# Audio-visual feature selection and reduction for emotion classification

*Sanaul Haq, Philip J.B. Jackson and James Edge*

Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, Guildford, UK

s.haq@surrey.ac.uk, p.jackson@surrey.ac.uk, j.edge@surrey.ac.uk

## Abstract

Recognition of expressed emotion from speech and facial gestures was investigated in experiments on an audio-visual emotional database. A total of 106 audio and 240 visual features were extracted and then features were selected with Plus $l$-Take Away $r$ algorithm based on Bhattacharyya distance criterion. In the second step, linear transformation methods, principal component analysis (PCA) and linear discriminant analysis (LDA), were applied to the selected features and Gaussian classifiers were used for classification of emotions. The performance was higher for LDA features compared to PCA features. The visual features performed better than audio features, for both PCA and LDA. Across a range of fusion schemes, the audio-visual feature results were close to that of visual features. A highest recognition rate of 53 % was achieved with audio features, 98 % with visual features, and 98 % with audio-visual features selected by Bhattacharyya distance and transformed by LDA.[1]

**Index Terms**: emotion recognition, multimodal feature selection, principal component analysis

## 1. Introduction

Emotion recognition is a growing field in developing friendly human-computer interaction systems. Human communication consists of two channels: the verbal channel, that carries the message, and the non-verbal channel, that includes information about the emotional state of the person. To convey the message correctly, both verbal and nonverbal information is necessary. There are two kinds of theory to describe emotions: discrete theory [1] is based on existence of universal basic emotions which vary in number and types, and dimensional theory [2, 3] classify emotions in two or more dimensional space. The most widely used basic emotions are anger, fear, happiness, sadness, surprise and neutral. This work is based on the discrete theory of emotion.

Speech databases of different types are recorded for investigation of emotion, some are natural while others are acted or elicited. Natural speech databases consist of recordings from people's daily life, e.g. Belfast Naturalistic Database [4] consist of 239 clips from TV programs and interviews of 100 male and female speakers. Acted databases consist of recordings from actors, e.g. Berlin Database of Emotional Speech (EMO-DB) [5], which consists of recordings from 10 speakers in 7 emotions. The Hebrew emotional speech database [6] is an elicited database, which consist of recordings from 40 subjects in 6 emotions. As both audio and visual modalities contribute to express emotions, for this work, we recorded an audio-visual database from a male actor in seven emotions.

Facial expression and speech characteristics contribute information to assist with emotion recognition. The important speech features for emotion recognition are prosodic and voice quality. The prosodic features consist of pitch, intensity and duration, while voice quality features are represented in spectral energy distribution, formants, Mel Frequency Cepstral Coefficients (MFCCs), jitter and shimmer. These features are identified as important both at utterance level [7, 8, 9, 10] and at frame level [11, 12, 13, 14]. The emotion recognition from facial expressions is performed by extracting forehead, eye-region, cheek and lip features [15, 16, 17, 18]. Both audio and visual modalities are important for emotion recognition and recently researchers are working on fusion of these two modalities to improve the performance of emotion recognition systems. Based on previous research, we extracted 106 audio features related to pitch, energy, duration and spectral envelope, and 240 visual features by placing markers on forehead, eye-regions, cheeks and lips. The feature extraction was performed at utterance level.

Appropriate feature selection is essential for achieving good performance with both global utterance level and instantaneous features. Luengo et al. [7] achieved comparable performance with top 6 global level prosodic features compared to 86 prosodic features. Lin and Wei [12] reported higher recognition rate for 2 prosodic and 3 voice quality instantaneous level features selected by the Sequential Forward Selection (SFS) method from fundamental frequency ($f_0$), energy, formants, MFCCs and Mel sub-band energies features. Kao and Lee [13] found that frame level features were better than syllable and word level features. The best performance was achieved with an ensemble of three levels feature. Schuller et al. [19] halved the error rate with 20 global pitch and energy features compared to that of 6 instantaneous pitch and energy features. Chen, Huang and Cook [15] proposed multimodal emotion recognition system. The facial features consisted of 27 features related to eyes, eyebrows, furrows and lips and acoustic features consisted of 8 features related to pitch, intensity and spectral energy. The performance of the visual system was better than the audio system, and the overall performance improved for the bimodal system. Busso et al. [16] performed emotion recognition using audio, visual and bimodal system. The audio system used 11 prosodic features selected by the Sequential Backward Selection (SBS) technique and visual features were obtained by first tracking 102 markers on the face and then applying PCA to each of the five parts of face: forehead, eyebrow, low eye, right cheek and left cheek. The visual system performed better than the audio system and the highest performance was achieved with the bimodal system. Along similar lines, we first extracted 106 audio and 240 visual features at utterance level and then feature selection was performed with Plus $l$-Take Away $r$ algorithm based on Bhattacharyya distance criterion [20].

