Fusion of Audio and Visual Cues for Laughter Detection

Stavros Petridis
Department of Computing
Imperial College London
London, UK
sp104@doc.ic.ac.uk

Maja Pantic
Department of Computing
Imperial College London, UK
EEMCS, Univ. Twente, NL
m.pantic@imperial.ac.uk

ABSTRACT

Past research on automatic laughter detection has focused mainly on audio-based detection. Here we present an audiovisual approach to distinguishing laughter from speech and we show that integrating the information from audio and video channels leads to improved performance over single-modal approaches. Each channel consists of 2 streams (cues), facial expressions and head movements for video and spectral and prosodic features for audio. We used decision level fusion to integrate the information from the two channels and experimented using the SUM rule and a neural network as the integration functions. The results indicate that even a simple linear function such as the SUM rule achieves very good performance in audiovisual fusion. We also experimented with different combinations of cues with the most informative being the facial expressions and the spectral features. The best combination of cues is the integration of facial expressions, spectral and prosodic features when a neural network is used as the fusion method. When tested on 96 audiovisual sequences, depicting spontaneously displayed (as opposed to posed) laughter and speech episodes, in a person independent way the proposed audiovisual approach achieves over 90% recall rate and over 80% precision.

Categories and Subject Descriptors

I.5.4 [**Computing Methodologies**]: Pattern Recognition—*Applications*; J.m [**Computer Applications**]: Miscellaneous

General Terms

Algorithms

Keywords

Audiovisual data processing, laughter detection, non-linguistic information processing

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
CIVR'08, July 7–9, 2008, Niagara Falls, Ontario, Canada.
Copyright 2008 ACM 978-1-60558-070-8/08/07 ...$5.00.

1. INTRODUCTION

In human - human interaction, information is communicated between the parties through various channels. Speech is usually the dominant channel. However, spoken words are highly person and context dependant [7], so speech recognition and extraction of semantic information is a very challenging task for machines [24]. Other cues like facial expressions, head gestures, hand gestures and non-linguistic vocalizations play an important role in communication as well. Although several works address the problem of recognising facial expressions and / or head gestures, the recognition of non-linguistic vocalisations is a quite unexplored area. Scherer [20] defines non-linguistic vocalizations as very brief, discrete, nonverbal expressions of affect in both face and voice. Schroeder [21] has demonstrated that people can recognize emotions with high accuracy just by hearing to such vocalizations which in turn reveals their significance in human communication. An attractive property of this kind of communication is that it is not as heavily person-dependent as speech is and, at the same time, it encodes rich information about the other person's emotional state [19].

One of the most important non-linguistic vocalizations is laughter, which is reported to be the most often annotated paralinguistic event occurring frequently in recorded natural speech [22]. Laughter is a very good indicator of someone's mental state since people very often express their emotion through laughter. Therefore, laughter is a powerful affective and social signal.

In human - computer interaction (HCI), automatic detection of laughter can be used as a useful cue for detecting the user's affective state and facilitating affect-sensitive human-computer interfaces [14]. Another application of laughter detection is the identification of semantically meaningful events in meetings such as topic change or jokes. Also, such a detector can be useful for the detection of non-speech in automatic speech recognition as well as for content-based video retrieval.

Few works have been recently reported on automatic laughter detection. The main characteristic of these studies is that only audio information is used, i.e., visual information carried by facial expressions of the observed person is ignored. Most of these studies use Hidden Markov Models (HMMs) as the classification tool (just as is the case in automatic speech recognition), [24]. This is mainly due to the ability of HMMs to represent the temporal characteristics of the phenomenon. Existing approaches to laughter detection include the work of Lockerd and Mueller [12], who used HMMs and spectral coefficients, the work of Cai et al. [3], who used

HMMs with Mel-Frequency Cepstral Coefficients (MFCCs) and perceptual features, and the work of Campbell et al. [4], who used phonetic features and HMMs to detect four types of laughter. Another approach is that of Kennedy and Ellis [11], who trained Support Vector Machines (SVM) with MFCCs and delta MFCCs. The most extensive study in this area was made by Truong and Leeuwen [22], who compared the performance of different auditory frame and utterance level features using different classifiers and different combination schemes. To the best of our knowledge, there are only three approaches which use audiovisual information [10],[18], [16]. Ito et al. [10] built an image-based laughter detector based on spatial locations of facial feature points and an audio-based laughter detector based on MFCC features. The output of the two detectors are combined with an AND operator to yield the final classification for an input sample. They attained 80% average recall rate using 3 sequences of 3 subjects in a person dependent way. Reuderink [18] used visual features based on principal components analysis and RASTA-PLP features for audio processing. Gaussian mixture models and support vector machines were used as classifiers which were fused on decision level obtaining an equal error rate of 14.2%. In our previous work [16], we compared decision level and feature level fusion with audio- and video-only laughter detection. We used spectral features and displacements of the tracked facial points as the audio and visual features respectively. Both fusion approaches outperformed single-modal detectors, achieving on average 84% recall in a person-independent test.

