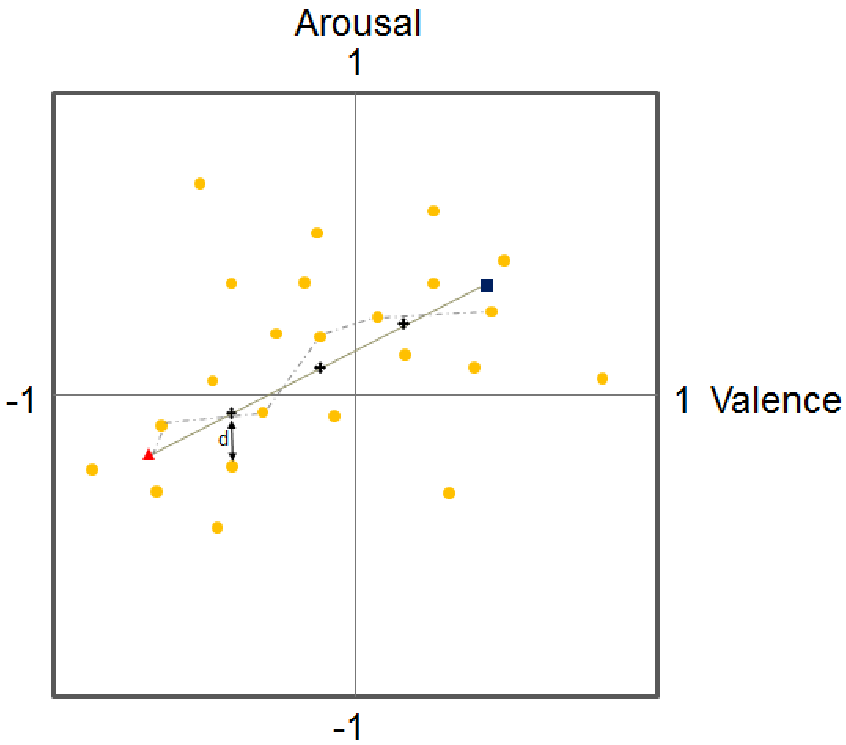


Figure 24: Songs (red dot) that are played on the cheer-up mode are selected based on their distances to the straight line between input (black dot) and target (blue dot) locations in formed line.

Then the selected songs that will be selected for the cheer-up mode are the one that follows the dashed line which were inferred using the shortest Euclidean distance.

2.11 Beagleboard XM

Beagleboard is a low-power, low-cost single-board embedded system produced by Texas Instruments. It is designed with open source software development in mind. Its purpose is to be used as an educational platform of open source hardware and software. The BeagleBoard XM has a very small footprint (82.55 x 82. 55 mm) and has a 1GHz ARM cortex-A8 core, 512 MB RAM. Data and OS are stored on a microSD card. Some peripheral connections the board embedded are: HDMI, USB OTG, 4 USB ports, microSD/MMC card slot, Stereo in and out jacks, RS-232 port.. a picture of the Beagleboard and its characteristics are shown in Figure 25. BeagleBoard has a laptop like performance and is ideal for signal processing involved in the emotion based music player. Beagle Board small footprint and processing capabilities are suitable for embedded robotic applications.

This platform was chosen because its characteristics are very similar to the ones uses in common embedded systems used in everyday life like the smartphones. Its size, low cost and low power consumption make it ideal for personal robots applications. The Beagelboard can support several operating systems like Ubuntu or Angstrom, but the Android operating system was used due to is current popularity in portable devices. Android operating system is based on a Linux kernel and was designed to be very memory and performance efficient and is highly tuned to limitations of small hardware [35]. Android is available as open source for developers. Applications are written in java programming language using Android SDK (Software Development Kit) which was the tool used to program the application of the emotion-based system that runs in the Beagle Board.

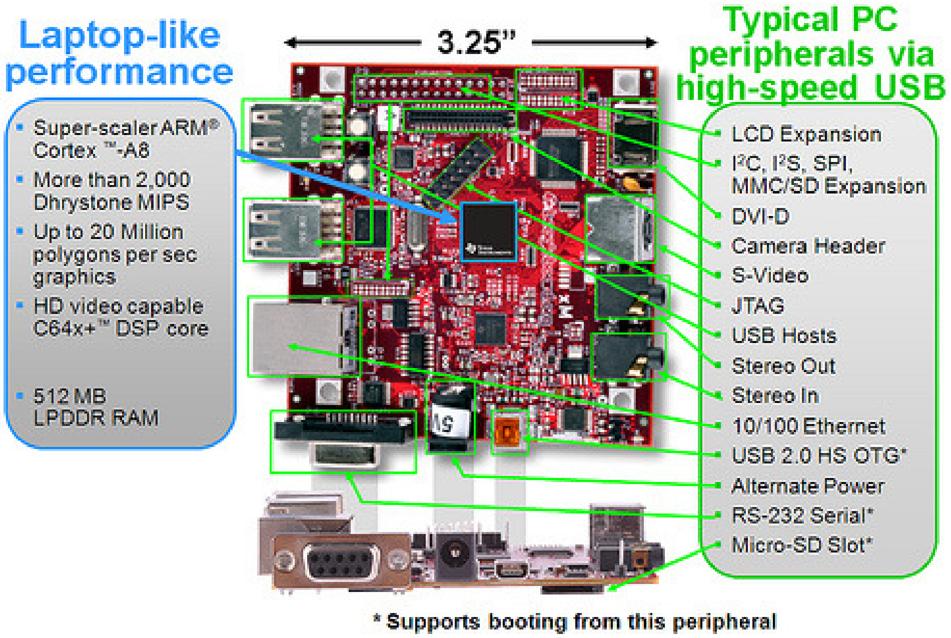

Laptop-like performance
Super-scaler ARM® Cortex ™-A8
More than 2,000 Dhrystone MIPS
Up to 20 Million polygons per sec graphics
HD video capable C64x+™ DSP core
512 MB LPDDR RAM
3.25"
Typical PC peripherals via high-speed USB
LCD Expansion
I²C, I²S, SPI, MMC/SD Expansion
DVI-D
Camera Header
S-Video
JTAG
USB Hosts
Stereo Out
Stereo In
10/100 Ethernet
USB 2.0 HS OTG*
Alternate Power
RS-232 Serial*
Micro-SD Slot*
* Supports booting from this peripheral

Figure 25: Beagleboard XM characteristics [13].

Android has many versions each of them adding more features and capabilities. The Android version used in this work was the 2.3 also known as Gingerbread. Figure 25 shows the Android architecture which shows the major components of this operating system.