The choice of classifier can also significantly affect the recognition accuracy. Gaussian Mixture Model (GMM), Hidden Markov Model (HMM) and Support Vector Machine

(SVM) are widely used classifiers in the field of emotion recognition. Luengo et al. [7] reported 92.3 % recognition rate for SVM classifier compared to 86.7 % for Gaussian classifier with same set of features. Borchert et al. [10] reported accuracy of 74.0 % for 7 classes using SVM and AdaBoost classifiers for speaker dependent case and 70.0 % for speaker independent case. Lin and Wei [12] achieved 99.5 % recognition rate for 5-state HMM and 5 best features. Schuller et al. [19] achieved 86.8 % accuracy with 4 component GMM for 7 emotions compared to 77.8 % for 64-state continuous HMM. Busso et al. [16] achieved recognition rate of 70.9 % with audio features and 85.0 % with visual features for 4 emotions using SVM classifier. An improved performance of 89.0 % was achieved for the fusion of two modalities at feature level and at decision level. Song, Chen and You [17] reported 85.0 % accuracy for 7 emotions with HMM classifier using both audio and visual features. As a simpler technique that is functionally related to these state-of-the-art GMM and HMM systems, we used single Gaussian classifiers for emotion classification. The feature extraction was performed in two steps, feature selection and then feature reduction. The following sections in this paper present our method, classification experiments, discussion, conclusions and future work.

## 2. Method

We performed the emotion recognition from audio and visual modalities in four steps. Firstly, audio features (prosodic and spectral) and visual features (marker locations on the face) were extracted, then feature selection was performed. In the third step, linear transformation methods, PCA and LDA, were applied to the selected features. Finally, Gaussian classifiers were used for classification between different emotion classes. The block diagram of our method is shown in Fig. 1.

### 2.1. Database

The database of 120 utterances was recorded from an actor with 60 markers painted on his face, reading sentences in seven emotions ($N = 7$): anger, disgust, fear, happiness, neutral, sadness and surprise. Recordings consisted of 15 phonetically-balanced TIMIT sentences per emotion: 3 common, 2 emotion specific and 10 generic sentences that were different for each emotion. The 3 common and 2 emotion specific sentences were recorded in neutral emotion, which resulted 30 sentences for neutral emotion. Emotion and sentence prompts were displayed on a monitor in front of actor during the recordings. The 3dMD dynamic face capture system provided colour video and Beyer dynamics microphone signals. The sampling rate was 44.1 kHz for audio and 60 fps for video. The 2D video of frontal face of the actor was recorded with one colour camera.

### 2.2. Feature extraction

#### 2.2.1. Audio features

A total of 106 utterance-level audio features were extracted related to fundamental frequency ($f_0$), energy, duration and spectral envelope. The audio feature extraction using Speech Filing
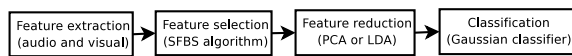


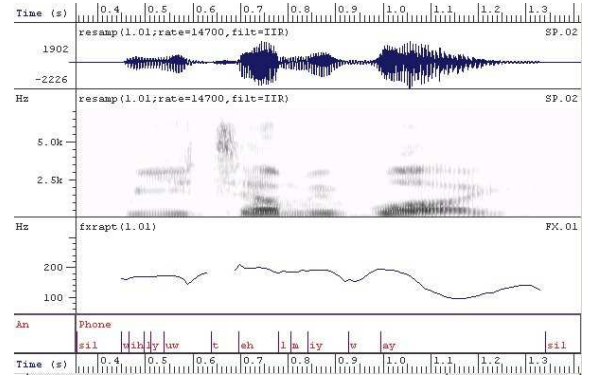Figure 1: *Block diagram of our experimental method.*



Figure 2: *Illustration of audio feature extraction using Speech Filing System (from top): waveform, spectrogram, pitch track and phone annotations.*

System software [21] is shown in Fig. 2.

