

SPEAKER-DEPENDENT AUDIO-VISUAL EMOTION RECOGNITION

Sanaul Haq and Philip J.B. Jackson

Centre for Vision, Speech and Signal Processing, University of Surrey, UK

{s.haq, p.jackson}@surrey.ac.uk



Abstract

Speaker-dependent audio-visual emotion recognition was performed using an English database (4 male actors, 7 emotions). A total of 106 audio and 240 visual features were extracted, and feature selection was performed with Plus l -Take Away r algorithm based on Bhattacharyya distance criterion. The PCA and LDA were applied to the selected features for feature reduction, and Gaussian classifiers were used for classification. Higher recognition performance was achieved for the visual and audio-visual features compared to the audio features.

1 Introduction

Human communication employs

- verbal
- non-verbal: facial expressions, and tone of voice

Important audio and visual features for emotion recognition

- Audio features: fundamental frequency, energy, duration, spectral energy, formants, MFCCs, jitter, and shimmer at utterance level [5, 9, 10], and at frame level [7, 6, 1].
- Visual features: forehead, eye-region, cheek and lip.

Since the fusion of audio and visual information has yielded higher recognition performance [4, 2, 8], therefore we combined the two modalities at decision level for audio-visual affect recognition.

2 Method

We performed the emotion recognition from audio and visual modalities in four steps (Fig. 1).

- 1. Feature extraction:** A total of 106 utterance-level audio features were extracted related to fundamental frequency, energy, duration and spectral envelope (Fig. 4 (left)); 240 visual features were obtained from 2D marker co-ordinates (Fig. 4 (right)).
- 2. Feature selection:** The top 40 features were selected for each of the audio and visual modality by Plus l -Take Away r algorithm [3] using Bhattacharyya distance criterion (Fig. 5).
- 3. Feature reduction:** PCA and LDA transformations were applied to the selected features (Fig. 6).
- 4. Classification:** Gaussian classifiers were used for classification between different emotions.

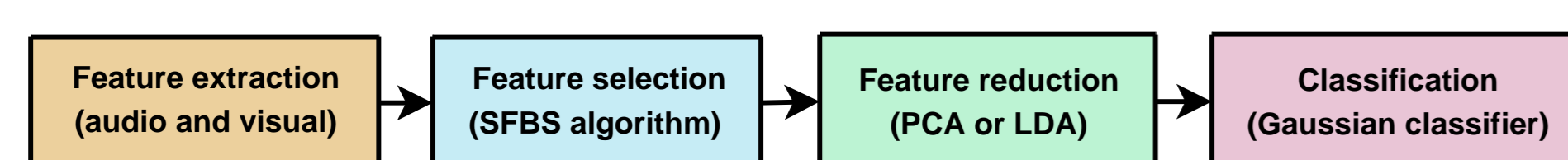


FIGURE 1: Block diagram of our experimental method.

Database:

- A total of 480 utterances from 4 male actors with 60 markers painted on actors' face.
- Seven emotions: anger, disgust, fear, happiness, neutral, sadness and surprise (Fig. 2).
- 15 phonetically-balanced TIMIT sentences per emotion: 3 common, 2 emotion specific and 10 generic sentences.

- 3dMD dynamic face capture system, video sampled at 60 fps and audio at 44.1 kHz.



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

Data Evaluation:

- Human tests provide a bench mark for the recognition accuracy.
- Three types of clips: audio, visual and audio-visual.
- Data evaluated by 10 subjects (5 native English speakers, 5 female).
- Subjects play clips and select from one of the 7 emotions on a paper sheet.
- 4 emotion categories: Displeased (anger, disgust), Gloomy (fear, sadness), Excited (happiness, surprise) and Neutral (neutral).

