Proceedings of the Pakistan Academy of Sciences 52 (1): 15–25 (2015)
Copyright © Pakistan Academy of Sciences
ISSN: 0377 - 2969 (print), 2306 - 1448 (online)

Pakistan Academy of Sciences

Research Article

Speaker-Dependent Human Emotion Recognition in Unimodal and Bimodal Scenarios

Sanaul Haq[1], Tariqullah Jan[2*], Muhammad Asif[1], Amjad Ali[3], and Naveed Ahmad[4]

[1]Department of Electronics, University of Peshawar, Pakistan
[2]Department of Electrical Engineering, University of Engineering & Technology, Peshawar, Pakistan
[3]Department of Electrical Engineering, Sarhad University of Science and Information Technology, Peshawar, Pakistan
[4] Department of Computer Science, University of Peshawar, Peshawar, Pakistan

Abstract: This paper presents an efficient technique of human emotion recognition using the audio and visual modalities for the speaker-dependent scenario. To achieve better emotion classification different audio and visual features were extracted. The feature selection was performed using the Plus l-Take Away r algorithm based on two criteria: Mahalanobis distance and KL-divergence. The feature selection was followed by feature reduction using PCA and LDA, and classification using the Gaussian classifier. Emotion classification was performed using both the unimodal and bimodal approaches. In the bimodal approach, audio and visual features were fused at two levels: feature and decision. The emotion classification performance comparable to humans was achieved on the SAVEE database for the unimodal and bimodal scenarios.

Keywords: Human emotion recognition, audio and visual feature selection, Plus l-Take Away r algorithm, feature reduction, PCA, LDA, classification, fusion of modalities

1. INTRODUCTION

In recent years, humans have developed advanced technologies which have made the human-computer interaction more attractive for humans. Primarily, human beings use speech to communicate with each other, but the complete message cannot be conveyed only through verbal content. Some additional information are required to completely understand the message, these includes vocalized emotions, facial expressions, hand gestures and body language [1, 2]. From human's point of view, the human-machine interaction will be more interesting if machines can recognize human emotions and are able to respond accordingly [3]. On the other hand, communication will be more reliable if machines are able to understand human emotions [4].

Automatic emotion recognition covers a wide range of applications including automobile systems, customer services systems, health care, and game and film industries [5]. In recent years, researchers from different disciplines have taken interest in the field of human emotion recognition and significant progress has been made in several areas including recording of emotional databases, feature extraction, feature selection, and classification [5, 6].

Previous studies have mainly focused on unimodal approaches (e.g., speech, facial expressions) for emotion classification. The modalities have been analyzed independently and the interrelation between different modalities has not been explored. In actual fact, speech and facial expressions are highly correlated, and emotions and linguistic content influence the relationship between these two modalities [7]. Humans utilize both speech and gesture to express their emotions, and for this reason an ideal emotion recognizer should be based on both verbal and non-verbal

Received, October 2014; Accepted, January 2015

*Corresponding author: Tariqullah Jan; Email: tariqullahjan@nwfpuet.edu.pk information [1, 8].

One of the factors upon which the reliability of an emotion recognizer depends is the quality of emotional data used to build the emotion recognition system. Examples of audio databases are the AIBO corpus and Berlin database [9, 10]. Visual databases include the Cohn-Kanade and MMI databases [11, 12]. The audio-visual databases are Facial Motion Capture database, GEMEP, and Belfast Naturalistic database [7, 13, 14]. The emotional databases are either acted or natural.

Audio and visual features of various types have been analyzed to achieve better emotion classification results. Important audio features are pitch, formants, duration, spectral energy, and Mel frequency cepstral coefficients (MFCCs). These features have been used both at utterance-level [15] and frame-level [16]. The vision-based techniques are mainly based on facial expressions, since face plays a vital role in conveying emotions. Facial features can be subdivided into two broad categories: geometric and appearance [17]. The technique [17] is based on geometric features, while [18] uses appearance features.

In the field of pattern recognition, feature selection and reduction techniques are used to remove the uninformative, redundant and noisy information. These techniques improve the classification accuracy and computational efficiency. The feature search techniques used in the field of emotion recognition include sequential floating forward selection [19], genetic algorithms [20] and best-first [21]. Feature reduction techniques include PCA and LDA [22, 23].

In any pattern recognition problem, the choice of classifier is very important. These classifiers include hidden Markov model [16], neural network [24], support vector machine [25], and adaptive boosting [26]. Multimodal approaches have been adopted to improve the emotion classification by fusion of data at feature [2], decision [27], and model [28] levels.

In the field of emotion recognition, most research is based on unimodal approaches (e.g., audio or visual), and less progress has been made in terms of multimodal approaches. This research aims to achieve better classification accuracy using the bimodal approach with appropriate feature selection and reduction techniques. The following sections present the SAVEE database, method, experimental results, discussion and conclusion.

2. SAVEE DATABASE

We used Surrey Audio-Visual Expressed Emotion (SAVEE) database [29] for our analysis. The database consists of data from four British male speakers in Ekman's six basic emotions (anger, fear, disgust, sadness, happiness and surprise) plus neutral [30]. The text material consisted of 15 phonetically-diverse sentences for each of the six emotions and 30 sentences for the neutral. The distribution of sentences in this way resulted in 120 utterances per actor and 480 utterances in total. A subject with four different emotions (anger, fear, happiness, and surprise) from the SAVEE database is shown in Fig. 1.