In this paper, we present our further research on audiovisual discrimination of laughter episodes from speech episodes. Our research on an audiovisual approach rather than an audio-only approach to laughter recognition is mainly driven by research on audiovisual speech recognition that reported improved performance over audio-only speech recognition [6], [17]. At this point we would like to remark that we use spontaneous (as opposed to posed) displays of laughter and speech episodes from the audiovisual recordings of the AMI meeting corpus [13]. We focus on person-independent recognition which makes the task of laughter detection even more challenging. We compare the performance of audiovisual laughter detectors where different combinations of auditory and visual cues are used for detection. We experimented with the audio channel consisting of two streams (cues): spectral features (PLP) and prosodic features (pitch and energy), and the visual channel consisting of two streams as well: head movements and facial expressions. We used decision level fusion to investigate which streams carry useful information for laughter detection. The most important cues were found to be facial expressions and spectral features. Pitch and energy can sometimes be beneficial as well, whereas the head movements do not seem to be of importance. Another significant finding is that a simple linear function is sufficient for the fusion of audio and visual cues. Our results show that audiovisual laughter detection outperforms single-modal (audio / video only) laughter detection, attaining over 90 % recall.

2. DATASET

Posed expressions may differ in visual appearance, audio profile, and timing from spontaneously occurring behavior. For example, spontaneous smiles are smaller in amplitude, longer in total duration, and slower in onset and offset time

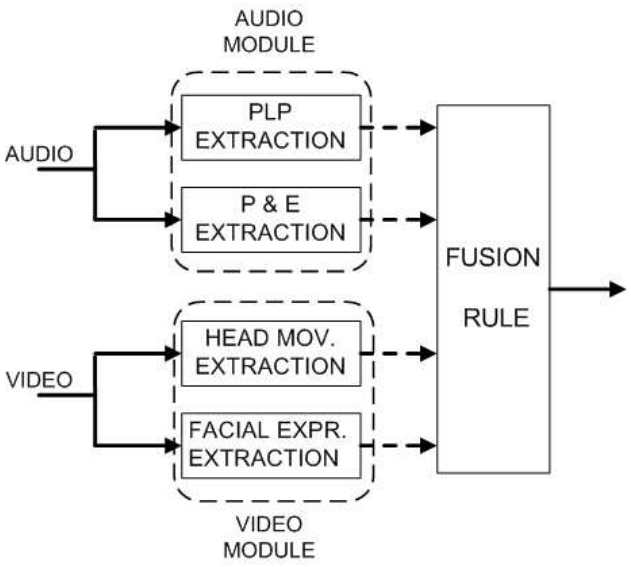


Figure 1: Feature extraction for audiovisual laughter detection

than posed smiles [23]. Therefore it is not surprising that tools trained and tested on posed expressions suffer a significant degradation in their performance when applied to spontaneous expressions. This is the reason we only use spontaneous expressions in all the experiments.

The AMI Meeting Corpus is an ideal dataset for our task since it consists of 100 hours of meetings recordings where people show a huge variety of spontaneous expressions. We only used the close-up video recordings of the subject's face (720 x 576 pixels, 25 frames per second) and the related individual headset audio recordings (16 kHz). The language used in the meetings is English and the speakers are mostly non-native speakers. For our experiments we used seven meetings (IB4001 to IB4011) and the relevant recordings of eight participants (6 young males and 2 young females) of Caucasian origin with or without glasses and no facial hair. An example of laughter from the dataset we use together with the tracking points is shown in Fig. 2.

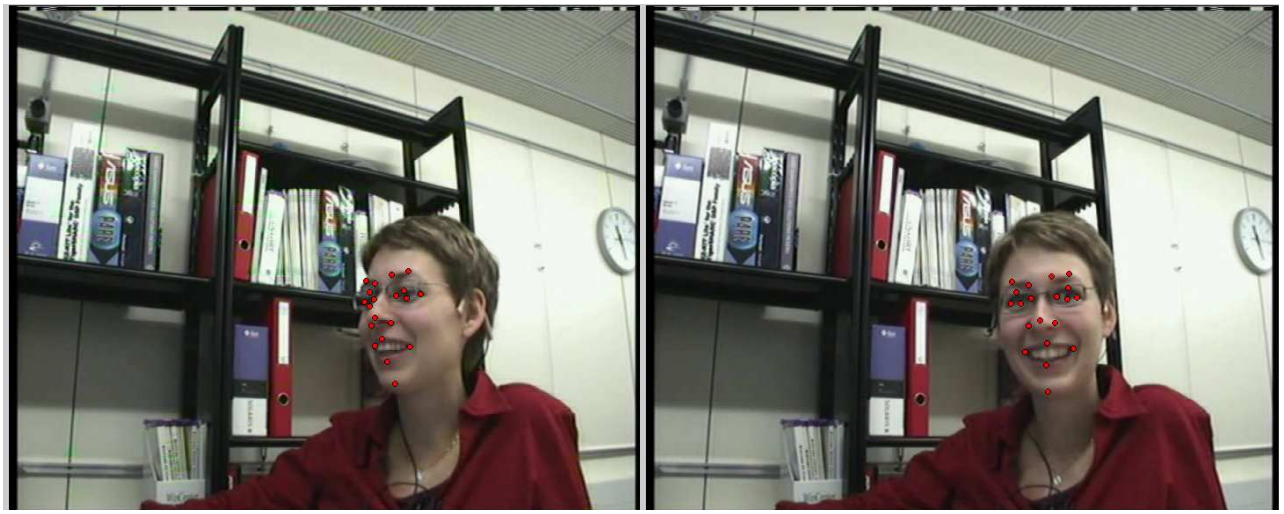
All laughter and speech segments were pre-segmented based on audio. Initially, laughter segments were selected based on the annotations provided with the AMI Corpus. After examining the extracted laughter segments we only kept those that do not co-occur with speech and laughter is clearly audible. Speech segments were also determined by the annotations provided with the AMI Corpus. We selected those that do not contain long pauses between two consecutive words. In total, we used 40 audio-visual laughter segments, 5 per person, with a total duration of 58.4 seconds (with mean duration $\mu = 1.46$ seconds and standard deviation $\sigma = 1.09$ seconds) and 56 audio-visual speech segments with a total duration of 118.08 seconds (with mean duration $\mu = 2.11$ seconds and standard deviation $\sigma = 1.09$ seconds).

