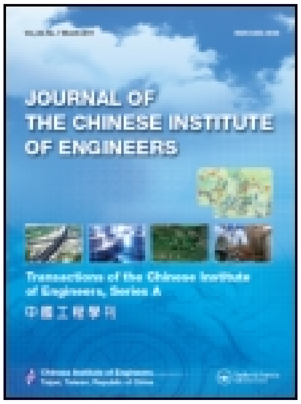

Journal of the Chinese Institute of Engineers

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/tcie20

Speech signal-based emotion recognition and its application to entertainment robots

Kai-Tai Song [a] , Meng-Ju Han [a] & Shih-Chieh Wang [a]

[a] Department of Electrical and Computer Engineering , National Chiao Tung University , Hsinchu , Taiwan , Republic of China

Published online: 19 Dec 2012.

To cite this article: Kai-Tai Song , Meng-Ju Han & Shih-Chieh Wang (2014) Speech signal-based emotion recognition and its application to entertainment robots, Journal of the Chinese Institute of Engineers, 37:1, 14-25, DOI: 10.1080/02533839.2012.751330

To link to this article: http://dx.doi.org/10.1080/02533839.2012.751330

Journal of the Chinese Institute of Engineers, 2014
Vol. 37, No. 1, 14–25, http://dx.doi.org/10.1080/02533839.2012.751330

Taylor & Francis
Taylor & Francis Group

Speech signal-based emotion recognition and its application to entertainment robots

Kai-Tai Song*, Meng-Ju Han and Shih-Chieh Wang

Department of Electrical and Computer Engineering, National Chiao Tung University, Hsinchu, Taiwan, Republic of China

(*Received 8 November 2011; final version received 19 August 2012*)

By recognizing sensory information, through touch, vision, or voice sensory modalities, a robot can interact with people in a more intelligent manner. In human–robot interaction (HRI), emotion recognition has been a popular research topic in recent years. This paper proposes a method for emotion recognition, using a speech signal to recognize several basic human emotional states, for application in an entertainment robot. The proposed method uses voice signal processing and classification. Firstly, end-point detection and frame setting are accomplished in the pre-processing stage. Then, the statistical features of the energy contour are computed. Fisher's linear discriminant analysis (FLDA) is used to enhance the recognition rate. In the final stage, a support vector machine (SVM) is used to complete the emotional state classification. In order to determine the effectiveness of emotional HRI, an embedded system was constructed and integrated with a self-built entertainment robot. The experimental results for the entertainment robot show that the robot interacts with a person in a responsive manner. The average recognition rate for five emotional states is 73.8% using the database constructed in the authors' lab.

Keywords: emotion recognition; speech signal processing; human–robot interaction

1. Introduction

The ability to recognize a user's emotion plays an important role in many human–robot interaction (HRI) scenarios. In recent years, speech-based HRI has gained increasing attention in the development of intelligent service and domestic robotic systems (Sato *et al*. 2006). The emotion communicator, Kotohana, developed by (NEC 2006) is a successful example of vocal emotion recognition. Kotohana is a flower-shaped terminal equipped with Light Emitting Diodes (LEDs). It can recognize a visitor's emotional speech and respond with a color display to convey the interaction. The terminal responds in a lively manner to the detected emotional state, via color variation in the flower. For human beings, speech provides an important communicative cue during social interaction. It also reveals emotion. A human emotion recognition system based on voice signals allows more natural interaction between a human and a robot (Ogawa and Watanabe 2001, Breazeal 2003).

Several methods have been reported for the design of emotion recognition systems that use speech signals. Most methods employ vocal features, including the statistics of fundamental frequency, energy contour, duration of silence, and voice quality (Kwon *et al.* 2003). In order to improve the recognition rate when more than two emotional categories are to be classified, Nwe *et al.* (2003) used short-time log frequency power coefficients to represent speech signals and a discrete hidden Markov model as the classifier. Based on the assumption that the pitch contour has a Gaussian distribution, Hyun *et al.* (2005) proposed a Bayesian classifier for emotion recognition in speech information. They reported that the zero value of a pitch contour causes errors in the Gaussian distribution and proposed a non-zero-pitch method for speech feature extraction. Pao and Chen (2003) used 16-bit linear predictive coding and 20 Mel-frequency cepstral coefficients (MFCC) to identify the emotional state of a speaker. Five emotional categories were classified by using the minimum-distance method and the *nearest mean* classifier. Neiberg *et al*. (2006) modeled the pitch feature by using standard MFCC and MFCC-low, which is calculated between 20 and 300 Hz. Their experiments showed that MFCC-low outperformed the pitch features.

You *et al.* (2006a) indicated that the effectiveness of principal component analysis (PCA) and linear discriminant analysis (LDA) is limited by their underlying assumption that the data are in a linear subspace. For non-linear structures, these methods fail to detect the real number of degrees of freedom of the data, so they proposed the method of Lipschitz embedding (You *et al.* 2006b). The method is not limited by an underlying assumption that the data belong to a linear subspace, so it can analyze the speech signal in more practical situations. Schuller *et al*. (2004) considered an initially large

*Corresponding author. Email: ktsong@mail.nctu.edu.tw set of more than 200 features; they ranked the statistical features according to LDA results and selected important features by ranking statistical features. Chuang and Wu (2004) showed that the contours of the fundamental frequency and energy are not smooth. In order to remove discontinuities in the contour, they used the Legendre polynomial technique to smooth the contours of these features. Their feature extraction procedures firstly estimated the fundamental frequency, energy, formant 1 (F1), and zero-crossing rate (ZCR). From these four features, the feature values are transformed to 33 statistical features. PCA was then used to select 14 principal components from these 33 statistical features, for the analysis of emotional speech. Busso *et al*. (2009) indicated that gross fundamental frequency contour statistics, such as mean, maximum, minimum, and range, are more emotionally prominent than features that describe the shape of the fundamental frequency. Using psychoacoustic harmony perception from music theory, Yang and Lugger (2010) proposed a new set of harmony features for speech emotion recognition. They reported improved recognition by the use of harmony parameters and state-of-the-art features.

For robotics applications, Li *et al.* (2008) developed a prototype chatting robot, which can communicate with a user in a speech dialogue. The recognition of the speech emotion of a specific person was successful for two emotional categories. Kim *et al*. (2009) focused on the speech emotion recognition for a thinking robot. They proposed a speaker-independent feature, namely the ratio of a spectral flatness measure to a spectral center, to solve the problem of diverse interactive users. Similarly, Park *et al*. (2009) also studied the issue of service robots interacting with diverse users who are in various emotional states. Acoustically, similar characteristics between emotions and variable speaker characteristics, caused by different users' style of speech, may degrade the accuracy of speech emotion recognition. They proposed a feature vector classification for speech emotion recognition, to improve performance in service robots.

The related works provide many powerful tools for emotion recognition using speech signals. However, several crucial problems still exist in practical robotic applications. Firstly, a robust speech signal acquisition system must be built on the front end of the design. It is also required that the robot is equipped with a stand-alone system for realistic HRI. One of the greatest challenges in emotion recognition for robotic applications is the performance required for nature and daily life environments. This paper proposes an emotion recognition system that uses a human natural speech signal to classify the five emotional categories. This study aims to develop a voice-based emotion recognition system for an entertainment robot. In such a robotic application, fast response to natural speech signals is required. The use of the statistical features of fundamental frequency and energy parameters to extract the emotion features is proposed. Furthermore, a 'DSP' stands for 'Digital Signal Processing' based embedded speech processing system is designed and implemented, in order to integrate the emotion recognition system with the entertainment robot.