The sampling rate for audio data was 44.1 kHz, while that for video was 60 fps. To extract facial features, each actor's frontal face was painted with 60 markers. The SAVEE database was evaluated at utterance level by 20 subjects (10 male, 10 female). Each actor's data were evaluated by 10 subjects. The classification accuracy for seven emotion classes averaged over four actor's data and 10 evaluators was 66.5 % (Standard Error (SE): 2.5) for audio, 88.0 % (SE: 0.6) for visual, and 91.8 % (SE: 0.1) for audio-visual data.

3. METHOD

The speaker-dependent emotion classification was performed using a three steps method: feature extraction, feature selection and reduction, and classification, as shown in Fig. 2.

3.1 Feature Extraction

For emotion classification in the speaker-dependent scenario, features were extracted at utterance-level consisting of 106 audio and 240 visual features. The details of the audio and visual features are given below.

3.1.1 Audio Features

The audio features were related to pitch (f_0), duration, energy and spectral envelop. These features were extracted using the Speech Filing System [31] and HTK [32].

Pitch Features: The fundamental frequency (f_0) was extracted using the Speech Filing System. Features related to f_0 contour were minimum and maximum of mel frequency; mean and standard deviation of first and second Gaussian of mel frequency; minimum, maximum, mean and standard deviation of mel frequency first order difference.

Duration Features: Phone labels were used to extract duration features. The phone labeling was performed in two steps: first HTK was used for automatic labeling of the audio, and second the Speech Filing System was used to correct the automatic phone labels. The extracted 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 phone duration ratio, speech rate, voiced-speech-to-sentence duration ratio, and unvoiced-speech-to-sentence duration ratio.

Energy Features: The extracted features were mean and standard deviation of total log energy; minimum, maximum, range, mean and standard deviation of normalised 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; minimum, maximum, range, mean and standard deviation of first order difference of normalised energies in the original speech signal and speech signal in the same frequency bands.

Spectral features: The spectral envelope features were mean and standard deviation of 12 MFCCs.

The range ($Range$), mean (μ) and standard deviations (σ) of different audio features were calculated using the following relations

$$Range = max - min \tag{1}$$

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2}$$

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2} \tag{3}$$

where max, min and n denote the maximum value, minimum value, and number of samples of a feature x, respectively.

3.1.2 Visual Features

The visual features were extracted by painting markers on actors' forehead, eyebrows, cheeks, lips and jaw. The markers were automatically tracked and normalized relative to a reference point at the bridge of the nose [7]. Finally, 240 visual features were obtained from the 2D marker coordinates as the mean and standard deviation of the marker coordinates.

The mean (μ_x) and standard deviations (σ_x) of marker x coordinates were computed by

$$\mu_x = \frac{1}{m}\sum_{j=1}^{m} x_j \tag{4}$$

$$\sigma_x = \sqrt{\frac{1}{m-1}\sum_{j=1}^{m}\left(x_j - \mu_x\right)^2} \tag{5}$$

The mean (μ_y) and standard deviations (σ_y) of marker y coordinates were calculated using the relations

$$\mu_y = \frac{1}{m}\sum_{j=1}^{m} y_j \tag{6}$$

$$\sigma_y = \sqrt{\frac{1}{m-1}\sum_{j=1}^{m}\left(y_j - \mu_y\right)^2} \tag{7}$$

where m denotes the number of marker x and y coordinates.

3.2 Feature Selection and Reduction

For the speaker-dependent emotion classification, we adopted a two-step process: feature selection with Plus l-Take Away r algorithm, followed by feature reduction with PCA and LDA.

3.2.1 Feature Selection

The Plus l-Take Away r algorithm is a feature search method based on some criterion function [33]. It combines the sequential forward selection (SFS) and sequential backward selection (SBS) algorithms to achieve better results. At each step, l numbers of features are included to the current feature set and r numbers of features are discarded. The process continues until the required feature set size is achieved. The feature search was performed with l = 2 and $r = 1$, i.e., one feature was added at each step. We used this algorithm for feature selection based on two different criteria: Mahalanobis distance and KL-divergence [34]. These distance measures have been used as dissimilarity measures in different applications including speaker recognition [34], emotion recognition [35] and texture retrieval [36]. Feature selection was performed over the full range of audio, visual and audio-visual feature sets.

Mahalanobis Distance: It defines the similarity between two classes [37]. For two normally distributed classes ω_i and ω_j , Mahalanobis distance is defined as

$$d_{Mah} = \sqrt{(\mu_i - \mu_j)^T (P_i\Sigma_i + P_j\Sigma_j)^{-1} (\mu_i - \mu_j)} \qquad (8)$$

where μ_i and μ_j are the means, Σ_i and Σ_j are the covariances, and P_i and P_j are the prior probabilities of classes ω_i and ω_j, respectively.

Kullback-Leibler (KL) divergence measure: It provides the dissimilarity between two classes based upon information theory [38]. For two normally distributed classes ω_i and ω_j, the KL-divergence is defined as

$$J_{Div} = \frac{1}{2} tr\left[(\Sigma_i^{-1} + \Sigma_j^{-1})(\mu_i - \mu_j)(\mu_i - \mu_j)^T\right]$$
$$+\tfrac{1}{2} tr[(\Sigma_i - \Sigma_j)(\Sigma_j^{-1} - \Sigma_i^{-1})] \qquad (9)$$

where tr denotes the matrix trace operation. The KL-divergence relation consists of two terms. The first term defines the difference between two classes based on class means, while the second term provides the difference using the covariance matrices.