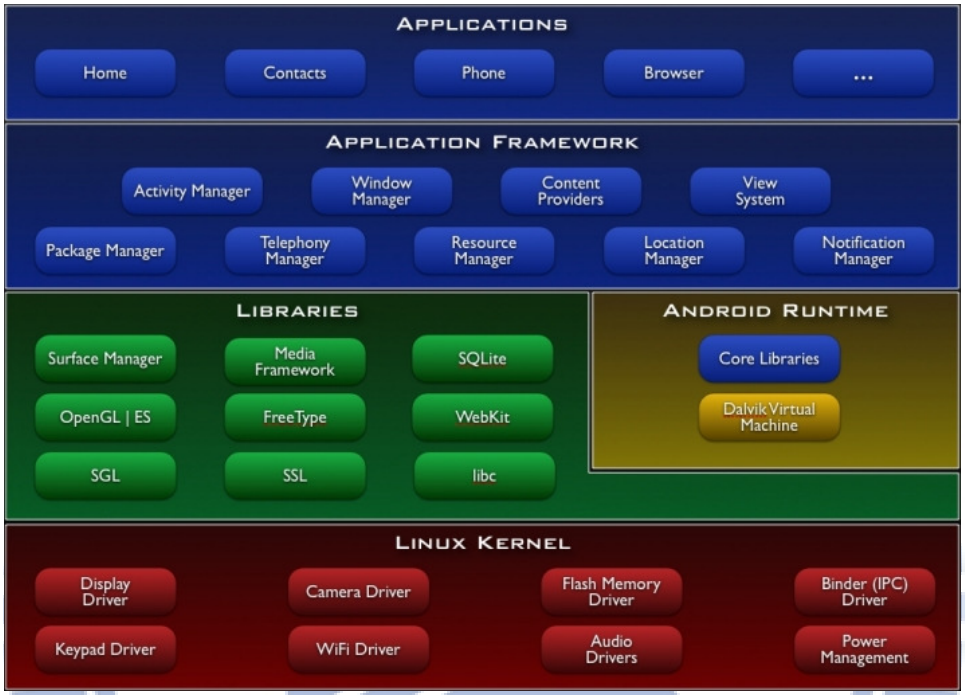


Figure 26: Android Architecture [35].

III. Experiments and Results

The system has been tested by a series of experiments where three kinds of databases have been used. One proprietary database was built in order to be able 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 will not be altered too much. The other two referenced databases were 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, at last the experiments and the results made with them will be discussed.

3.1 Emotional speech databases

Since in these work, emotional recognition from the speech is treated as a pattern recognition problem where a classifier (neural network) is used and 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 improve mapping of the input utterances on the two-dimensional emotional plane. Previous works on emotional speech recognition like [36] and [37]; often build their same databases and some of them where the recording conditions are more professional also made them available online. Most of the time the recorded emotions are few basic ones like anger, happiness, boredom, sadness and surprise. Speech databases can be built using natural speech where daily life dialogs or recordings can be used, but this is too difficult to accomplish since the controllability of the recording cannot be guaranteed and also the emotions of the individuals involved in the recording could be influenced by the recording process. Another common method to build emotional speech databases is use professional actors or normal people that can try to speak phrases emulating a determinate emotion. This method is the most used and also it was adopted in the databases used in this work.

3.1.1 Proprietary database

This database was built in order to train the system and also in order to be used for testing on real persons. This is because the recording conditions during training and validation phases are expected to be similar, thus a recognition error must be lower. For this database ten lab mates at the Intelligent System and Control Integration (ISCI) lab were invited to utter some short phrases using acted emotional content. The speakers were males between 20 and 30 years old. Six basic emotions where used: anger, boredom, happiness, neutral, sadness and surprise. Six invited person 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 XM embedded system and an electret microphone. The setting sampling frequency was 8 kHz and 16 bit per sample.

3.1.2 ISCI Lab database

The Intelligent System and Control Integration lab at National Chiao Tung University in Taiwan have done previous research on emotional recognition from speech. For that purposes they built before an emotional speech database. There are five categories of emotional speech in the database: happiness, sadness, surprise, anger and neutral. For each category, there are three kinds of different sentences. In order to express the emotion in natural way, each subject was asked to narrate expressive sentences in Chinese to imitate actual interactive scenario. The database includes various emotional utterances from five persons. Each person spoke three different utterances ten times per emotion category. So there are 150 utterances per speaker and there are totally 750 utterances in this database. The recordings were done in a quite environment using a Digital signal processing board from Texas Instruments and a microphone. The sampling rate where set at 8 kHz and 8 bits per sample.

3.1.3 Berlin Database

As a project funded by the DFG (German Research Community) a database of emotional speech [38] 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 800 sentences (seven emotion * ten actors * ten sentences * some second versions) but after a manual selection just phrases where emotion impression was the same for listeners selecting them were left. So in total there are about 500 utterances. Recordings were done using a Sennheiser MKH 40P 48 microphone and a Tascam DA-P1 portable DAT recorder. 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.

3.3 Off-Line Experiments

To test the performance of the proposed system, off-line experiments were performed were the databases were 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. Although the expected output of the AV mapping module is arousal and valence values, this values are also changed to corresponding emotion categories for comparison purposes with other works using emotion categories as output. This conversion is done by using the emotional plane in Figure 18 and the Euclidean distance equation in (17). For each of the obtained arousal and valence values from the validation database, the Euclidean distance to the target emotion categories used for training (Figure 18) is calculated and the arousal and valence pair is classified on to an emotion category according to which emotion category location has the shortest distance. This process is shown graphically in Figure 27:

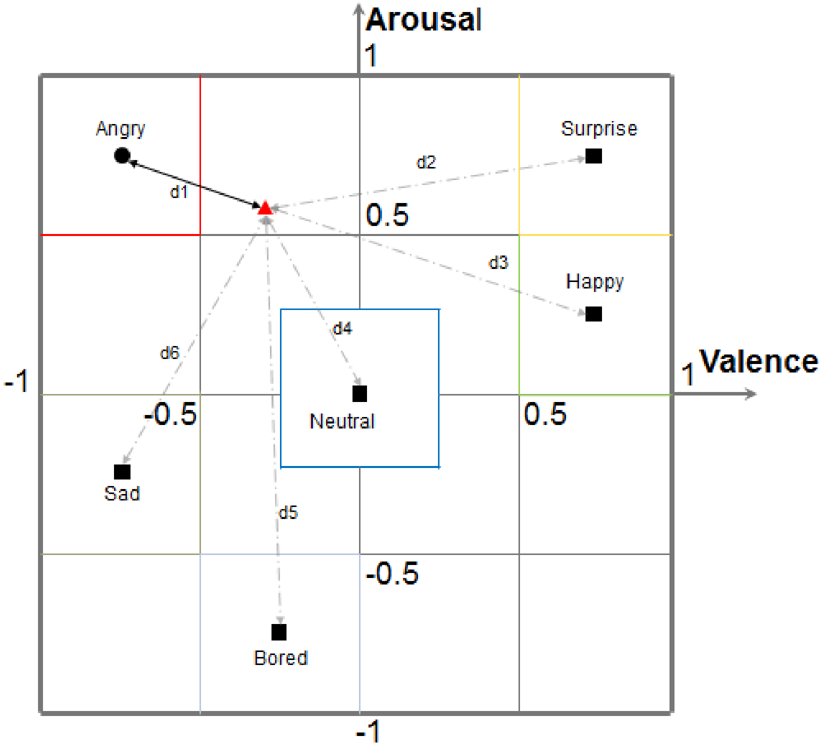


Figure 27: Arousal and valence pair classified as emotion category method.

where an obtained arousal and valence pair (red triangle) is classified as “anger” since the Euclidean distance d1 between it and the anger location in the emotional plane (black dot) is the shortest one compared to the distances d2, d3, d4, d5 and d6 to the others emotion categories locations in the emotional plane (black squares).

Once all of the obtained arousal and valence values have been classified in emotion categories, a confusion matrix is calculated in where all the validation utterances supposed emotion categories are compared to the obtained emotion categories. This information is tabulated in order to get a percentage value that indicates how well the inputted utterances have been classified as emotions. Conversion is done in order to compare the reliability of the system, but it is necessary to clarify that the desired characteristic is to be able to recommend a song that best matches the inputted emotional speech and this is better done in the emotional plane where an inputted utterance is expected to have a continuous location in the emotional plane and that every inputted utterance will never be the same so the difference can be expected to be inferred from obtained arousal and valence values.

