Expert Systems with Applications 41 (2014) 5858–5869

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

A new approach of audio emotion recognition

Chien Shing Ooi [a,*], Kah Phooi Seng [b], Li-Minn Ang [b], Li Wern Chew [c]

[a] *Department of Computer Science & Networked System, Sunway University, 46150 Petaling Jaya, Malaysia*
[b] *School of Engineering, Edith Cowan University, WA 6027, Australia*
[c] *Intel Microelectronics (M) Sdn. Bhd., 11900 Pulau Pinang, Malaysia*

ARTICLE INFO

Keywords:
Audio emotion recognition
RBF neural network
Prosodic features
MFCC feature

ABSTRACT

A new architecture of intelligent audio emotion recognition is proposed in this paper. It fully utilizes both prosodic and spectral features in its design. It has two main paths in parallel and can recognize 6 emotions. Path 1 is designed based on intensive analysis of different prosodic features. Significant prosodic features are identified to differentiate emotions. Path 2 is designed based on research analysis on spectral features. Extraction of Mel-Frequency Cepstral Coefficient (MFCC) feature is then followed by Bi-directional Principle Component Analysis (BDPCA), Linear Discriminant Analysis (LDA) and Radial Basis Function (RBF) neural classification. This path has 3 parallel BDPCA + LDA + RBF sub-paths structure and each handles two emotions. Fusion modules are also proposed for weights assignment and decision making. The performance of the proposed architecture is evaluated on eNTERFACE'05 and RML databases. Simulation results and comparison have revealed good performance of the proposed recognizer.

1. Introduction

Speech signals can rapidly deliver information or messages by human. Audio emotion recognition is a way to identify the emotional state of human from these speech signals. It is very useful for many applications such as safety in automotive (Nass et al., 2005), diagnosis tool (Edwards, Jackson, & Pattison, 2002), customer satisfaction assessments in call centers (Petrushin, 1999), etc.

In previous research, audio emotion recognition has been studied on different aspects. One of the aspects is the investigation on the emotion representation of audio features. Audio features such as pitch (Busso, Lee, & Narayanan, 2009; Devillers, Vidrascu, & Layachi, 2010), log energy, zero crossing rate (Chien Hung, Ping Tsung, & Chen, 2010; Chih-Chang, Chien-Hung, Ping-Tsung, & Chen, 2010), spectral features (Wong & Sridharan, 2001), voice quality (Lugger & Bin, 2007), jitter(Xi et al., 2007), etc. have been discovered useful in emotion recognition. However, it is insufficient to classify emotions correctly with only single type of audio features due to similarities may in certain emotions. Another aspect of the related research is based on the classification techniques. Few efforts have been reported that different types of classifiers such as Support Vector Machine (SVM) (Hu, Xu, & Wu, 2007; Morrison, Wang, De Silva, & Xu, 2005), neural network (Bhatti, Yongjin, & Ling, 2004; Bulut, Lee, & Narayanan, 2008; Khanchandani & Hussain, 2009; Nicholson, Takahashi, & Nakatsu, 2000; Petrushin, 1999) and Hidden Markov Model (HMM) (Bhaykar, Yadav, & Rao, 2013; Kammoun & Ellouze, 2006; Tin Lay, Say Wei, & De Silva, 2003; Yi-Lin & Gang, 2005; Zeng, Tu, Pianfetti, & Huang, 2008) are integrated in their systems. However, accuracy of classification is rather low especially when more than two number of emotion enclosed in their systems.

Recent trends in research of audio emotion recognition emphasized the use of combination of different features to achieve improvement in the recognition performance. System and prosodic features represent mostly mutually exclusive information of the speech signal. Therefore, these features are complementary in nature to each other. Combination of complementary features is expected to improve the intended performance of the system. For instance, researcher Yeh, Pao, Lin, Tsai, and Chen (2011) developed a system to recognize 5 emotions recently using up to 128 audio features. Their design used their exclusive segmentation method and feature selections method to recognize emotions every significant portion of the input signal. However, their system is not language-independent. Only Mandarin utterances were considered in their system. Another recent effort that utilizes combination of complementary features is reported by Lee, Mower, Busso, Lee, and Narayanan (2011). They used two databases, IEMOCAP (Busso et al., 2009) and AIBO to build a model consists of multiple layers of binary classifications. There are 384 audio features

* Corresponding author. Tel.: +60 125064977.
E-mail addresses: ocshing@gmail.com, 11057221@imail.sunway.edu.my (C.S. Ooi).

http://dx.doi.org/10.1016/j.eswa.2014.03.026
0957-4174/© 2014 Elsevier Ltd. All rights reserved.

(inclusive several statistical coefficients) extracted in their system such as zero crossing rate, root-mean-square energy, voice quality, pitch, MFCC. SVM classifier was used in each layer to distinguish two different emotions. This approach is language-independent and able to boost the accuracy due to binary classification. However, less than five emotions can be recognized. Wu and Liang (2011) also designed an architecture to extract various prosodic features such as pitch, duration, intensity, formants, and MFCC on affective speech based on semantic labels, and classified using Gaussian Mixture Models (GMM), SVM and neural network. Despite good recognition rate obtained using multi-layer classification scheme in their research, only 4 emotions (neutral, happy, angry and sad) were considered and own non-standard databases were used in performance evaluation.

In this paper, a new architecture of audio emotion recognition is presented. Different prosodic and spectral features are analyzed and useful features are identified to assist the design of this architecture. The universal six emotions such as Happy, Angry, Sad, Disgust, Surprise and Fear are considered in this paper. The proposed architecture has two main paths. The first path has an Audio Features Analyzer to extract different audio features and an audio feature-level fusion module. The Audio Features Analyzer is designed based on intensive research analyses on prosodic audio features. Prosodic audio features such as pitch, log-energy, zero-crossing rate (ZCR) and Teager Energy Operator (TEO) are found to be useful. On the other hand, the second path is designed after useful spectral audio features are identified. This path consists of MFCC features extraction followed by three parallel sub-paths for three sets of emotion groups. They are Emotion Group 1 (Angry and Happy), Emotion Group 2 (Sad and Disgust) and Emotion Group 3 (Surprise and Fear). An audio decision-level fusion is also proposed to fuse the information from both audio paths 1 and 2. A weight assignment mechanism is also designed. A decision making mechanism is also included in the fusion module to decide the final emotion.

This paper is organized as follow: a brief review of speech features, classification techniques, and databases are given in Section 2. Section 3 provides the details about the proposed architecture of audio emotion recognition. Experimental to evaluate the performance of the proposed system and results are presented in Section 4. Finally, the conclusion is presented in Section 5.

2. Brief review

This section provides a brief review on some important speech features and processing techniques, classification techniques and widely used databases for speech emotion recognition.

2.1. Speech features

Different speech features represent different speech information, e.g. emotion, speaker, in highly overlapped manner. These have motivated intensive research of audio emotion recognition in discovering the significant manner of the speech features on specific emotions. Speech features can be classified into 3 groups: vocal tract system, prosodic, and excitation source features.

Vocal Tract System features usually can be extracted from a short segment of speech signals. This kind of features represents the distribution of energy of a range of speech frequency. There were also some research works based on various vocal tract features and their combination. For instance, another vocal tract feature called Log Frequency Power Coefficients (LFPC) has been used by Tsang-Long, Yu-Te, Jun-Heng, and Pei-Jia (2006) along with Perceptual Linear Prediction (PLP), MFCC, and LPCC to recognize emotions such as Angry, Happy, Sad, Bored and Neutral. Highest recognition rate obtained from their experiments is 84.2% on their own Mandarin database. Linear prediction cepstral coefficients (LPCC) and MFCC which are two popular spectral features were also used by Nwe, Foo, and De Silva (2003a, 2003b) to classify the universal six emotions. Using their own database called Burmese-Mandarin corpus, LPCC and MFCC, respectively provided 56.1% and 59% of average classification rate in their experiment. In one of the recent literatures, Krishna Kishore and Krishna Satish (2013) used Sub-band based Cepstral Parameter (SBC) and MFCC to recognize six emotions (i.e. Angry, Fear, Happy, Sad, Disgust and Neutral) on SAVEE database. Their best achieved result is 79% of recognition rate. Another recent effort from Bhaykar et al. (2013) experimented the performance of MFCC feature alone on speaker dependent and speaker independent situation. With IITKGP-SESC (Koolagudi, Maity, Kumar, Chakrabarti, & Rao, 2009) and IITKGP-SEHSC (Koolagudi, Reddy, Yadav, & Rao, 2011) databases, Angry, Disgust, Fear, Happy, Neutral, Sarcastic, and Surprise are the seven emotions used in their effort. The reported results showed that although speaker dependent case could score 89.20%, speaker independent case only obtained 48.18% of recognition rate.