3. SYSTEM OVERVIEW

As an audiovisual approach to laughter detection is investigated in this study, information is extracted simultaneously from the audio and visual channel. The visual channel



(a) Frame 2 (b) Frame 42



(c) Frame 82 (d) Frame 122



(e) Frame 162 (f) Frame 176

Figure 2: Example of a laughter episode, from the AMI corpus, with illustrated facial point tracking results

is divided into 2 streams (cues): face and head movements as shown in Fig. 1. Similarly, the audio channel is divided into 2 streams as well: spectral (PLP) prosodic features (pitch and energy). Details on how the audio and visual features are extracted are presented in sections 4 and 5. To recognize whether an input audio and / or visual sample is an episode of speech or an episode of laughter a neural network classifier is used. Neural networks were used as the classifier since they are able to learn non-linear function from examples. However, any learning algorithm which can learn complex functions from examples such as Support Vector Machines is expected to perform equally well. Finally, each cue is modelled by a neural network and then different combinations of those cues are integrated using decision level fusion in order to achieve audiovisual laughter detection as shown in Fig. 1. A simple linear function (SUM rule) is compared to non-linear functions modelled by neural networks for the task of audiovisual fusion.

In order to compare the performance of different laughter detectors using different combinations of cues, the following two measures are used.

F1 Measure: Recall and precision are two commonly used rates for measuring the quality of binary classification problems. Recall describes the completeness of the retrieval. It is defined as the portion of the positive examples retrieved by the classifier over the total number of existing positive examples (including the ones not retrieved by the classifier). Precision describes the actual accuracy of the retrieval, and is defined as the portion of the actual positive examples that exist in the total number of examples retrieved as positive by the classifier. While recall and precision rates can be individually used to determine the quality of a classifier, it is often more convenient to have a single measure to do the same assessment. The F_α measure combines the recall and precision rates in a single equation:

$$F_\alpha = \frac{(1+\alpha) \times precision \times recall}{\alpha \times precision + recall} \quad (1)$$

where α defines how recall and precision will be weighted. In the case that recall and precision are evenly weighted then the F1 measure is defined where $\alpha = 1$.

AUC: By varying the threshold on the output of the classifier we get the Receiver Operating Characterestic (ROC) curve which is a plot of the true positive rate vs the false positive rate. A commonly used measure for comparing different ROC curves is the Area Under the ROC Curve (AUC).

4. AUDIO MODULE

The audio module is responsible for the audio channel processing, as shown in Fig. 1. It extracts features from the audio signal on a frame-by-frame basis which are then used by the classification algorithm. Two different categories of features are used in this study: 1) spectral features which are derived from a special type of spectrum representation (cepstrum, which is the result of taking the fourier transform of the spectrum) of the audio signal, and 2) and non-spectral features which are based on other properties of speech / laughter such as prosody. The spectral features considered are the Perceptual Linear Prediction (PLP) coefficients . The second category of features includes pitch and energy features.

Spectral Features: Spectral or cepstral features, such as PLP features [9], have been successfully used for speech recognition. They have been also successfully used for laughter detection as well. Truong and Leeuwen [22], for example, reported a higher success rate in automatic laughter detection when using PLP features than other non-spectral features. The PLP signal analysis technique is based on the short-term spectrum of the signal, subsequently modified by several psychophysically based spectral transformations [9].