For every database used, after the training phase, a different FFBP neural network structure (see Figure 21) was obtained. The configuration of the structure is shown on Table 2:

Table 2: Number of nodes per layer for the neural network architectures used according the speech database used.

Number of nodes per layer				
Dataset	**Input layer**	**Hidden layer 1/2**	**Output layer**	**Total nodes**
Proprietary	11	19/8	2	40
ISCI Lab	11	23/7	2	43
EmoDB	11	18/7	2	38

The first test was performed using the proprietary database. The proprietary database has a total of 456 utterances. Since every speaker spoke different emotional phrases for each of the six emotion categories, about 75% of the phrases were taken for training and the other 25% were used for validation. In this way the training dataset has 348 utterances and the validation one 72 utterances. The result is plotted in the two dimensional emotional plane in Figure 28 and the corresponding confusion matrix in Table 3. As seen in Figure 28, the distribution of arousal and valence pairs with boredom (b), surprise (f) and neutral (c) emotional content fall 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. Confusion matrix in Table 3 shows that arousal and valence values with neutral emotional content are classified in most of the cases and the same is valid for surprise. On the contrary, values with happiness and sadness emotional category have lower classification rate since it seems that many utterances are classified as having emotional content of the closest emotional category such as boredom or neutral instead of sadness emotional content. Values with happiness emotional content are very close to the ones having surprise emotional content. The overall recognition result is quite acceptable since it reaches 68% and only values with sadness emotional content has a lower classification rate.

The second test was performed using the ISCI lab database. This database was also used for similar research in emotion recognition in the past. The dataset has a total of 750 utterances. Every speaker uttered 3 differences phrases for every emotion category. Every phrase was uttered 10 times. So in this database the training set is composed of the first 9th utterances of every phrase for a total of 675 utterances. Each 10th utterance of every phrase is taken for the validation set for a total of 75 utterances. The results are plotted in the emotional plane in Figure 29 and the confusion matrix in Table 4. Distribution of values in Figure 29 shows that in overall, inputted utterances detected arousal and valence values are located much closer to the target emotions and the distribution are closer in the case of anger emotional content. Happiness and surprise has a worst emotional classification since some samples are confused with neutral emotional content since these categories are located much closer between them. Confusion matrix in Table 4 confirms that values with anger emotional content has the best classification rate but values with surprise emotional content are confused with happiness and neutral and hence the lower classification rate. It is possible that results using ISCI Lab database are better than those using the proprietary database since in the first one, a single phrase is uttered several times and recording conditions were better monitored.

The third experiment was done using the Berlin emotional speech database (Emo-DB). This database is quite popular and has been referenced in other works on emotional speech recognition. Emotion is acted and also short utterances are used so it was used as mean to compare system architecture performance. The Emo-DB has almost 500 utterances comprising seven emotions (anger, boredom, disgust, anxiety, happiness, sadness, neutral), for fairness in the comparison result, disgust and anxiety were taken out. At the end there are a total of 416 emotional speech utterances. The works in [39] and [40] are used for comparison. They use 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. In this work this strategy is adopted. So a total of 389 utterances are used as training set and 27 as the validation set. The results are plotted in the two dimensional emotional plane in Figure 30 and the corresponding confusion matrix in Table 5. Results from Figure 30 and Table 5 shows that arousal and valence values has an acceptable overall classification for all emotions categories since classification rate almost reaches 60% or above except from values with sadness emotional content, since sadness is confused with boredom because their values are located very closed to each other. Results from Table 5 shows that although there are many utterances and recording is professionally controlled, results using the proprietary and ISCI Lab databases are better maybe because utterances in German database are not well distributed since there are different number of phrases for each emotion category.

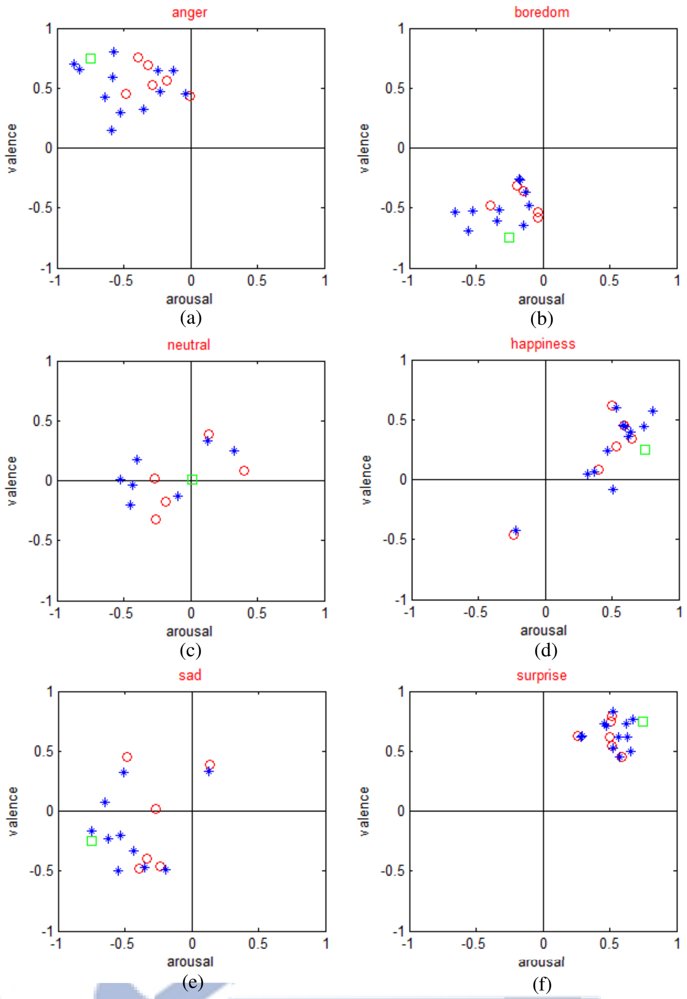
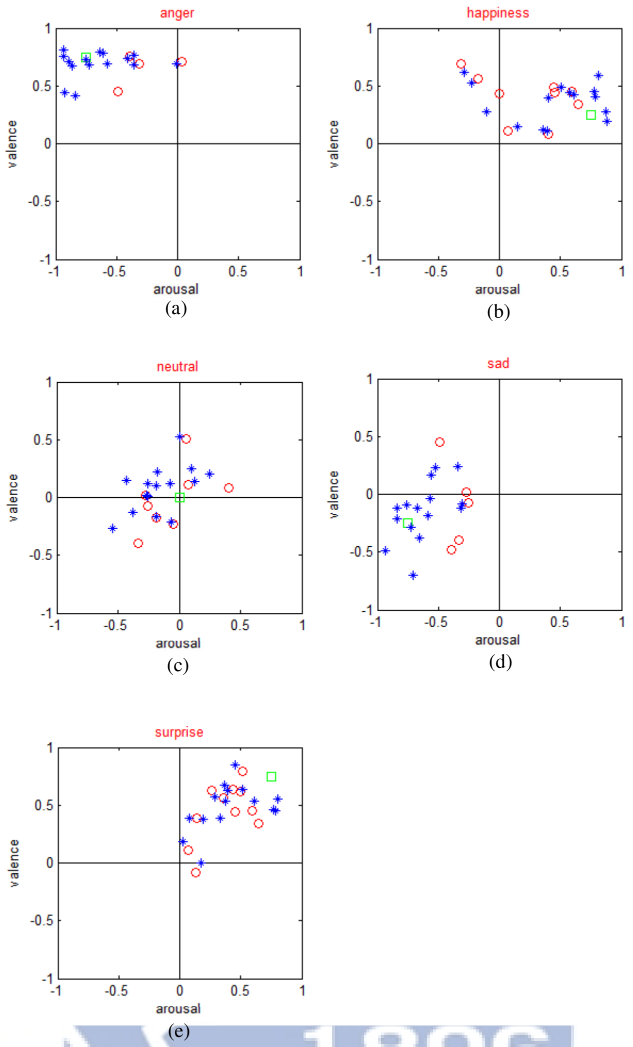


Figure 28: System output using proprietary database. Green square indicates target emotional value, blue asterisk indicates the inputted utterance detected AV values and the red circle indicates the system proposed song.