Features were normalized prior to applying the feature selection by using the Z-norm (i.e., mean subtraction and division by standard deviation). The Mahalanobis distance and KL-divergence measure provide the separability between two classes. For m number of classes the separability measure is obtained by averaging it over all binary combinations of the classes.

3.2.2 Feature Reduction

Statistical methods can be used to reduce the dimensionality of a feature set. PCA technique [39] is used to extract the important characteristics of high-dimensional data and to remove the uninformative and noisy data. The LDA method [40] provides the separation between classes based on the ratio of between-class variance to within-class variance. For feature reduction, we applied both PCA and LDA as linear transformation techniques to the selected features.

3.3 Classification and Fusion of Modalities

The Gaussian classifier utilizes the Bayes decision theory for classification. It is assumed that the class-conditional probability $p(x|\omega_i)$ have Gaussian distribution for each class ω_i. The Bayes decision rule is described as

$$i_{Bayes} = arg \max_i P(\omega_i|x) =$$
$$arg \max_i p(x|\omega_i)P(\omega_i) \qquad (10)$$

where $P(\omega_i|x)$ denotes the posterior probability, and $P(\omega_i)$ defines the prior class probability. A single component Gaussian was used to model each emotion class ω_i using a diagonal covariance matrix.

The audio-visual emotion classification was performed by the fusion of modalities at feature-level and at decision-level.

4. EXPERIMENTAL RESULTS

Experiments were performed both in the unimodal and bimodal scenarios. In unimodal scenario, experiments were conducted using the audio or visual features. In the case of bimodal scenario, we investigated the audio-visual fusion both at feature-level and at decision-level. Each speaker data were divided into four sets, and experiments were conducted with four different training and testing sets. The results were averaged over four tests. For each experiment, three sets were used for training and one set for testing. Finally, the results were averaged over four speakers' data.

4.1 Audio Emotion Classification

The average classification accuracies achieved for seven emotions using the features selected by Mahalanobis distance and KL-divergence are plotted in Fig. 3a. In general, the features selected by Mahalanobis distance performed better than those by KL-divergence. The LDA-transformed features performed better for up to 50 selected features, after which its performance deteriorated. For the LDA-transformed features, the best classification accuracy with Mahalanobis distance was 61 % (SE: 7.5) using 25 selected features, and with KL-divergence it was 55 % (SE: 6.0) using 40 selected features. In the case of PCA-transformed features, the classification performance improved with an increasing number of selected features. The best classification performance for the Mahalanobis distance was 54 % (SE: 8.0) using 95 selected features, and for the KL-divergence it was 56 % (SE: 7.6) using 70 selected features. In general, the LDA technique performed better for fewer selected features (up to

