Speaker-Dependent Audio-Visual Emotion Recognition

Sanaul Haq and Philip J.B. Jackson

Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, UK
{s.haq, p.jackson}@surrey.ac.uk

Abstract

This paper explores the recognition of expressed emotion from speech and facial gestures for the speaker-dependent case. Experiments were performed on an English audio-visual emotional database consisting of 480 utterances from 4 English male actors in 7 emotions. A total of 106 audio and 240 visual features were extracted and features were selected with Plus l-Take Away r algorithm based on Bhattacharyya distance criterion. 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. The performance was higher for LDA features compared to PCA features. The visual features performed better than the audio features and overall performance improved for the audio-visual features. In case of 7 emotion classes, an average recognition rate of 56 % was achieved with the audio features, 95 % with the visual features and 98 % with the audio-visual features selected by Bhattacharyya distance and transformed by LDA. Grouping emotions into 4 classes, an average recognition rate of 69 % was achieved with the audio features, 98 % with the visual features and 98 % with the audio-visual features fused at decision level. The results were comparable to the measured human recognition rate with this multimodal data set. [1]

Index Terms: audio-visual emotion, data evaluation, linear transformation, speaker-dependent

1 Introduction

In human-to-human interaction, emotions play an important role by allowing people to express themselves beyond the verbal domain. To convey a message correctly, paralinguistic information plays an important role in addition to linguistic information [1]. The linguistic channel carries the textual content of a message, and paralinguistic channel carries other information including body language, facial expressions, pitch and tone of voice, health and identity.

Emotion recognition is a growing area of research to enhance human-computer interaction systems. There are two theories to describe emotions: discrete theory [2] is based on existence of universal basic emotions which vary in number and types, and dimensional theory [3, 4] classify emotions in two or more dimensional space. The most widely used basic emotions are anger, disgust, fear, happiness, sadness, surprise and neutral. This work is based on the discrete theory of emotion.

Speech databases of different types have been recorded for investigation of emotion, some natural while others acted or elicited. Natural speech databases consist of recordings from people's daily life, e.g. Belfast Naturalistic Database [5] consists 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) [6] consists of recordings from 10 speakers in 7 emotions. The Hebrew emotional speech database [7] is an elicited database, which consists 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 four English male actors in seven emotions in a controlled environment.

Both facial expression and speech characteristics give information to assist with emotion recognition. Higher performance is reported for their multilevel fusion [8]. The important speech features for emotion recognition are pitch, intensity, duration, spectral energy distribution, formants, Mel Frequency Cepstral Coefficients (MFCCs), jitter and shimmer. These features are identified as important both at utterance level [9, 10, 11, 12] and at frame level [13, 14, 15, 16]. The emotion recognition from facial expressions is performed by extracting forehead, eye-region, cheek and lip features [17, 18, 19, 1]. We extracted audio features related to pitch, energy, duration and spectral envelope, and visual features by painting markers on forehead, eye-regions, cheeks and lips. The feature extraction was performed at the utterance level.

Appropriate feature selection is essential for achieving good performance with both global utterance level and instantaneous frame level features. Lin and Wei [14] 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. In emotion recognition from multilevel features [15], it was found that frame level features were better than syllable and word level features. The best performance was achieved with an ensemble of three feature levels. In phoneme based emotion recognition, it was found that some phonemes were more important than others, particularly semivowels and vowels [20]. Schuller et al. [21] halved the error rate with 20 global pitch and energy features compared to that of 6 instantaneous pitch and energy features. Some researchers proposed multimodal emotion recognition [17]. Their facial features consisted of 27 features related to eyes, eyebrows, furrows and lips, and the 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. [18] 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 the visual features were obtained from 102 markers on the face by applying PCA to each of the five parts of face: forehead, eyebrow, low eye, right cheek and left cheek. The visual system performed better

[1] Thanks to Kevin Lithgow, James Edge, Joe Kilner, Darren Cosker, Nataliya Nadtoka, Samia Smail, Idayat Salako, Affan Shaukat and Aftab Khan for help with the data capture, evaluation and as subjects, to James Edge for his marker tracker, to Adrian Hilton for use of his equipment, and to Univ. Peshawar, Pakistan for funding.

than the audio system and the overall performance improved with the bimodal system. The emotion recognition in noisy conditions [22] improved with noise and speaker adaptation which is further improved by addition of feature selection. The experiments on audio-visual data showed that the performance for both audio and visual features improved with feature selection, and combining the two modalities before feature selection further improved the performance. In a similar way, we first extracted audio and visual features at utterance level and then feature selection was performed with Plus l-Take Away r algorithm based on Bhattacharyya distance criterion [23].