Table 3: Confusion matrix for the emotion categories obtained using proprietary database.

	Anger	Boredom	Happiness	Neutral	Sadness	Surprise
Anger	**66.7**	0.0	0.0	25.0	8.3	0.0
Boredom	0.0	**75.0**	0.0	16.7	8.3	0.0
Happiness	0.0	8.3	**58.3**	16.7	0.0	16.7
Neutral	0.0	0.0	0.0	**75.0**	25.0	0.0
Sadness	8.3	25.0	0.0	16.7	**50.0**	0.0
Surprise	0.0	0.0	16.7	0.0	0.0	**83.3**

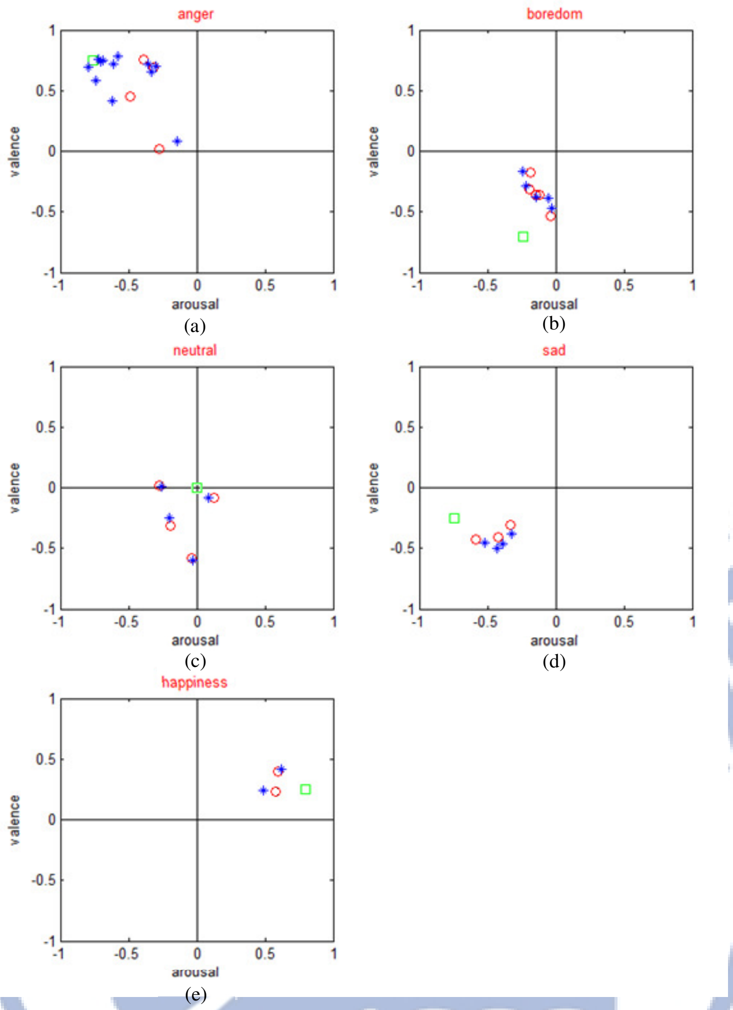


(a) (b) (c) (d) (e)

Figure 29: System output using ISCI Lab database. Green square indicates target emotional value, blue asterisk indicates the inputted utterance detected AV values and the red circle indicates the system proposed song.

Table 4: Confusion matrix for the emotion categories obtained using ISCI Lab database.

	Anger	Happiness	Neutral	Sadness	Surprise
Anger	**93.3**	0	6.7	0	0
Happiness	6.7	**60.0**	26.7	0	6.7
Neutral	0	0	**86.7**	13.3	0
Sadness	6.7	0	20.0	**80.0**	0
Surprise	0	20.0	26.7	0	**53.3**



(a) (b)

(c) (d)

(e)

Figure 30: System output using EMODB database. Green square indicates target emotional value, blue asterisk indicates the inputted utterance detected AV values and the red circle indicates the system proposed song.

Table 5: Confusion matrix for the emotion categories obtained using Emo-DB database.

	Anger	Boredom	Happiness	Neutral	Sadness
Anger	**62.2**	15.0	8.7	8.7	5.5
Boredom	13.6	**70.4**	2.5	7.4	6.2
Happiness	12.9	12.9	**64.3**	7.1	2.9
Neutral	16.7	10.3	11.5	**57.7**	3.8
Sadness	8.3	36.7	6.7	6.7	**41.7**

(d)

Results obtained from the emotional databases plotted in the emotional plane, shows that an input utterance is adequately placed in the emotional plane since for a certain emotion category,   the obtained arousal and valence pairs (blue asterisks) are located in the area where the target emotion where supposed to be . This desire characteristic is useful to later select songs based on the arousal and valence values detected. This music selection process result is seen also in the emotional planes where the songs to be proposed by the system are also plotted (red circle). Like that just the song with the closest emotional content is selected. From the confusion matrices, it is also seen that the database to use is very important since for every one of them, the results are quite different, and it has to be related to the subjects who participate in the recordings, the kind of phrases uttered and the spoken language. Comparing the proprietary database and the ISCI lab databases, both using the Chinese mandarin language, is seen that there is a better performance of the system by using the ISCI lab one. This is thought to be due to the kind of phrases used. For instance in ISCI lab database, one phrase is uttered ten times by same speaker and also is known that the setting up of the experiment was better controlled. Recognition rate obtained using Emo-DB database has a slightly lower recognition rate value than the both the proprietary and ISCI lab maybe because the number of utterances for training the neural network is less and also because the non-uniformity of the database (the number of utterances for every emotion category is not the same). Yet the recognition rate is quite similar to that of the proprietary database.

In [39] and [40] the authors used the Emo-DB in order to test their speech recognition systems. In both works focus is to improve recognition rate in emotion recognition from speech using different kinds of features. Comparison of the confusion matrices obtained on the offline test using the Emo-DB shows that recognition depends highly on the features extracted from the speech. In [39] overall recognition rate is higher than [40] and this work, but comparing speaker independent row in [40] and this work shows similarity in the results although better results are obtained in these work. Features chosen in this work are very common features used in emotion recognition and also their complexity implementation is less that that used in [39] and [40], this was the desired choice since an embedded system was the target platform. The confusion matrices of the entire off-line test realized in this work and those of [39] and [40] are combined in Table 6. It is seen that overall recognition rate using Emo-DB is quite comparable to [39] and [40] probing the feasibility of the proposed architecture. Still it has to be noticed that mapping on the emotional plane is difficult to classify in discrete emotions since similar emotion content can merge together as in bored neutral and sad or happy and surprise, but still this does not mean that output is incorrect since the main purpose is to allow different emotional content to be mapped in the emotional plane to infer some differentiation that will be used to choose right songs and this has been accomplished.