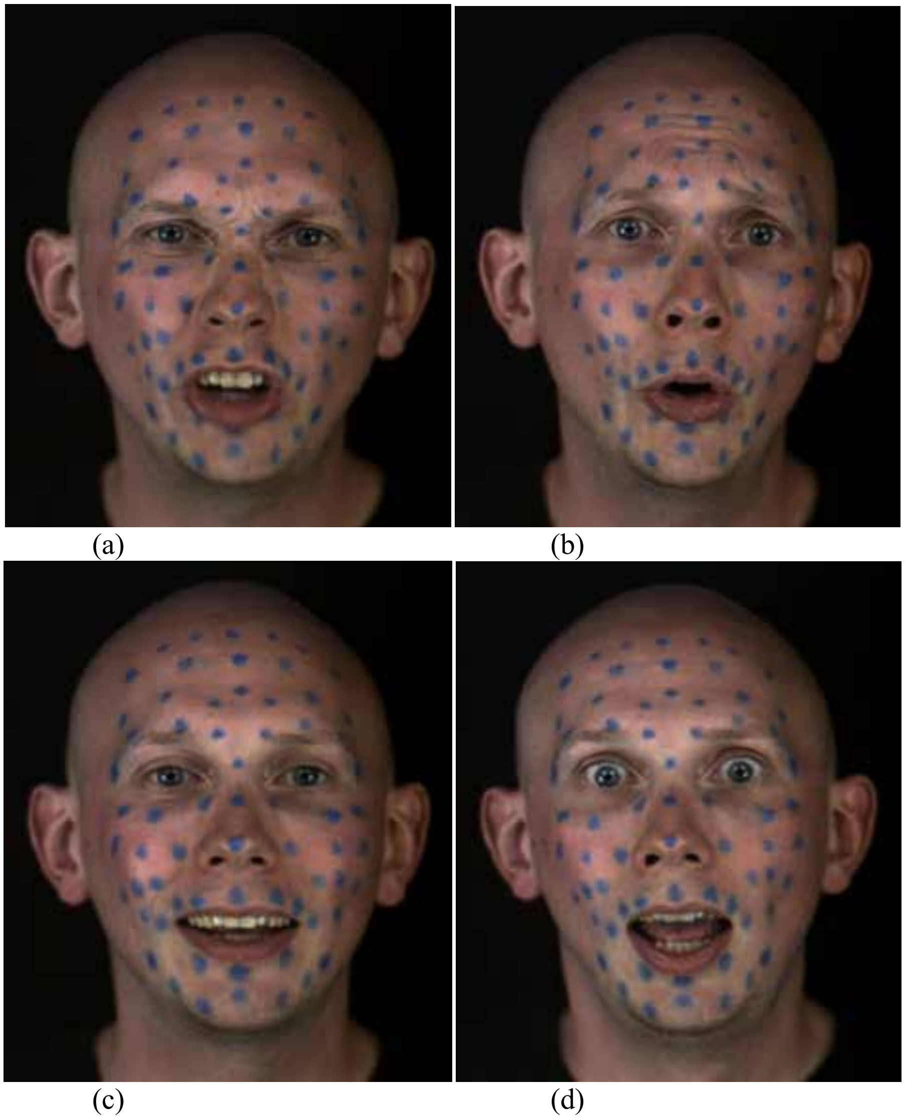


Fig. 1. Different expressed emotions of a subject from SAVEE database: (a) anger, (b) fear, (c) happiness, and (d) surprise.

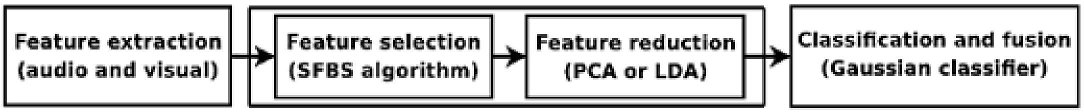
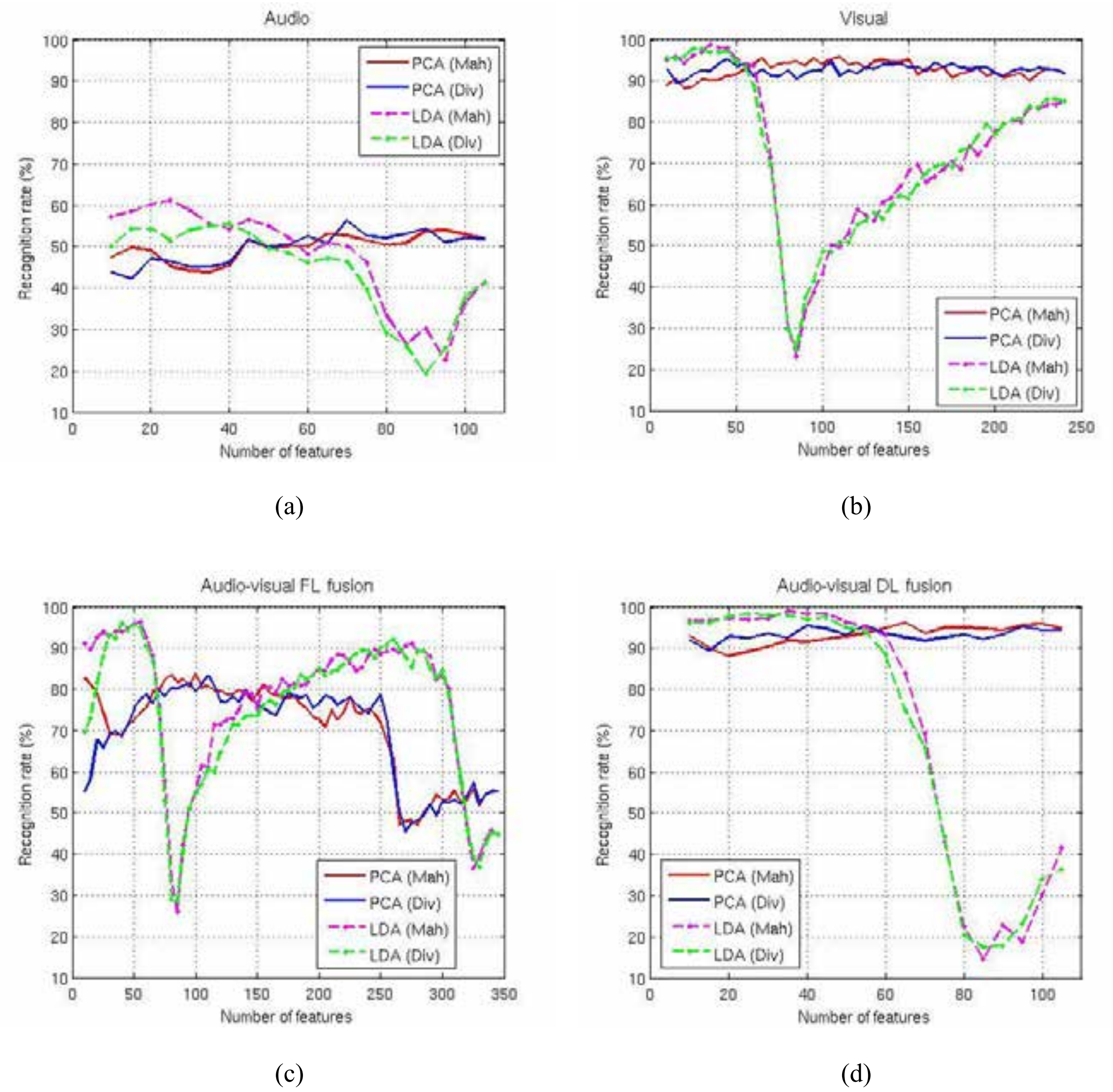

Feature extraction (audio and visual)
Feature selection (SFBS algorithm)
Feature reduction (PCA or LDA)
Classification and fusion (Gaussian classifier)

Fig. 2. Block diagram of emotion classification method for the speaker-dependent scenario.