**Pitch features:** The fundamental frequency ($f_0$) extraction was performed with Speech Filing System software [21] by RAPT algorithm. The following features were extracted from $f_0$ contour: Mel freq. minimum, Mel freq. maximum, mean and standard deviation of first and second Gaussian of Mel freq., minimum of Mel freq. first order difference, maximum of Mel freq. first order difference, mean of Mel freq. first order difference, standard deviation of Mel freq. first order difference.

**Energy features:** Firstly, the signal was filtered in bands using Butterworth filter (order 9) and then energy was calculated at frame level using Hamming window of 25ms with a step size of 10ms. The following energy features were extracted: mean and standard deviation of total log energy; mean, standard deviation, minimum, maximum and range of normalized energies in the original speech signal and speech signal in the frequency bands 0-0.5 kHz, 0.5-1 kHz, 1-2 kHz, 2-4 kHz and 4-8 kHz; mean, standard deviation, minimum, maximum and range of first order difference of normalized energies in the original speech signal and speech signal in the same frequency bands.

**Duration features:** Manual phone labels were used to extract duration features, which were based on listening assisted by waveform and spectrogram. The extracted duration features were: voiced speech duration, unvoiced speech duration, sentence duration, average voiced phone duration, average unvoiced phone duration, voiced-to-unvoiced speech duration ratio, average voiced-to-unvoiced speech duration ratio, speech rate (phone/s), voiced-speech-to-sentence duration ratio, unvoiced-speech-to-sentence duration ratio.

**Spectral features:** The spectral envelope features were extracted using HTK software [22], at utterance level: mean and standard deviation of 12 MFCCs, $C_1, ..., C_{12}$.

#### 2.2.2. Visual features

The visual features were created by painting 60 frontal markers on the face of the actor. The markers were painted on forehead, eyebrows, low eyes, cheeks, lips and jaw. After data capture the markers were manually labelled for the first frame of a sequence and tracked for the remaining frames using a marker tracker. The tracked marker $x$ and $y$ coordinates were normalized. Each marker's mean displacement from the bridge of the nose was subtracted. In the last step, 240 visual features were obtained from 2D marker coordinates which consisted of mean and standard deviation of the adjusted marker coordinates. The
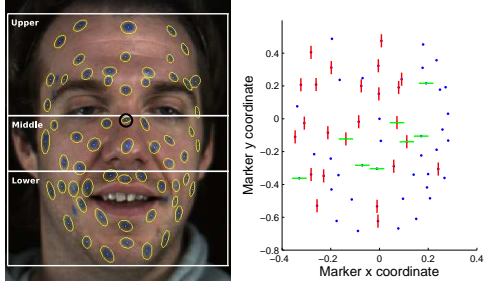
Figure 3: *Example video data (left) with overlaid tracked marker locations. The marker on the bridge of the nose (encircled in black) was taken as a reference. Figure on the right shows top 40 visual features for a neutral frame, where horizontal line (green) shows mean value of x-coordinate, and vertical line (red) shows mean value of y-coordinate of a selected marker. The dot (blue) shows marker location.*

markers were divided into three main groups, as in Busso and Narayanan [18]: upper, middle, and lower face regions, shown in Fig. 3 (left). The upper face region includes the markers above the eyes in the forehead and eyebrow area. The lower face region contains the markers below the upper lip, including the mouth and jaw. The middle face region contains the markers in the cheek area between the upper and lower face regions.

### 2.3. Feature selection

The feature selection was performed using a standard algorithm based on a discriminative criterion function. This process helps to remove uninformative, redundant or noisy features. The Plus $l$-Take Away $r$ algorithm [23] is a feature search method based on some distance function that uses both SFS and SBS algorithms. The SFS algorithm is a bottom up search method where one feature is added at a time. First the best feature is selected and then the function is evaluated for combination with the remaining candidates and the best new feature is added. The problem with the SFS algorithm is that once a feature is added (which may become unhelpful later as the feature set grows), it cannot be removed. The SBS on the other hand is a top down process. It starts from complete feature set and at each step the worst feature is discarded such that the reduced set gives maximum value of the criterion function. The SBS gives better results but is computationally more complex.

Sequential forward backward search offers benefits of both SFS and SBS, via Plus $l$-Take Away $r$ algorithm. At each step, $l$ features are added to the current feature set and $r$ features are removed. The process continues until the required feature set size is achieved. We used this algorithm to select from full feature sets (audio, visual, and audio-visual), with Bhattacharyya distance as a criterion [20]. The distribution of classes was assumed to be Gaussian. The feature search was performed with $l=2$ and $r=1$, i.e. one feature was added at each step. The top 40 audio features were obtained by selecting 6 pitch, 18 energy, 6 duration, and 10 spectral features. The top 40 audio features are listed in Table 1. The top 40 visual features were obtained by selecting 14 upper face, 14 middle face, and 12 lower face features. The top 40 visual features are shown in Fig. 3 (right).