Table 6: Average emotion recognition rates comparison table.

	Proprietary	ISCI	Emo-DB	Work in[39] using Emo-DB	Work in [40] using Emo-DB
Anger	66.7	**93.3**	62.2	86.1	66.1
Boredom	75.0	-	70.4	**84.8**	56.8
Happiness	58.3	**60.0**	64.3	52.7	40.1
Neutral	75.0	**86.7**	57.7	52.9	36.7
Sadness	50.0	80.0	41.7	**87.6**	48.4
Surprise	**83.3**	53.3	-	-	-
Average	68.05	74.6	59.3	**72.8**	49.6

3.4 On-Line Experiment

An online experiment has been carried out in order to test system performance on a real scenario. Figure 31 shows a picture of the set-up used for the online experiments. It is composed of the Beagleboard and a touchscreen 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. Otherwise a song will play just once and just a green dot will be plotted on the emotional plane. The user can stop the music whenever he desires just by pressing the stop button again. 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.

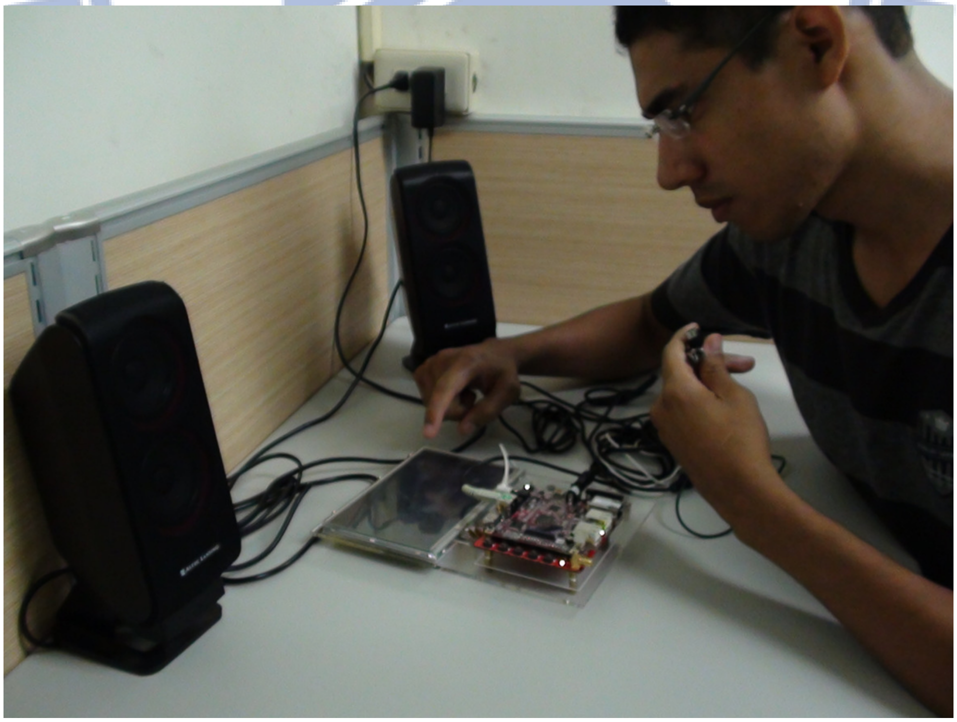
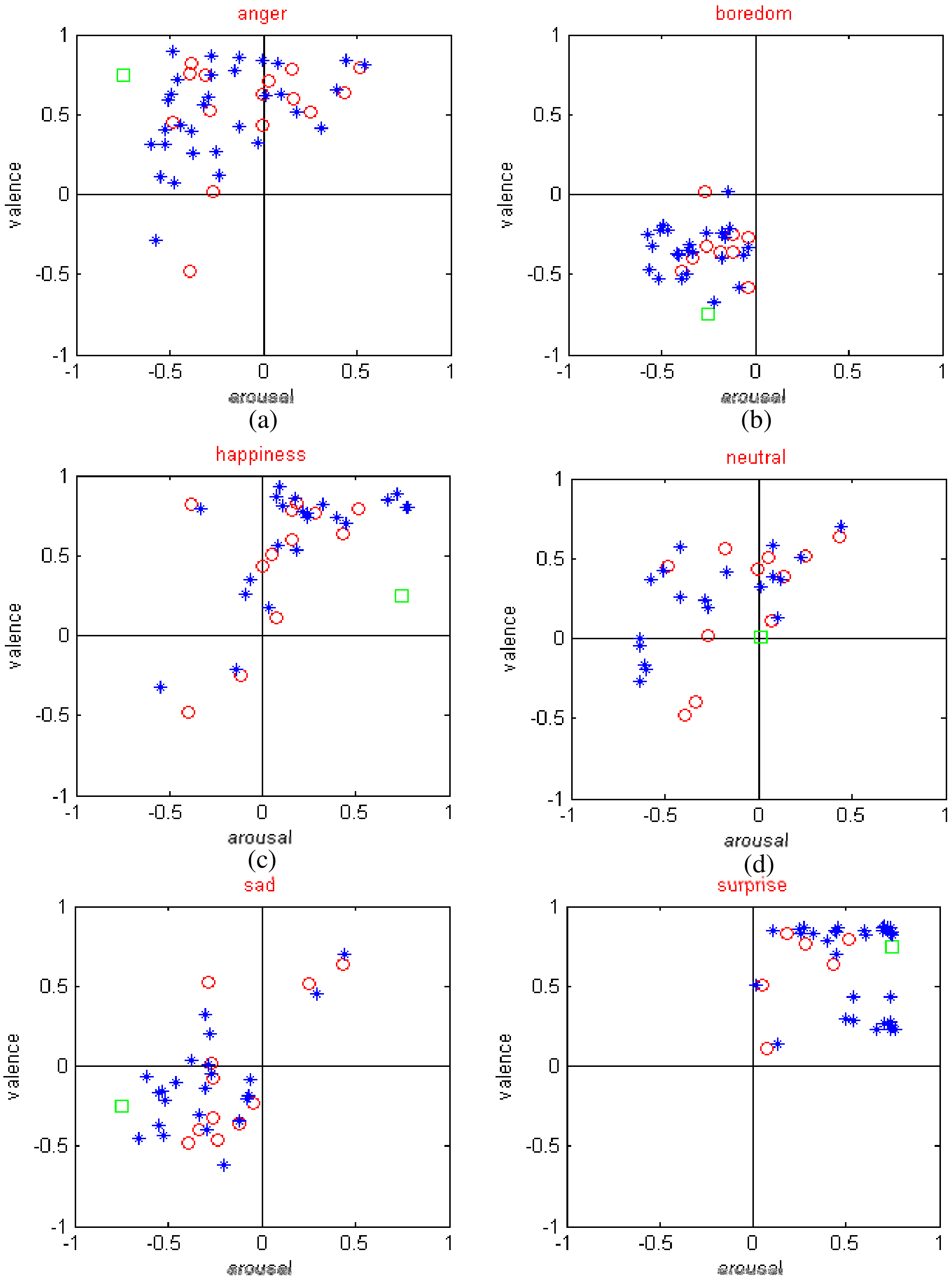


Figure 31: Emotional Speech recognition based music player system.