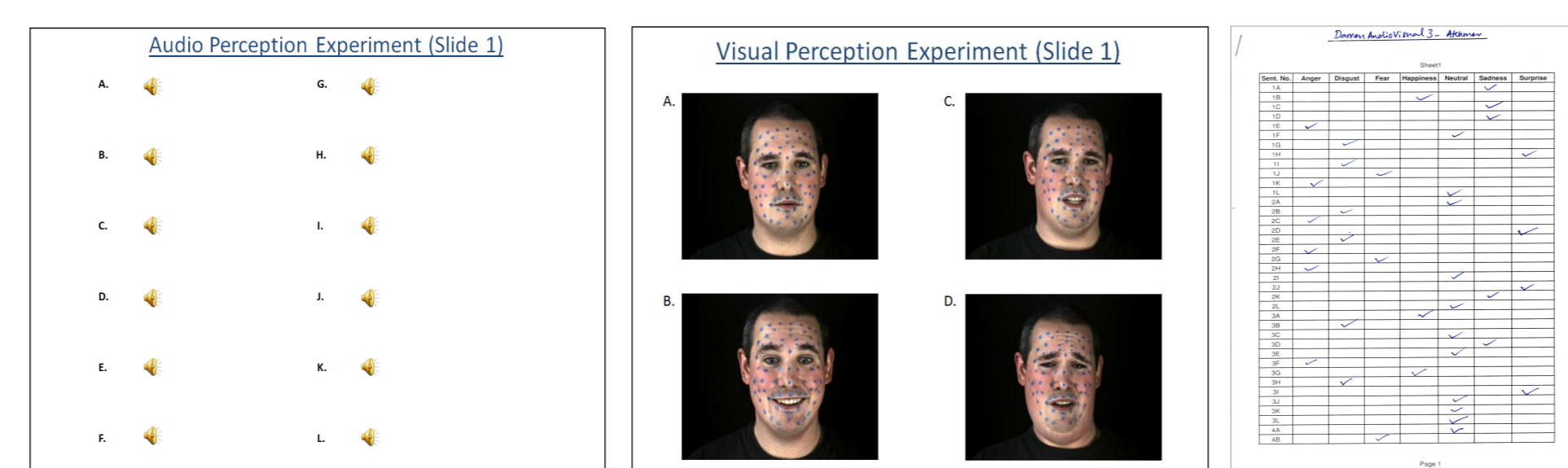


FIGURE 3: Perception experiments (from left): audio, visual, and response sheet.