Table 1: Top 40 audio features selected using Bhatt. criterion

| Feature | Description |
|---|---|
| Pitch | mean and standard deviation of first and second Gaussian of Mel freq., minimum and standard deviation of Mel freq. first order difference. |
| Energy | mean and standard deviation of total log energy, standard deviation of normalized energies in the original speech signal and the speech signal in freq. bands 1-2 kHz, and 4-8 kHz, minimum of normalized energies in the original speech signal and the speech signal in freq. band 4-8 kHz, maximum of normalized energies in the speech signal in freq. bands 0-0.5 kHz, 1-2 kHz, and 2-4 kHz, range of normalized energy in the speech signal in freq. band 0.5-1 kHz, mean of normalized energies first order difference in the original speech signal and the speech signal in freq. bands 0.5-1 kHz, 1-2 kHz, and 4-8 kHz, standard deviation of normalized energies first order difference in the speech signal in freq. bands 1-2 kHz, and 4-8 kHz, minimum of normalized energies first order difference in the original speech signal and the speech signal in freq. band 1-2 kHz, maximum of normalized energy first order difference in the speech signal in freq. band 4-8 kHz. |
| Duration | Voiced phone duration, unvoiced phone duration, sentence duration, voiced-to-unvoiced speech duration ratio, voiced-speech-to-sentence duration ratio, unvoiced-speech-to-sentence duration ratio. |
| Spectral | mean of MFCCs: $C_1, C_2, C_5, C_8, C_9$, standard deviation of MFCCs: $C_5, C_7, C_8, C_{11}, C_{12}$. |

### 2.4. Feature reduction

The dimensionality of a feature set can be reduced by using statistical methods to maximize the relevant information preserved. This can be done by applying a linear transformation, $x = Wz$, where $x$ is a feature vector in the reduced feature space, $z$ is the original feature vector, and $W$ is the transformation matrix. PCA [24] is widely used to extract essential characteristics from high dimensional data sets and discard noise, while LDA [25] maximizes the ratio of between-class variance to within-class variance to optimize separability between classes. The PCA and LDA methods involve feature centering and whitening, covariance computation and eigen decomposition. We applied both PCA and LDA as linear transformation techniques for feature reduction.

### 2.5. Classification

A Gaussian classifier uses Bayes decision theory where the class-conditional probability density $p(x|\omega_i)$ is assumed to have Gaussian distribution for each class $\omega_i$. The Bayes decision rule is described as

$$i_{\text{Bayes}} = \arg\max_i P(\omega_i|x) = \arg\max_i p(x|\omega_i)P(\omega_i) \quad (1)$$

where $P(\omega_i|x)$ is the posterior probability, and $P(\omega_i)$ is the prior class probability. We used single Gaussian classifiers (1-mix) to represent $p(x|\omega_i)$ for emotion recognition experiments.

## 3. Experiment and results

We performed three sets of emotion recognition experiments. First, audio feature sets were obtained by first selecting the top 40 audio features using Plus $l$-Take Away $r$ algorithm based on
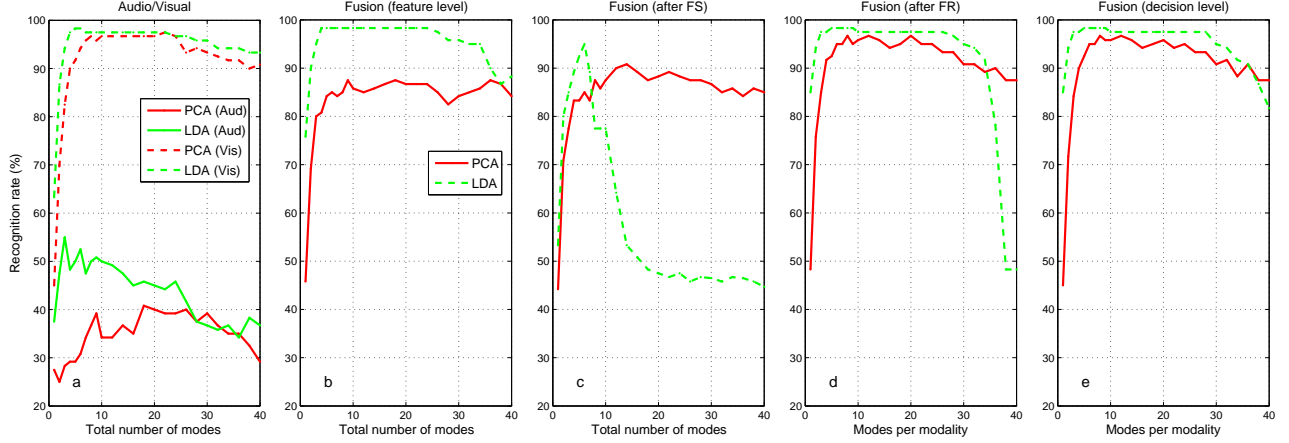
Figure 4: *Classification accuracy (%) with (a) audio, (a) visual, and audio-visual features: (b) fused at feature level, (c) fused after feature selection, (d) fused after feature reduction, and (e) fused at decision level.*