(a) (b)

(c) (d)

Fig. 3. Average classification accuracy (%) achieved for seven emotions on SAVEE database for the speaker-dependent scenario. The audio, visual, and audio-visual features were selected with the Mahalanobis (Mah) distance and KL-divergence (Div). Here FL: feature-level, DL: decision-level.

50), while the PCA method achieved better results for larger sets of selected features. For the Mahalanobis distance criterion, the performance of LDA-transformed features was better than PCA-transformed features. In the case of KL-divergence, the classification performance of PCA-transformed and LDA-transformed features was comparable.

In the case of Mahalanobis distance, the best 25 selected features included 4 pitch, 16 energy and 5 MFCC features, and the top 95 selected features consisted of 8 pitch, 58 energy, 10 duration and 19 MFCC features. For KL-divergence, the top 40 selected features included 5 pitch, 33 energy, 1 duration and 1 MFCC features, and the best 70 selected features consisted of 8 pitch, 45 energy, 5 duration and 12 MFCC features. In general, the energy features were found to be most important, followed by MFCC, pitch and duration.

4.2 Visual Emotion Classification

Recognition rates for seven emotions using the visual features selected by Mahalanobis distance and KL-divergence are plotted in Fig. 3b. The LDA-transformed features performed better for up to 60 selected features, and from that point onwards the performance dropped sharply. The classification accuracy recovered beyond 100 selected features but was lower than the best performance. For the LDA-transformed features, the performance of features selected by Mahalanobis distance and KL-divergence was quite close. The best classification performance for the Mahalanobis distance was 99% (SE: 1.2) using 35 selected features, and for the KL-divergence it was 98 % (SE: 2.0) using 30 selected features. For the PCA-transformed features, the best recognition rate for Mahalanobis distance was 96 % (SE: 2.2) using 110 selected features, and for KL-divergence it was 95 % (SE: 2.4) using 45 selected features. The overall performance of LDA-transformed features was higher for fewer selected features (up to 60), while the PCA-transformed features performed better for larger sets of selected features. The performance of LDA-transformed features was better than the PCA-transformed features for both the Mahalanobis distance and KL-divergence criteria.

In the case of Mahalanobis distance, the top 35 selected features included 11 mean of marker x coordinates, 8 mean of marker y coordinates, 12 standard deviation of marker x coordinates and 4 standard deviation of marker y coordinates. The best 110 selected features consisted of 26 mean of marker x coordinates, 28 mean of marker y coordinates, 33 standard deviation of marker x coordinates and 23 standard deviation of marker y coordinates. For KL-divergence, the best 30 selected features included 6 mean of marker x coordinates, 11 mean of marker y coordinates, 9 standard deviation of marker x coordinates and 4 standard deviation of marker y coordinates. The top 45 selected features consisted of 9 mean of marker x coordinates, 12 mean of marker y coordinates, 14 standard deviation of marker x coordinates and 10 standard deviation of marker y coordinates. In general, standard deviation of marker x coordinates were the best features, followed by mean of marker y coordinates, mean of marker x coordinates, and standard deviation of marker y coordinates. The forehead and eyebrow areas were more discriminative as compared to the cheek area and lower part of the face.

4.3 Audio-Visual Emotion Classification

The audio-visual emotion classification was performed by fusion of two modalities at feature-level and decision-level. The results were obtained for seven emotions using the features selected by Mahalanobis distance and KL-divergence. The decision-level fusion was performed using the equal numbers of selected audio and visual features. For decision-level fusion, the probabilities obtained for the audio and visual modalities were multiplied with equal weighting to obtain the audio-visual classification result.

4.3.1 Feature-Level Fusion

The average classification accuracies for the audio-visual fusion at feature-level are plotted in Fig. 3c. For the feature-level fusion, the best result achieved with the LDA-transformed features for Mahalanobis distance was 97 % (SE: 2.2) using 55 selected features, and for KL-divergence it was 96% (SE: 2.5) using 40 selected features. In the case of PCA-transformed features, the best classification score for Mahalanobis distance was 84% (SE: 4.3) with 100 selected features, and for KL-divergence it was 83 % (SE: 4.7) with 110 selected features.

For the Mahalanobis distance, the best 55 selected features included 19 audio and 36 visual features. The audio features consisted of 2 pitch, 13 energy and 4 MFCC, and visual features included 11 mean of marker x coordinates, 7 mean of marker y coordinates, 9 standard deviation of marker x coordinates and 9 standard deviation of marker y coordinates. The top 100 features selected by Mahalanobis distance consisted of 6 audio and 94 visual features, where audio features included 1 pitch, 4 energy and 1 MFCC, and visual features consisted of 24 mean of marker x coordinates, 19 mean of marker y coordinates, 33 standard deviation of marker x coordinates and 18 standard deviation of marker y coordinates. For the KL-divergence criterion, the top 40 selected features included 9 audio and 31 visual features, where audio features consisted of 4 pitch, 2 energy, 1 duration and 2 MFCC, and visual features included 10 mean of marker x coordinates, 4 mean of marker y coordinates, 8 standard deviation of marker x coordinates and 9 standard deviation of marker y coordinates. The best 110 features selected by KL-divergence consisted of 3 audio and 107 visual features, where audio features included 2 energy and 1 MFCC, and the visual features consisted of 33 mean of marker x coordinates, 19 mean of marker y coordinates, 31 standard deviation of marker x coordinates and 24 standard deviation of marker y coordinates. The proportion of audio and visual features selected with feature-level fusion indicates that the visual features were more discriminative as compared to the audio.