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. [9] reported 92.3 % recognition rate for SVM classifier compared to 86.7 % for Gaussian classifier with same set of features. Borchert et al. [12] 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 [14] achieved 99.5 % recognition rate with 5-state HMM using 5 best features. Schuller et al. [21] achieved 86.8 % accuracy with 4 component GMM for 7 emotions, compared to 77.8 % with 64-state continuous HMM. Busso et al. [18] achieved recognition rate of 70.9 % with audio features and 85.0 % with visual features for 4 emotions using SVM classifier. The performance improved to 89.0 % for the fusion of two modalities at feature level and at decision level. Song, Chen and You [19] 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 creation of features was performed in three steps: feature extraction, feature selection and feature reduction. A preliminary version of our work has been presented in [24]. Here, we have extended our tests to 4 speakers and compare results with 7 and 4 emotion classes against human performance. The following sections in this paper present our method, classification experiments, discussion, conclusions and future work.

2 Method

We performed the emotion recognition from the audio and visual modalities in four steps. The audio features (prosodic and spectral) and the visual features (marker locations on the face) were extracted, and feature selection was performed. In the next 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

An audio-visual emotional database was recorded from four English male actors in seven emotions: anger, disgust, fear, happiness, neutral, sadness and surprise. The database consists of 120 utterances per actor, which gave 480 sentences in total. The data were recorded by painting 60 markers on the face of actor for extraction of visual features. 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 text prompts were displayed on a

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

monitor in front of actor during the recordings. The data were captured in 3D vision laboratory during different time of the year. The 3dMD dynamic face capture system [25] provided 2D frontal colour video and Beyer dynamics microphone signals. The sampling rate was 44.1 kHz for audio and 60 fps for video. The data capture is shown in Fig. 2.

2.2 Human evaluation experiments

The main purpose of performing the human evaluation experiments was to check the quality of the recorded emotional data. Human tests provide a bench mark for evaluation of recognition accuracy. The recordings were evaluated by 10 subjects, of which 5 were native English speaker and the remaining were non-native but had lived in UK for more than a year. In some studies it was found that female experience emotion more intensively than men [26], to avoid gender biasing, half of the evaluators were female. Three types of human evaluation experiments were designed: audio, visual, and audio-visual. Slides in Microsoft Powerpoint software were used to give audio, visual and audio-visual clips of each utterance. There were 120 utterances for each of the four actors. Each of the audio, visual and audio-visual clips were divided into 10 groups, resulting 12 clips per group. There were 12 clips per slide for the audio data, so 10 slides per actor. There were 4 clips per slide for the visual and audio-visual data, i.e. 3 slides per group, so 30 slides per actor for each of the visual and audio-visual data. The data were randomized to remove systematic bias from the responses of human evaluators. For the 10 evaluators, ten different sets were created for each of the audio, visual, and audio-visual data per actor by using Balanced Latin Square method [27]. The subjects were trained by using slides that contained three facial expression pictures, two audio files, and a short movie clip for each of the emotion. One facial expression picture per emotion and all the audio clips were taken from our database. The movie clips were taken from different movies. Subjects were not given additional speaker-dependent training, although some of the actors were known to some of them. The subjects were asked to play audio, visual and audio-visual clips and select from one of the seven emotions on a paper sheet. The responses were averaged over 10 subjects for each actor audio, visual, and audio-visual data. The results for 7 emotions are shown in Table 1, and for 4 emotion classes are shown in Table 3. It was found that the visual data were easy to recognize compared to the audio, yet the overall performance improved by combining the two modalities.