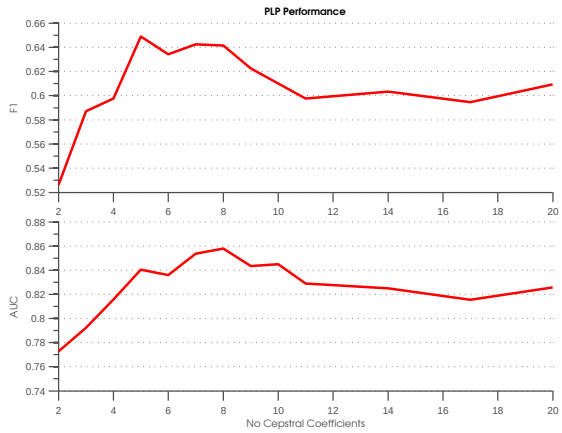


Figure 3: F1 measure and AUC as a function of the number of PLP coefficients used

The use of 13 PLP coefficients is very popular in literature [22]. However, Kennedy and Ellis [11] reported that the same level of performance is achieved with just 6 coefficients. In order to investigate the influence of the number of coefficients on laughter detection we experimented with different number of coefficients. The performance measures used were the F1 measure and the AUC. The results obtained are shown in Fig. 3. As can be seen in terms of the F1 measure and the AUC, 5 and 8 coefficients give the best performance respectively. In both cases, 7 PLP coefficients achieve the second best performance, so we decided to use 7 coefficients in this study. This result is consistent with the finding of Kennedy and Ellis [11]. In [16] four different frame rates were investigated, 25, 50, 75 and 100 frames per second (fps) with 50% overlap. With the exception of 25 fps which resulted in poor performance, the rest resulted in comparable performance. Therefore, the frame rate of 50 fps was selected, i.e., the window size is 40ms and the step-size 20ms. In addition to the 7 PLP coefficients, their delta features were calculated as well. The deltas are calculated by a linear regression over a short neighborhood around a spectral feature. The slope of the fitted line represents the derivative of the spectral feature and therefore can capture some local temporal characteristics which seem to be very important for interpretation of human behavioural cues, as argued in psychological literature (e.g., [19]). So in total 14 features are computed per frame.

Prosodic features: The two most commonly used prosodic features in studies of emotion detection are pitch and energy [24]. Both of them have also been used in previous works on audio-only laughter detection [22]. Pitch is the perceived fundamental frequency of a sound. While the actual fundamental frequency can be precisely determined through physical measurement, it may differ from the perceived pitch. Bachorowski et al. [1] found that the mean pitch in both male