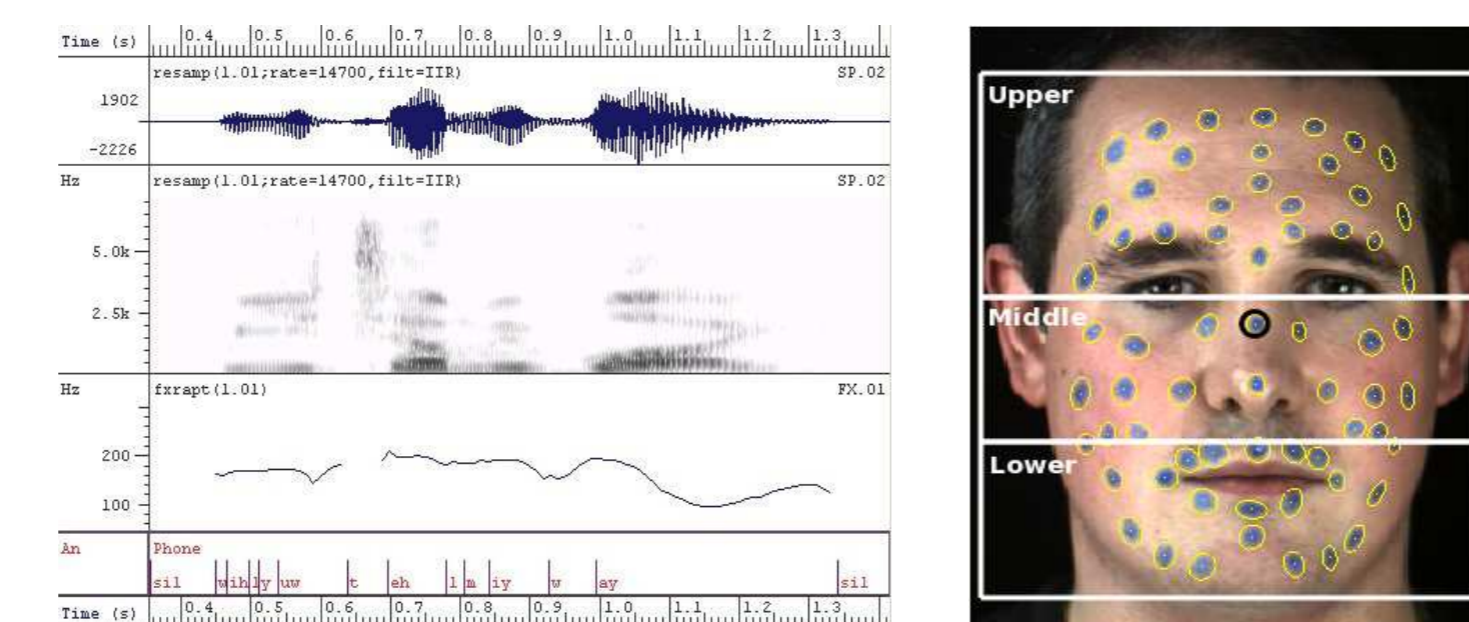


FIGURE 4: Audio feature extraction with SFS software (left), and video data (right) with overlaid tracked marker locations. The reference marker was on the bridge of the nose (black circle).

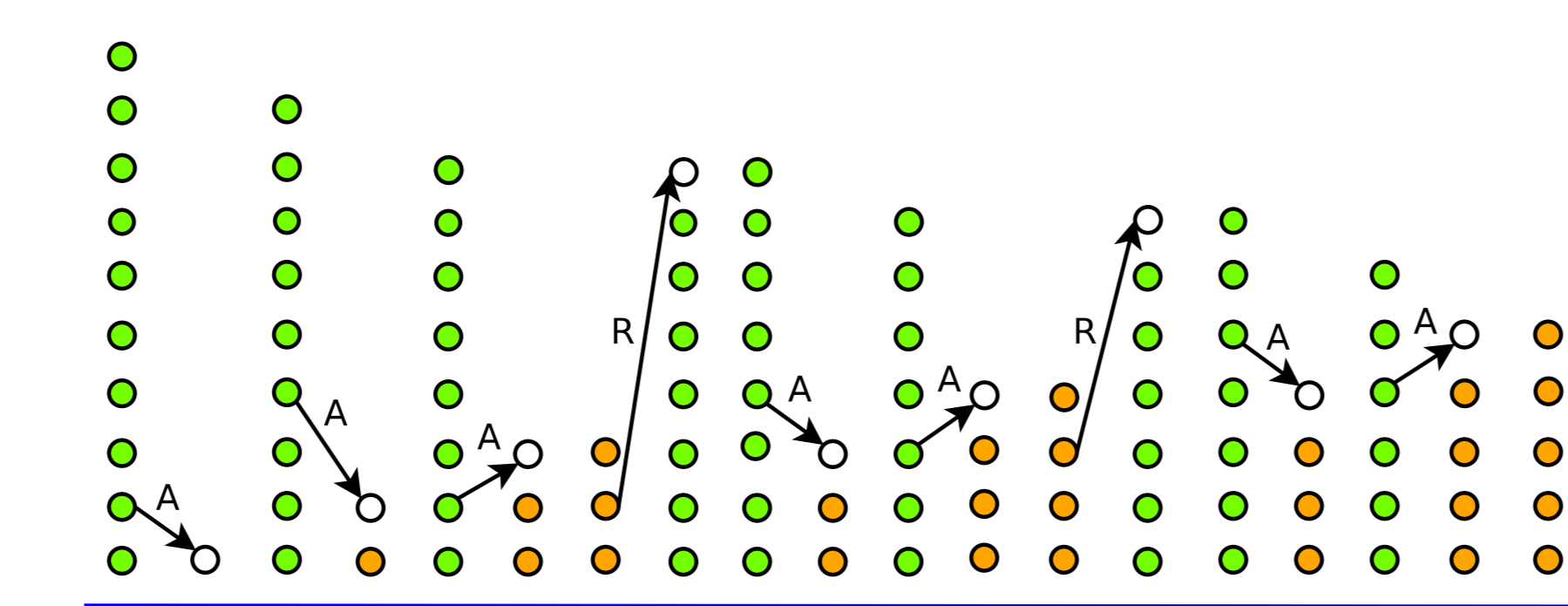


FIGURE 5: Feature selection with Plus l -Take Away r ($l=2, r=1$) algorithm. The green circles show the original features, and orange ones are the selected features.