Figure 2: *Facial markers placed on four subjects with expressions (from left): Displeased (anger, disgust), Gloomy (fear, sadness), Excited (happiness, surprise) and Neutral (neutral).*

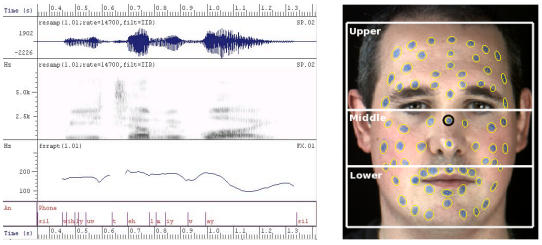

Figure 3: *Audio feature extraction with Speech Filing System software (left), and video data (right) with overlaid tracked marker locations. The reference marker was on the bridge of the nose (black circle).*

2.3 Feature extraction

2.3.1 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 System software [28] is shown in Fig. 3 (left).

Pitch features: The fundamental frequency (f_0) extraction was performed with Speech Filing System software [28] by RAPT algorithm. The following features were extracted from f_0 contour: minimum and maximum of Mel frequency; mean and standard deviation of first and second Gaussian of Mel frequency; minimum and maximum of Mel frequency first order difference; mean and standard deviation of Mel frequency first order difference.

Energy features: The energy features were extracted by first filtering the signal in different bands using Butterworth filter (order 9) and then energy was calculated at frame level using Hamming window of 25 ms with a step size of 10 ms. 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: Semi-automated phone labels were used to extract duration features. The phone labeling was performed in two steps: first the automatic labeling of the data was performed with HTK software [29], and secondly Speech Filing System software [28] was used to correct the automatic phone labels based on listening assisted by waveform and spectrogram. Suitably trained ASR systems can fully automate the alignment of labels. We focus on the use of derived features for emotion recognition. 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 at utterance level using HTK software [29]: mean and standard deviation of 12 MFCCs, $C_1, ..., C_{12}$.

2.3.2 Visual features

Facial coordinates were extracted by painting 60 frontal markers on the face of each actor. The markers were painted on the forehead, eyebrows, cheeks, lips and jaw. After data capture, 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 by subtracting the mean displacement from the bridge-of-the-nose reference and rotating markers for correction of head pose [1]. For deployment, our system assumes that tracked facial coordinates are available, either from markers or by means of computer vision. Our interest is in identifying what facial information provides good features for emotion recognition. Finally, 240 visual features were obtained from 2D marker coordinates as mean and standard deviation of the adjusted marker coordinates. Markers were divided into three groups, as by Busso and Narayanan [1]: upper, middle and lower face regions, shown in Fig. 3 (right). The upper region includes markers above the eyes in the forehead and eyebrow area. The lower region contains markers below the upper lip, including the mouth and jaw. The middle region covers the cheek area between the upper and lower regions.

2.4 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 [30] is a feature search method based on some criterion function that uses both Sequential Forward Selection (SFS) and Sequential Backward Selection (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 and visual), with Bhattacharyya distance as a criterion [23]. 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 visual features were obtained by selecting 14 upper face, 14 middle face, and 12 lower face features.

2.5 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 [31] is widely used to extract essential characteristics from high dimensional data sets and discard noise, while LDA [32] maxi-

mizes the ratio of between-class variance to within-class variance to optimize separability between classes. We applied LDA by using covariance of all training data rather than between-class variance, in order to compare LDA with PCA at different number of features. 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. In our earlier experiments on an audio emotional database EMO-DB [6], we achieved higher performance when PCA was applied to the selected feature set, but for LDA higher performance was achieved with the full feature set. For comparison of PCA and LDA techniques, we used feature selection followed by feature reduction.

2.6 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 ω_i. The Bayes decision rule is described as

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

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 $p(x|\omega_i)$ with a diagonal covariance matrix for emotion recognition experiments, which give simple emotion models for our speaker-dependent task.