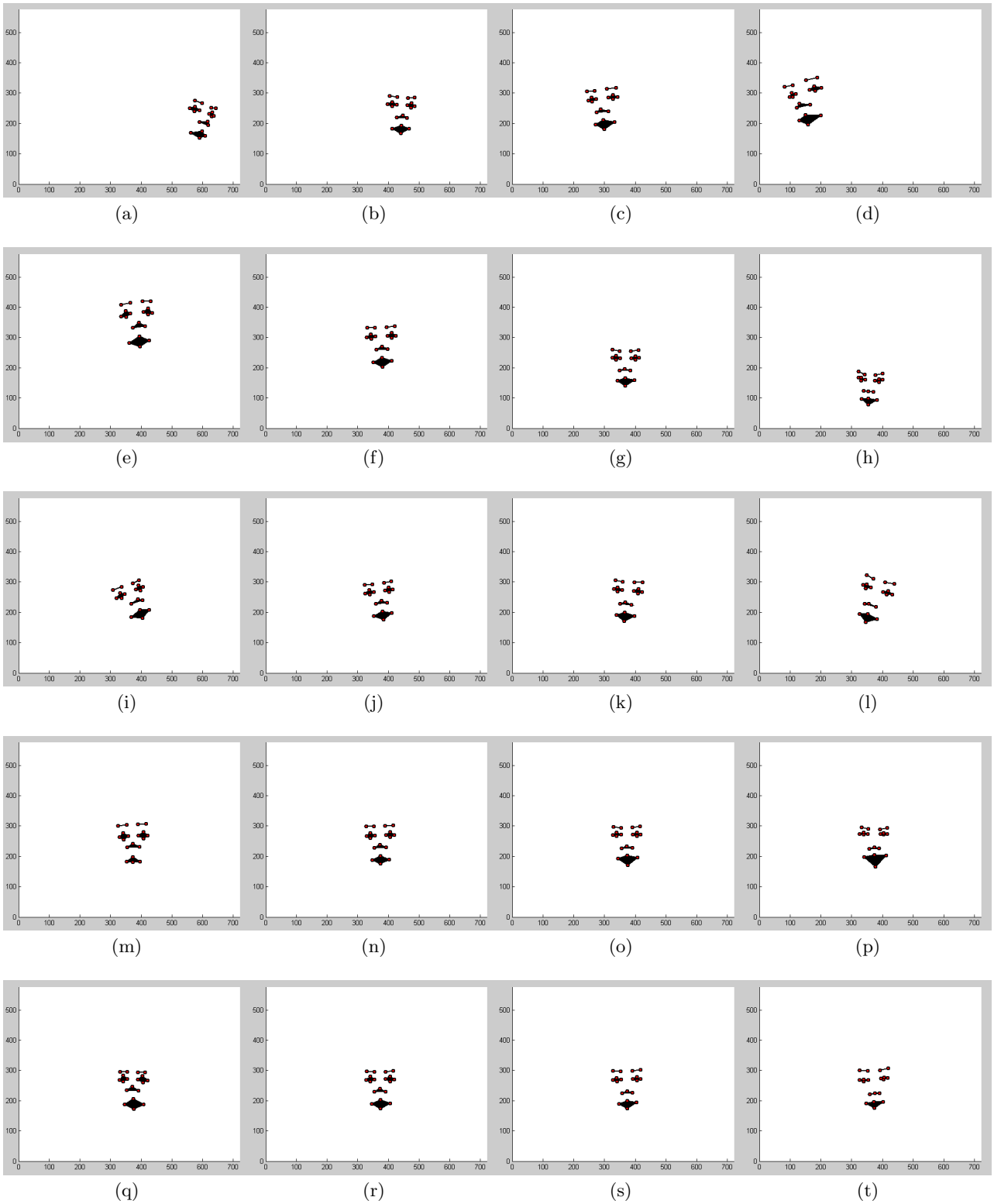


Figure 4: Principal Component Analysis - 1st row: Effect of varying b_1, 2nd row: Effect of varying b_2, 3rd row: Effect of varying b_3, 4th row: Effect of varying b_7, 5th row: Effect of varying b_8

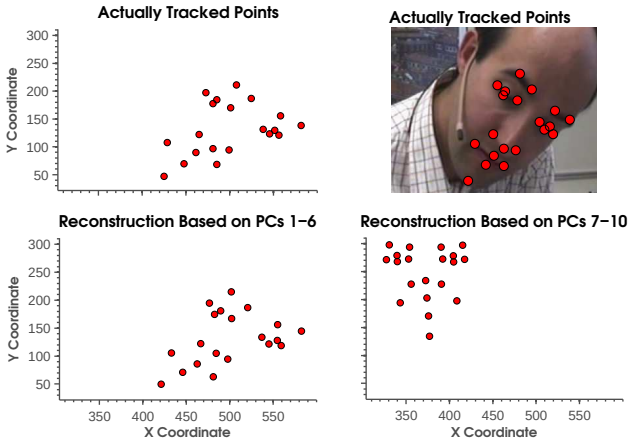


Figure 5: PCA analysis of facial point tracking. Upper row: actually tracked facial points. Bottom row: (left) 20 facial points after they have been reconstructed using the first 6 principal components, (right) 20 facial points after they have been reconstructed using principal components 7 to 10.

and female laughter was higher than in modal speech. Pitch was computed in each frame using the same algorithm as Praat [2].The energy feature used is the Root-Mean-Square (RMS) energy. In addition, the deltas of energy and pitch were computed resulting in 4 features. Those features are extracted in the same frame rate as the PLP coefficients, i.e. every 20ms over a window of 40ms.

5. VIDEO MODULE

The video module is responsible for the visual channel processing as shown in Fig. 1. The first step in this module is to track some characteristic facial points which then they will be used for feature extraction. Then a Point Distribution Model (PDM) is learnt with the aim of decoupling the head movement from the facial expressions. The features are extracted from the visual channel on a frame-by-frame basis which are then used by the classification algorithm. Two different categories of features are used in this study: 1) features which correspond to the head movements, and 2) features which correspond to the facial expressions.

5.1 Tracking

To capture the facial expression dynamics, we track 20 facial points, as shown in Fig. 2 and Fig. 5, in the video segments. These points are the corners / extremities of the eyebrows (2 points), the eyes (4 points), the nose (3 points), the mouth (4 points) and the chin (1 point). To track these facial points we used the Patras - Pantic particle filtering tracking scheme [15], applied for tracking color-based templates centered around the facial points to be tracked. The points were manually annotated in the first frame of an input video and tracked for the rest of the sequence. Hence, for each video segment containing K frames, we obtain a set of K vectors containing 2D coordinates of the 20 points tracked in K frames.

5.2 Decoupling of Head and Face

While speaking and especially while laughing, people tend to exhibit large head movements. It is even more so in the case of our data since we use recordings of naturalistic (spontaneous) expressions rather than deliberately displayed episodes of speech and laughter. Since we are interested in seperating facial expression configurations (relevant to speech and laughter episodes) from head movements, we need to distinguish between changes in the location of facial points caused by facial expressions and changes caused by rigid head movements. In other words, we wish to decouple head movements from facial expressions so we can investigate the effect of each cue seperately. To do so we use a similar approach with Gonzalez-Jimenez and Alba-Castro [8] in which Principal Component Analysis (PCA) is used for decoupling, skipping the alignment of the shapes in order to capture the head movement as well. This approach is based on PDMs [5] and has also been used in [18].