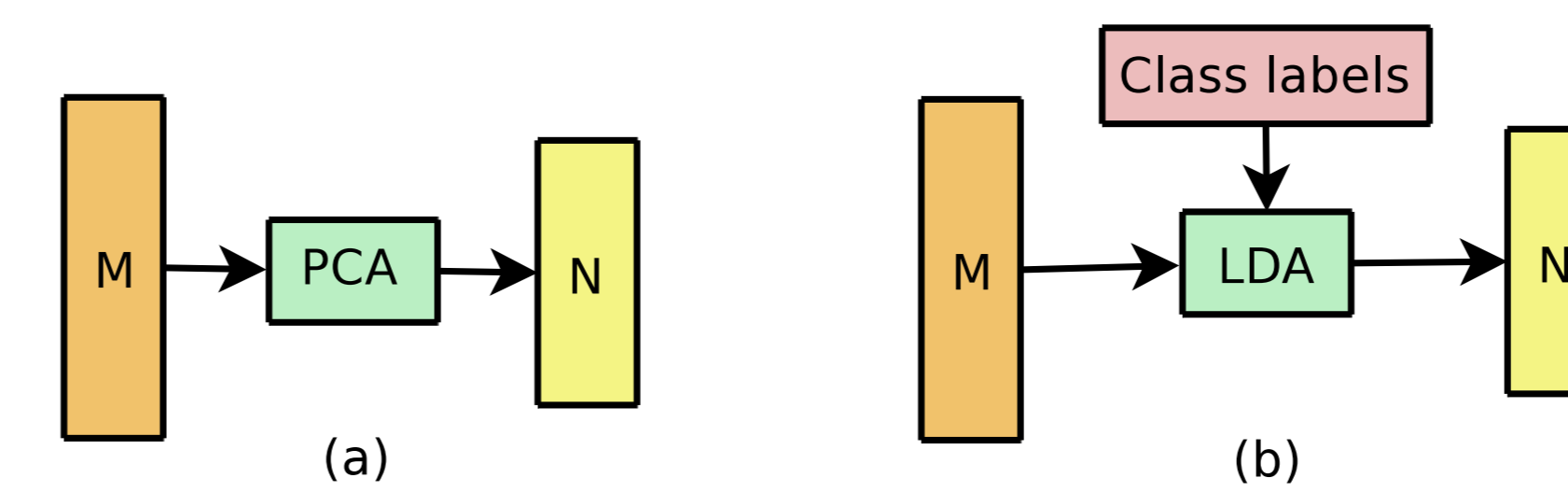


FIGURE 6: Feature reduction techniques: (a) Principal Component Analysis, and (b) Linear Discriminant Analysis where $N < C$, where C is the number of classes.

3 Experimental results

Three sets of emotion recognition experiments were performed

1. Audio
2. Visual
3. Audio-visual

The audio-visual experiments were performed by combining the two modalities at decision level (product) with equal weighting (Fig. 7).