The rest of this paper is organized as follows. Section 2 describes the proposed emotion recognition system, which uses a natural voice signal. In Section 3, an implementation of the system on the self-built entertainment robot is presented. Section 4 presents several interesting experimental results for the proposed system. Section 5 gives conclusions and directions for further research.

2. The proposed emotion recognition system

An embedded speech processing system was designed and produced for real-time speech signal acquisition and processing. Figure 1 shows the block diagram of the proposed emotion recognition system. Speech signals are acquired from a microphone. Using a speech signal pre-processing procedure, the speech voice frames are determined by end-point detection (Han *et al.* 2007). In the speech feature extraction stage, the fundamental frequency and energy features of a speech frame are extracted to represent the speech signal of interest. After obtaining the features of speech frame, Fisher's linear discriminant analysis (FLDA) is utilized to transfer feature values to a suitable space (Kim *et al*. 2002). The feature values in the transferred space represent significant emotional traits and improve the recognition rate. Finally, a hierarchical support vector machine (SVM) classifies the emotional categories. In order to simplify the design of the emotion recognition system for an entertainment robot, it is assumed that each sentence corresponds to only one emotional category. The detailed design of the emotion recognition system is presented in the following section.

2.1. *Speech signal pre-processing*

Before extracting the features of the speech signal for recognition, a voice signal pre-processing stage separates speech frames from the acquired signal. In this design, pre-processing consists of analog to digital conversion, end-point detection, and frame signal separation.

2.1.1. *Signal normalization*

Speech signals acquired from the microphone are analog voltage signals. Through amplification and sampling, the analog voltage signal is converted to digital, in a discrete form. Based on the sampling theorem, a sampling

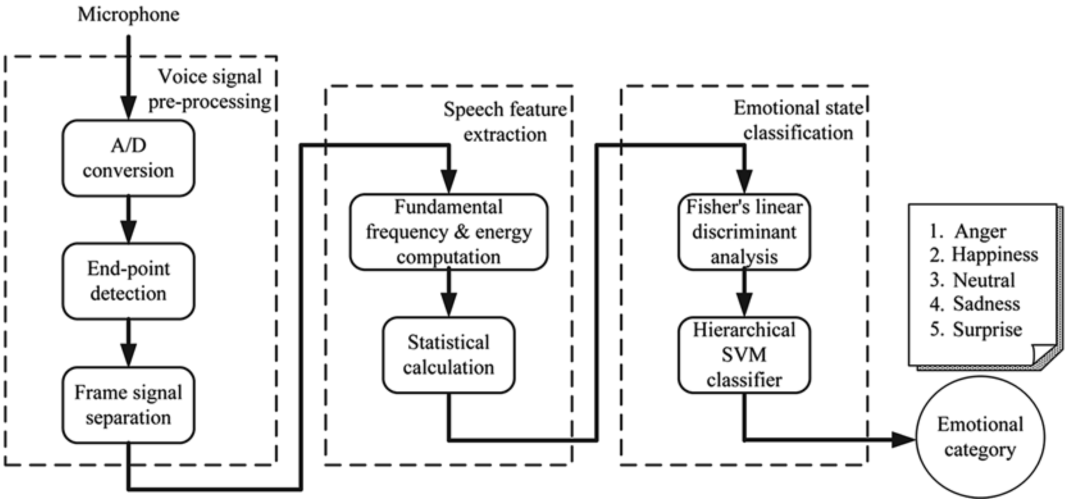
Microphone
Voice signal pre-processing
A/D conversion
End-point detection
Frame signal separation
Speech feature extraction
Fundamental frequency & energy computation
Statistical calculation
Emotional state classification
Fisher's linear discriminant analysis
Hierarchical SVM classifier
1. Anger
2. Happiness
3. Neutral
4. Sadness
5. Surprise
Emotional category

Figure 1. Block diagram of the proposed emotion recognition system.

frequency is set to be more than twice the bandwidth of the input signals, in order to avoid signal distortion. In general, the spectrum of human speech is less than 4 kHz. The sampling frequency is set to 8 kHz, in this study. Furthermore, a normalization scheme is used to reduce the influence of constantly changing input signals. The normalized speech signal is obtained such that,

$$x(n) = \frac{x_{\text{ori}}(n)}{x_{\max}} \quad n = 1, 2, \ldots, N \tag{1}$$

$$x_{\max} = \max\,(x_{\text{ori}}(n)) \quad n = 1, 2, \ldots, N \tag{2}$$

where $x(n)$ represents the normalized speech signal, $x_{\text{ori}}(n)$ represents the original speech signal, and $x_{\max}$ is the maximum value in the sequence, $x_{\text{ori}}(n)$. By dividing with $x_{\max}$, as shown in Equation (1), the amplitudes of whole speech signal are normalized between -1 and 1.

2.1.2. *End-point detection*

In order to extract the emotional features in a voice, a frame size must first be determined for the digitized speech signal. Short-time energy, which is an acoustic feature that correlates the sampled amplitude in each voice frame, is calculated such that

$$E(k) = \sum_{m=0}^{N-1} |x(n+m)|, \tag{3}$$

where $E(k)$ is the short-time energy in the kth frame, $x(n)$ represents the normalized speech signal, and N is the frame size. The starting and terminal thresholds are then determined for the voice frame, to determine the starting and terminal points, respectively, by using empirical rules. Once the value of $E(k)$ is greater than the starting threshold, the starting point is determined. However, the terminal point is determined when the value of $E(k)$ is smaller than the terminal threshold. Hence a frame size, N, is determined as the real speech signal. As shown in Figure 2, the starting and terminal points of a speech frame are determined by the starting threshold and the terminal threshold, respectively.

The ZCR is then used for audio frame setting. ZCR is a basic acoustic feature. It is equal to the number of zero crossings of the waveform within a given frame. Here the ZCR is defined as the number of times which the speech signals cross the zero value origin of the y-coordinates. In general, the ZCR of non-speech and environmental noise is lower than that of human speech (Yan 2002). The ZCR is calculated such that,

$$Z(k) = \frac{1}{2}\sum_{m=0}^{N-1} |\text{sgn}(x(n+m)) - \text{sgn}(x(n+m-1))|, \tag{4}$$

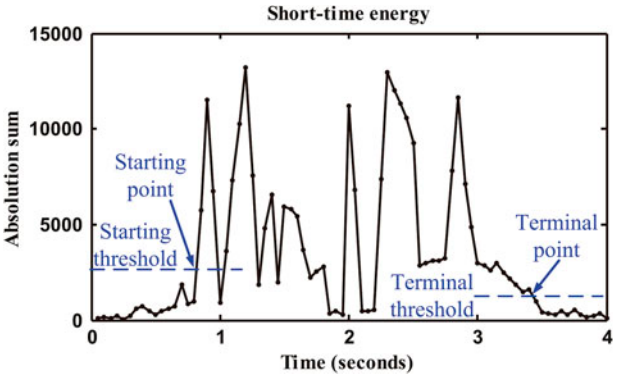

Figure 2. Energy of a speech signal.

$$\mathrm{sgn}[x(n)] = \left\{ \begin{matrix} 1 & if & x(n) \geq 0 \\ -1 & if & x(n) < 0 \end{matrix} \right\}, \qquad (5)$$

where $Z(k)$ is the ZCR of the kth frame. In practice, the short-time energy is used to estimate the starting and terminal points of the whole speech segment, wherein the speech voice occurs. Then, the ZCR is used to find the real speech signal more precisely. As shown in Figure 3, the real speech signal is determined by the ZCR threshold.