First, we concatenate the (x, y) coordinates of the 20 tracking points in a 40-dimensional vector. Then we use PCA to extract 40 principal components (PCs) for all the frames in the dataset. PCA is defined as an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance of the data comes to lie on the 1st coordinate (i.e., 1st PC), the 2nd greatest variance on the 2nd coordinate, and so on. Given that in our dataset head movements account for most of the variation in the data, lower-order PCs are expected to reflect rigid-movement aspects of the data while higher-order PCs are expected to retain non-rigid-movement (facial expression) aspects of the data. To test this assumption, we computed the PCs for the whole dataset and then reconstructed the position of the points in each frame by using different combinations of the PCs with the help of the following equations:

$$b = (x - \overline{x})P \tag{2}$$

$$x \approx \hat{x} = \overline{x} + bP^T \tag{3}$$

where P contains N out of the 40 eigenvectors and b is a N-dimensional vector. With the help of equation 2 we can compute the shape parameters b and then the face can be reconstructed using equation 3.

As can be seen from Fig. 4 and Fig. 5, it seems that indeed the lower-order PCs reflect rigid-movement aspects of the data, while the higher-order PCs reflect facial expression aspects of the data. The same has been reported by Gonzalez-Jimenez and Alba-Castro [8]. We can further investigate what is captured by each PC by varying the elements of shape parameters b one at a time between 3 standard deviations from the mean value and leaving all other elements at zero. In this way we can vary the shape $\hat{x}$ using equation 3. The variance of the i^{th} parameter, b_i, is given by its corresponding eigenvalue λ_i , so each b_i takes values in the range of $\pm 3\sqrt{\lambda_i}$. The variation corresponding to the i^{th} parameter, b_i, is called the i^{th} mode of the model [5]. By visual inspection we can identify which PCs contain facial expression information (non-rigid motion) and which contain head pose information (rigid motion).

Head: Modes 1, 2, and 3 are shown in Fig. 4a - 4l. We see that the 1st, 2nd, and 3rd PCs correspond to horizontal movement, vertical movement and rotation respectively. Similarly, the 4th, 5th and 6th modes correspond to scale and left / right head yaw respectively. In other words the first 6 PCs contain the head pose information. Therefore, we use the first 6 shape parameters, i.e. b_1 to b_6, as the head features.

Cues	Features Used	F1	AUC
Video Only			
Face	b_7 to b_{10}	0.83	0.945
Head	b_1 to b_6	0.525	0.652
Face + Head	b_1 - b_6 + b_7 - b_{10}	0.753	0.937
Audio Only			
Spectral Features	7 PLP Coef. + 7 Deltas	0.69	0.886
Prosody	Pitch + Energy + 2 Deltas	0.534	0.761
Spectral + Prosodic Features	14 Spectral + 4 Prosodic	0.668	0.879

Table 1: F1 measure and AUC for audio- and video-only laughter detector. The fusion function for the two audio / video cues is the SUM rule

Facial Expressions: Modes 7 and 8 are shown in Fig. 4m - 4t. We see that the 7th and 8th modes correspond to mouth opening and closing respectively. In our study, we use shape parameters from 7 to 10, i.e. b_7 to b_{10} as the facial expressions features. Higher modes result in almost no visible expressions and therefore, they are not used in further processing.

Ideally, we would like that the first 6 shape parameters contain only rigid head motion information whereas the other shape parameters (7-10) contain only non-rigid facial motion. However, this is not true and depends on the training data used to built the PDM. This is shown, for example, in Fig. 4a - 4d where apart from the horizontal head movement subtle facial expressions are also present. The same problem is reported in [8] so their training set was augmented with artificial symmetric samples.

6. EXPERIMENTAL STUDIES

In order to investigate which cues carry most of the useful information, we conducted several experimental studies by combining different audio and visual cues for audiovisual laughter detection as well as for audio-only- and video-only-based laughter detection. In all the experiments we performed leave-one-subject-out cross validation, using in every validation fold all samples of one subject as test data and all other samples as training data. The results given in Tables 1, 2, 3 and Fig. 6 represent an average of the results obtained for each fold. In this way it is guaranteed that the obtained results are subject independent. In each cross validation fold, all the features used for training are z-normalized to a mean $\mu = 0$ and standard deviation $\sigma = 1$. Then, the obtained μ and σ are used to z-normalize the features in the test fold. The training and testing of the classifiers is performed on a video / audio frame-level basis. ROC curves, F1 measures, recall and precision rates are used as the performance measures. The following sections describe our results in single-modal and audiovisual laughter detection.

6.1 Single-modal laughter detection

In this set of experiments the detector of laughter vs speech uses information extracted only from one modality, video or audio. The audio-based detector uses two neural networks trained using 14 features for PLP and 4 for pitch and energy respectively, extracted in each frame at 50FPS. Similarly, the video-based detector uses two neural networks trained using 6 features for head and 4 for face, extracted in each frame at 25 FPS. In both cases the SUM rule is used as the integration function. The performance of each channel is shown in the 5th and 9th rows respectively in Table 1.

In order to investigate the performance of each cue seperately, we trained four classifiers using one cue at a time. The results are shown in Table 1. As can be seen, the best performing cues for video and audio are the face and the spectral features respectively. Head movements and prosodic features perform much worse, an indication that they do not contain much useful information for laughter detection when used alone. Even when they are used together with the face and the spectral features respectively they result in degraded performance compared to face and spectral features alone.