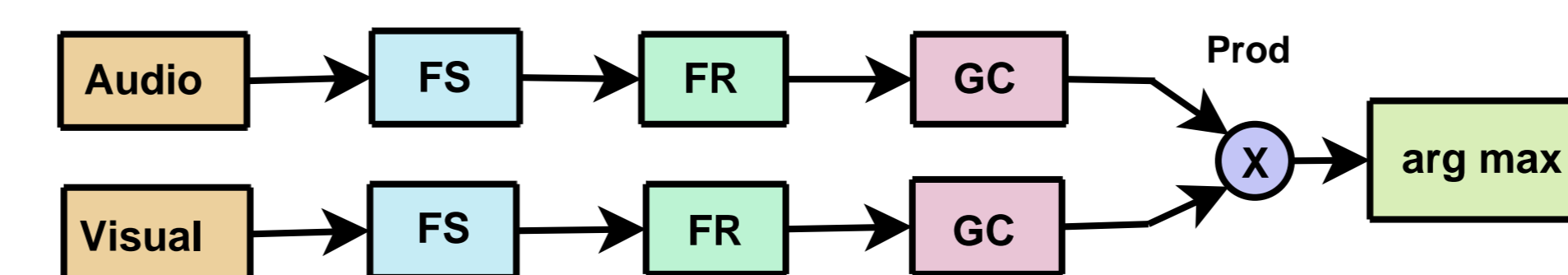


FIGURE 7: Diagram of audio and visual fusion at decision level.