In this online test, the FFBP neural network with the proprietary database was used. Four different speakers of ages between 20 and 30 were invited to utter emotional speech as the one used in the proprietary database. Every speaker uttered ten phrases for most of the emotion categories (some uttered more phrases than others). At the end there were a total of 183 utterances inputted to the system. The experiment was done in a quite environment in a laboratory using the Beagleboard hardware. The arousal and valence values outputted are plotted in Figure 32. The figure shows that the detected arousal and valence values for the input utterances (marked as blue asterisks) cluster nearby the target emotional values in most of the samples. For inputted utterance with boredom, neutral, surprise and sadness emotional content is easy to observe that the system has placed them nearby the target area (see Figure 32 (b) (d) (e) (f)). Inputted utterances having happiness and anger emotional content does not fall to close to the target value but still they are mapped in a nearby area (see Figure 32 (a) (c)). In the anger case it seems that some utterances are mapped as having more neutral-sad emotional content and in the case of happiness, some utterances are mapped by the system as having more positive emotional content like surprise. Still, the online test result also shows that emotion mapping on the dimensional plane is possible and the target of selecting a song based on arousal and valence values obtained can be achieved. The output screen of the system after one of the inputted speech utterances has been processed is shown in Figure 33. 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. In this case the user has set-up the cheer-up target value as 0.61 for valence and 0.53 for arousal the number of songs to be listened until cheer-up target value has been reached has been set as 10. It can be seen that the last song to be played has a positive emotional content. Songs names are also shown on the screen.

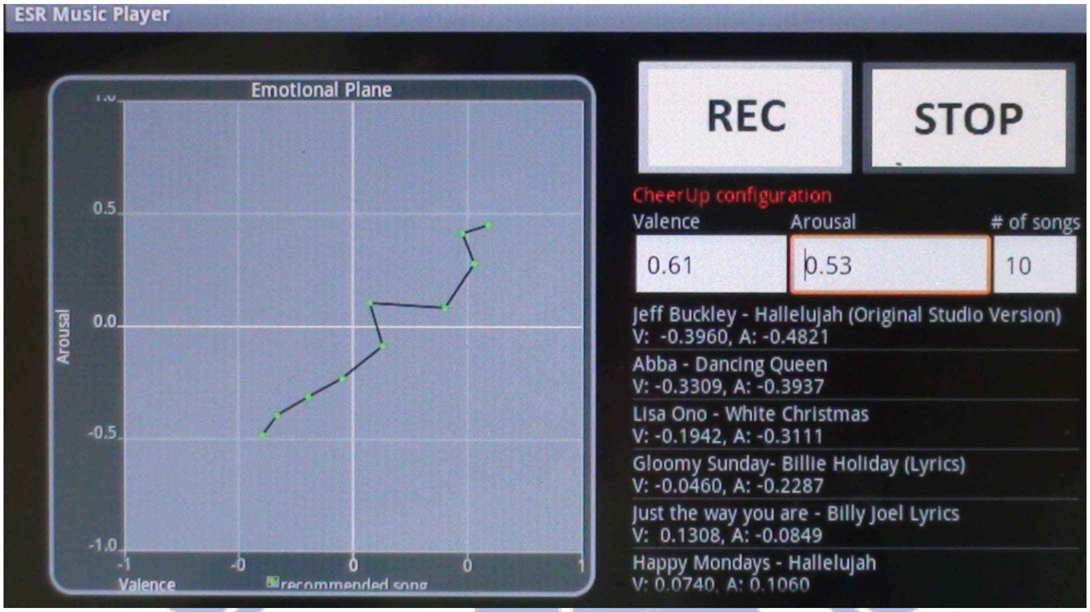


(a) (b)

(c) (d)

(e) (f)

Figure 32: system output for the online experiment. Green square indicates target emotional value, blue asterisk indicates the inputted utterance detected AV values and the red circle indicates the system proposed song.



CheerUp configuration

Valence	Arousal	# of songs
0.61	0.53	10

Jeff Buckley - Hallelujah (Original Studio Version)
V: -0.3960, A: -0.4821

Abba - Dancing Queen
V: -0.3309, A: -0.3937

Lisa Ono - White Christmas
V: -0.1942, A: -0.3111

Gloomy Sunday- Billie Holiday (Lyrics)
V: -0.0460, A: -0.2287

Just the way you are - Billy Joel Lyrics
V: 0.1308, A: -0.0849

Happy Mondays - Hallelujah
V: 0.0740, A: 0.1060

Figure 33: A screen capture of the music player.

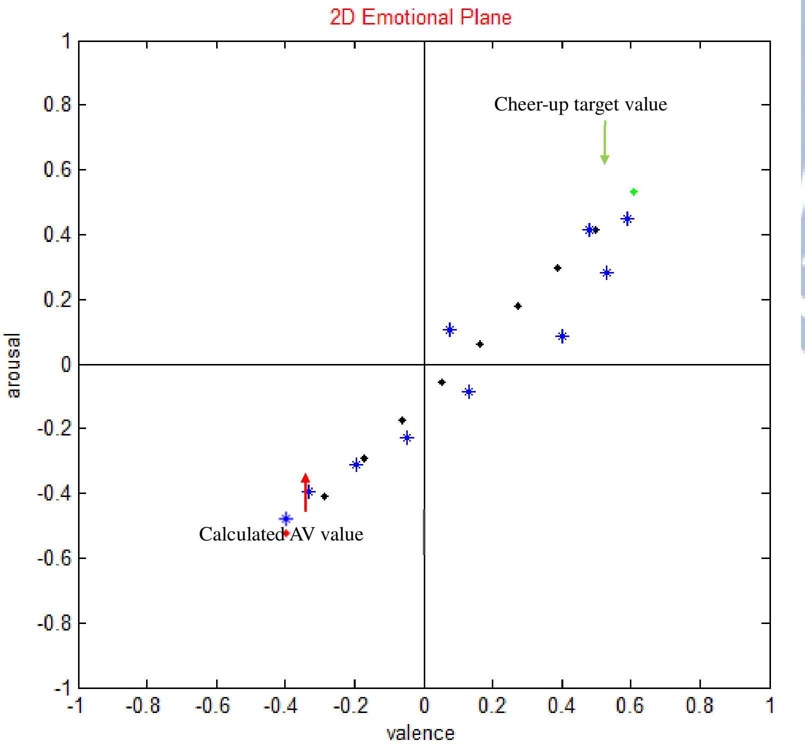


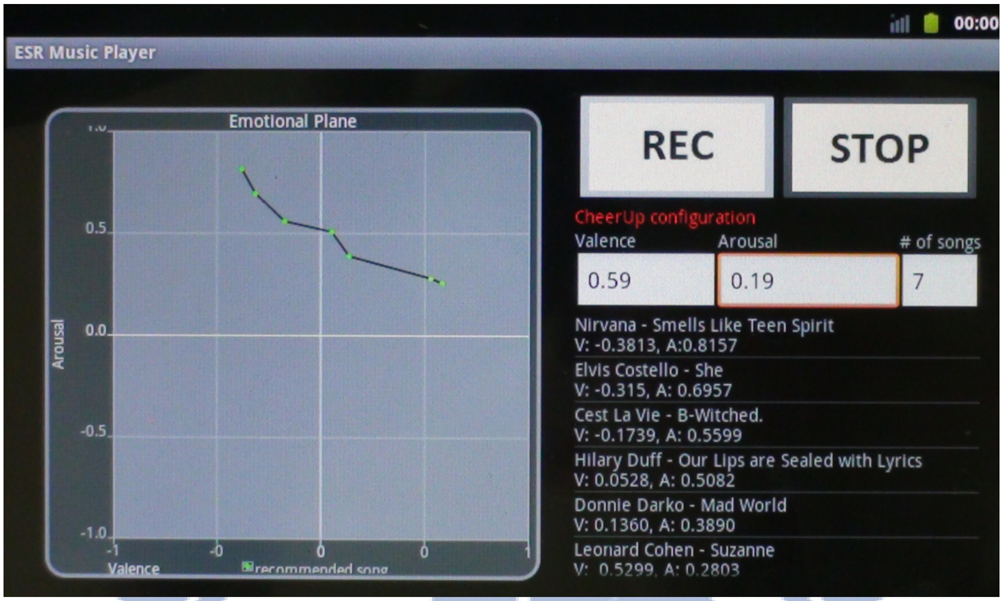
Figure 34: Example of the system output for an input utterance whit bored emotional content. Songs played are blue numbered asterisks.