In this design, ZCR and short-time energy are both used to detect the starting and terminal points of non-speech. Figure 4 shows the four rules to find the real human speech signal:

(1) If $E(k)$ is lower than the terminal threshold, it belongs to non-speech.
(2) If $E(k)$ is higher than the starting threshold, the starting point of the human speech signal is determined.
(3) If $E(k)$ is lower than the starting threshold and $Z(k)$ is higher than the ZCR threshold of the ZCRs, this is determined as the starting point of the human speech signal.
(4) If $E(k)$ is lower than the terminal threshold, after the starting point, it is determined that this is the terminal point of the human speech signal.

Using the above rules, the starting and terminal points of speech signals are determined. The boundary of real human speech is also determined. Figure 5 shows an example of the end-point detection.

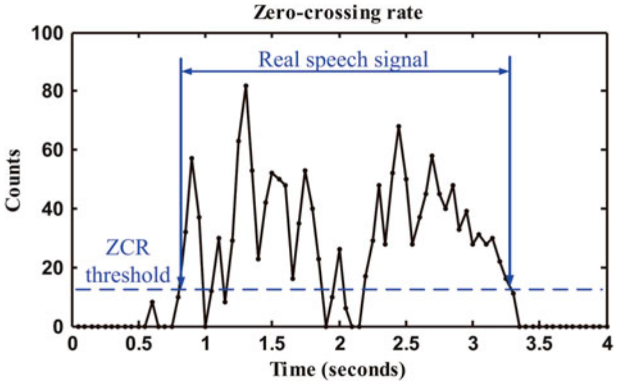

Figure 3. ZCR of a speech signal.

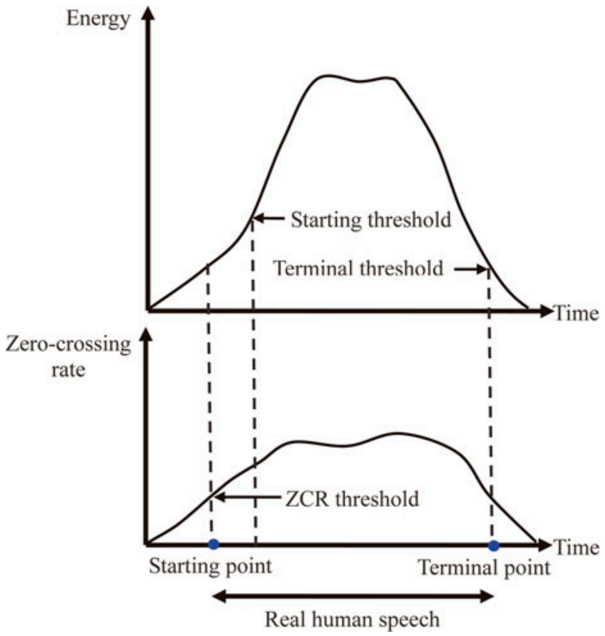

Figure 4. Example of real human speech detection.

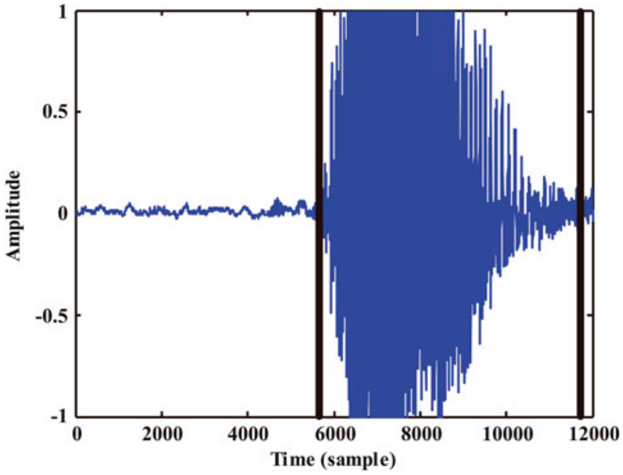

Figure 5. An example of end-point detection.

2.1.3. *Frame separation*

After obtaining the end-points of the actual human speech signal, suitable presentation of the speech signal is required, before the feature extraction step. In order to reduce the variation between adjacent frames, the overlapping part of the signal is used to avoid discontinuity. This study uses a Hamming window to emphasize the medium signal and to restrain both side signals (Gold and Morgan 2000). Figure 6 shows the frame-signal separation using a Hamming window. It can be seen that there are overlaps between frames. The Hamming window is represented such that

$$\mathrm{Window}(n) = \left\{ \begin{matrix} 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) & n = 0, 1, \ldots, N-1 \\ 0 & \text{otherwise} \end{matrix} \right\}. \qquad (6)$$

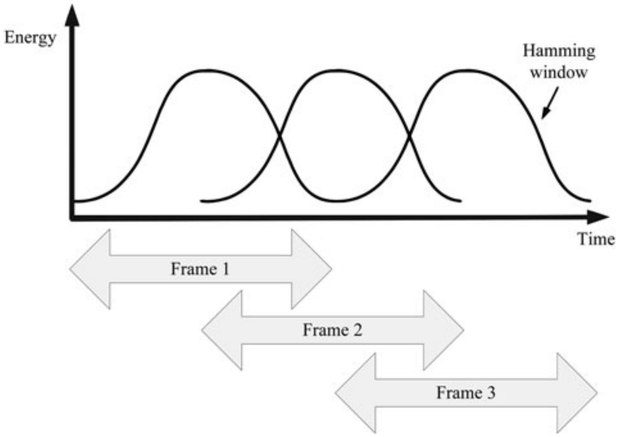

Figure 6. Frame-signal separation using a Hamming window.

where N is the length of the frame and n is the sample point in a frame. Figure 7 shows the procedure for speech signal extraction in each frame. Figure 7(a) shows an example of an original speech signal in a frame. Figure 7(b) depicts the Hamming window. Figure 7(c) is the extracted result for the original speech signal multiplied by the Hamming window. This study uses the first 128 samples to determine the energy threshold values and then divides a frame into several 32 ms periods, for further feature extraction.

2.2. *Feature extraction*

After the speech signal is obtained for each frame, useful features are extracted from the speech signal. In this work, the contours of the fundamental frequency and energy (Schuller *et al*. 2004) are used for human emotion recognition.