4.3.2 Decision-Level Fusion

The average classification accuracies for audio-visual fusion at decision-level are plotted in Fig. 3d. For the decision-level fusion, the best accuracy achieved with the LDA-transformed features for Mahalanobis distance was 99 % (SE: 1.0) with 35 selected features, and for KL-divergence it was 98% (SE: 1.3) with 25 selected features. For the PCA-transformed features, the best classification performance achieved for Mahalanobis distance was 96 % (SE: 2.5) using 65 selected features, and for KL-divergence it was 95% (SE: 1.6) using 55 selected features.

For the Mahalanobis distance criterion, the best 35 audio features consisted of 4 pitch, 22 energy, 1 duration and 8 MFCC, and the corresponding 35 visual features included 11 mean of marker x coordinates, 8 mean of marker y coordinates, 12 standard deviation of marker x coordinates and 4 standard deviation of marker y coordinates. The top 65 audio features selected by Mahalanobis distance consisted of 7 pitch, 42 energy, 7 duration and 9 MFCC, while the corresponding 65 visual features included 19 mean of marker x coordinates, 13 mean of marker y coordinates, 24 standard deviation of marker x coordinates and 9 standard deviation of marker y coordinates. For the KL-divergence criterion, the best 25 audio features consisted of 4 pitch, 16 energy, 2 duration and 3 MFCC, and the corresponding 25 visual features included 8 mean of marker x coordinates, 5 mean of marker y coordinates, 6 standard deviation of marker x coordinates and 6 standard deviation of marker y coordinates. The top 55 audio features selected by KL-divergence consisted of 7 pitch, 37 energy, 5 duration and 6 MFCC, while the corresponding 55 visual features consisted of 15 mean of marker x coordinates, 14 mean of marker y coordinates, 16 standard deviation of marker x coordinates and 10 standard deviation of marker y coordinates.

The overall performance of decision-level fusion was better than feature-level fusion for both the Mahalanobis distance and KL-divergence criteria. In the case of LDA-transformed features, the performance of the two fusion methods was quite close, but for the PCA-transformed features the decision-level fusion performed much better than the feature-level fusion. Among the selected audio features, the energy features were the most important, followed by MFCC, pitch and duration. The most important visual features were standard deviation of marker x coordinates, followed by mean of marker x coordinates, mean of marker y coordinates and standard deviation of marker y coordinates.

5. DISCUSSION

Human and machine classification performance on the SAVEE database for the speaker-dependent task indicated similar patterns of accuracy for the audio, visual and audio-visual (decision-level fusion) modalities. The comparison of human and machine performance is given in Table 1. In general, the visual modality performed better than the audio, and the overall performance improved when the two modalities were combined at decision-level. An overall improvement was observed in the standard error between visual and audio-visual modalities for both Mahalanobis distance and KL-divergence criteria, although no significant improvement was observed in the classification accuracy. The possible reason was that a high level of accuracy was obtained for the visual modality alone and no further improvement was obtained when the visual modality was combined with the audio modality.

Table 1. Average classification accuracy in percentage (± standard error) for seven emotions on SAVEE database achieved by human and machine. The audio, visual, and audio-visual (decision-level fusion) features were selected with Mahalanobis distance and KL-divergence measure, followed by linear transformation with PCA and LDA.

Modality	Human	Machine			
		Mahalanobis distance		KL-divergence measure	
		PCA	LDA	PCA	LDA
Audio	67 ± 2.5	54 ± 8.0	61 ± 7.5	56 ± 7.6	55 ± 6.0
Visual	88 ± 0.6	96 ± 2.2	99 ± 1.2	95 ± 2.4	98 ± 2.0
Audio-Visual	92 ± 0.1	96 ± 2.5	99 ± 1.0	95 ± 1.6	98 ± 1.3

The overall performance of LDA-transformed features was better than the PCA-transformed features for the features selected by both Mahalanobis distance and KL-divergence criteria. In the case of LDA-transformed features, the Mahalanobis distance performed much better than the KL-divergence for the audio modality, while for visual and audio-visual modalities the performance of two criteria was comparable. For the PCA-transformed features, the performance of Mahalanobis distance was comparable to KL-divergence. The overall performance of Mahalanobis distance was better than KL-divergence for both the LDA-transformed and PCA-transformed features.

For the Mahalanobis distance and KL-divergence criteria, the LDA-transformed features performed better for a small number of selected features (approximately up to 50), while the performance of PCA-transformed features improved with an increasing number of selected features. In general, the decision-level fusion performed better than feature-level fusion. The overall performance of the two fusion methods was comparable for the LDA-transformed features. In the case of PCA-transformed features, the decision-level fusion performed much better than feature-level fusion.

Differences existed between the classification accuracies of human and machine. The possible reasons are differences in the training data, i.e., the machine was trained/tested in a speaker-dependent scenario but humans were adapted to a small amount of data and evaluation was performed in a speaker-independent fashion, the task was discrete emotion classification, and the expressed emotions may have been lacking in naturalness.