Figure 34 shows the emotion cheer up strategy used to select the songs to be played in the example of figure 33. The blue asterisks are the suggested songs that will be played. The red dot is the detected arousal and valence 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 the user would like to listen in the cheer-up strategy and are shown just for reference. Table 7 shows the obtained arousal and valence values for the songs that will be played and the name of them which in this case are ten songs.

Figure 35 shows another result screen for an utterance with anger emotional content. In this case the cheer-up target values have been set as 0.59 for valence and 0.19 for arousal. The number of songs to play during cheer-up has been set as seven. Figure 36 shows the corresponding emotion cheer-up strategy plot and Table 8 shows the songs to be played and the corresponding arousal and valence values.

Table 7: Proposed songs and corresponding AV in example of Figure 33.

Song	valence	arousal	song name
1	-0.3960	-0.4821	Jeff Buckley - Hallelujah
2	-0.3309	-0.3937	Abba - Dancing Queen
3	-0.1942	-0.3111	Lisa Ono - White Christmas
4	-0.0460	-0.2287	Gloomy Sunday- Billie Holiday
5	0.1308	-0.0849	Just the way you are - Billy Joel Lyrics
6	0.0740	0.1060	Happy Mondays - Hallelujah
7	0.4022	0.0840	Vanessa Carlton ~ A Thousand Miles
8	0.5299	0.2803	Leonard Cohen - Suzanne
9	0.4775	0.4146	The Rose - Janis Joplin
10	0.5907	0.4503	Krystal Meyers - Anticonformity



CheerUp configuration

Valence	Arousal	# of songs
0.59	0.19	7

Nirvana - Smells Like Teen Spirit
V: -0.3813, A:0.8157

Elvis Costello - She
V: -0.315, A: 0.6957

Cest La Vie - B-Witched.
V: -0.1739, A: 0.5599

Hilary Duff - Our Lips are Sealed with Lyrics
V: 0.0528, A: 0.5082

Donnie Darko - Mad World
V: 0.1360, A: 0.3890

Leonard Cohen - Suzanne
V: 0.5299, A: 0.2803

Figure 35: A screen capture of the music player to an utterance with anger emotional content.

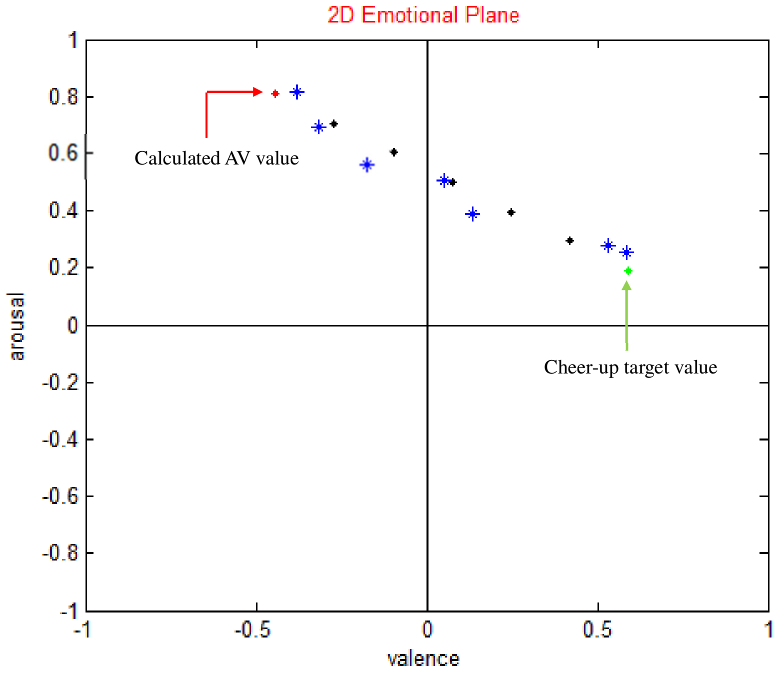


Figure 36: Example of the system output for an input utterance whit anger emotional content. Songs played are blue numbered asterisks.

Table 8: Proposed songs and corresponding AV in example of Figure 35.

Song	arousal	valence	song name
1	-0.3813	0.8157	Nirvana - Smells Like Teen Spirit
2	-0.3151	0.6957	Elvis Costello - She
3	-0.1739	0.5599	Cest La Vie - B-Witched
4	0.0528	0.5082	Hilary Duff - Our Lips are Sealed with Lyrics
5	0.1360	0.3890	Donnie Darko - Mad World\
6	0.5299	0.2803	Leonard Cohen - Suzanne
7	0.5838	0.2569	All I have to do is dream

The cheer-up strategy was also evaluated by a questionnaire survey of 10 different people. In this survey, subjects are asked what they think about the songs proposed in cheer-up mode using the mentioned two examples for a fixed cheer up target value. In the first example, the input speech utterance was supposed to have bored emotional content and the second one angry emotional content. Then each subject was asked to listen to the songs proposed by the cheer-up strategy (songs were trimmed to about 10 seconds) and then write down the degree of agreement to the proposed songs. The result of this evaluation is shown in Table 9:

Table 9: Result of questionnaire survey on emotion cheer-up strategy

Degree of agreement	Percentage %
totally agree	10
somewhat agree	70
neither agree nor disagree	10
somewhat disagree	10
totally disagree	0

In general, people agree to the systems proposed songs during cheer-up mode since the highest value for degree of agreement is for “somewhat agree” with 70%.

The output screens obtained from some emotional speech in the online test and the result from Table 9 shows that the system can rely on the emotional mapping to successfully select a song that shares user emotional content. 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. This functionality could be used in robotic applications where personal robots like a robot pet can make use of the emotional mapping and cheer-up strategies to keep the user in a good mood. Other application could be cellphones or fixed phones where background music could be played according to what the speaker is talking.

IV. Conclusion and Future Work

An emotional speech based music player has been proposed and implemented using and embedded system platform targeting for personal robotics. 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. Using a referenced database of songs, which arousal and valence values has been manually annotated by several users, the system can successfully 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.

In this work, energy pitch and formants features extracted from speech signals were selected for low complexity implementation and result show that they can be used to detect emotional content in the speech. Neural network architecture was designed for mapping speech to arousal and valence values. The performance was tested using 3 different emotional speech databases. Three off-line tests, the online test and the evaluation survey probed the feasibility of the proposed system. Obtained arousal and valence values were converted to emotional categories in order to compare the performance of the system to other works. Performed test shows that an overall recognition rate of 59.24% is good result compared to that of 73.5% and 49.12% in [39] and [40] respectively. A questionnaire survey further shows that the 80% subjects somewhat or totally agree with the songs selected by proposed cheer-up strategy based on the emotional model.