3 Experimental results

Three sets of emotion recognition experiments were performed: audio, visual, and audio-visual. The experiments were performed with 7 and 4 emotion classes. The 4 emotion classes were obtained by grouping together some of the emotions due to their confusion based on human evaluation experiments. The following 4 emotions were obtained: Displeased (anger, disgust), Gloomy (fear, sadness), Excited (happiness, surprise) and Neural(neutral). In the audio and visual experiments, first the top 40 feature sets were selected. The linear transformation methods were applied to the selected feature sets, and finally single component Gaussian classifiers were used for classification. The audio-visual experiments were performed by combining the two modalities at decision level. The data were divided into four sets in a jack-knife procedure. Each round, three sets were used for training and one set for testing. The experiments were repeated for four different rounds of training and testing sets, and the results averaged.

3.1 Audio experiments

In audio experiments, the top 40 audio features were selected with Plus l-Take Away r algorithm based on Bhattacharyya distance criterion. The feature reduction techniques, PCA and LDA, were applied in the next stage. The classification experiments were performed for seven and four emotion classes with single Gaussian classifiers. The best results with LDA and PCA for 7 emotions are shown in Table 2, and for 4 emotions are shown in Table 4.

The LDA features performed better than PCA features for both 7 and 4 emotion classes. The best overall performance for 7 emotions was obtained with LDA 6 components, and for 4 emotions with LDA 3 components (i.e., N-1). The comparison of Table 1 and Table 2 shows that for 7 emotions the recognition rate achieved with LDA was close to human except for the actors KL and JK. The average recognition rate for LDA was 56 %, and for PCA was 50 % compared to 67 % by human. The best overall result of 59 % was achieved with LDA 4 components. In case of 4 emotion classes, the recognition rate for LDA was close to human except for the actor JK. The average recognition rate over all subjects for LDA was 69 %, and for PCA was 62 % compared to 76 % by human. 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.

3.2 Visual experiments

The top 40 visual features were selected with Plus l-Take Away r algorithm using Bhattacharyya distance as a criterion function. The PCA and LDA were then applied to the selected feature set and finally single component Gaussian classifier was used for classification. The best results with LDA and PCA for 7 emotions are shown in Table 2, and for 4 emotions in Table 4.

The recognition rates for LDA features were higher compared to PCA features. The best overall performance for 7 emotions was obtained with LDA 6 components, and for 4 emotions with LDA 3 components. For the 7 emotion classes, both LDA and PCA performed higher than human recognition rate. The average recognition rate for LDA was 95 %, and for PCA was 92 % compared to 88 % by human. In case of 4 emotions, the recognition rate for LDA was higher than human but was lower for PCA. The average recognition rate for LDA was 98 %, and for PCA was 90 % compared to 91 % by human. The top 40 Bhattacharyya features consisted of 14 features from each of the upper and middle face regions, and 12 features from the lower face region.

3.3 Audio-visual experiments

The audio-visual experiments were performed by combining the two modalities at decision level. The block diagram for audio-visual experiments is shown in Fig. 4. The top 40 audio and top 40 visual features were selected, and then feature reduction techniques were applied. The probability for each of the emotions was calculated for the audio and visual features separately and were multiplied with equal weighting to get the final result. The experiments were performed with single Gaussian classifiers. The best results with LDA and PCA for 7 emotions are shown in Table 2, and for 4 emotions are shown in Table 4.

In audio-visual experiments, the LDA performance was higher than PCA, as in case of audio and visual experiments. The best overall performance for 7 emotions was obtained with LDA 6 components, and for 4 emotions with LDA 3 components. In case of 7 emotions, both LDA and PCA performed better than human. The average recognition rate for LDA was 98 %, and for PCA was 93 %, compared to 92 % by human. For the 4 emotions, LDA performed better than human but the PCA performance was lower. The average recognition rate for LDA was 98 %, and for PCA was 92 %, compared to 95 % by human. The overall classification performance was higher for the audio-visual features compared to the audio and visual features.