2.2.1. *Fundamental frequency and short-time energy*

Several methods can be used to extract the fundamental frequency from a speech signal (Jelinek 1999). In this design, the contour of the fundamental frequency is obtained using an autocorrelation function. The fundamental frequency is determined by the maximum

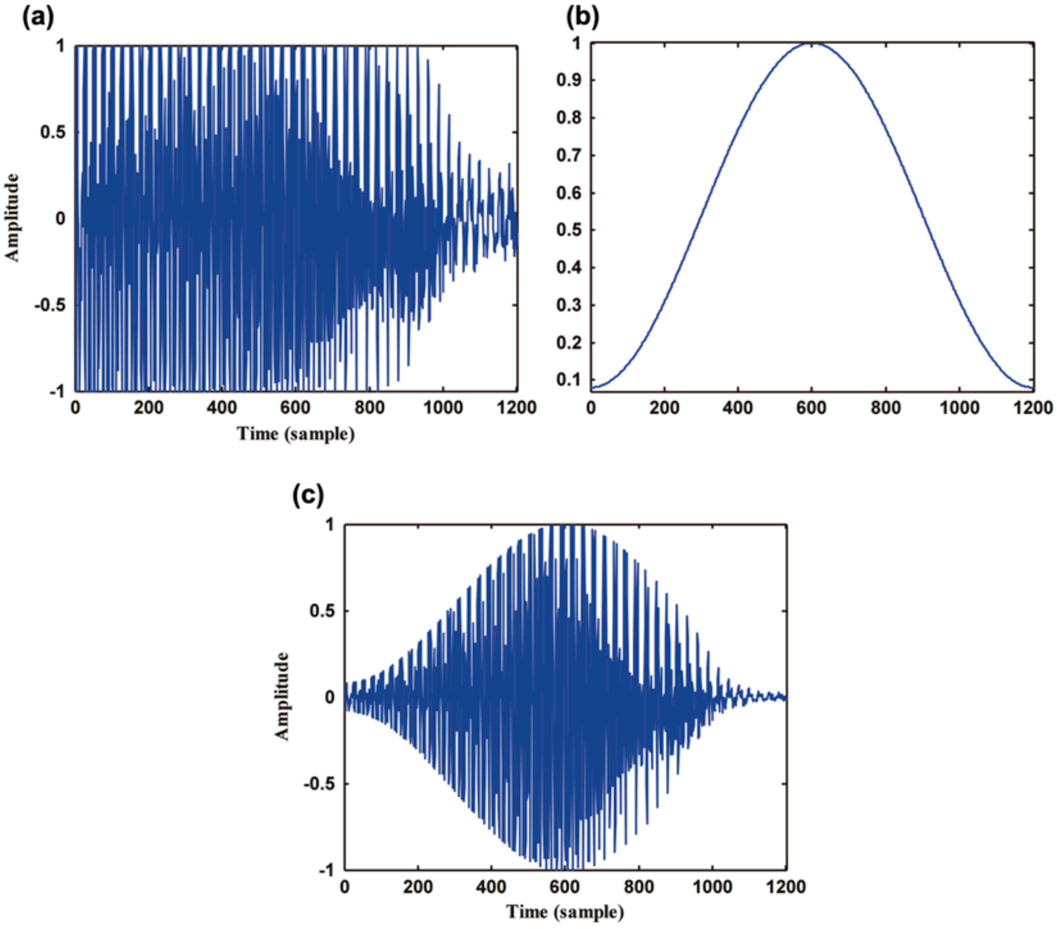

Figure 7. Procedure for speech signal extraction in each frame. (a) A frame of original speech signal. (b) Hamming window. (c) Result of original speech signal multiplied by Hamming window.

autocorrelation value. The autocorrelation function is defined such that

$$R(d) = \sum_{d=0}^{N-1-d} x(n) \cdot x(n+d), \qquad (7)$$

where d is the shifting parameter. The value of d that maximizes $R(d)$ over a specified range is selected as the period of the fundamental frequency of the sample points. Figure 8 shows the original time response of the speech signal and results for feature extraction of the fundamental frequency. The energy contour is obtained by calculating the short-time energy of each frame in Equation (3).

2.2.2. *Statistical features*

After obtaining the short-time energy and fundamental frequency values, the statistical values of fundamental frequency and energy features are calculated, including average, standard deviation, maximum, minimum, and median. These statistical values are listed in Table 1. Fourteen statistical values are defined in this study. Based on observation, the selected statistical features are sufficient to express variations in emotion and produce satisfactory results.

Table 1. Definition of statistical speech features.

Fundamental frequency ($F0$)	Energy
(1) Average of $F0$	(9) Average energy
(2) Standard deviation of $F0$	(10) Standard deviation of energy
(3) Maximum of $F0$	(11) Maximum energy
(4) Minimum of $F0$	(12) Median energy
(5) Median of $F0$	(13) Average of energy derivation
(6) Average of $F0$ derivation	(14) Standard deviation of energy derivation
(7) Standard deviation of $F0$ derivation	
(8) Maximum of $F0$ derivation	

2.3. *Emotional state classification*

After obtaining the statistical features from the speech signal, a suitable classification procedure is required to recognize the emotion categories. In order to increase the discriminability of the feature values, FLDA (Mika *et al.* 1999) is used to find a suitable subspace to discriminate the emotional categories. An SVM (Vapnik 1995, Cristianini and Shawe-Taylor 2000) has been an effective method for designing recognition systems. This study uses both FLDA and SVM to classify the emotional categories.

2.3.1. *Fisher's linear discriminant analysis*

FLDA is a popular method for pattern recognition, to find a linear combination of features which separate two or more classes of objects. It projects the original high-dimensional data onto a low-dimensional space. All of the classes are well separated by maximizing the Raleigh quotient (Duda and Hart 1973). In FLDA, one assumes that there are r training sample vectors, given by $\{ts_i\}_{i=1}^{r}$, for p classes: $C_1, C_2, \ldots, C_p$, and r_j samples for the jth class, such that

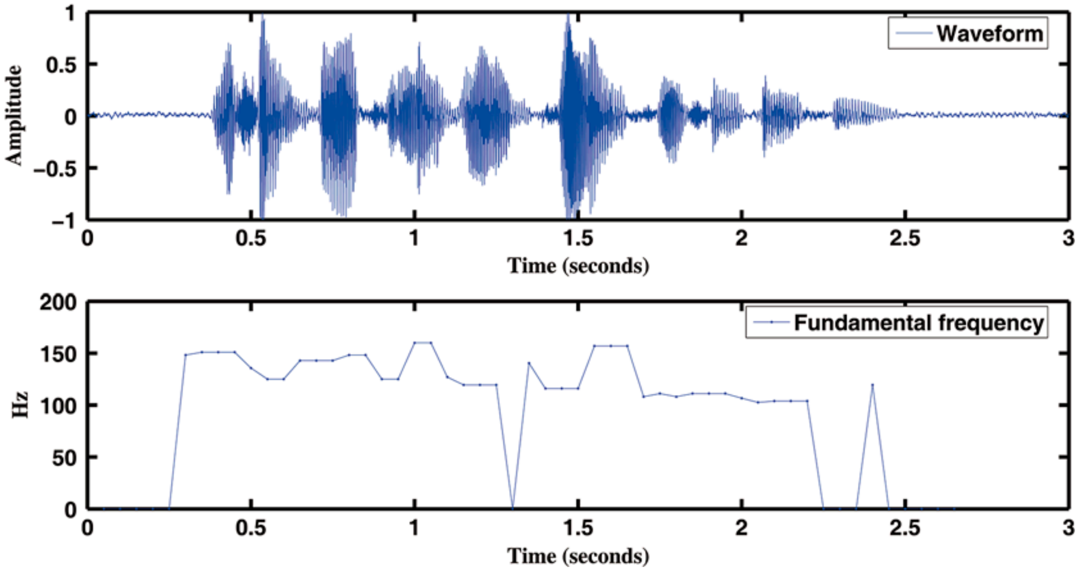

Figure 8. The original time response of the speech signal and the results for feature extraction of the fundamental frequency.