Among the prosodic features, pitch (or fundamental frequency) information is the most widely used for determining emotions (Busso et al., 2009; Devillers et al., 2010). It can well discriminate emotions compared to other features. Besides pitch, it was also reported (Kammoun & Ellouze, 2006) that log energy is also one of the most considered parameters of to evaluate speaking styles and emotions. Experiments using log-energy feature in Kammoun and Ellouze (2006) was reported that Angry emotion can be distinguished from Fast, Lombard, Question, Slow and Soft emotions using SUSAS database (Hansen & Bou-Ghazale, 1997). Another prosodic feature, zero crossing rate (ZCR) of speech signal was also a good parameter that related to emotions in the previous research (Chien Hung et al., 2010; Chih-Chang et al., 2010), stated that angry emotion has the higher mean value than happy emotion due to the higher frequency of vibration present in speech signals. Formants also could represent emotions based on the previous research, especially on the first and second position (Goudbeek, Goldman, & Scherer, 2009). For instance, it was reported by Pribil and Pribilova (2012) that Angry has highest value in formant frequency compared to Joy and Sad, while Sad has the lowest value. In literature, very few attempts (Cummings & Clements, 1995; Ling, Hu, & Wang, 2005) have been made to explore the excitation source information for developing any of the speech systems. Thus excitation source features are not reviewed here and they are not considered in our research.

2.2. Classification

Different classification methods have been developed for speech-related application such as speech recognition, emotion classification, speaker verification, etc. The classification methods used in audio emotion recognition typically can be divided into linear and nonlinear classifications. Linear classification performs the classification by making a decision based on weighted linear combination of the object characteristics, while non-linear classification is based on non-linear weighted combination of object characteristics. Non-linear classifiers are more widely used and effective in classifying the overlapped emotional characteristics of different emotions.

One of the most popularly classification methods for audio emotion recognition is HMM (Bhaykar et al., 2013; Kammoun & Ellouze, 2006; Tin Lay et al., 2003; Yi-Lin & Gang, 2005; Zeng et al., 2008). It is based on probability algorithm to model sequential data. Neural Network has also been widely applied in audio emotion recognition. It can be divided into 3 categories which are Recurrent Neural Network (RNN) (Wei & Guanglai, 2009), Multi-Layer Perception (MLP) Neural Network (Lu & Wei, 2004), and RBF Neural Network (Chen, Cowan, & Grant, 1991). Examples of existing research works which utilize neural network in emotion recognition field can be found in (Bhatti et al., 2004; Bulut et al., 2008; Khanchandani & Hussain, 2009; Nicholson et al., 2000; Petrushin, 1999). Support Vector Machines (SVM) classifiers are also utilized extensively in many studies that related to audio emotion recognition which can be found in Schuller (2011) and Schuller et al. (2010). Table 1 shows the comparison of these three classification methods.

2.3. Databases

The performance and robustness of the recognition systems will be easily affected it is not well-trained with suitable database (Ayadi, Kamel, & Karray, 2011). Therefore, it is essential to have sufficient and suitable speeches in the database to train the emotion recognition system and subsequently evaluate the performance of the system. Table 1 lists some databases which are available to public. Standard databases such as Emo-DB, eNTERFACE'05, and RML emotion database are frequently used by researchers and will be used in our research in later sections. Their details are provided as follow (see Table 2).

Berlin emotional database (Emo-DB) (Burkhardt, Paeschke, Rolfes, Sendlmeier, & Weiss, 2005) is a collection of actor based simulated audio emotion database in German language. There are 5 male and 5 female actors contributed in preparing the database. The emotions recorded in the database are anger, boredom, disgust, fear, happiness, neutral and sadness. Ten linguistically neutral sentences are chosen for database construction. Eight hundred and forty (840) utterances of Emo-DB are used in this work. Eight speakers' speech data is used for training the models and remaining 2 speakers' speech data is used for validating the trained models with the approach of leave-one-speaker-out cross-validation, during training and testing.

The eNTERFACE'05 (Martin, Kotsia, Macq, & Pitas, 2006) is an audio-visual emotion database which contains 1170 utterances. The utterances were produced by 42 subjects (34 male and 8 female) from 14 different nations. The emotions included in this English spoken database are Happy, Angry, Disgust, Sad, Surprise and Fear. The emotions are the reactions of the subjects after listening to six different short stories. Each subject was required to read five phrases based on their reactions to each situation.

RML emotion database (Wang & Guan, 2008) is an audio-visual emotion database which consists of 720 videos from 8 subjects. Videos of these subjects were recorded in English, Mandarin, Urdu, Punjabi, Persian, and Italian languages. The emotions covered in this database are the six universal emotions as well. Each subject is provided a list of emotional sentences and was instructed to act naturally with the respective emotions.

3. Audio features

In this section, analyses on audio features and results are presented. The new architecture of audio emotion recognizer in section IV will be designed based on the research findings from this section. In our analyses, the database consists of 180 samples (or 30 samples each emotions) of utterances randomly chosen from eNTERFACE'05 (Martin et al., 2006) and Emo-DB (Burkhardt et al., 2005) database. Utterances from these two databases are mixed to increase the robustness or relevance of the analysis results.

3.1. Extraction of pitch features

In this section, the time-domain or autocorrelation based pitch extraction method (Xufang, O''Shaughnessy, & Nguyen, 2007), is used. This pitch extraction method calculates the distance between the zero crossing points of the signal. Speech signal, $x(m)$ in time domain is firstly divided into n number of frames by windowing, $w(n)$ and is denoted as $s(m)$.Then the pitch can be obtained from its periodicity, R:

$$R(k) = \sum_{m=0}^{L-k-1} s(m)s(m+k) \tag{1}$$

where L denotes the window length, and k refers to the representation of pitch period of a peak. Mean values for pitch features shown in Fig. 1(a). It reveals great discrimination between the six emotions, especially Angry and Sad which, respectively obtain the highest and lowest amplitudes. To further observe the behavior of the pitch feature, Hamming window is applied to the averaged pitch signals for each emotion. The setting of Hamming window (Wang & Guan, 2008) is using the window of 512 points with 256 points or 50% of overlapping. Results are plotted in Fig. 1(b). Angry emotion still gives the highest amplitude, while Sad emotion gives the lowest. Based on autocorrelation pitch values in Fig. 1(a) and averaged pitch values in Fig. 1(b) of 6 emotions, it can be concluded that Angry and Happy emotions have higher pitch, while Sad and Fear emotions have lower pitch.

3.2. Extraction of log-energy features

Energy estimation indicates the amplitude of the signal at an interval. Energy-based features can be used for determining the emotional states of speeches. As reviewed in Section 2, log-energy features are commonly used in audio based emotion recognition. Log-energy indicates the total squared amplitude in a segment of speech. This feature is simply the amount of normalized power or volume in the signal. According to (Nwe et al., 2003a, 2003b) the simple calculation of logarithm of energy or log-energy can be formulated as:

Table 1
Comparison of HMM, SVM and Neural Network classifiers.

Classifier	Advantages	Usage examples
HMM	Effective statistical grounding (Yi-Lin & Gang, 2005)	Guan and Xie (2013) treated the prosodic features and spectral features separately on 6 universal emotions and combined all features using information fusion. HMM classifier was subsequently used for decision making.
SVM	Simple implementation (Yao et al., 2005), great data-dependent generalization bounds (Steinwart & Christmann, 2008)	Schuller et al. (2010) accumulated a set of 1,406 generated audio features and post-processed them using low-pass filter. SVM classifier was subsequently used to classify them.
Neural network	Simple implementation (Yao et al., 2005), best handled on small dataset (Schuller, Batliner, Steidl, & Seppi, 2011)	Yongjin and Ling (2004) generated 55 audio features and performed feature selection to further choosing the important number of features. Three-layer feed forward neural network based on back propagation was used to classify the six universal emotions.

Table 2
List of database for audio emotion recognition.

Database	Availability	Size	Elicitation method	Emotion types
MPEG-4 (Schuller et al., 2005)	No	2440 Utterances, 35 subjects	Movies (English)	6 Universal emotions + Neutral
Berlin emotional database (Emo-DB) (Burkhardt et al., 2005)	Yes	800 Utterances, 10 subjects	Professional Actors (German)	Anger, joy, sadness, disgust, boredom, neutral
Emotional Prosody Speech and Transcript (LDC) (Liberman, Davis, Grossman, Martey, & Bell, 2002)	Yes, Commercial	More than 2500 utterances, 7 subjects	Professional Actors (English)	Neutral, Disgust, Panic, Anxiety, Hot Anger, Cold Anger, Despair, Sadness Elation, Happy, Interest, Boredom, Shame, Pride, and Contempt
eNTERFACE'05 database (Martin et al., 2006)	Yes	1170 Videos, 42 subjects	Nonprofessional actors	6 Universal emotions
RML database (Wang & Guan, 2008)	Yes	720 Videos, 8 subjects	Nonprofessional actors (English, Mandarin, Urdu, Punjabi, Persian, and Italian)	6 Universal emotions