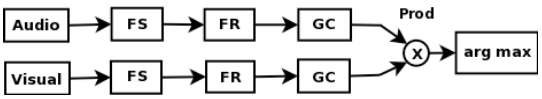

Figure 4: *Diagram of audio and visual fusion at decision level.*

Table 1: *Average human classification accuracy (%) with 7 emotion classes, over 10 participants. Mean is also averaged over 4 subject data with 95 % confidence interval (CI) based on standard error (n=40).*

Human	KL	JE	JK	DC	Mean (±CI)
Audio	53.2	67.7	71.2	73.7	66.5 ± 2.5
Visual	89.0	89.8	88.6	84.7	88.0 ± 0.6
Audio-visual	92.1	92.1	91.3	91.7	91.8 ± 0.1

Table 2: *Average system classification accuracy (%) with 7 emotion classes, over 4 jack-knife tests, with LDA 6 and PCA 10. Mean is averaged over 4 subjects with standard error CI (n=16).*

LDA 6	KL	JE	JK	DC	Mean (±CI)
Audio	35.0	62.5	56.7	70.8	56.3 ± 6.7
Visual	96.7	92.5	92.5	100	95.4 ± 1.6
Audio-visual	97.5	96.7	95.8	100	97.5 ± 0.8
PCA 10	**KL**	**JE**	**JK**	**DC**	**Mean (±CI)**
Audio	30.8	51.7	55.0	62.5	50.0 ± 6.0
Visual	91.7	87.5	89.2	98.3	91.7 ± 2.1
Audio-visual	92.5	86.7	94.2	98.3	92.9 ± 2.1

Table 3: *Average human classification accuracy (%) with 4 emotion classes, over 10 participants. Mean is also averaged over 4 subject data with 95 % confidence interval based on standard error (n=40).*

Human	KL	JE	JK	DC	Mean (±CI)
Audio	63.2	80.9	79.2	82.0	76.3 ± 2.4
Visual	90.6	97.2	90.0	87.4	91.3 ± 1.1
Audio-visual	94.4	98.3	93.5	94.5	95.2 ± 0.6

Table 4: *Average system classification accuracy (%) with 4 emotion classes, over 4 jack-knife tests, with LDA 3 and PCA 7. Mean is averaged over 4 subjects with standard error CI (n=16).*

LDA 3	KL	JE	JK	DC	Mean (±CI)
Audio	60.0	75.8	58.3	80.0	68.5 ± 4.8
Visual	97.5	100	94.2	100	97.9 ± 1.2
Audio-visual	97.5	100	94.2	100	97.9 ± 1.2
PCA 7	**KL**	**JE**	**JK**	**DC**	**Mean (±CI)**
Audio	50.0	58.3	65.8	72.5	61.7 ± 4.3
Visual	80.8	98.3	93.3	88.3	90.2 ± 3.3
Audio-visual	83.3	97.5	95.0	93.3	92.3 ± 2.7

4 Discussion

In the audio, visual, and audio-visual experiments, LDA performed better than PCA. Higher performance was achieved with the visual and audio-visual features compared to the audio features. In seven-emotion classification, the overall performance of both PCA and LDA was lower than human for the audio features, but was higher than human for the visual and audio-visual features. The best overall results were obtained with LDA 6 features, and PCA 10 components. The averaged recognition rates over 4 actors are plotted in Fig. 5. The recognition rate with LDA was lower than human for the audio features, but higher recognition rates were achieved for the visual and audio-visual features. For 4 emotion classes, the best performance was achieved with LDA 3 features, and PCA 7 components. Average results for 4 actors are plotted in Fig. 6. The recognition rate with audio was lower than human for both PCA and LDA, but was higher than human with the visual and audio-visual features for LDA. The PCA results were close to human for the visual and audio-visual features. The performance of LDA 3 components was comparable to human and even higher for the visual and audio-visual features. The higher performance of machine compared to human for the visual and audio-visual data was due to the difference in training data, i.e. machine was trained on a large part of data but humans were trained on a small amount of data, the task was discrete emotion classification, and the emotions may not be properly acted.

While we expect the feature extraction, selection and reduction techniques to transfer to a speaker-independent task, our preliminary trials show a significant reduction in performance. We will be looking at appropriate methods of speaker normalization. There are also opportunities to improve the overall classification accuracy by using more sophisticated schemes, such as GMM or SVM.