$$r = \sum_{j=1}^{p} r_j. \tag{8}$$

Let μ be the mean of all of the training samples, such that

$$\mu = \frac{1}{r} \sum_{i=1}^{r} ts_i, \tag{9}$$

and μ_j be the mean of the jth class, such that

$$\mu_j = \frac{1}{r_j} \sum_{r_i \in C_j} ts_i, \tag{10}$$

where the within-class scatter matrix $\boldsymbol{S_W}$ and the between-class scatter matrix $\boldsymbol{S_B}$ are defined as follows:

$$S_W = \sum_{i=1}^{r_j} \sum_{j=1}^{p} (ts_i - \mu_j)(ts_i - \mu_j)^T, \tag{11}$$

$$\boldsymbol{S_B} = \sum_{j=1}^{p} r_j(\mu_j - \mu)(\mu_j - \mu)^T. \tag{12}$$

The goal is to find a transform vector $\boldsymbol{w}$ such that the Raleigh quotient is maximized. The Raleigh quotient is defined such that

$$q = \frac{\boldsymbol{w^T S_B w}}{\boldsymbol{w^T S_W w}}, \tag{13}$$

here, $\boldsymbol{w}$ can be defined by solving a generalized eigen problem, as specified by $\boldsymbol{S_B w} = \lambda \boldsymbol{S_W w}$, where λ is a generalized eigenvalue. An $L \times M$ matrix, $\boldsymbol{W}$, can be found to transform the original L-dimensional data into a M-dimensional space. It is expected that the p classes can be well separated in this M-dimensional space. In this work, M is selected as 12 from the practical test. Since voice signals are noisy and direction sensitive, the FLDA is used to efficiently discriminate the speech features. In this study, each emotional sentence is represented as 14 statistical features which are listed in Table 1. Then, these 14 statistical features are projected into a subspace by using the transformation matrix, $\boldsymbol{W}$, to obtain new 12 feature values. Afterward each emotional sentence is transformed into 12 feature values for recognition.

2.3.2. *Hierarchical SVM classifier*

SVM is a two-class classifier for a set of related supervised learning methods that analyze the data and recognize patterns. The SVM model represents examples as points in space. It determines a hyperplane, so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category, based on the side of the gap on which they fall. In this study, five classes of emotional categories are classified by using SVM. In order to utilize this two-class classifier to classify five categories, a hierarchical SVM is adopted (Han *et al.* 2007). The hierarchical SVM classifier is illustrated in Figure 9. In this design, five emotional states are categorized. An SVM hyperplane can distinguish two categories. Therefore a four-stage classifier is developed, as shown in Figure 9. Each stage determines one emotional state from the two and the selected one proceeds to the next stage, until a final emotional state is determined. When unknown emotional speech is imported into the SVM, as shown in Figure 9, the SVM first classifies neutral vs. happiness and then classifies anger vs. surprise. After these stages, the corresponding results are further classified at the next stage. For example, the results of the first and second stage classifiers are assumed to be happiness and surprise (shown as 1 and 2 in Figure 9). At the third stage, the classifier determines the unknown data as surprise or sadness. If the classification result is surprise (shown as 3), then the classifier determines that the unknown data is happiness or surprise, in the final stage (shown as 4). The system eventually produces a recognition result.

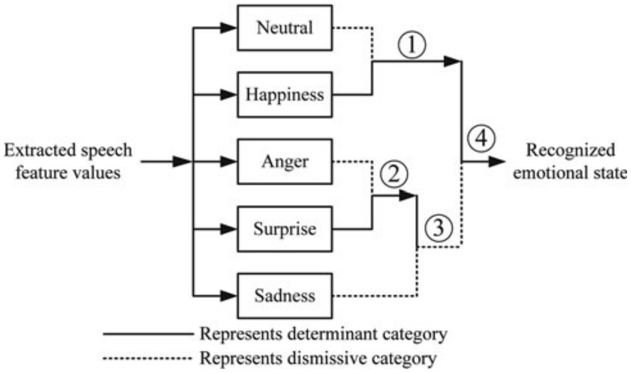
Extracted speech feature values
Neutral
Happiness
Anger
Surprise
Sadness
①
②
③
④
Recognized emotional state
Represents determinant category
Represents dismissive category

Figure 9. Structure of the hierarchical SVM classifier.

3. Implementation of the emotion recognition embedded system

The developed algorithms were implemented on a DSP-based embedded system (Andrian and Song 2005), to facilitate the experimental study of an entertainment robot. The embedded system consists of a microphone and a DSK6416 DSP board from Texas Instruments. The selection of the DSK6416 as the main processing unit is because of its high performance in fixed-point calculation, with a 1 GHz clock rate. Figure 10 shows the

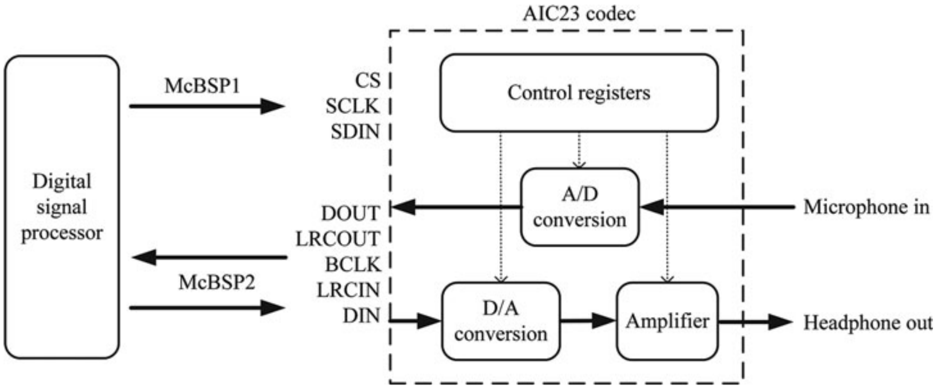

Figure 10. The TMS320C6416 DSK codec interface.

TMS320C6416 DSK codec interface (Texas Instruments Inc. 2003, 2004). The DSK uses a Texas Instruments AIC23 stereo codec for input and output of audio signals. The codec samples an analog signal from a microphone and converts the signal into digital data, so that it can be processed by the DSP. The DSP chip and codec communicate via two serial channels; one controls the codec's internal configuration registers and the other is responsible for digital audio samples. As shown in Figure 10, the McBSP1 is used as the unidirectional control channel; the McBSP2 is used as the bi-directional audio data channel. The codec has a 12 MHz system clock. The internal sample rate subdivides the 12 MHz clock to generate common frequencies, including 48, 44.1, and 8 kHz; a frequency of 8 kHz is selected to sample the user's speech signal, in this study. As a user speaks into the microphone, the embedded system acquires speech signals and begins to recognize the user's emotional state. The recognition results are transmitted via RS-232 serial link to a host computer (PC), where intelligent responses are generated to react to the received speech signal.

In order to test the emotion recognition system in practical scenarios of HRI, the embedded speech processing system is integrated within the self-built entertainment robot. The control architecture of this robot is depicted in Figure 11. The DSP-based system is installed at the back of the entertainment robot. Seven Radio Controlled (RC) servos are used to control the movement of the ears, head, hands, and legs of the entertainment robot. A motor servo controller, from Pololu Robotics and Electronics Inc. (2012), controls the RC servos in the robot. The DSP-based emotion recognition system estimates emotion categories and determines, in real time, a suitable response for the entertainment robot. Figure 12 shows an interaction scenario between a user and the entertainment robot. Some interesting studies (Valin *et al*. 2007, Nakajima *et al*. 2010) have utilized microphone arrays to avoid using a headset. Their methods improve the speech recognition system to cope with noise and direction sensitivity problems. In this study, we focus on the integration of emotional speech recognition algorithm and entertainment robot. In order to reduce the influence of the sound of robot motion or surrounding interference, a headset is used in the experiments, as shown in Figure 12.