Cues	Recall	Precision
Face	0.846	0.815
PLP	0.702	0.679
Face + PLP (SUM)	0.871	0.861
Face + PLP + Head (SUM)	0.853	0.878
Face + PLP (NN)	0.9	0.824
Face + PLP + P & E (NN)	0.919	0.835

Table 3: Recall and precision rates for the best performing audio (2nd row), visual (1st row) and combination of audio and visual cues (rows 3 - 6)

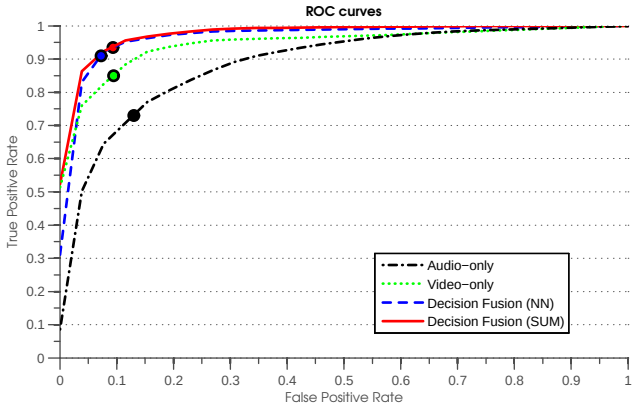


Figure 6: ROC curves for audio-, video-only and decision level fusion. The markers indicate the operating point of each detector which corresponds to the recall and precision rates presented in Tables 1, 2 and 3

6.2 Audiovisual laughter detection

In this set of experiments we use information extracted from both modalities, video and audio. Each cue within each modality is modelled by a neural network and then the outputs of the networks are fused using either another neural network or the SUM rule. We tried all the possible combinations of audio and visual cues and the results are shown in Table 2. The best two combinations are shown in bold for both rules. The spectral features and the face are between the best performing combinations no matter which rule is used. This is not surprising since they are the best performing cues in each modality as shown in section 6.1. In the

	NN		SUM rule	
Cues	F1 Measure	AUC	F1 Measure	AUC
Face + PLP	**0.861**	**0.967**	**0.866**	**0.976**
Face + P & E	0.85	0.958	0.843	0.962
Head + PLP	0.647	0.875	0.657	0.888
Head + P & E	0.554	0.737	0.545	0.792
Face + PLP + P & E	**0.875**	**0.967**	0.839	0.97
Head + PLP + P & E	0.641	0.873	0.687	0.895
Face + Head + PLP	0.832	0.957	**0.865**	**0.978**
Face + Head + P & E	0.742	0.936	0.803	0.96
Face + Head + PLP + P & E	0.847	0.958	0.847	0.975

Table 2: F1 measure and AUC for the audiovisual laughter detector based on different combinations of the audio and visual cues.

case of the SUM rule, the combination Face + Head + PLP is the second best combination and achieves almost the same performance with Face + PLP. This means that the addition of the head does not add any truly useful information that can be exploited by the SUM rule. In fact, as can be seen from Table 2, the addition of head movements in the feature set tends to degrade the detector's performance. This is an indication that head movements are not important for distinguishing between laughter and speech. On the other hand the use of prosodic features in addition to Face + PLP adds information which can be successfully exploited by the neural network resulting in the best overall performance. In fact, from Table 2 we see that, in general, the addition of pitch and energy to the feature set tends to slightly increase the detector's performance. It is interesting that the detector's performance improves when pitch and energy is used together with the visual cues but not when used with the spectral features. For example, compare the 7th and 9th rows of Table 1 with the 3rd row of Table 1 and 4th row of Table 3. This example illustrates the benefits of fusing audio and visual cues since complementary information carried by the audio and video channels can be exploited.

The best performing combination of Face + Head + P & E is an example where the use of a nonlinear function for audiovisual fusion is clearly advantageous over a linear function. However, we see that the performances of the neural network and the SUM rule are comparable. This implies that the use of a nonlinear function modelled by a neural network does not help much. In other words, a linear function is a very good option for the fusion of audio and visual cues.

Table 3 shows the recall and precision rates for the best performing cues in audio, video and audiovisual laughter detection. Fig. 6 shows the ROC curves for the audio-, video-only and audiovisual detectors. The markers indicate the operating point of each detector which corresponds to the recall and precision rates presented in Tables 1, 2 and 3. The results clearly indicate that integrating the information from audio and video leads to an improved performance over single-modal approaches. This is consistent with the results presented in [16].

From Table 3 we also see that neural networks achieve a very high recall rate, over 0.9, in the expense of a lower precision value. On the other hand the SUM rule achieves a balance between recall and precision rates. This is also true for the other combinations of audio and visual cues shown in Table 2.

7. CONCLUSIONS

In this paper we proposed a (semi-)automated audiovisual system for distinguishing laughter from speech episodes. To the best of our knowledge, this is the first study investigating which combination of audiovisual cues leads to the best performance. Initial results suggest that visual information is more important than auditory information for distinguishing laughter from speech. This is a significant finding, especially since visual information has been largely neglected so far in studies on non-linguistic vocal outbursts. The results also suggest that integrating the information from audio and video leads to an improved reliability over single-modal approaches. It is also interesting that a linear function performs very well as the integration function for audiovisual fusion. In this study, only decision level fusion was only investigated. So, when it comes to the level at which the two channels should be integrated, further research is needed.