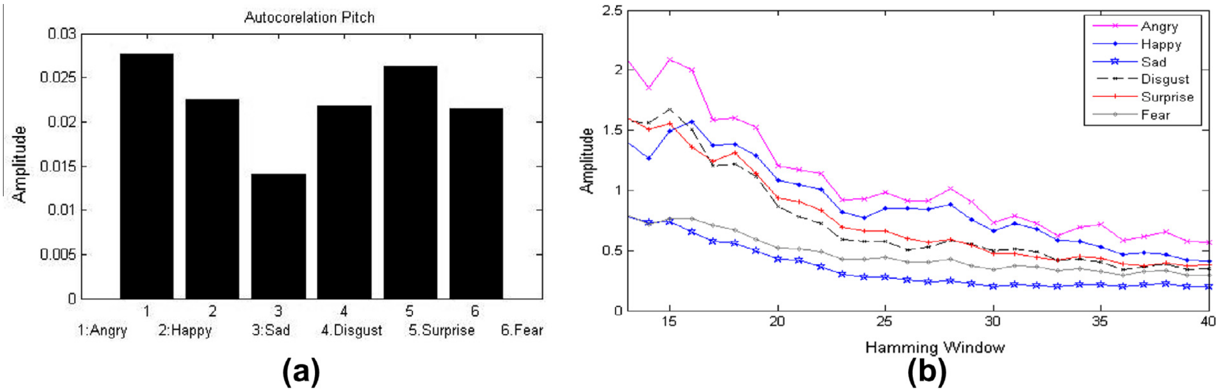

Fig. 1. (a) Mean of Pitch Amplitude for 6 emotions, and (b) Hamming Windowed amplitude of Pitch feature for 6 emotions.

$$E = \log_{10}\left(\sum_{i=1}^{N} x^2\right) \tag{2}$$

where N denotes the number of frames and x denotes the sample of speech. The bar chart in Fig. 2(a) shows Angry, Happy and Sad have lower amplitude of energy compared to other 3 emotions. Surprise gives the highest amplitude (in dB) of Log Energy. Log-energy feature discriminates Surprise, Disgust, and Fear emotions from Angry, Happy and Sad emotions.

3.3. Extraction of zero rate crossing (ZCR) features

ZCR calculates the weighted average of the number of times the speech signal changes sign within a particular time window. In another words, ZCR counts one as the signal changes from positive to negative or otherwise. The equation of ZCR can be shown as follow:

$$\mathrm{sgn}\{x\} = \begin{cases} 1 & \text{if} \quad x(n) \geqslant 0 \\ -1 & \text{if} \quad x(n) < 0 \end{cases} \tag{3}$$

$$w\{n\} = \begin{cases} \frac{1}{2N} & \text{if} \quad 0 \leqslant n \leqslant N-1 \\ -1 & \text{if} \quad \text{otherwise} \end{cases} \tag{4}$$

and N refers to the total number of samples and n refers to the current sample. Fig. 2(b) shows the zero-crossing rate of each emotion. The result reveals the zero-crossing rate feature can discriminate Angry, Happy, and Fear emotions from Sad, Disgust, and Surprise emotions.

3.4. Extraction of teager energy operator (TEO) features

The Teager Energy Operator (TEO) is used for computing the energy of the signal in a nonlinear manner. The nonlinear operator measures the energy in the system that generated the signal rather than the energy of the signal itself. It was stated that the nonlinear component changes appreciably between different emotional speeches (Gao, Chen, & Su, 2007). A simple nonlinear energy-tracking operator for speech signal is given by the following equation.

$$\Psi[s(n)] = s^2(n) - s(n+1)s(n-1) \tag{5}$$

where $\Psi[s(n)]$ is the Teager Energy operator (TEO), and $s(n)$ is the sampled speech signal.

The analysis on TEO feature is to study the possibility to use TEO feature to discriminate one or two emotions from other emotions. As shown in Fig. 2(c), the mean value of TEO features for Disgust emotion is clearly differentiated from other emotions especially Sad, Surprise and Fear. This analysis has confirmed that TEO can be used to differentiate Disgust emotion from other 5 emotions.

3.5. Extraction of mel-frequency cepstral coefficients (MFCC)

MFCC is the cepstral coefficients derived from a mel-scale frequency filter-bank (Jothilakshmi, Ramalingam, & Palanivel, 2009). In computations of MFCC, speech signal is firstly divided into multiple frames of equal duration. The frames overlap each other to preserve the continuity of the speech signal. Each frame is subsequently multiplied with a Hamming Window so that the continuity of the left and right side of the frame can be increased or

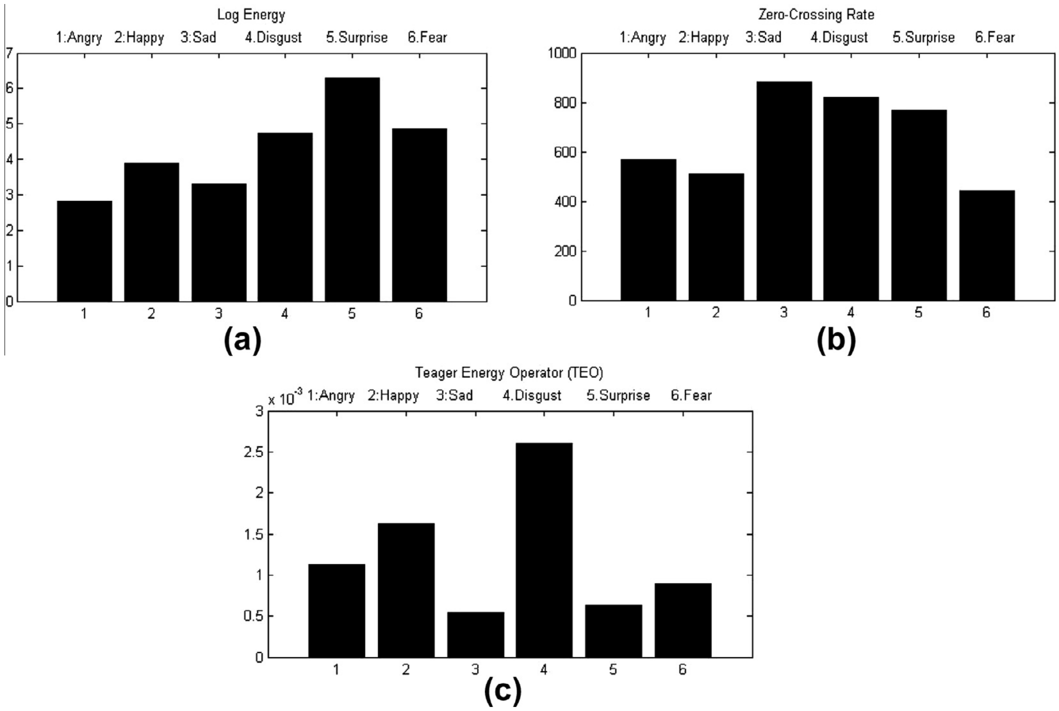

Fig. 2. Mean values of (a) log-energy, (b) Zero-Crossing Rate, and (c) TEO feature.

maintained. The Hamming Window, $w(n)$ used in MFCC is defined with the equation:

$$w(n) = 0.54 - 0.46\cos\left(2\pi\frac{n}{N}\right), \quad 0 \leqslant n \leqslant N \tag{6}$$

$$Y(n) = X(n) \times w(n) \tag{7}$$

where $Y(n)$ is the output signal, $X(n)$ is the input signal, N is the number of samples in each frame. After multiplying with the Hamming Window, it is necessary to convert to the frequency domain because the signal features are difficult to observe in the time domain. Thus the frames are then converted into the frequency domain by using Fast Fourier Transform (FFT) (Merz, 1983). Triangular filters are subsequently used to compute a weighted sum of filter spectral components so that the output of process approximates to a Mel scale (Jothilakshmi et al., 2009). Then, each filter output is the sum of its filtered spectral components. The result of the scaling process is used to compute the Mel spectrum using the Eq. (8):

$$Mel(n) = 2595 \times \log_{10}\left[1 + \left(\frac{n}{700}\right)\right] \tag{8}$$

The computed results can be observed in vector form as MFCC coefficients after Discrete Cosine Transform (DCT) (Ahmed, Natarajan, & Rao, 1974). The DCT results in the most signal energy being compacted in the first 15 coefficients.

In our analysis of MFCC features, PCA and LDA techniques are used to project MFCC features of all 6 emotions into different subspaces to reduce the dimensionalities and discriminate them. Fig. 3(a) shows the result of this projection. PCA is a data reduction method based on the weighted sum of the principal components. LDA is used to further discriminate the emotion features to their respective classes. Our next analysis hopes to verify MFCC features can be further differentiated after projecting into the PCA + LDA subspace, if there is lesser number of emotion classes. Thus, the number of emotion classes will be decreased from 6 to 2. Emo-DB and eNTERFACE'05 databases are used in this analysis. 80 samples for each emotion are used to show the pattern of emotion scattering in PCA + LDA subspace. Results are presented in the following figures: Fig. 3(a) (6 emotion classes), Fig. 3(b) (5 emotion classes), Fig. 3(c) (4 emotion classes), Fig. 3(d) (3 emotion classes), and Fig. 4 (2 emotion classes). Analysis results shown in Fig. 3 reveals that MFCC features are easier to be discriminated in PCA + LDA subspace when less number of emotions is considered. Thus, to increase the performance of emotion classification, a parallel structure with 3 sub-paths with two emotion classes each will be considered in our design in Section 4.

Based on the findings from previous analysis of audio features in Section 3, 6 emotions are grouped into 3 groups: (i) Group 1 – Angry & Happy, (ii) Group 2 – Sad & Disgust, and (iii) Group 3 - Surprise & Fear. Angry and Happy emotions in Group 1 are not easily discriminated using pitch, log-energy, zero-crossing rate and TEO features. Sad and Disgust emotions in Group 2 also give the same problem in zero-crossing rate and pitch features. Surprise and Fear in Group 3 can be clearly differentiated using log-energy and TEO features.