5 Conclusions

In classification tests on the British English audio-visual emotional database, for the speaker-dependent system, a good recognition rate was achieved, comparable to human. The LDA outperformed PCA with the top 40 features selected by Bhattacharyya distance. Results show that both audio and visual information were useful for emotion recognition, although visual features performed much better here, perhaps because the subjects were more expressive facially compared to their voice. The energy and MFCC features were identified as most important audio features, although pitch and duration features also contributed. Most important visual features involved the mean marker y-coordinate, i.e. vertical movement of face regions was key for emotion recognition. For 7 emotion classes, average recognition rates of 56 %, 95 % and 98 % were achieved for the audio, visual and audio-visual features with LDA 6 components, compared to 67 %, 88 %, and 92 % by humans. For the four class case, average recognition rates of 69 %, 98 % and 98 % were achieved for the audio, visual and audio-visual features with LDA 3 components, compared to 76 %, 91 %, and 95 % by humans. A possible reason for higher performance of system was that the evaluators were not given equal opportunity to exploit the speaker-specific characteristics of the data. Some potential remains for improving performance

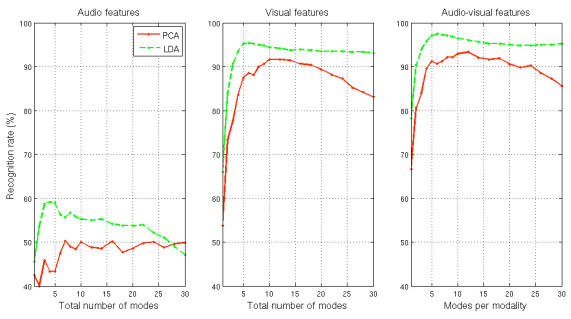

Figure 5: *Average recognition rate (%) over 4 actors with audio, visual, and audio-visual features for 7 emotion classes.*

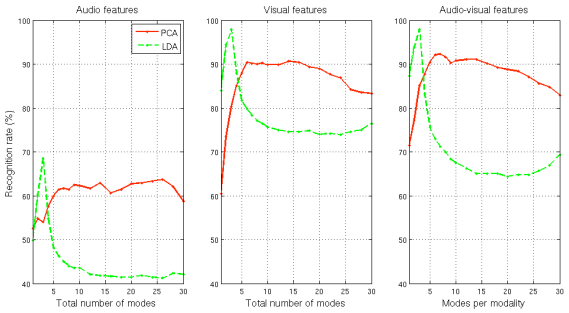

Figure 6: *Average recognition rate (%) over 4 actors with audio, visual, and audio-visual features for 4 emotion classes.*

in the audio modality. Future work involves speaker-independent experiments, using more data and other classifiers, such as GMM or SVM.

References

[1] C. Busso and S. Narayanan, "Interrelation between speech and facial gestures in emotional utterances: A single subject study," *IEEE Transactions on ASLP*, vol. 15, no. 8, pp. 2331–2347, 2007.

[2] A. Ortony et al., "What's Basic About Basic Emotions?" *Psychological Review*, vol. 97, no. 3, pp. 315–331, 1990.

[3] K. Scherer, "What are emotions? and how can they be measured?" *Social Science Information*, vol. 44, no. 4, pp. 695–729, 2005.

[4] J. Russell, L. Ward, and G. Pratt, "Affective Quality Attributed to Environments: A Factor Analytic Study," *Environment and Behaviour*, vol. 13, no. 3, pp. 259–288, 1981.

[5] E. Douglas-Cowie, R. Cowie, and M. Schroeder, "A New Emotional Database: Considerations, Sources and Scope," in *Proc. ISCA Workshop Speech and Emotion: A conceptual framework for research*, 2000, pp. 39–44.

[6] F. Burkhardt et al., "A Database of German Emotional Speech," in *Proc. Interspeech*, 2005, pp. 1517–1520.

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

[8] G. Chetty et al., "A multilevel fusion for audiovisual emotion recognition," in *Proc. AVSP*, 2008, pp. 115–120.

[9] I. Luengo, E. Navas et al., "Automatic Emotion Recognition using Prosodic Parameters," in *Proc. Interspeech*, 2005, pp. 493–496.