8. ACKNOWLEDGMENTS

The research leading to these results has been funded in part by the EU IST Programme Project FP6-0027787 (AMIDA) and the EC's 7th Framework Programme [FP7/2007-2013] under grant agreement no 211486 (SEMAINE).

9. REFERENCES

[1] J. A. Bachorowski, M. J. Smoski, and M. J. Owren. The acoustic features of human laughter. *Journal-Acoustical Society of America*, 110(1):1581–1597, 2001.

[2] P. Boersma and D. Weenink. Praat: doing phonetics by computer (version 4.3.01) (www.praat.org)). Technical report, 2005.

[3] R. Cai, L. Lu, H.-J. Zhang, and L.-H. Cai. Highlight sound effects detection in audio stream. In *Multimedia and Expo, 2003. International Conference on*, volume 3, pages III–37–40 vol.3, 2003.

[4] N. Campbell, H. Kashioka, and R. Ohara. No laughing matter. In *European conference on speech communciation and technology:; Interspeech*, pages 465–468, 2005.

[5] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models - their training and application. *Computer Vision and Image Understanding*, 61(1):38, 1995.

[6] S. Dupont and J. Luettin. Audio-visual speech

modeling for continuous speech recognition. *Ieee Transactions on Multimedia*, 2(3):141–151, 2000.

[7] G. Furnas, T. Landauer, G. L., and S. Dumais. The vocabulary problem in human-system communication. *Communications of the ACM*, 30(11):964–972, 1987.

[8] D. Gonzalez-Jimenez and J. L. Alba-Castro. Toward pose-invariant 2-d face recognition through point distribution models and facial symmetry. *IEEE Transactions on Information Forensics and Security*, 2(3):413–429, 2007.

[9] H. Hermansky. Perceptual linear predictive (plp) analysis of speech. *Journal of the Acoustical Society of America*, 87(4):1738–1752, 1990.

[10] A. Ito, W. Xinyue, M. Suzuki, and S. Makino. Smile and laughter recognition using speech processing and face recognition from conversation video. In *International Conference on Cyberworlds, 2005*, page 8 pp., 2005.

[11] L. Kennedy and D. Ellis. Laughter detection in meetings. In *NIST ICASSP 2004 Meeting Recognition Workshop*, 2004.

[12] A. Lockerd and F. Mueller. Lafcam: Leveraging affective feedback camcorder. In *CHI '02 extended abstracts on Human factors in computing systems*, pages 574–575, 2002.

[13] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, and V. Karaiskos. The ami meeting corpus. In *International conference on methods and techniques in behavioral research; Proceedings of measuring behaviour 2005*, pages 137–140, 2005.

[14] M. Pantic, A. Pentland, A. Nijholt, and T. Huang. Human computing and machine understanding of human behavior: A survey. In *Multimodal interfaces; ICMI '06*, pages 239–248, 2006.

[15] I. Patras and M. Pantic. Particle filtering with factorized likelihoods for tracking facial features. In *International conference on automatic face and gesture recognition*, pages 97–104, 2004.

[16] S. Petridis and M. Pantic. Audiovisual discrimination between laughter and speech. In *International Conference on Acoustics Speech and Signal Processing*, pages 5117–5120, 2008.

[17] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior. Recent advances in the automatic recognition of audiovisual speech. *Proceedings of the Ieee*, 91(9):1306–1326, 2003.

[18] B. Reuderink. Fusion for audio-visual laughter detection. *MS Thesis, University of Twente, the Netherlands*, 2007.

[19] J. A. Russell, J. A. Bachorowski, and J. M. Fernandez-Dols. Facial and vocal expressions of emotion. *Annual Review of Psychology*, 54:329–349, 2003.

[20] K. Scherer. Affect bursts. In S. van Goozen, N. van de Poll, and J. Sergeant, editors, *Emotions: Essays on emotion theory*, pages 161–193. 1994.

[21] M. Schroder, D. Heylen, and I. Poggi. Perception of non-verbal emotional listener feedback. In R. Hoffmann and H. Mixdorff, editors, *Speech prosody*, pages 1–4, 2006.

[22] K. P. Truong and D. A. van Leeuwen. Automatic discrimination between laughter and speech. *Speech Communication*, 49(2):144–158, 2007.

[23] M. F. Valstar, H. Gunes, and M. Pantic. How to distinguish posed from spontaneous smiles using geometric features. In *Proc. of the ACM Int. Conf. on Multimodal Interfaces*, pages 38–45, 2007.

[24] Z. Zeng, M. Pantic, G. Roisman, and T. Huang. A survey of affect recognition methods: Audio, visual and spontaneous expressions. *IEEE Trans. Pattern Analysis and Machine Intelligence, accepted for publication*, 30, 2008.