As shown in Fig. 4 for Group 1, Fig. 5 for Group 2 and Fig. 6 for Group 3, 2 emotions in a group are apparently more separated. Right strategy to utilize the above prosodic and spectral features in this section can lead to a robust design of audio emotion recognizer in Section 4.

4. The proposed audio emotion recognizer

This section presents a new architecture of audio emotion recognizer. This architecture is illustrated as shown in Fig. 7. It is divided into two paths after input audio signal is pre-processed by VAD technique. The main purpose of using VAD technique is to eliminate the background noise and segment out the non-speech portions of the audio signal. VAD technique used short-time energy and short-time zero-crossing rate (Yang, Tan, Ding, Zhang, & Gong, 2010). Firstly, speech signal, $x(m)$ in time domain is divided into n number of frames. Short-time energy (STE) detection is performed to determine energy within each frame or segmented voice signal. Subsequently, Short-time zero-crossing

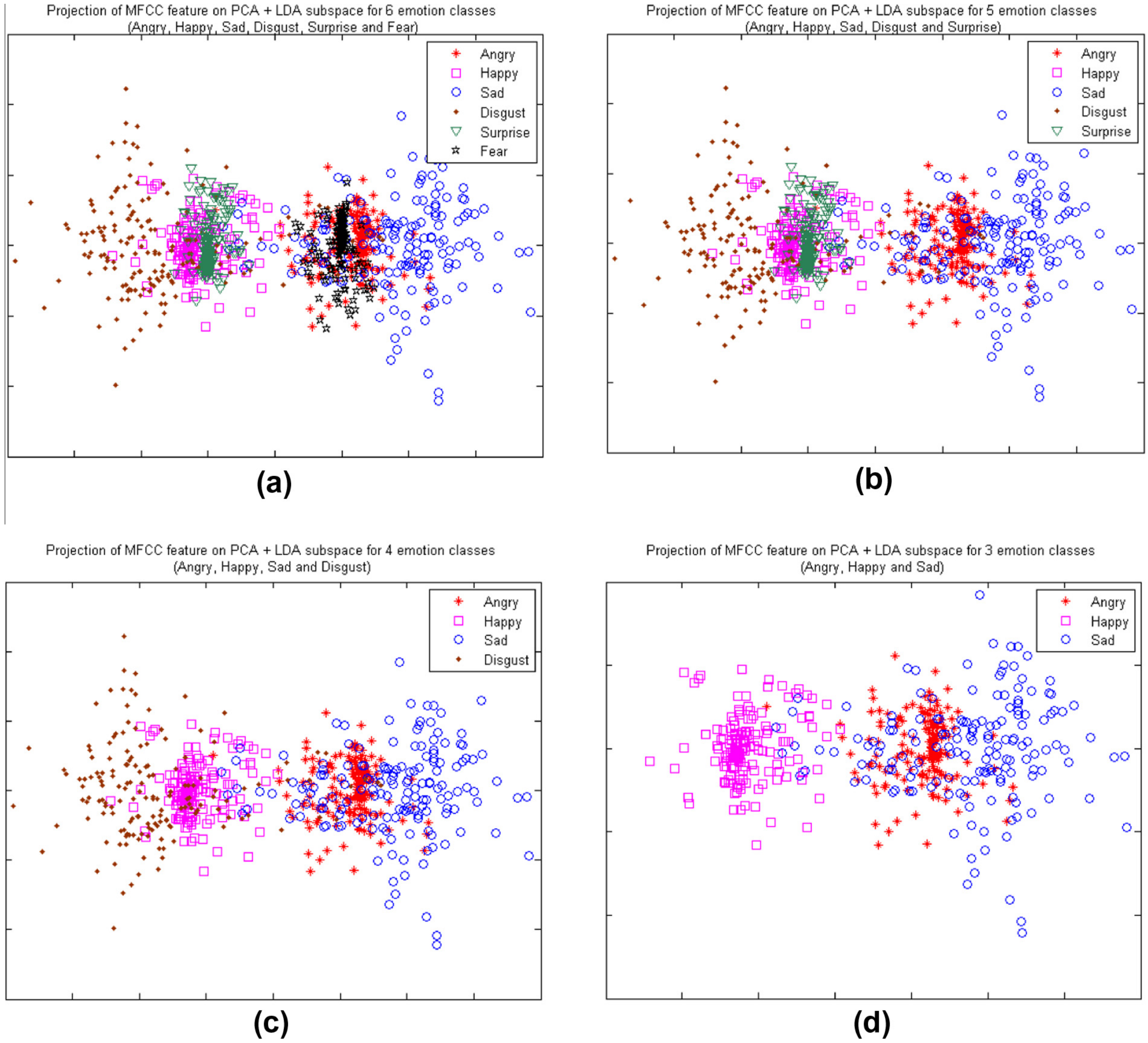

Fig. 3. Projection of MFCC feature on PCA + LDA subspace for (a) 6 emotions, (b) 5 emotions, (c) 4 emotions, and (d) 3 emotions.

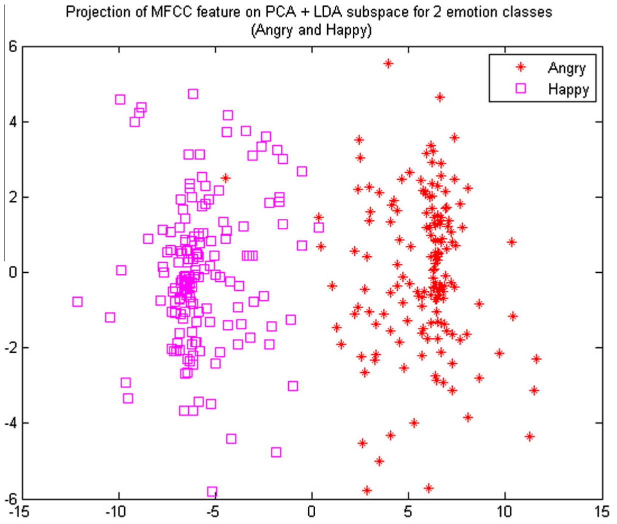

Fig. 4. Projection of MFCC feature on PCA + LDA subspace for Group 1 emotions (Angry and Happy).

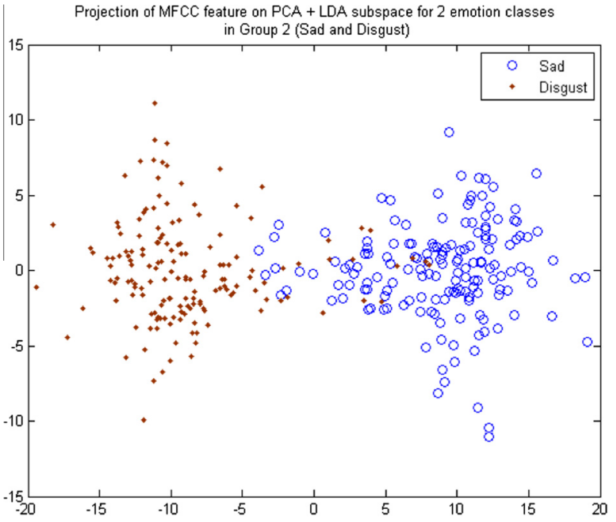

Fig. 5. Projection of MFCC feature on PCA + LDA subspace for Group 2 emotions (Sad and Disgust).

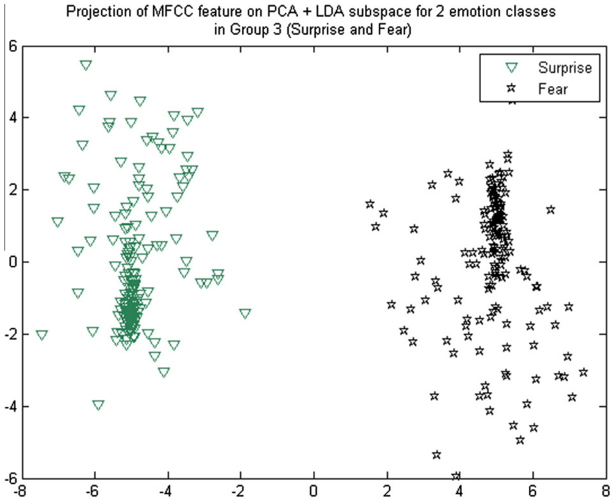

Fig. 6. Projection of MFCC feature on PCA + LDA subspace for Group 3 emotions (Surprise and Fear).

rate (STZCR) which is calculated from the weighted average of the number of times the speech signal changes sign within a particular time window. The computed energy and zero crossing rates are compared to determine the existence of speech in the signal. The unwanted signal frames (silence or unvoiced) will be then segmented out.