6. CONCLUSIONS

For the speaker-dependent task using the SAVEE database, the average recognition rates comparable to humans were achieved for the audio, visual and audio-visual modalities. A baseline system consisting of feature selection using the Plus l-Take Away *r* algorithm, feature reduction using the PCA and LDA, and classification using a Gaussian classifier was tested and high performance was achieved.

The visual modality performed better than the audio, and the overall accuracy improved for the bimodal scenario. The LDA-transformed features performed better than the PCA-transformed features for the Mahalanobis distance and KL-divergence. The decision-level fusion performed better than the feature-level fusion. The important audio features were the energy features, followed by spectral, pitch and duration features, while for the visual modality, features from the forehead and eyebrow areas were found most discriminative. In general, the visual features were more discriminative compared to the audio features.

The overall best results for the unimodal and bimodal scenarios were achieved with the features selected by Mahalanobis distance criterion. The best classification accuracy for the audio modality was 61 % (SE: 7.5) using 25 selected features, for the visual modality it was 99 % (SE: 1.2) using 35 selected features, and for the bimodal scenario (decision-level fusion) it was 99 % (SE: 1.0) using 35 selected features per modality. These results were achieved with 6 LDA-transformed features using a Gaussian classifier.

In Future we will extend the speaker-dependent emotion classification to the speaker independent scenario. The feature extraction, feature selection and reduction, and classification techniques will be transferred to the speaker-independent task. For this purpose, additional audio and visual features will be extracted, and appropriate speaker normalization techniques will be used. To achieve better classification performance, we will adopt more sophisticated schemes such as SVM for the classification.

7. REFERENCES

1. Sebe, N., I. Cohen & T.S. Huang. Multimodal Emotion Recognition. In: *Handbook of Pattern Recognition and Computer Vision*. World Scientific (2005).
2. Kessous, L., G. Castellano & G. Caridakis. Multimodal Emotion Recognition in Speech-Based Interaction using Facial Expression, Body Gesture and Acoustic Analysis. *Journal on Multimodal User Interfaces* 3(1): 33–48 (2010).
3. Picard, R.W. *Affective Computing*. MIT Press, Cambridge (1997).
4. Cowie, R., E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz & J.G. Taylor. Emotion Recognition in Human-Computer Interaction. *IEEE Signal Processing Magazine* 18(1): 32–80 (2001).
5. Zeng, Z., M. Pantic, G.I. Roisman & T.S. Huang. A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31(1): 39–58 (2009).
6. Haq, S. & P.J.B. Jackson. Multimodal Emotion Recognition. In: *Machine Audition: Principles, Algorithms and Systems* IGI Global Publisher, USA (2010).
7. Busso, C. & S.S. Narayanan. Interrelation Between Speech and Facial Gestures in Emotional Utterances: A Single Subject Study. *IEEE Transactions on Audio, Speech and Language Processing* 15(8): 2331–2347 (2007).
8. Pantic, M., N. Sebe, J.F. Cohn & T. Huang. Affective Multimodal Human Computer Interaction. In: *Proceedings of ACM International Conference on Multimedia*, Hilton, Singapore, p. 669–676 (2005).
9. Batliner, A., C. Hacker, S. Steidl, E. Noth, S. D'Arcy, M. Russel & M. Wong. "You Stupid Tin Box" - Children Interacting with the AIBO Robot: A Cross-Linguistic Emotional Speech Corpus. In *Proc. International Conference on Language Resources and Evaluation*, Lisbon, Portugal, p. 171–174 (2004).
10. Burkhardt, F., A. Paeschke, M. Rolfes, W. Sendlmeier & B. Weiss. A Database of German Emotional Speech. In *Proc. Interspeech*, Lisbon, Portugal, p. 1517–1520 (2005).
11. Kanade, T., J. Cohn & Y. Tian. Comprehensive Database for Facial Expression Analysis. In *Proc. IEEE International Conference on Face and Gesture Recognition*, p. 46–53, Grenoble, France (2000).
12. Pantic, M, M. Valstar, R. Rademaker & L. Maat. Web-Based Database for Facial Expression Analysis. In *Proc. ACM Int'l Conf. Multimedia and Expo*, Amsterdam, The Netherlands, p. 317–321 (2005).
13. Bänziger, T., H. Pirker & K.R. Scherer. GEMEP-GEneva Multimodal Emotion Portryals: A Corpus for the Study of a Multimodal Emotional Expressions. In *Proc. LREC Workshop on Corpora for Research on Emotion and Affect*, Genova, Italy, p. 15–19 (2006).
14. Douglas-Cowie, E., N. Campbell, R. Cowie & P. Roach. Emotional Speech: Towards a New Generation of Databases. *Speech Communication* 40(1-2): 33–60 (2003).
15. Haq, S. & P.J.B. Jackson. Speaker-Dependent Audio-Visual Emotion Recognition. In *Proc. International Conference on Auditory-Visual Speech Processing*, Norwich, UK, p. 53–58 (2009).
16. Lin, Y. & G. Wei. Speech Emotion Recognition Based on HMM and SVM. In *Proc. International Conference Machine Learning and Cybernetics*, Guangzhou, China, p. 4898–4901 (2005).
17. Pantic, M. & M.S. Bartlett. Machine Analysis of Facial Expressions. In: *Face Recognition.* I-Tech Education and Publishing, Vienna, Austria, p. 377–416 (2007).
18. Bartlett, M.S., G. Littlewort, M. Frank, C. Lainscsek, I. Fasel & J. Movellan. Fully Automatic Facial Action Recognition in Spontaneous Behavior. In *Proc. International Conference on Automatic Face and Gesture Recognition*, Southampton, UK, p. 223–230 (2006).
19. Haq, S., P.J.B. Jackson & J.D. Edge. Audio-Visual Feature Selection and Reduction for Emotion Classification. In *Proc. International Conference on Auditory-Visual Speech Processing*, Tangalooma, Australia, p. 185–190 (2008).
20. Casale, S., A. Russo & S. Serranoa. Multistyle Classification of Speech Under Stress using Feature Subset Selection Based on Genetic Algorithms. *Speech Communication* 49(10-11): 801–810 (2007).
21. Gunes, H. & M. Piccardi. Affect Recognition from Face and Body: Early Fusion vs. Late Fusion. In *Proc. IEEE International Conference on Systems, Man, and Cybernetics*, Hawaii, USA, p. 3437–3443 (2005).
22. Busso, C., Z. Deng, S. Yildirim, M. Bulut, C.M. Lee, A. Kazemzadeh, S. Lee, U. Neumann & S. Narayanan. Analysis of Emotion Recognition using Facial Expressions, Speech and Multimodal