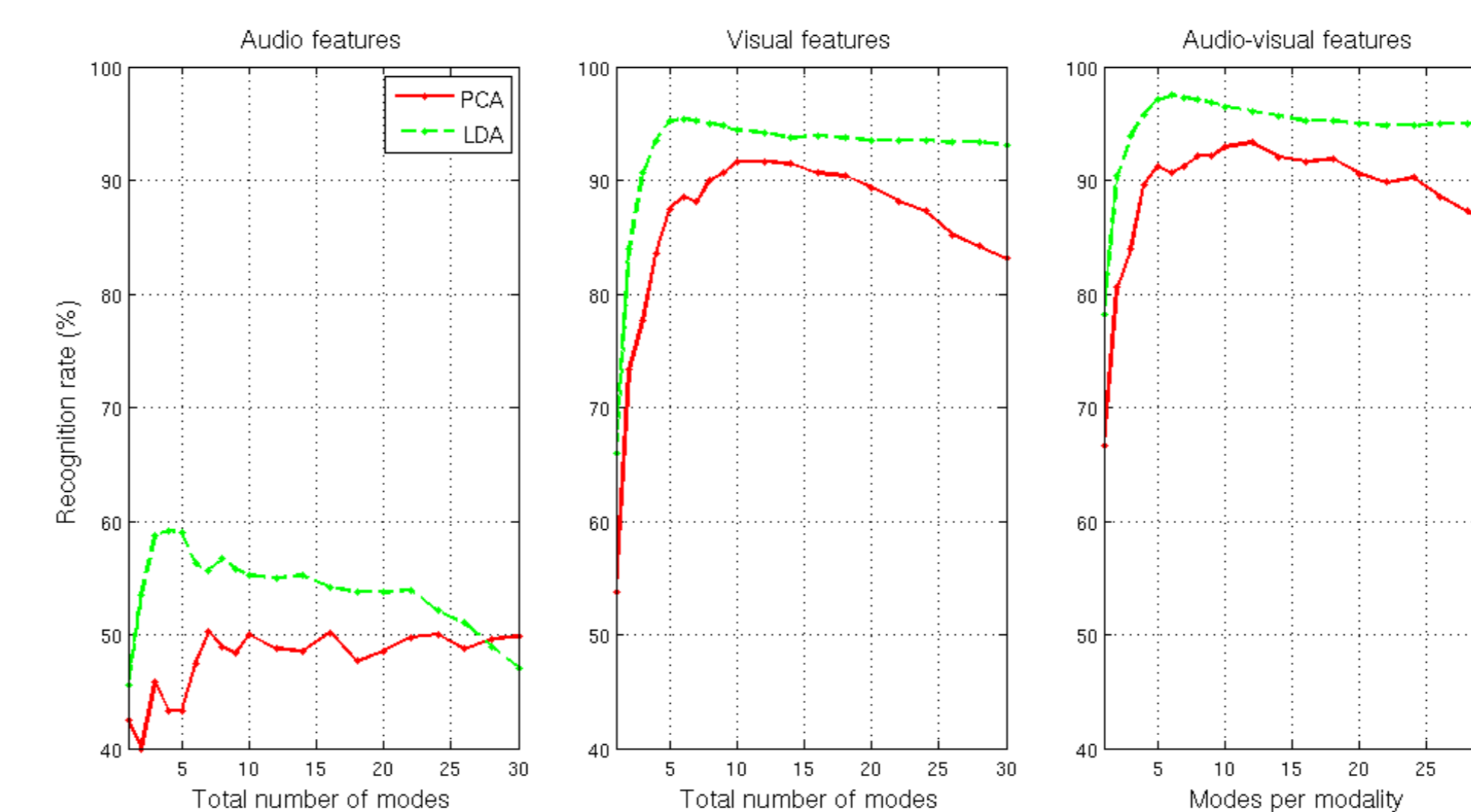


FIGURE 8: Average recognition rate (%) over 4 actors with audio, visual, and audio-visual features for 7 emotion classes.