Results from the present work shows that there are some aspects of the system that can be further improved in order to increase emotional mapping and also implementation for use in a useful system like in a pet robot:

- A more powerful embedded platform could also make possible the use of more powerful algorithms to improve system performance thus improving the user-robot interaction.
- Using other sensors like a video camera can allow the use of image recognition to have other means of emotion recognition and allow better music recommendation even if the user is not speaking.
- A better microphone with better sensibility and noise rejection can be included in order to avoid being very close to the device to speak.
- To improve music recommendation, more songs can be added to the actual music database; furthermore, music emotion recognition technology can be added in order to allow the user to load his personalized music database.
- Adding more emotional related features, use of a different neural network can improve mapping in the emotional plane and robustness in speaker independent mode.
- A better data set that has far more speech utterances and maybe natural language could improve speaker independency recognition. Here manual annotation of arousal and valence values by many speakers could also improve performance since every utterance used for training could have a better emotional representation in the dimensional plane.

References

[1] Ch. Wan and L. Liu, “Research of Speech Emotion Recognition Based on Embedded System,” in *Proc. Of IEEE International conf. on Computer Science and Education,* Cape Town, South Africa, August, 2010, pp. 1129-1133.

[2] Y. Huang, X. Xu and G. Zhang, “Speech Emotion Recognition Research Based on the Stacked Generalization Ensemble Neural Network for Robot Pet,” in *Proc. Of Chinese Conference on pattern recognition,* China, November 2009, pp. 1-5.

[3] D. Ververidis and C. Kotropoulos, “Emotional Speech Recognition: Resources, Features, and Methods,” in *ELSEVIER Speech Communication*, vol. 48, no. 9, pp. 1162-1181, September 2006.

[4] Z. Xiao, *Recognition of Emotions in Audio Signals*, Ph.D. Dissertation, Ecole Doctorale Informatique et Information pour la Société, Lyon, France, 2008. Available: http://www.google.com [Accessed Apr. 2011].

[5] M. Han, J. Hsu, K. T. Song and F. Chang “A New Information Fusion Method for Bimodal Robotic Emotion Recognition,” in *Proc. Of IEEE International conference on Systems, Man and Cybernetics*, Montreal, Quebec, Canada, October 2007, pp. 2656-2661.

[6] K. Jang and O. Kwon, “Speech Emotion Recognition for Affective Human-Robot Interaction,” in *Proc. of international conference on Speech and Computer,* St. Petersburg, Russia, June 2006, pp. 25-29.

[7] B. Schuller, G. Rigoll, S. Can and H. Feussner, “Emotion Sensitive Speech Control for Human-Robot Interaction in Minimal Invasive Surgery,” in *Proc. Of IEEE International Symposium on Robot and Human Interactive Communication,* Munich, Germany, August 2008, pp.453-458.

[8] S. Dornbush, K. Fisher, K. McKay, A.Prikhodko and Z. Segall, “ XPOD &#8211; 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.

[9] Y. H. Yang, Y. C. Lin, H. T. Cheng and H. Chen, “Mr. Emo: Music Retrieval in the Emotion Plane,” in *Proc. Of ACM International Conference on Multimedia,* Vancouver,

Canada, 2008, pp1003-1004.

[10] V. A. Petrudhin, “Emotion in Speech: Recognition and Application to Call Centers,” in *Proc. Of International conference on Artificial Neural Networks,* Edinburgh, England, 1999, pp.7-10.

[11] C. M. Thibeault, O. Sessions, P. H. Goodman and F. C. Harris Jr., “Real-Time Emotional Speech Processing for Neurobotics Applications,” in *Proc. Of International Conference on Computer Aplications in Industry and Engineering,* Las Vegas, NV, 2010, pp.239-244.

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

[13] (2011, May 5). *BeagleBoard-XM* [Online]. Available: http://www.beagleboard.org

[14] 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(2), pp.1097–1108, February 1993.

[15] P. Oudeyer, “The Production and Recognition of Emotions in Speech: Features and Algorithms,” *International Journal in Human-Computer Studies,* vol. 59/1-2, pp. 157-183, special issue on Affective Computing, 2003.

[16] A. Iliev, *Emotion Recognition Using Glottal and Prosodic Features,* Ph.D. Thesis, University of Miami, Florida, 2009.

[17] (2011, September 20). *Acoustic Phonetics* [Online]. Available: http://www.kfs.oeaw.ac.at/content/blogsection/26/396/

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

[19] G. Saha, “A New Silence Removal and Endpoint Detection Algorithm for Speech and speaker recognition applications,” in *Proc. Of National Conference on Comunications,* India, 2005, pp.291-295.

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

Hall, 2001.

[21] M. Mansoorizadeh and N. M. Charkari, “Speech Emotion Recognition: Comparison of Speech Segmentation Approaches,” in *Proc of International conference on Knowledge Technology*, Mashhad, Iran, 2007, pp.133-136.

[22] 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, pp. 18-28, 2010.

[23] J. Sidorova, *Speech Emotion Recognition*, Ph.D. Thesis, Universitat Pompeu Fabra, Barcelona, Spain, 2007.

[24] 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, 2010, pp. 18-28.

[25] C. Lee, *2011 Short Course on Digital Speech Processing and Applications*, unpublished, June 2011.

[26] D. Morrison, R. Wang and L. C. De Silva, “Spoken Affect Classification Using Neural Networks,” in *Proc of IEEE International Conference on Granular Computing* , Beiging, China, July 2005, pp. 583- 586.

[27] C. Breazeal and L. Aryananda, “Recognition of Affective Communicative Intent in Robot-Directed Speech, ” *Autonomous Robots*, vol. 12, pp. 83–104, 2002.

[28] Ch. 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, 2006.

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

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

[31] Y. Yoshitomi, “Effect of Sensor Fussion for Recognition of Emotional States Using

Voice, Face Image and Thermal Image of Face," *Proc. Of International workshop on robot and human interactive communication*, Osaka, Japan, pp. 1186-1197, pp. 178-183, 2000.

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

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

[34] Y. H. Yang, Y. F. Su, Y. Ch. Lin and H. Chen, "Music emotion recognition: the role of individuality," in *Proc. Of the International workshop on Human-centered Multimedia,* Augsburg, Bavaria, Germany, 2007, pp.13-22.

[35] (2011, May 10). *Andorid developers* [Online]. Available: http://developer.android.com/guide/basics/what-is-android.html

[36] P. Reddy, "Gender Based Emotion Recognition System for Telegu Rural Dialects Using Hidden Markov Models," *Journal of Computing,* Vol. 2, No. 6, pp.94-98, 2010.

[37] S. Wu, T. H. Falk and W. Y .Chan, "Automatic Recognition of Speech Emotion Using Long-Term Spectro-Temporal Features," in *Proc. Of the International conference on Digital Signal Processing,* Santorini, Greece, 2009, pp.1-6.

[38] F. Burkhardt, A. Paeschke, M. Rolfes, W.F. Sendlmeier and B. Weiss, "A Database of German Emotional Speech," in *Proc. Of the International Speech Communication Association,* Lisboa, Italy, 2005, pp1517-1520.

[39] B. Yang and M. Lugger, "Emotion Recognition From Speech Signals Using New Harmony Features," in *Signal Processing,* vol. 90, No. 5, pp. 1415-1423, 2010.

[40] Y. Huang, G. Zhang, X. Li and F. Da, "Improved Emotion Recognition with Novel Global Utterance Level Features," in *Applied Mathematics & information Sciences,* vol. 5, No. 2, pp. 147-153, 2011.