For the Path 1, the output from the VAD is fed into an audio feature analyzer. This path consists of two main parts: (i) Audio Feature Analyzer (ii) Audio Feature-level Fusion. The function of this audio feature analyzer is to extract and analyze the audio features such as pitch, log-energy, zero ZCR, and TEO features. Those audio features are then passed to a module called audio feature-level fusion. In this module, there are sets of rules to decide the right emotion. These sets of rules are designed based on the analyses from Section 3. For Path 2, it can be noted that the pre-processed audio signal is passed to next stage to extract its MFCC features. Research analysis in Section 3 has shown that emotions are well discriminated when smaller number of emotions is involved. The six emotions are grouped into 3: (i) Emotion Group 1 – Angry & Happy, (ii) Emotion Group 2 – Sad & Disgust, (iii) Emotion Group 3 – Surprise & Fear. A decision-level fusion module is designed to fuse outcomes from both paths.

4.1. Path 1 – audio feature analyzer

Fig. 8 illustrates the block diagram of Path 1. This path consists of the Audio Feature Analyzer that input the pre-processed audio data and output a decision based on the analysis of the audio features. The Audio Feature Analyzer is designed based on findings of our analysis on audio features Section 3. The purpose of the Audio Feature Analyzer is to further assist the decision making in the final decision-level fusion module after paths 1 and 2 are merged.

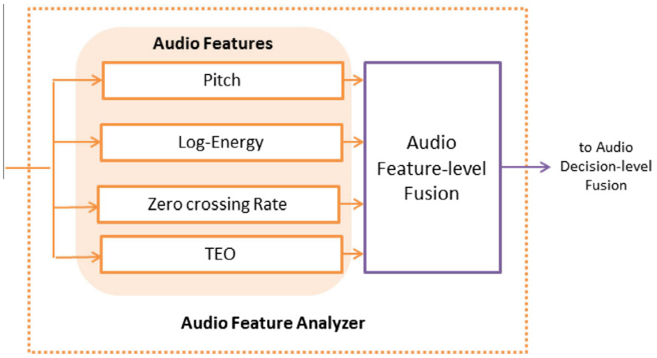

Fig. 8. Block diagram of the proposed audio feature analyzer and feature-level fusion.

The analyzer firstly extracts those important audio features such as pitch, log-energy, zero crossing rates and TEO features, as identified in Section 3. After obtaining values for those features, analysis will be carried out. Audio feature-level fusion module will process and decide with the high possibility of their corresponding emotion group using sets of rules. These sets of rules are designed based on the research analyses in Section 3. As concluded in Section 3, each of these 4 audio features gives good discriminations between certain emotion(s). TEO feature can be used to discriminate disgust emotions from the other five emotions. On the other hand, ZCR feature can be used to distinguish Sad, Disgust and Surprise from the remaining three emotions. Normally Angry and Surprise emotions have the higher pitch values while Sad and Fear emotions have the lower values. Log-energy feature can be used to distinguish Surprise emotion because this emotion gives highest value amongst all 6 emotions.

4.2. Path 1 – audio features-level fusion

After all computations of pitch, log-energy, zero-crossing rate and TEO features, they are fed into the feature-level fusion module. This module mainly used to fuse the results of the computed values of 4 features based on their respective threshold values. Its fusion scheme has sets of rules to process the information and decide which emotion group in Path 2 should be emphasized. In

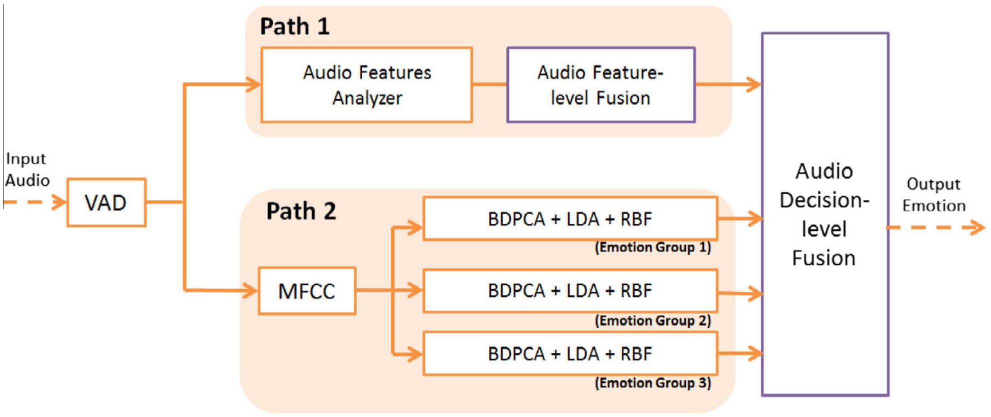

Fig. 7. Architecture of proposed audio emotion recognizer in Audio Visual Emotion Recognition.

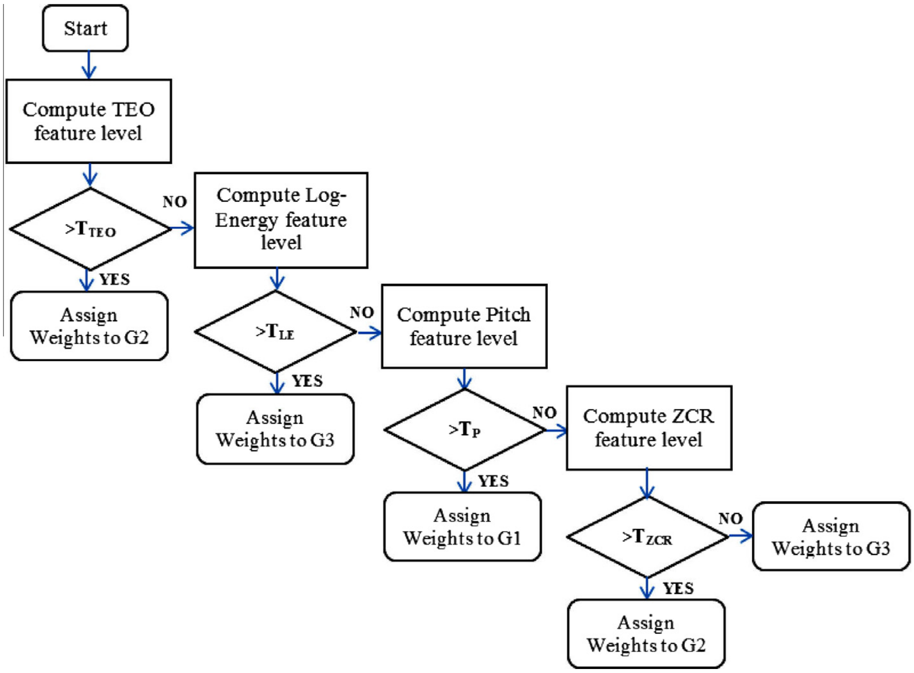

Fig. 9. Flowchart of Feature-level Fusion Module.

another words, given three outputs from three emotion groups in path 2, this path 1 will assist by informing which output from path 2 is more important and has strong influence on final decision.

Fig. 9 presents a flowchart of audio features-level fusion module. The operation of fusion module is based on threshold values that set for the corresponding features. These values are pre-set based on the analyses presented in Section III. As shown in the Fig. 2, TEO feature has highest value for Disgust emotion. Disgust, Surprise, and Fear emotions can be clearly discriminated from Happy, Angry and Sad emotions using log-energy. As shown in Fig. 1(a), Sad can be clearly differentiated from Angry and Happy emotions based on pitch feature. ZCR feature can be used as well to discriminate Fear from Sad emotion.

The flow of the fusion starts by comparing the value of actual or computed TEO value with the threshold level of TEO feature, T_{TEO}. If the computed value is higher than the pre-set level of TEO feature, the emotion Disgust has the highest probability and heavier weight is assigned to the output of the Emotion Group 2 (Disgust/Sad) from path 2 in the module of Audio Decision Level Fusion. If the value of TEO is low, then the actual or computed log-energy value is then compared to log-energy's threshold level. Heavier weight is assigned to the output of the Emotion Group 1 (Surprise/Fear) if the computed value exceeds the threshold level. If the computed log-energy value is below the threshold, T_{LE}, then the computed value of pitch feature is compared with the threshold pitch level, T_P. Analysis in Section III has confirmed Sad and Fear emotions normally give lower pitch values compared with other emotions. Thus, the high pitch value has higher possibility to discriminate Angry and Happy emotions from the remaining emotions, while the low pitch value has higher possibility to discriminate Sad and Fear emotions from others. As mentioned in the Section III, Angry and Happy emotions are grouped together as Emotion Group 1 in Path 2. If the computed pitch value is below the threshold, then the zero-crossing rate will be examined. Heavy weight will be assigned to the output of Emotion Group 2 from path 2 if the obtained zero-crossing rate is below threshold value, T_{ZCR}, while Heavy weight will be assigned to the output of Emotion Group 3 if higher zero-crossing rate is obtained.

4.3. Path 2 – MFCC, BDPCA + LDA feature extraction techniques and RBF neural network

MFCC is one of most influential audio features for audio emotion recognition. It extracts the significant emotion components from the speech audio data and represents them according to a Mel-Frequency scale which is identical to the behavior of the human ear. The computations or steps to obtain this feature can be referred back to Section 3. Reason to group 6 emotions into 3 has also been presented in Section 3.