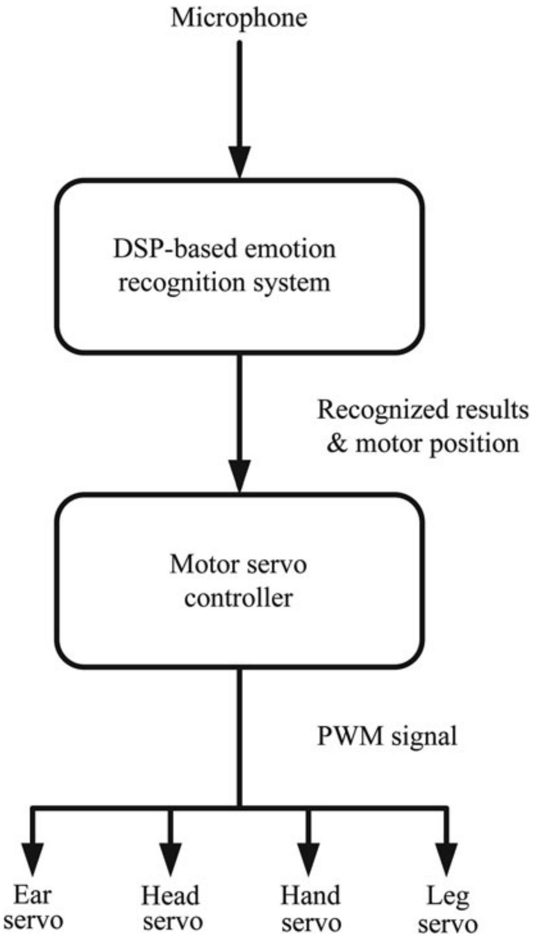

Figure 11. Control architecture of the entertainment robot.

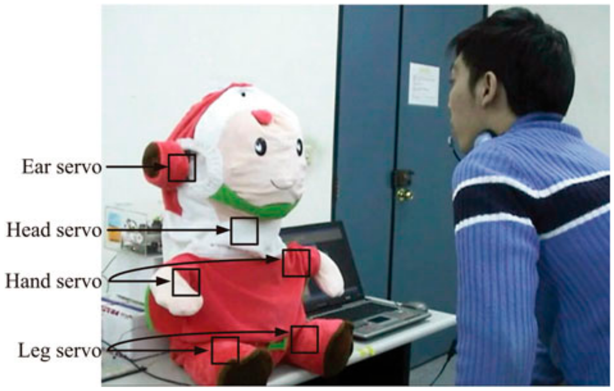

Figure 12. Interaction scenario between a user and the entertainment robot.

4. Experimental results

The performance of the proposed emotional voice recognition system was evaluated using a self-built database. Furthermore, experimental study of the proposed system was performed by integrating the DSP-based system into an entertainment robot.

4.1. *Experiments using the self-built database*

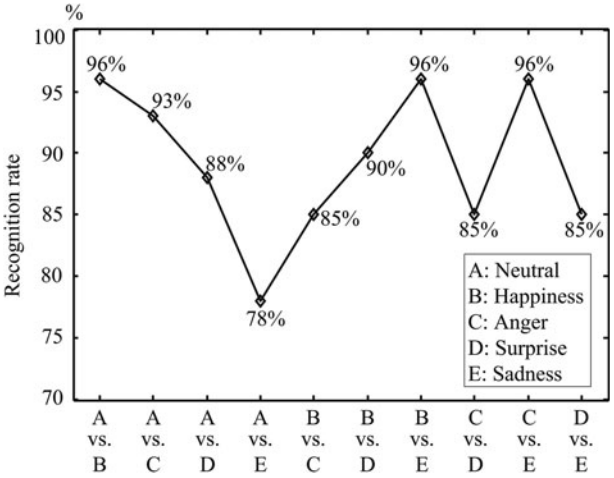

Figure 13. Experimental results of recognition rate for any two emotional categories.

The proposed emotion recognition system was tested using a speech database built in the ISCI lab of National Chiao Tung University. 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 a natural way, each subject was asked to narrate expressive sentences, in Chinese, to imitate an actual interactive scenario. Table 2 lists the meaning of each sentence, in English. Currently, the database includes various emotional utterances from five persons. Each person recorded each sentence six times, so there are 90 utterances per emotion category and 450 utterances in total, in this database. In the following experiments, 45 data samples were randomly selected as training data for each emotional category and the other 45 data samples were used as test data. Part of the voice clips of the database can be found in ISCI Lab (2011a).

Table 2. Meaning of sentence content for five emotional categories.

Emotional category	Content of sentence
Anger	1. How can you do that without my agreement? 2. It's none of your business 3. What you are doing is wrong!
Happiness	1. It's almost new year! 2. I will go abroad on vacation tomorrow 3. I won the lottery!
Neutral	1. It's a sunny day 2. I have something to do later 3. Are you hungry?
Sadness	1. My cat is lost 2. I got a cold 3. Everything went without a hitch today
Surprise	1. Are you serious? 2. I can't believe that it really happened 3. Ah! My notebook is lost

In order to compare the effect of speech features, fundamental frequency, and short-time energy features, the emotion is first evaluated between any two emotional categories. Figure 13 shows the experimental results of the SVM classification of any two emotional categories. There are 10 combinations of any two emotional expressions. It is seen that the recognition rates for these nine combinations are higher than 85%. The recognition rate of neutral vs. sadness (A vs. E) is the lowest, mainly due to the small prosodic variation between neutral and sad speech utterances. The other recognition rates lie between 85 and 96%. The average recognition rate is 89.2%. This indicates that the proposed statistical features can represent emotional characteristics properly.

The hierarchical SVM classifier (shown as Figure 9) was then employed to recognize five emotional categories. In the experiments, the SVM classifier was trained using a set of 45 data samples for each emotional category. These 45 data samples came from five persons, with each person contributing three samples of each emotional sentence. The other 45 data samples were tested for recognition of the emotion category. The test results are presented in Table 3. The average recognition rate for the five emotional expressions is 73.78%. It is noted that anger can be classified as surprise. It is due to the similar speech rates and tones of these two kinds of sentences in the self-built database. Moreover, the accent

Table 3. Experimental results of recognizing five emotional categories.

Output input	Anger	Happiness	Neutral	Sadness	Surprise	Recognition rate (%)
Anger	30	0	3	4	8	66.67
Happiness	1	37	3	4	0	82.22
Neutral	1	6	35	3	0	77.78
Sadness	0	4	6	30	5	66.67
Surprise	5	2	1	3	34	75.56
Average						73.78

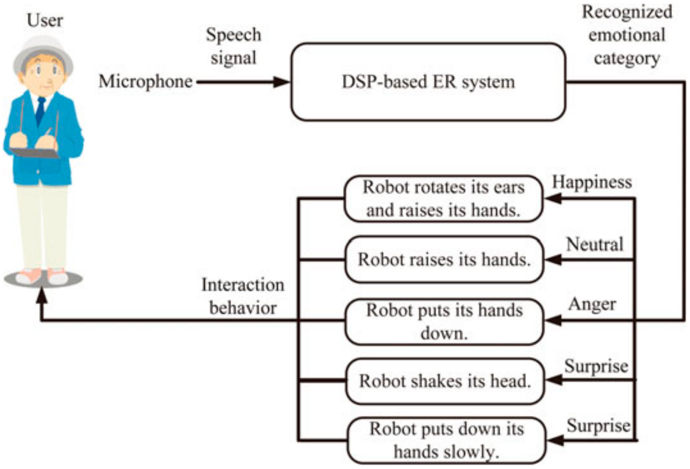

Figure 14. Block diagram of the emotional interaction system.