Bhattacharyya distance, then applying feature reduction techniques, PCA and LDA. In the second, visual feature sets were obtained by first selecting the top 40 visual features using Plus *l*-Take Away *r* algorithm based on Bhattacharyya distance, then applying feature reduction techniques, PCA and LDA. Thirdly, audio-visual experiments were performed by fusion of audio and visual features at different stages. Experiments were performed with single component Gaussian classifiers. The data were divided into six sets in a jack-knife procedure. Each round, five sets were used for training and one set for testing. The experiments were repeated for six different rounds of training and testing sets, and the results averaged.

### 3.1. Audio experiments

In these experiments, the top 40 audio features were selected. The feature reduction techniques, PCA and LDA, were applied in the next stage. The classification experiments were performed for seven emotions with single Gaussian classifiers. The results are plotted in Fig. 4a.

Higher recognition rates were achieved with LDA features compared to PCA features. The highest PCA recognition rate of 40.8 % was achieved with 18 features which contained 92.5 % energy. A recognition rate of 52.5 % was achieved with 6 LDA features. Energy and MFCCs were identified as the most important features for emotion recognition, although pitch and duration features also contributed. The top 40 Bhattacharyya features consisted of 18 energy, 10 MFCCs, 6 pitch and 6 duration features. The recognition rate was higher for anger and neutral, and lower for disgust and fear. The disgust, fear, and sadness emotions were confused with neutral, and happiness with surprise. While this level of performance is disappointing and unsuitable for applications, it is still three or four times above chance.

### 3.2. Visual experiments

The top 40 visual features were selected, and PCA and LDA were applied to the selected feature sets. The classification experiments were performed with single Gaussian classifiers. The results are plotted in Fig. 4a.

The recognition rates for LDA features were higher compared to PCA features. The highest recognition rate of 97.5 %

was achieved with 22 PCA features which contained 99.9 % energy. The maximum recognition rate of 98.3 % was achieved with 6 LDA features. The top 40 Bhattacharyya features consisted of 14 features from each of the upper and middle face regions, and 12 features from lower face region. A recognition rate of 100 % was achieved for anger, disgust, fear, and happiness. The surprise emotion had the lowest recognition rate due to some confusion with anger. So, at 98 %, the performance of the visual emotion classification is substantially improved to a useful level.

### 3.3. Audio-visual experiments

The audio-visual experiments were performed by combining the two modalities at feature level, after feature selection, after feature reduction, and at decision level. The block diagram for different audio-visual experiments are shown in Fig. 5.

#### 3.3.1. Fusion at feature level

All audio and visual features were grouped together to get a total of 346 audio-visual features. The top 40 audio-visual features were selected, PCA and LDA were applied to the selected feature sets, and classification experiments were performed with single Gaussian classifiers. The results are plotted in Fig. 4b.

The recognition rates for LDA features were higher compared to PCA features. The highest PCA recognition rate of 87.5 % was achieved with 9 features which contained 90.3 % energy. The maximum recognition rate of 98.3 % was achieved with 6 LDA features. A recognition rate of 100 % was achieved for anger, disgust, happiness, neutral, and sadness with LDA features.

#### 3.3.2. Fusion after feature selection

The top 40 audio and top 40 visual features selected were grouped together. The linear transformation methods, PCA and LDA were then applied. Single Gaussian classifiers were used for classification in the last step. The results are plotted in Fig. 4c.