In Path 2, MFCC features are extracted first. These features are passed to three parallel paths as shown in Fig. 10. The two data reduction methods, BDPCA and LDA are used in all three sub-paths in Path 2. In this subsection, the algorithm of both BDPCA and LDA are presented. Besides, each 2-class classification is performed using RBF neural classifier. BDPCA aims to improve the performance by extending the scatter matrix calculations of a 2D MFCC matrix in 2DPCA approach to two directions, row and column separately. Given a training set $[X = X_1, \ldots, X_K]$, K is the number of the training samples and the size of each MFCC matrix is $p \times q$ (where p denotes the number of coefficients and q denotes the number of frames).

(1) By representing the ith MFCC matrix X_i as a p set of $1 \times q$ row vectors, the row total scatter (Jian, Zhang, Frangi, & Jing-Yu, 2004) can be expressed as:

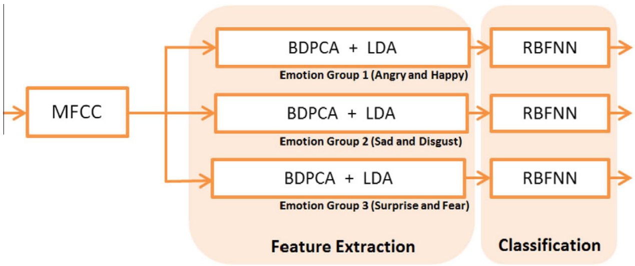
MFCC
BDPCA + LDA
Emotion Group 1 (Angry and Happy)
BDPCA + LDA
Emotion Group 2 (Sad and Disgust)
BDPCA + LDA
Emotion Group 3 (Surprise and Fear)
RBFNN
RBFNN
RBFNN
Feature Extraction
Classification

Fig. 10. Block diagram of Path 2 in proposed audio emotion recognizer.

$$S_t^{row} = \frac{1}{Kp}\sum_{i=1}^{K}\sum_{j=1}^{p}\left(x_i^j - \bar{x}^j\right)^T\left(x_i^j - \bar{x}^j\right) = \frac{1}{Kp}\sum_{i=1}^{K}\left(X_i - \overline{X}\right)^T\left(X_i - \overline{X}\right) \tag{9}$$

where x_i^j denote j-th row of sample X_i and $\bar{x}^j$ denote the jth row of mean matrix $\overline{X}$, respectively.

(2) Similarly as row representation in Step 1, i^{th} MFCC matrix X_i that represents as a q set of $p \times 1$ column vectors is used for calculating its column total scatter:

$$S_t^{col} = \frac{1}{Kq}\sum_{i=1}^{K}\sum_{k=1}^{q}\left(x_i^k - \bar{x}^k\right)\left(x_i^k - \bar{x}^k\right)^T = \frac{1}{Kq}\sum_{i=1}^{K}\left(X_i - \overline{X}\right)\left(X_i - \overline{X}\right)^T \tag{10}$$

where x_i^k denote kth column of sample X_i and $\bar{x}^k$ denote the k-th column of mean matrix $\overline{X}$.

(3) The row eigenvectors and column eigenvectors are subsequently calculated based on their respective total scatter matrix. The obtained row and column eigenvectors are then sorted corresponding to their d_{row} and d_{col} largest eigenvalues of S_{row} and S_{col}, respectively. This step is to construct the bi-directional projection matrices. Therefore, the resulting row projection matrix, W_r and column projection matrix, W_c are obtained in the dimensions of p by d_{row} and q by d_{col}, respectively.

(4) The projection can be applied to input MFCC matrix X to extract its significant components:

$$Y = W_c^T X W_r \tag{11}$$

The obtained BDPCA-reduced MFCC feature is then rearranged to K numbers of row vectors to perform the discriminant analysis with LDA. LDA is normally conducted by computing the maximum between-class distance and minimum within-class distance is maximized (Sharkas & Elenien, 2008). The steps of obtaining LDA features:

(1) By calculating the mean, m_j of the MFCC for each class, the within-class scatter matrix, S_W and between-class scatter matrix, S_B can be calculated according to (12) and (13), respectively (Sharkas & Elenien, 2008).

$$S_W = \sum_{j=1}^{C}\sum_{i=1}^{N}\left(y_i^j - m_j\right)\left(y_i^j - m_j\right)^T \tag{12}$$

$$S_B = \sum_{j=1}^{C}(m_j - m)(m_j - m)^T \tag{13}$$

where C is the number of class, N is the number of training MFCC for each class, and m is the mean for overall training input from BDPCA, Y.

(2) The largest eigenvalues, W and its eigenvector, λ are then computed from both S_B and S_W matrix, as define in (14) (Zuo, Zhang, Yang, & Wang, 2006):

$$S_B W = \lambda S_W W \tag{14}$$

The Best Projection Matrix, W_{BPM} is then computed to find the best separated space:

$$W_{BPM} = \arg\max \frac{|W^T S_B W|}{|W^T S_w W|} \tag{15}$$

where $W^T S_B W$ and $W^T S_W W$ is the scatter of transformed feature vectors of S_B and S_W, respectively.

(3) The feature vector from PCA or the MFCC matrix can then be projected into the LDA space, and the discriminated trained data, P can be obtained (Sharkas & Elenien, 2008):

$$P = W_{BPM}^T Y. \tag{16}$$

RBF Neural Network (Orr, 1999) acts as the classifier for each sub-path. The construction of its model is based on:

(1) At the input of each neuron in input layer, activation of hidden units is calculated based on the distance, d_i between the input vector, p_i and center of the neuron, μ_j.

$$d_i = \|p_i - \mu_j\| \tag{17}$$

where i denotes the number of input samples and j denotes the number of hidden units.

(2) When the distance obtained in (17) is smaller than the spread width or width of the receptive field, σ_j in the input space, the appreciable value can be obtained by radial basis function ϕ_j:

$$\phi_j(\mathrm{x}) = \exp\left(-\frac{\|p - \mu_i\|}{2\sigma_i^2}\right) \tag{18}$$

where $p = [p_1, p_2, p_3, \ldots, p_n]$.

(3) By applying the basis function to the distance, output of the neuron can be formed in the output layer. Linear regression is then can be performed to predict the targeted outputs, $\boldsymbol{y}$ using the equation below:

$$output = \sum_{i=1}^{j} w_j \phi_j(\mathrm{x}) \tag{19}$$

where w_j are the weights between the hidden and output layers. Weight updated algorithm can be found in (Park & Sandberg, 1993).

4.4. Audio decision-level fusion

There are two different fusion modules in this architecture. Audio Features-Level Fusion module from Path 1 has been presented in the previous subsection. Another module, called Audio Decision-Level Fusion functions as the final decision maker after compiling information received from Paths 1 and 2.

As shown in Fig. 11, this module collects information from two paths. The Audio Feature-Level Fusion module in path 1 assists via the Weights Assignment Mechanism to assign weights as W_{Group1}, W_{Group2}, and W_{Group3} to 3 outputs (3 emotion groups) from Path 2. As discussed in previous subsection, Path 2 first extracts MFCC features. Then MFCC features are passed to the parallel structure with 3 BDPCA + LDA + RBF sub-paths. BDPCA + LDA are used for dimensionality reduction and feature discrimination, while RBF acts as an classifier in each sub-path. Each set of the emotion groups consists of two emotions. They are Group 1 (Angry & Happy), Group 2 (Sad & Disgust) and Group 3 (Surprise & Fear). Therefore, path 2 gives 3 outputs due to three sets of emotion groups or 3 sub-paths. Based on W_{Group1}, W_{Group2}, and W_{Group3}, the final decision is made by considering the output with the heaviest weight. For example, if the heaviest weight is found in W_{Group1}, then the output of $(\text{BDPCA + LDA + RBF})_{Group1}$ from Path 2 will be taken. If sub-path's RBF classifier of Group 1 in Path 2 gives '*Happy*', then the final decision of the emotion recognizer is '*Happy*'.

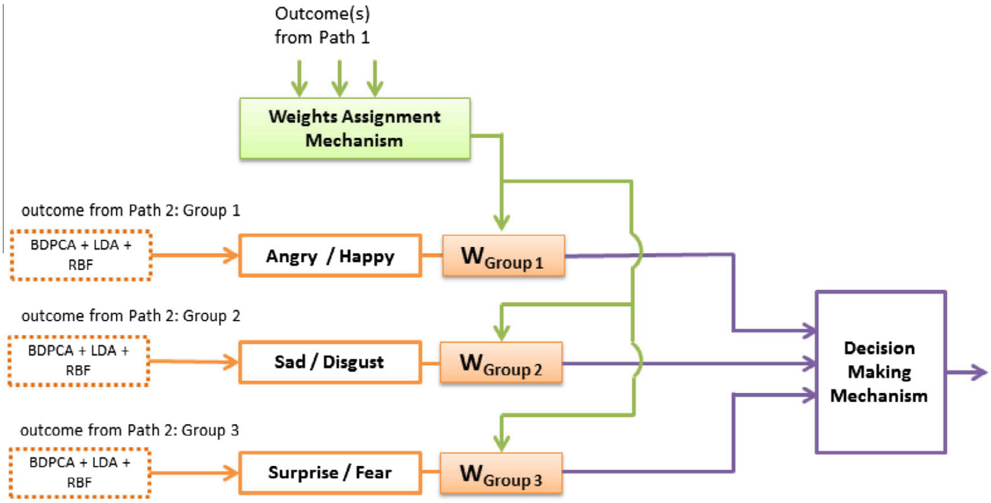

Fig. 11. Decision-level fusion module.