Information. In *Proceedings of International Conference on Multimodal Interfaces*, State College, PA, USA, p. 205–211 (2004).

23. Kim, E.H., K.H. Hyun & Y.K. Kwak. Improvement of Emotion Recognition from Voice by Separating of Obstruents. In *Proc. IEEE International Symposium on Robot and Human Interactive Communication*, Hatfield, UK, p. 564–568 (2006).
24. Petrushin, V.A. Emotion in Speech: Recognition and Application to Call Centers. In *Proc. Artificial Neural Networks in Engineering*, St. Louis, Missouri, USA, p. 7–10 (1999).
25. Ashraf, A.B., S. Lucey, J.F. Cohn, T. Chen, Z. Ambadar, K. Prkachin, P. Solomon & B.J. Theobald. The Painful Face – Pain Expression Recognition Using Active Appearance Models. In *Proc. ACM International Conference on Multimodal Interfaces*, Nagoya, Japan, p. 9–14 (2007).
26. Bartlett, M.S., G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Recognizing Facial Expression: Machine Learning and Application to Spontaneous Behavior. In *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, San Diego, CA, USA, p. 568–573 (2005).
27. Zeng, Z., J. Tu, M. Liu, T.S. Huang, B. Pianfetti, D. Roth, and S. Levinson. Audio-Visual Affect Recognition. *IEEE Transactions on Multimedia*, 9(2):424–428 (2007).
28. Petridis, S. & M. Pantic. Audiovisual Discrimination Between Laughter and Speech. In *Proc. IEEE International Conference Acoustics, Speech, and Signal Processing*, Las Vegas, Nevada, USA, p. 5117–5120 (2008).
29. Jackson, P. & S. Haq. *Surrey Audio-Visual Expressed Emotion (SAVEE) Database*. http://personal.ee.surrey.ac.uk/Personal/P.Jackson/SAVEE/References.html (2014).
30. Ekman, P., W.V. Friesen, M. O'Sullivan, A. Chan, I. Diacoyanni-Tarlatzis, K. Heider, R. Krause, W.A. LeCompte, T. Pitcairn, P.E. Ricci-Bitti, K. Scherer & M. Tomita. Universals and Cultural Differences in the Judgments of Facial Expressions of Emotion. *Journal of Personality and Social Psychology* 53(4):712–717 (1987).
31. Huckvale. M. *Speech Filing System*. http://www.phon.ucl.ac.uk/resource/sfs/ (2014).
32. Young, S., G. Evermann, D. Kershaw, D. Moore, J. Odell, D. Ollason, V. Valtchev & P. Woodland. *Hidden Markov Model Toolkit*. http://htk.eng.cam.ac.uk/ (2014).
33. Kittler, J. Feature Set Search Algorithms. In: *Pattern Recognition and Signal Processing*. p. 41–60. Sijthoff & Noordoff International Publishers, The Netherlands (1978).
34. Campbell, J.P. Speaker Recognition: A Tutorial. *Proceedings of the IEEE*, 85(9):1437–1462 (1997).
35. Wang, Y. & L. Guan. Recognizing Human Emotional State From Audiovisual Signals. *IEEE Transactions on Multimedia* 10(5):936–946 (2008).
36. Comaniciu, D., P. Meer & D. Tyler. Dissimilarity Computation Through Low Rank Corrections. *Pattern Recognition Letters* 24(1-3):227–236 (2003).
37. Mahalanobis, P.C. On the Generalised Distance in Statistics. In *Proc. National Institute of Sciences of India*, vol. 2, p. 49–55 (1936).
38. Kullback, S. *Information Theory and Statistics*. Dover Publications, New York (1968).
39. Shlens, J. *A Tutorial on Principal Component Analysis*. Systems Neurobiology Laboratory, Salk Institute for Biological Studies, La Jolla (2005).
40. Duda, R.O., P.E. Hart & D.G. Stork. *Pattern Classification*. John Wiley & Sons, USA, (2001).