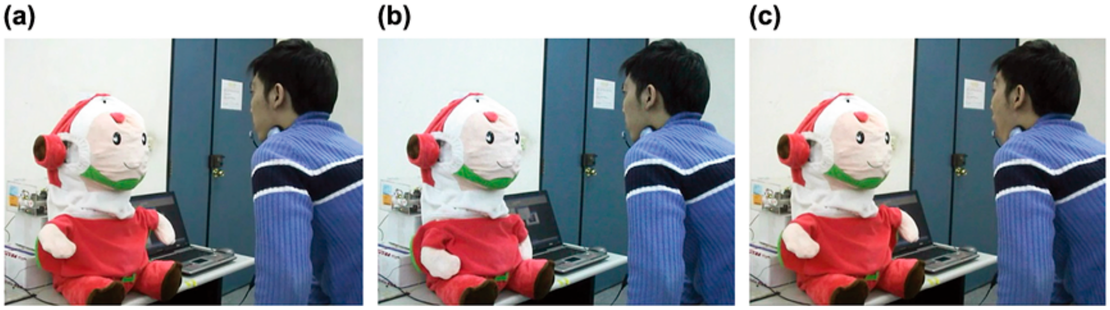

Figure 15. Interactive response of the robot as the user says, 'I am angry!' (a) The robot puts down its hands to portray fear. (b) The robot continues to put down its hands to the lowest position. (c) The robot raises its hands back to the original position.

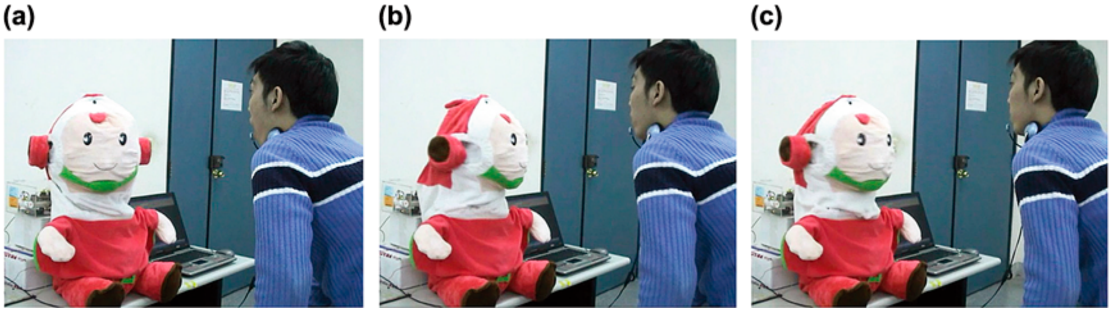

Figure 16. Interactive response of the robot, when the user speaks in a surprised tone. (a) The robot shakes its head to the right. (b) The robot shakes its head to the left. (c) The robot puts its head back to the original position.

and noise of the voice influence the classification that results a lot. We will take these factors into consideration in future work.

4.2. *Experiments with the entertainment robot*

In this study, we aim to develop an entertainment robot suitable as a children's toy. In such a robotic application, fast response to natural speech signal is required. Therefore, a simple entertainment robot is built to verify the proposed natural speech signal emotion recognition algorithm. The complete emotion recognition system was integrated into the self-constructed entertainment robot. Figure 14 shows a block diagram of the implemented interaction control system on the robot.

In the experiment, a user speaks in front of the robot, as shown in Figure 12. After acquiring the speech signals, the emotion recognition system begins to process the audio information. When no human speech is detected, the robot manifests a bored behavior by turning its head to look around. When a user says 'hello' to the robot, with neutral emotion, the robot raises its hands to respond to the user. If a happy emotion from the user is detected, the robot rotates its ears and raises its hands to show a happy gesture. When the user expresses anger to the robot, the robot puts its hands down to portray fear. However, the robot shakes its head if surprise is detected from the user. Figure 15 shows the interaction responses of the robot when the user said, 'I am angry!' with an angry tone. After that, the user used a surprise tone to the robot. As shown in Figure 16, the robot shook its head in response to the recognized emotional state. The experimental results verify that the proposed emotion recognition system allows the robot to interact with a user in a natural and friendly manner. A video clip of the experimental results can be found in ISCI Lab (2011b). In the future, a fast system will be further studied to recognize human's emotional speech and interact in a more human-like manner. Some suitable psychological findings will also be considered to apply to the emotional robotic system.

5. Conclusions

In this paper, an emotion recognition system based on natural speech signals was designed and implemented for an entertainment robot. The emotion recognition system developed classifies the five emotional categories, in real time. Experimental results using an entertainment robot show that the robot can interact with a user in a responsive manner, using the developed speech signal recognition system. Using a database built in the ISCI lab, the proposed system achieves an average recognition rate of 73.8% for five emotional states. Because the voice signal must be acquired using the embedded system, it is difficult to establish a benchmark to evaluate the developed recognition algorithm. The recognition rate can be improved further. In the future, a method to extract key phrases in an utterance will be investigated, to increase the recognition rate. The emotional state can be estimated more directly from the speech signal, than from the extracted statistical features of the whole voice frame.

Acknowledgments

This work was partly supported by grants from Industrial Technology Research Institute, A341XS6M00 and Ministry of Economic Affairs, 95-EC-17-A-04-S1-054.

Nomenclature

BCLK	I^2S serial-bit clock
C_i	*i*th class of training sample vectors
CS	control port input latch/address select
d	shifting parameter of autocorrelation function
DIN	I^2S format serial data input to the sigma-delta stereo DAC
DOUT	I^2S format serial data output from the sigma-delta stereo ADC
$E(\mathrm{k})$	short-time energy in *k*th frame
$F0$	fundamental frequency
L	dimension of original data
LRCIN	I^2S DAC-word clock signal
LRCOUT	I^2S ADC-word clock signal
M	dimension of projected subspace
N	speech frame size
p	number of class of training sample vectors
q	defined Raleigh quotient
$R(d)$	autocorrelation function
r	number of training sample vectors of FLDA procedure
r_j	number of samples for the *j*th class
SCLK	control port serial data clock
SDIN	control port serial data input
$\boldsymbol{S_W}, \boldsymbol{S_B}$	within-class and between-class scatter matrices
ts	training sample vector of FLDA procedure
$\boldsymbol{W}$	matrix to transform the original data into a subspace
$\boldsymbol{w}$	transform vector
Window	Hamming window
x	normalized speech signal
x_{ori}	original speech signal
$x_{\max}$	maximum value in original speech signal
$Z(k)$	zero-crossing rate in *k*th frame
μ	mean of the entire training samples
μ_j	mean of the *j*th class

References

Andrian, H., and Song, K.T., 2005. Embedded CMOS imaging system for real-time robotic vision. *In*: *Proceedings of IEEE/RSJ international conference on intelligent robots and systems*, 2–6 August 2005 Alberta. Canada: IEEE, 3694–3699.

Breazeal, C., 2003. Emotive qualities in lip-synchronized robot speech. *Advanced robotics*, 17 (2), 97–113.