Table 3
Comparison of recognition results for audio emotion recognizers on eNTERFACE'05 database.

Audio emotion recognition system	Classification method	Recognition rate (%)
Audio emotion recognition system (Schuller, 2011)	Support Vector Machines (SVM)	62.80
Audio emotion recognition system (Schuller et al., 2010)	Support Vector Machines (SVM)	72.00
The proposed audio emotion recognition system	RBF Neural Network	75.89

Table 4
Comparison of recognition results for audio emotion recognizers on RML database.

Audio emotion recognition system	Classification method	Recognition rate (%)
Audio emotion recognition system with KPCA (Guan & Xie, 2013)	Hidden Markov Model (HMM)	38.00
Audio emotion recognition system with KCCA(Guan & Xie, 2013)	Hidden Markov Model (HMM)	52.00
Audio emotion recognition system (Wang et al., 2012)	Hidden Markov Model (HMM)	64.00
Proposed audio emotion recognition system	RBF Neural Network	68.57

5. Experimental results

The performance of the proposed audio emotion recognition system in Section 4 is evaluated on eNTERFACE'05 and RML databases. Comparisons with existing systems on these databases are conducted in our experiments.

5.1. eNTERFACE'05 Database

In this simulation, the performance of the proposed audio emotion recognizer is compared with that of the audio emotion recognition system or recognizers reported in (Björn Schuller, 2011; Björn Schuller et al., 2010). The eNTERFACE'05 database is chosen to evaluate their performances. All six universal emotions (i.e. Angry, Happy, Sad, Disgust, Surprise, and Fear) are considered in this simulation. There are 42 subjects selected from the database and each subject has 5 sentences for each emotion. Thus, each subject group consists of 30 samples. Leave-One-Subject-Group-Out strategy is used to evaluate performance of both systems. In another words, the simulation is conducted 42 times as each time one subject group is taken out for testing while the remaining subject groups are used for training. The simulation result is tabulated in Table 3 and the recognition rates reported in Schuller et al. (2010) and Schuller (2011) are also included in this table.

As shown in Table 3, recognizers proposed by Schuller (2011) and Schuller et al. (2010) have scored the recognition rates of 62.80% and 72%, respectively. On the other hand, the proposed audio emotion recognizer manages to obtain the recognition rate of 75.89%, which is approximately 3.89% and 13.09% higher than the rates obtained by recognizers in Schuller et al. (2010) and Schuller (2011), respectively.

5.2. RML database

In this simulation, the performance of the proposed audio emotion recognizer is evaluated using the RML database (Wang & Guan, 2008). Audio utterances in this database are extracted from video sequences for all six universal emotions. In this experiment, 400 samples are randomly selected from the database and 75% of them are used for training while the remaining 25% is used for testing. The simulation results are presented in Table 4. The performance of our proposed recognizer is also compared with that of the recognition systems in Guan and Xie (2013) and Wang, Guan, and Venetsanopoulos (2012).

It is important to note that the system in Wang et al. (2012) has used more than 100 projected dimensionality of combined features to achieve its optimum performance which is approximately 64%. Our proposed audio emotion recognition system only requires less than half of their calculations which are approximately 40 projected dimensionality of MFCC feature and 4 prosodic features in overall. Besides, as reported by Wang et al. (2012) the recognition systems require more computation due to calculations on statistic and variations of both prosodic and spectral features. When low (approximately 40) projected dimensionality of combined features are involved in audio recognition system (Guan & Xie, 2013), systems with KPCA and KCCA have only achieved the rate of 38% and 52%, respectively. The proposed audio emotion recognizer is more efficient and manages to achieve higher recognition rate.

6. Conclusion

This paper has presented a new architecture of audio emotion recognition system. Intensive research analyses on audio features have been performed. Audio features such as log-energy, pitch, TEO, ZCR, and MFCC have been identified as significant audio features and used to design the new emotion recognizer. Audio emotion recognizer which has the 2-paths structure effectively processes the significant prosodic and spectral features in parallel. Feature-level and decision-level fusion modules have also been proposed at the final stage to assist the decision making. The performance of the proposed audio visual emotion recognition system is evaluated on eNTERFACE'05 and RML databases. Comparisons with existing audio emotion recognition systems on those databases have also included. Simulation results and comparisons have revealed the good performance of the proposed audio emotion recognition system. For future works, other modalities such as facial expression can be incorporate into the design of the proposed audio emotion recognition for certain applications. This paper has emphasized on six universal emotions and it is possible to extend the design for more non-universal emotions such as depress and stress. To do so, audio prosodic features for those emotions need to be analyzed. Then, new sets of rules can be designed to take both prosodic and spectral features into the considered. Besides that, speech recognition can be integrated to the design to recognize some words or short sentences which are normally found in the speech with particular emotion.

References

Ahmed, N., Natarajan, T., & Rao, K. R. (1974). Discrete cosine transform. *IEEE Transactions on Computers, 100*(1), 90–93.

Ayadi, M. E., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition, 44*(3), 572–587.

Bhatti, M. W., Yongjin, W., & Ling, G. (2004). A neural network approach for human emotion recognition in speech. Paper presented at the Circuits and Systems, 2004, ISCAS '04, Proceedings of the 2004 International Symposium on (23–26 May).

Bhaykar, M., Yadav, J., & Rao, K. S. (2013). Speaker dependent, speaker independent and cross language emotion recognition from speech using GMM and HMM. Paper presented at the Communications (NCC), 2013 National Conference on (15–17 Feb).

Bulut, M., Lee, S., & Narayanan, S. (2008). Recognition for synthesis: automatic parameter selection for resynthesis of emotional speech from neutral speech. Paper presented at the Acoustics, Speech and Signal Processing, 2008, ICASSP 2008, IEEE International Conference on.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., & Weiss, B. (2005). A database of german emotional speech. Paper presented at the Proceedings of the Interspeech 2005, Lissabon, Portugal.

Busso, C., Lee, S., & 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.

Chen, S., Cowan, C. F. N., & Grant, P. M. (1991). Orthogonal least squares learning algorithm for radial basis function networks. *IEEE Transactions on Neural Networks, 2*(2), 302–309.

Chien Hung, C., Ping Tsung, L., & Chen, O. T. C. (2010). Classification of four affective modes in online songs and speeches. Paper presented at the 19th Annual Wireless and Optical Communications Conference (WOCC) (14–15 May).

Chih-Chang, C., Chien-Hung, C., Ping-Tsung, L., Chen, O.T. (2010). Affective understanding of online songs and speeches. Paper presented at the 53rd IEEE International Midwest Symposium on Circuits and Systems (MWSCAS) (1–4 Aug.).

Cummings, K. E., & Clements, M. A. (1995). Analysis of the glottal excitation of emotionally styled and stressed speech. *The Journal of the Acoustical Society of America, 98*, 88.

Devillers, L., Vidrascu, L., & Layachi, O. (2010). Automatic detection of emotion from vocal expression. In K. R. Scherer, T. Bänziger, & E. B. Roesch (Eds.), *A blueprint for affective computing: a sourcebook* (pp. 232–244). New York, United States: Oxford University Press.

Edwards, J., Jackson, H. J., & Pattison, P. E. (2002). Emotion recognition via facial expression and affective prosody in schizophrenia: a methodological review. *Clinical Psychology Review, 22*(6), 789–832.

Gao, H., Chen, S., & Su, G. (2007). Emotion classification of mandarin speech based on TEO nonlinear features. Paper presented at the Eighth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing.

Goudbeek, M., Goldman, J. P., & Scherer, K. R. (2009). Emotion dimensions and formant position. Paper presented at the Interspeech.

Guan, L., & Xie, Z. (2013). Multimodal information fusion of audio emotion recognition based on kernel entropy component analysis. *International Journal of Semantic Computing, 07*(01), 25–42. http://dx.doi.org/10.1142/S1793351X13400023.

Hansen, J. H. L., & Bou-Ghazale, S. (1997). Getting started with SUSAS: A speech under simulated and actual stress database. Paper presented at the Proceedings of EUROSPEECH '97.

Hu, H., Xu, M. -X., & Wu, W. (2007). GMM supervector based SVM with spectral features for speech emotion recognition. Paper presented at the Acoustics, Speech and Signal Processing, 2007, ICASSP 2007, IEEE International Conference on.

Jian, Y., Zhang, D., Frangi, A. F., & Jing-Yu, Y. (2004). Two-dimensional PCA: A new approach to appearance-based face representation and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 26*(1), 131–137.

Jothilakshmi, S., Ramalingam, V., & Palanivel, S. (2009). Unsupervised speaker segmentation with residual phase and MFCC features. *Expert Systems with Applications, 36*(6), 9799–9804. http://dx.doi.org/10.1016/j.eswa.2009.02.040.

Kammoun, M., & Ellouze, N. (2006). Pitch and energy contribution in emotion and speaking styles recognition enhancement. Paper presented at the IMACS Multiconference on Computational Engineering in Systems Applications (4–6 Oct.).