[10] D. Ververidis and C. Kotropoulos, "Emotional speech classification using Gaussian mixture models," in *Proc. ISCAS*, 2005, pp. 2871–2874.

[11] L. Vidrascu et al., "Detection of real-life emotions in call centers," in *Proc. Interspeech*, 2005, pp. 1841–1844.

[12] M. Borchert and A. Düsterhöft, "Emotions in Speech - Experiments with Prosody and Quality Features in Speech for Use in Categorical and Dimensional Emotion Recognition Environments," in *Proc. NLP-KE*, 2005, pp. 147–151.

[13] A. Nogueiras, A. Moreno et al., "Speech Emotion Recognition Using Hidden Markov Models," in *Proc. Eurospeech*, 2001, pp. 2679–2682.

[14] Y. Lin and G. Wei, "Speech Emotion Recognition Based on HMM and SVM," in *Proc. 4th Int. Conf. on Mach. Learn. and Cybernetics*, 2005, pp. 4898–4901.

[15] Y. Kao and L. Lee, "Feature Analysis for Emotion Recognition from Mandarin Speech Considering the Special Characteristics of Chinese Language," in *Proc. Interspeech*, 2006, pp. 1814–1817.

[16] D. Neilberg, K. Elenius et al., "Emotion Recognition in Spontaneous Speech Using GMMs," in *Proc. Interspeech*, 2006, pp. 809–812.

[17] C. Chen, Y. Huang, and P. Cook, "Visual/Acoustic emotion recognition," in *Proc. Int. Conf. on Multimedia & Expo*, 2005, pp. 1468–1471.

[18] C. Busso, Z. Deng, S. Yildirim et al., "Analysis of Emotion Recognition using Facial Expressions, Speech and Multimodal Information," in *Proc. ACM Int. Conf. on Multimodal Interfaces*, 2004, pp. 205–211.

[19] M. Song, C. Chen, and M. You, "Audio-visual based emotion recognition using tripled Hidden Markov Model," in *Proc. ICASSP*, vol. 5, 2004, pp. 877–880.

[20] V. Sethu, E. Ambikairajah, and J. Epps, "Phonetic and speaker variations in automatic emotion classification," in *Proc. Interspeech*, 2008, pp. 617–620.

[21] B. Schuller, G. Rigoll, and M. Lang, "Hidden Markov Model-Based Speech Emotion Recognition," in *Proc. ICASSP*, vol. 2, 2003, pp. 1–4.

[22] B. Schuller, M. Wimmer et al., "Detection of security related affect and behaviour in passenger transport," in *Proc. Interspeech*, 2008, pp. 265–268.

[23] J. Campbell, "Speaker Recognition: A Tutorial," in *Proc. IEEE*, vol. 85, no. 9, 1997, pp. 1437–1462.

[24] S. Haq, P. Jackson, and J. Edge, "Audio-visual feature selection and reduction for emotion classication," in *Proc. AVSP*, 2008, pp. 185–190.

[25] *3dMD 4D Capture System*. Online: http://www.3dmd.com, accessed on 3 May, 2009.

[26] M. Swerts and E. Krahmer, "Gender-related differences in the production and perception of emotion," in *Proc. Interspeech*, 2008, pp. 334–337.

[27] A. L. Edwards, *Experimental Design in Psychological Research*. New York: Holt, Rinehart and Winston, 1962.

[28] M. Huckvale, *Speech Filing System*. UCL Dept. of Phonetics & Linguistics, UK. Online: http://www.phon.ucl.ac.uk/resource/sfs/, accessed on 3 May, 2009.

[29] S. Young and P. Woodland, *Hidden Markov Model Toolkit*. Cambridge University Engineering Department, UK. Online: http://htk.eng.cam.ac.uk/, accessed on 3 May, 2009.

[30] C. Chen, *Pattern Recognition and Signal Processing*. Sijthoff & Noordoff, The Netherlands, 1978.

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

[32] R. Duda, P. Hart, and D. Stork, *Pattern Classication*. John Wiley & Sons, Inc. USA, Canada, 2001.