The recognition rates for LDA features were higher compared to PCA features. The highest recognition rate for PCA was 90.8 % with 14 features and for LDA was 95.0 % with 6
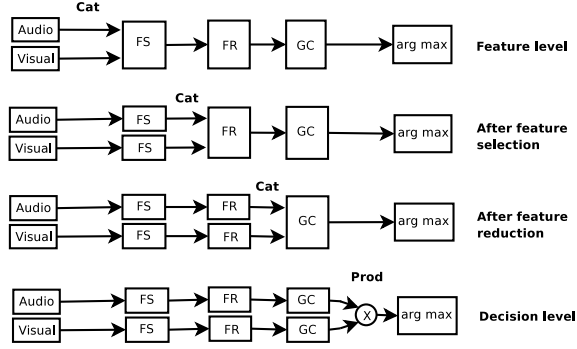
Figure 5: *Block diagram of audio-visual experiments which involve combining the two modality at different levels (from top): at feature level, after feature selection, after feature reduction, and at decision level.*

features. A recognition rate of 100 % was achieved for happiness with LDA features.

### 3.3.3. Fusion after feature reduction

In these experiments, the top 40 audio and top 40 visual features were selected, and then PCA and LDA were applied to the selected audio and visual features, separately. The audio and visual features were then combined to calculate the probability for each of the emotions. The classification experiments were performed with single Gaussian classifiers. The results are plotted in Fig. 4d.

The recognition rates for LDA features were higher compared to PCA features. The highest recognition rate with PCA was 96.7 % for 8 features per modality and with LDA was 98.3 % for 6 features per modality. A recognition rate of 100 % was achieved for anger, disgust, fear, and happiness for LDA.

### 3.3.4. Fusion at decision level

The top 40 audio and top 40 visual features were selected, feature reduction was applied to the selected audio and visual features, separately. The probability for each of the emotions was calculated for the audio and visual features separately and were multiplied to get the final result. The classification experiments were performed with single Gaussian classifiers. The results are plotted in Fig. 4e.

The recognition rates for LDA features were higher compared to PCA features. The highest recognition rate of 96.7 % was achieved with 8 PCA features per modality. A maximum recognition rate of 98.3 % was achieved with 6 LDA features per modality. A recognition rate of 100 % was achieved for anger, disgust, fear, and happiness with LDA features.

## 4. Discussion

In the audio-visual experiments, LDA performed better than PCA. The fusion after feature reduction, and at decision level performed better than fusion at feature level and after feature selection. A maximum recognition rate of 98.3 % was achieved with LDA, and 96.7 % with PCA. Some of the emotions were confused with others, like fear with sadness, neutral with happiness, and surprise with anger.

The highest recognition rates obtained by applying PCA and LDA to top 40 audio, visual and audio-visual features are

Table 2: *Maximum emotion classification scores (%) applying PCA and LDA to top 40 audio, visual and audio-visual Bhattacharyya features. The values show average recognition rate with standard error over 6 jack-knife tests.*

| Feature set | PCA | LDA |
|---|---|---|
| Audio features | $40.8 \pm 8.5$ (18 feat.) | $52.5 \pm 7.2$ (6 feat.) |
| Visual features | $97.5 \pm 3.3$ (22 feat.) | $98.3 \pm 3.6$ (6 feat.) |
| Audio-visual fusion (feature level) | $87.5 \pm 4.2$ (9 feat.) | $98.3 \pm 2.3$ (6 feat.) |
| Audio-visual fusion (after feature selection) | $90.8 \pm 3.0$ (14 feat.) | $95.0 \pm 5.1$ (6 feat.) |
| Audio-visual fusion (after feature reduction) | $96.7 \pm 2.1$ (8 feat.) | $98.3 \pm 3.3$ (6 feat.) |
| Audio-visual fusion (decision level) | $96.7 \pm 2.1$ (8 feat.) | $98.3 \pm 3.3$ (6 feat.) |

shown in Table 2. The LDA features performed better than PCA features for all three kinds of experiment. Higher performance was achieved with visual and audio-visual features compared to audio features.

In order to investigate the poor performance of the audio features, we performed some comparative experiments between our English database and the Berlin Database of Emotional Speech (EMO-DB) [5]. Our database consisted of 120 utterances from a male speaker, so we selected two male speakers (speaker number 11 and 15) data from EMO-DB to get 110 utterances in total. Both databases covered of seven emotions, but EMO-DB has boredom instead of surprise in the English database. A total of 106 audio features related to fundamental frequency, energy, duration and spectral envelope were extracted at utterance level from each database. The same experimental procedure was adopted for classification as in section 3.1. Results are plotted in Fig. 6.