Busso, C., Lee, S., and Narayanan, S., 2009. Analysis of emotionally salient aspects of fundamental frequency for emotion detection. *IEEE transactions on audio, speech, and language processing*, 17 (4), 582–596.

Chuang, Z.J., and Wu C.H., 2004. Emotion recognition using acoustic features and textual content. *In*: *Proceedings of IEEE international conference on multimedia and expo*, 27–30 June 2004 Taipei. Taiwan: IEEE, 53–56.

Cristianini, N. and Shawe-Taylor, J., 2000. *An introduction to support vector machines and other kernel-based learning methods*. New York, NY: Cambridge University Press.

Duda, R.O. and Hart, P.E., 1973. *Pattern classification and scene analysis*. New York, NY: Wiley.

Gold, B. and Morgan, N., 2000. *Speech and audio signal processing: processing and perception of speech and music*. New York, NY: John Wiley.

Han, M.J., Hsu, J.H., Song, K.T., and Chang, F.Y. 2007. A new information fusion method for SVM-based robotic audio-visual emotion recognition. *In*: *Proceedings of IEEE international conference on systems, man and cybernetics*, 7–10 October 2007 Montreal. Canada: IEEE, 2656–2661.

Hyun, K.H., Kim, E.H., and Kwak, Y.K., 2005. Improvement of emotion recognition by Bayesian classifier using non-zero-pitch concept. *In*: *Proceedings of 14th IEEE international workshop on robot and human unteractive communication*, 13–15 August 2005 Nashville, TN. USA: IEEE, 312–316.

ISCI Lab, intelligent system control integration laboratory, 2011a. *Emotional utterance voice clip* [online]. Available from: http://isci.cn.nctu.edu.tw/JCIE/VoiceClip/ [Accessed 1 August 2012].

ISCI Lab, intelligent system control integration laboratory, 2011b. *Experimental video clip* [online]. Available from: http://isci.cn.nctu.edu.tw/JCIE/VideoClip/ [Accessed 1 August 2012].

Jelinek, F., 1999. *Statistical methods for speech recognition*. Cambridge, MA: MIT Press.

Kim, H.C., Kim, D.J., and Bang, S.Y., 2002. Face recognition using LDA mixture model. *In*: *Proceedings of IEEE 16th international conference on pattern recognition*, 11–15 August 2002 Quebec. Canada: IEEE, 925–928.

Kim, E.H., Hyun, K.H., Kim, S.H., and Kwak, Y.K., 2009. Improved emotion recognition with a novel speaker-independent feature. *IEEE transactions on mechatronics*, 14 (3), 317–325.

Kwon, O.W., Chan, K., Hao, J., and Lee, T.W. 2003. Emotion recognition by speech signals. *In*: *Proceedings of 8th european conference on speech communication and technology (Eurospeech '03)*, 1–4 September 2003 Geneva. Switzerland: International Speech Communication Association, 125–128.

Li, C., Zhou, Q., Cheng, J., Wu, X., and Xu, Y. 2008. Emotion recognition in a chatting robot. *In*: *Proceedings of 2008 IEEE international conference on automation and logistics*, 1–3 September 2008 Qingdao. China: IEEE, 1452–1457.

Mika, S., Ratsch, G., Weston, J., Scholkopf, B., and Muller, K.R. 1999. Fisher discriminant analysis with kernels. *In*: 1999 *IEEE international workshop on neural networks for signal processing*, 23–25 August 1999 Madison. WI. USA: IEEE, 41–48.

Nakajima, H., Nakadai, K., Hasegawa, Y., and Tsujino, H., 2010. Blind source separation with parameter-free adaptive step-size method for robot audition. *IEEE transactions on audio, speech, and language processing*, 18 (6), 1476–1485.

NEC., 2006. *NEC's KOTOHANA - Emotion communicator* [online]. Available from: http://thefutureofthings.com/pod/1042/necs-kotohana-emotion-communicator.html [Accessed 1 August 2012].

Neiberg, D., Elenius, K., and Laskowski, K., 2006. Emotion recognition in spontaneous speech using GMMs. *In*: *Proceedings of 9th ISCA international conference on spoken language processing (INTERSPEECH 2006)*, 17–21 September 2006 Pittsburgh. PA. USA: International Speech Communication Association, 809–812.

Nwe, T.L., Foo, S.W., and De Silva, L.C., 2003. Speech emotion recognition using hidden Markov models. *Speech communication*, 41 (4), 603–623.

Ogawa, H. and Watanabe, T., 2001. InterRobot: speech-driven embodied interaction robot. *Advanced robotics*, 15 (3), 371–377.

Pao, T.L., and Chen, Y.T., 2003. Mandarin emotion recognition in speech. *In*: *Proceedings of IEEE workshop on automatic speech recognition and understanding*, 30 November–4 December 2003 St. Thomas. Virgin Islands: IEEE, 227–230.

Park, J.S., Kim, J.H., and Oh, Y.H., 2009. Feature vector classification based speech emotion recognition for service robots. *IEEE transactions on consumer electronics*, 55 (3), 1590–1596.

Pololu Robotics and Electronics Inc., 2012. *Pololu serial 8-servo controller* [online]. Available from: http://www.pololu.com/products/pololu/0727/ [Accessed 1 August 2012].

Sato, M., Sugiyama, A., and Ohnaka, S., 2006. Auditory system in a personal robot, PaPeRo. *In*: *Proceedings of IEEE international conference on consumer electronics*, 7–11 January 2006 Las Vegas, NV. USA: IEEE, 19–20.

Schuller, B., Rigoll, G., and Lang, M., 2004. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine – Belier network architecture. *In*: *Proceedings of IEEE international conference on acoustics, speech, and signal processing*, 17–21 May 2004 Montreal, Quebec. Canada: IEEE, 577–580.

Texas Instruments Inc., 2003. *TMS320C6416 DSK technical reference* [online]. Available from: http://c6000.spectrum-digital.com/dsk6416/V1/docs/dsk6416_TechRef.pdf [Accessed 1 August 2012].

Texas Instruments Inc., 2004. *TLV320AIC23B data manual* [online]. Available from: http://www.ti.com/lit/ds/symlink/tlv320aic23b.pdf [Accessed 1 August 2012].

Valin, J.M., Yamamoto, S., Rouat, J., Michaud, F., Nakadai, Kazuhiro, and Okuno, H.G., 2007. Robust recognition of simultaneous speech by a mobile robot. *IEEE transactions on robotics*, 23 (4), 742–752.

Vapnik, V., 1995. *The nature of statistical learning theory.* New York, NY: Springer.

Yan, K.M., 2002. *Development of a home robot speech recognition system*. Thesis (Master). National Chiao Tung University, Hsinchu, Taiwan.

Yang, B. and Lugger, M., 2010. Emotion recognition from speech signals using new harmony features. *Signal processing*, 90 (5), 1415–1423.

You, M, Chen, C., Bu, J., Liu, J., and Tao, J. 2006a. Emotional speech analysis on nonlinear manifold. *In*: *Proceedings of IEEE international conference on pattern recognition*, 20–24 August 2006 Hong Kong. China: IEEE, 91–94.

You, M, Chen, C., Bu, J., Liu, J., and Tao, J. 2006b. A hierarchical framework for speech emotion recognition. *In*: *Proceedings of IEEE international symposium on industrial electronics*, 9–13 July 2006 Montreal, Quebec. Canada: IEEE, 515–519.