Khanchandani, K., & Hussain, M. A. (2009). Emotion recognition using multilayer perceptron and generalized feed forward neural network. *Journal of Scientific and Industrial Research, 68*(5), 367.

Koolagudi, S., Maity, S., Kumar, V., Chakrabarti, S., & Rao, K. S. (2009). IITKGP-SESC: Speech database for emotion analysis. In S. Ranka, S. Aluru, R. Buyya, Y.-C. Chung, S. Dua, A. Grama, S. S. Gupta, R. Kumar, & V. Phoh (Eds.). *Contemporary computing* (40, pp. 48–492). Berlin Heidelberg: Springer.

Koolagudi, S. G., Reddy, R., Yadav, J., & Rao, K. S. (2011). IITKGP-SEHSC: Hindi speech corpus for emotion analysis. Paper presented at the Devices and Communications (ICDeCom), 2011 International Conference on (24–25 Feb.).

Krishna Kishore, K. V., & Krishna Satish, P. (2013). Emotion recognition in speech using MFCC and wavelet features. Paper presented at the Advance Computing Conference (IACC), 2013 IEEE 3rd International (22–23 Feb.).

Lee, C.-C., Mower, E., Busso, C., Lee, S., & Narayanan, S. (2011). Emotion recognition using a hierarchical binary decision tree approach. *Speech Communication, 53*(9), 1162–1171.

Liberman, M., Davis, K., Grossman, M., Martey, N., & Bell, J. (2002). *Emotional prosody speech and transcripts*. Philadelphia: Linguistic Data Consortium.

Ling, Z.-H., Hu, Y., & Wang, R.-H. (2005). A novel source analysis method by matching spectral characters of LF model with STRAIGHT spectrum. In J. Tao, T. Tan, & R. Picard (Eds.). *Affective computing and intelligent interaction* (3784, pp. 441–448). Berlin Heidelberg: Springer.

Lu, Y. Z., & Wei, Z. Y. (2004). Facial expression recognition based on wavelet transform and MLP neural network. Paper presented at the ICSP '04. 2004 7th International Conference on Signal Processing, 2004, Proceedings.

Lugger, M., & Bin, Y. (2007). The relevance of voice quality features in speaker independent emotion recognition. Paper presented at the Acoustics, Speech and Signal Processing, 2007, ICASSP 2007, IEEE International Conference on (15–20 April).

Martin, O., Kotsia, I., Macq, B., & Pitas, I. (2006). The eNTERFACE'05 audio-visual emotion database. Paper presented at the Proceedings of the 22nd International Conference on Data Engineering Workshops.

Merz, G. (1983). Fast Fourier transform algorithms with applications. In *Computational aspects of complex analysis* (pp. 249–278). Netherlands: Springer.

Morrison, D., Wang, R., De Silva, L. C., & Xu, W. L. (2005). Real-time spoken affect classification and its application in call-centres. Paper presented at the Third International Conference on Information Technology and Applications.

Nass, C., Jonsson, I. -M., Harris, H., Reaves, B., Endo, J., Brave, S., & Takayama, L. (2005). Improving automotive safety by pairing driver emotion and car voice emotion. Paper presented at the CHI'05 extended abstracts on Human factors in computing systems.

Nicholson, J., Takahashi, K., & Nakatsu, R. (2000). Emotion recognition in speech using neural networks. *Neural Computing & Applications, 9*(4), 290–296.

Nwe, T. L., Foo, S. W., & De Silva, L. C. (2003). Detection of stress and emotion in speech using traditional and FFT based log energy features. Paper presented at the Proceedings of the 2003 Joint Conference of the Fourth International Conference on Information, Communications and Signal Processing, 2003 and Fourth Pacific Rim Conference on Multimedia.

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

Orr, M. J. (1999). Matlab functions for radial basis function networks.

Park, J., & Sandberg, I. W. (1993). Approximation and radial-basis-function networks. *Neural Computation, 5*(2), 305–316.

Petrushin, V. (1999). Emotion in speech: Recognition and application to call centers. Paper presented at the Proceedings of Artificial Neural Networks in Engineering.

Pribil, J., & Pribilova, A. (2012). Formant features statistical analysis of male and female emotional speech in Czech and Slovak. Paper presented at the 35th International Conference on Telecommunications and Signal Processing (TSP) (3–4 July).

Schuller, B. (2011). Affective speaker state analysis in the presence of reverberation. *International Journal of Speech Technology, 14*(2), 77–87.

Schuller, B., Batliner, A., Steidl, S., & Seppi, D. (2011). Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication, 53*(9–10), 1062–1087.

Schuller, B., Reiter, S., Muller, R., Al-Hames, M., Lang, M., & Rigoll, G. (2005). Speaker independent speech emotion recognition by ensemble classification. Paper presented at the IEEE International Conference on Multimedia and Expo, 2005, ICME 2005.

Schuller, B., Vlasenko, B., Eyben, F., Wöllmer, M., Stuhlsatz, A., Wendemuth, A., et al. (2010). Cross-corpus acoustic emotion recognition: Variances and strategies. *IEEE Transactions on Affective Computing, 1*(2), 119–131.

Sharkas, M., & Elenien, M. A. (2008). Eigenfaces vs. fisherfaces vs. ICA for face recognition: A comparative study. Paper presented at the 9th International Conference on Signal Processing, 2008, ICSP 2008.

Steinwart, I., & Christmann, A. (2008). *Support vector machines*. New York: Springer-Verlag.

Tin Lay, N., Say Wei, F., & De Silva, L. C. (2003). Classification of stress in speech using linear and nonlinear features. Paper presented at the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003, Proceedings (6–10 April).

Tsang-Long, P., Yu-Te, C., Jun-Heng, Y., & Pei-Jia, L. (2006). Mandarin emotional speech recognition based on SVM and NN. Paper presented at the Pattern Recognition, 2006. ICPR 2006. 18th International Conference on.

Wang, Y., & Guan, L. (2008). Recognizing human emotional state from audiovisual signals. *IEEE Transactions on Multimedia, 10*(5), 936–946.

Wang, Y., Guan, L., & Venetsanopoulos, A. N. (2012). Kernel cross-modal factor analysis for information fusion with application to bimodal emotion recognition. *IEEE Transactions on Multimedia, 14*(3), 597–607. http://dx.doi.org/10.1109/tmm.2012.2189550.

Wei, W., & Guanglai, G. (2009). Online handwriting mongolia words recognition with recurrent neural networks. Paper presented at the ICCIT ‘09. Fourth International Conference on Computer Sciences and Convergence Information Technology.

Wong, E., & Sridharan, S. (2001). Comparison of linear prediction cepstrum coefficients and mel-frequency cepstrum coefficients for language identification. Paper presented at the Intelligent Multimedia, Video and Speech Processing, 2001. Proceedings of 2001 International Symposium on.

Wu, C.-H., & Liang, W.-B. (2011). Emotion recognition of affective speech based on multiple classifier using acoustic-prosodic information and semantic labels. *IEEE Transactions on Affective Computing, 2*(1), 10–21.

Xi, L., Jidong, T., Johnson, M. T., Soltis, J., Savage, A., Leong, K. M., & Newman, J. D. (2007). Stress and emotion classification using jitter and shimmer features. Paper presented at the IEEE International Conference on Acoustics, Speech and Signal Processing.

Xufang, Z., O''Shaughnessy, D., & Nguyen, M. -Q. (2007). A processing method for pitch smoothing based on autocorrelation and cepstral F0 detection approaches. Paper presented at the International Symposium on Signals, Systems and Electronics (July 30–Aug. 2).

Yang, X., Tan, B., Ding, J., Zhang, J., & Gong, J. (2010). Comparative study on voice activity detection algorithm. Paper presented at the International Conference on Electrical and Control Engineering (ICECE), 2010.

Yao, X. J., Panaye, A., Doucet, J. P., Chen, H. F., Zhang, R. S., Fan, B. T., et al. (2005). Comparative classification study of toxicity mechanisms using support vector machines and radial basis function neural networks. *Analytica Chimica Acta, 535*(1–2), 259–273.

Yeh, J.-H., Pao, T.-L., Lin, C.-Y., Tsai, Y.-W., & Chen, Y.-T. (2011). Segment-based emotion recognition from continuous Mandarin Chinese speech. *Computers in Human Behavior, 27*(5), 1545–1552. http://dx.doi.org/10.1016/j.chb.2010.10.027.

Yi-Lin, L., & Gang, W. (2005). Speech emotion recognition based on HMM and SVM. Paper presented at the Proceedings of 2005 International Conference on Machine Learning and Cybernetics (18–21 Aug.).

Yongjin, W., & Ling, G. (2004). An investigation of speech-based human emotion recognition. Paper presented at the Multimedia Signal Processing, 2004 IEEE 6th Workshop on (29 Sept.–1 Oct.).

Zeng, Z., Tu, J., Pianfetti, B. M., & Huang, T. S. (2008). Audio-visual affective expression recognition through multistream fused HMM. *IEEE Transactions on Multimedia, 10*(4), 570–577.

Zuo, W., Zhang, D., Yang, J., & Wang, K. (2006). BDPCA plus LDA: A novel fast feature extraction technique for face recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 36*(4), 946–953.