Higher recognition rates were achieved for EMO-DB compared to the English database for both PCA and LDA. For the English database, a maximum recognition rate of 40.8 % was achieved for 18 PCA features which contained 92.5 % energy. The recognition rate for LDA was 52.5 % with 6 features. For the EMO-DB a maximum recognition rate of 67.6 % was achieved for 30 PCA features which contained 98.9 % energy and the same recognition rate was achieved for 6 LDA features. We suggest that the reason of low recognition rates for English database was that the actor was not as expressive as in EMO-DB. Another important difference was the evaluation of EMO-DB by a panel of listeners to validate the expressed emotions.

Other researchers have reported higher accuracy with EMO-DB compared to our results. Borchert and Düsterhöft [10] achieved a recognition accuracy of 76.1 % with SVM, and 74.8 % with AdaBoost for speaker dependent case. The recognition accuracy was 70.6 % with SVM, and 72.1 % with AdaBoost for speaker independent case. A set of 63 features related to pitch, relative intensity, formants, spectral energy, HNR, jitter, and shimmer were used for classification. Schuller et al. [26] reported a recognition accuracy of 83.2 % for speaker independent case, and 95.1 % for speaker dependent case with SVM classifier. A set of 1,406 acoustic features related to pitch, energy, envelope, formants, MFCCs, HNR, jitter, and shimmer were extracted. The speaker normalization and feature selection was performed before classification, and SVM with linear kernel was used for classification. The focus of our work was to investigate the fusion of audio and visual features at different
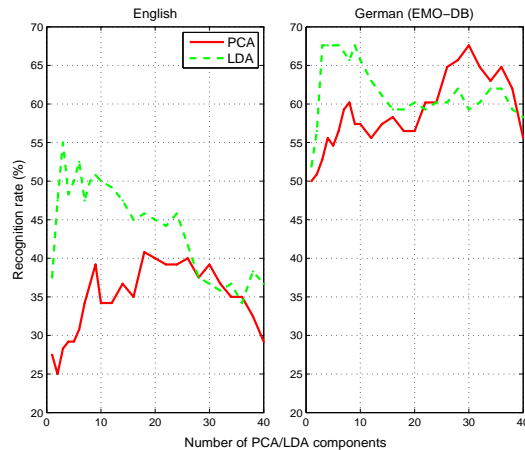
Figure 6: *Recognition rate (%) with PCA and LDA applied to top 40 audio Bhattacharyya features of English and German (EMO-DB) databases.*

stages. The low overall recognition rate in our case was due to use of small set of features and simpler classifier. These issues will be investigated in future.

# 5. Conclusions

In classification tests on the British English audio-visual emotional database, LDA outperformed PCA with the top 40 features selected by Bhattacharyya distance. Results show that both audio and visual information are useful for emotion recognition, although visual features performed much better here, perhaps because the actor was more expressive facially compared to his voice. The energy and MFCC features were identified as most important audio features for emotion recognition, although pitch and duration features also contributed. The important visual features were the mean value of $y$-coordinate of markers, i.e. vertical movement of face was more important for emotion classification. The best recognition rate of 98 % was achieved with 6 LDA features ($N - 1$) with audio-visual and visual features, whereas audio LDA scored 53 %. Maximum PCA results for audio, visual, and audio-visual features were 41 %, 98 %, and 97 % respectively. In audio experiments, the recognition rate was higher for anger and neutral, and lower for disgust and fear. The disgust, fear, and sadness emotions were confused with neutral, and happiness with surprise. In visual and audio-visual experiments, a recognition rate of 100 % was achieved for anger, disgust, fear, and happiness. The neutral was confused with happiness, and surprise with anger. Future work involves experiments with more subjects and other classifiers, like GMM and SVM. Another interesting area concerns the relationship between vocal and facial expressions of emotion.

# 6. References

[1] Ortony, A. and Turner, T.J., "What's Basic About Basic Emotions?", Psychological Review, 97(3):315-331, 1990.

[2] Scherer, K.R., "What are emotions? And how can they be measured?", Social Science Information, 44(4):695-729, 2005.

[3] Russell, J.A., Ward, L.M. and Pratt, G., "Affective Quality Attributed to Environments: A Factor Analytic Study", Environment and Behaviour, 13(3):259-288, 1981.