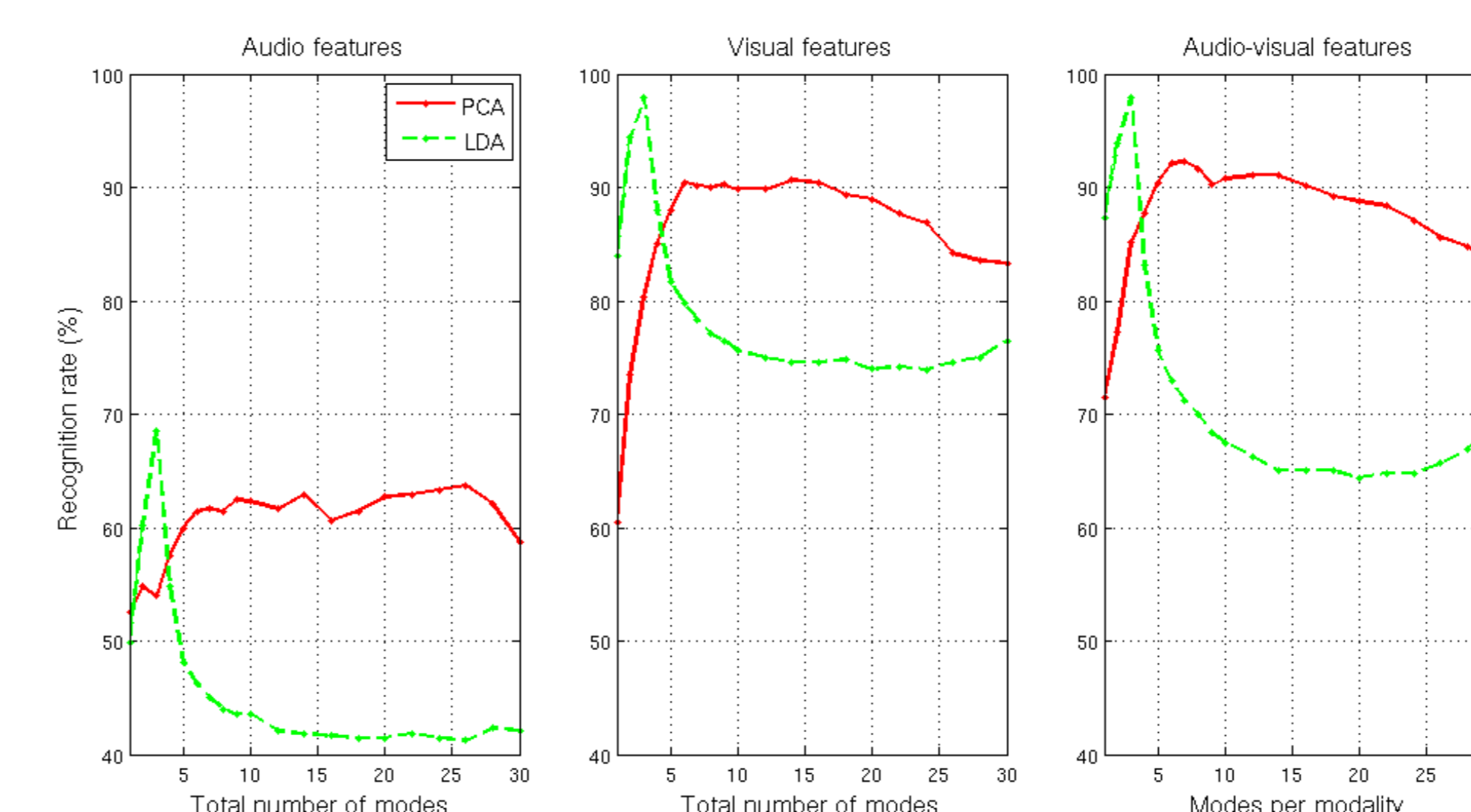


FIGURE 9: Average recognition rate (%) over 4 actors with audio, visual, and audio-visual features for 4 emotion classes.

TABLE 1: Average classification scores (%) over 4 actors achieved with the audio, visual and audio-visual data for 7 emotions.

Recogniser	Audio	Visual	Audio-visual
Human (10 subjects)	66.5 ± 2.5	88.0 ± 0.6	91.8 ± 0.1
Machine (LDA 6)	56.3 ± 6.7	95.4 ± 1.6	97.5 ± 0.8
Machine (PCA 10)	50.0 ± 6.0	91.7 ± 2.1	92.9 ± 2.1

TABLE 2: Average classification scores (%) over 4 actors achieved with the audio, visual and audio-visual data for 4 emotions.

Recogniser	Audio	Visual	Audio-visual
Human (10 subjects)	76.3 ± 2.4	91.3 ± 1.1	95.2 ± 0.6
Machine (LDA 3)	68.5 ± 4.8	97.9 ± 1.2	97.9 ± 1.2
Machine (PCA 7)	61.7 ± 4.3	90.2 ± 3.3	92.3 ± 2.7

4 Conclusions

In classification tests on a British English audio-visual emotional database:

- LDA outperformed PCA.
- Both audio and visual information are useful for emotion recognition.
- The energy and MFCC features were identified as the most important audio features for emotion recognition.
- The vertical movement of face was more important for emotion recognition, especially in eye and cheek regions.
- For the visual and audio-visual features, our system recognition performance was very close and even higher than humans.
- Future work is to perform speaker independent experiments with more data and by using other classifiers, e.g. GMM and SVM.

References

- [1] 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'05, Wuhan*, pages 147–151, 2005.
- [2] C. Busso and et al. Analysis of Emotion Recognition using Facial Expressions, Speech and Multimodal Information. In *Proc. ACM Int. Conf. on Multimodal Interfaces*, pages 205–211, 2004.
- [3] C.H. Chen. Pattern Recognition and Signal Processing. *Sijthoff & Noordhoff International Publishers, Netherlands*, 1978.
- [4] Y.K. Huang C.Y. Chen and P. Cook. Visual/Acoustic emotion recognition. In *Proc. Int. Conf. on Multimedia & Expo*, pages 1468–1471, 2005.
- [5] E. Navas I. Luengo and et al. Automatic Emotion Recognition using Prosodic Parameters. In *Proc. Interspeech, Lisbon*, pages 493–496, 2005.
- [6] Y. Kao and L. Lee. Feature Analysis for Emotion Recognition from Mandarin Speech Considering the Special Characteristics of Chinese Language. In *Proc. Interspeech, Pittsburgh*, pages 1814–1817, 2006.
- [7] Y. Lin and G. Wei. Speech Emotion Recognition Based on HMM and SVM. In *Proc. 4th Int. Conf. on Mach. Learn. & Cybernetics, Guangzhou*, pages 4898–4901, 2005.
- [8] C. Chen M. Song and M. You. Audio-visual based emotion recognition using tripled Hidden Markov Model. In *Proc. ICASSP*, 5:877–880, 2004.
- [9] D. Ververidis and C. Kotropoulos. Emotional speech classification using Gaussian mixture models. In *Proc. ISCAS, Kobe*, pages 2871–2874, 2005.
- [10] L. Vidrascu and L. Devillers. Detection of real-life emotions in call centers. In *Proc. Interspeech, Lisbon*, pages 1841–1844, 2005.