[4] Douglas-Cowie, E., Cowie, R. and Schroeder, M., "A New Emotional Database: Considerations, Sources and Scope", In Proc. of ISCA Workshop Speech and Emotion: A conceptual framework for research, Belfast, 39-44, 2000.

[5] Burkhardt, F., et al., "A Database of German Emotional Speech", In Proc. of Interspeech 2005, Lisbon, 1517-1520, 2005.

[6] Amir, N., Ron, S. and Laor, N., "Analysis of emotional speech corpus in Hebrew based on objective criteria", In Proc. of ISCA Workshop Speech and Emotion: A conceptual framework for research, Belfast, 29-33, 2000.

[7] Luengo, I., Navas, E., et al., "Automatic Emotion Recognition using Prosodic Parameters", In Proc. of Interspeech 2005, Lisbon, 493-496, 2005.

[8] Ververidis, D. and Kotropoulos, C., "Emotional speech classification using Gaussian mixture models", In Proc. of ISCAS 2005, Kobe, 2871-2874, 2005.

[9] Vidrascu, L., et al., "Detection of real-life emotions in call centers", In Proc. of Interspeech 2005, Lisbon, 1841-1844, 2005.

[10] Borchert, M. and Düsterhöft, A., "Emotions in Speech - Experiments with Prosody and Quality Features in Speech for Use in Categorical and Dimensional Emotion Recognition Environments", In Proc. of NLP-KE'05, Wuhan, 147-151, 2005.

[11] Nogueiras, A., Moreno, A., et al., "Speech Emotion Recognition Using Hidden Markov Models", In Proc. of Eurospeech 2001, Scandinavia, 2679-2682, 2001.

[12] Lin, Y. and Wei, G., "Speech Emotion Recognition Based on HMM and SVM", In Proc. of the 4th Int. Conf. on Mach. Learn. and Cybernetics, Guangzhou, 4898-4901, 2005.

[13] Kao, Y. and Lee, L., "Feature Analysis for Emotion Recognition from Mandarin Speech Considering the Special Characteristics of Chinese Language", In Proc. of Interspeech 2006, Pittsburgh, 1814-1817, 2006.

[14] Neilberg, D., Elenius, K., et al., "Emotion Recognition in Spontaneous Speech Using GMMs", In Proc. of Interspeech 2006, Pittsburgh, 809-812, 2006.

[15] Chen, C.Y., et al., "Visual/Acoustic emotion recognition", In Proc. of Int. Conf. on Multimedia and Expo, 1468-1471, 2005.

[16] Busso, C., et al., "Analysis of Emotion Recognition using Facial Expressions, Speech and Multimodal Information", In Proc. of the ACM Int. Conf. on Multimodal Interfaces, 205-211, 2004.

[17] Song, M., Chen, C., and You, M., "Audio-visual based emotion recognition using tripled Hidden Markov Model", In Proc. of Int. Conf. on ASSP, 5:877-880, 2004.

[18] Busso, C., and Narayanan, S., "Interrelation Between Speech and Facial Gestures in Emotional Utterances: A Single Subject Study", IEEE Transactions on ASLP, 2007.

[19] Schuller, B., Rigoll, G. and Lang, M., "Hidden Markov Model-Based Speech Emotion Recognition", In Proc. of ICASSP 2003, Hong Kong, 2:1-4, 2003.

[20] Campbell, J.P., "Speaker Recognition: A Tutorial", In Proc. of the IEEE, 85(9):1437-1462, 1997.

[21] Huckvale, M., "Speech Filing System", UCL Dept. of Phonetics & Linguistics, UK. Online: http://www.phon.ucl.ac.uk/resource/sfs/, accessed on 3 April 2008.

[22] Young, S. and Woodland, P., "Hidden Markov Model Toolkit", Cambridge University Engineering Department (CUED), UK. Online: http://htk.eng.cam.ac.uk/, accessed on 3 April 2008.

[23] Chen, C.H., "Pattern Recognition and Signal Processing", Sijthoff & Noordoff International Publishers, The Netherlands, 1978.

[24] Shlens, J., "A Tutorial on Principal Component Analysis", Systems Neurobiology Laboratory, Salk Institute for Biological Studies, La Jolla, 2005.

[25] Duda, R.O., Hart, P.E. and Stork, D.G., "Pattern Classification", John Wiley & Sons, Inc. USA, Canada, 2001.

[26] Schuller, B., Vlasenko, B., et al., "Comparing one and two-stage acoustic modeling in the recognition of emotion in speech", IEEE Workshop on ASRU, 596-600, 2007.