Audiovisual Laughter Detection Based on Temporal Features

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

Previous research on automatic laughter detection has mainly been focused on audio-based detection. In this study we present an audio-visual approach to distinguishing laughter from speech based on temporal features and we show that integrating the information from audio and video channels leads to improved performance over single- modal approaches. Static features are extracted on an audio/video frame basis and then combined with temporal features extracted over a temporal window, describing the evolution of static features over time. The use of several different temporal features has been investigated and it has been shown that the addition of temporal information results in an improved performance over utilizing static information only. It is common to use a fixed set of temporal features which implies that all static features will exhibit the same behaviour over a temporal window. However, this does not always hold and we show that when AdaBoost is used as a feature selector, different temporal features for each static feature are selected, i.e., the temporal evolution of each static feature is described by different statistical measures. 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 an F1 rate of over 89%.

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

ICMI'08, October 20–22, 2008, Chania, Crete, Greece.

1. INTRODUCTION

Interpersonal communications are regulated by audiovisual feedback provided by the involved parties. There are several channels through which the feedback can be provided with the most common being speech. However, spoken words are highly person and context dependant [5], so speech recognition and extraction of semantic information about the underlying intent is a very challenging task for machines [28]. Other channels which provide useful feedback in human-human interactions include facial expressions, head / hand gestures and non-linguistic vocalizations. While there are numerous works on automatic recognition of facial expressions and head/hand gestures, automatic recognition of non-linguistic vocalisations remains an unexplored area. Scherer [20] defines non-linguistic vocalizations as very brief, discrete, nonverbal expressions of affect in both face and voice. People are very good at recognizing emotions just by hearing such vocalizations [21], which suggests that rich information related to human emotions is encoded by these vocalizations. For example, laughter is a very good indicator of amusement and crying is a very good indicator of sadness.

One of the most important non-linguistic vocalizations is laughter, which is reported to be the most frequently annotated non-verbal behaviour in meeting corpora [10]. In the same work it is reported that 8.6% of the time when a person vocalizes in a meeting is spent on laughing. Laughter is a powerful affective and social signal since people very often express their emotion and regulate conversations by laughing [19]. In human - computer interaction (HCI), automatic detection of laughter can be used as a useful cue for detecting the user's affective state and conversational signals such as agreement and, in turn, facilitate affect-sensitive human-computer interfaces [13]. Also, semantically meaningful events in meetings such as topic change or jokes can be identified with the help of a laughter detector. In addition, such a detector can be used to recognize segments of non-speech in automatic speech recognition and 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. Existing approaches to laughter detection include the work of Lockerd and Mueller [11] and Cai et al. [1], who used spectral/cepstral coefficients and HMMs for laughter detection, the work of Campbell et al. [2], who used phonetic features and HMMs to detect four types of laughter, and the work of Kennedy and Ellis [9], who trained Support Vector Machines

(SVM) with Mel-Frequency Cepstral Coefficients (MFCCs) and delta MFCCs. The most extensive study in this area was made by Truong and Leeuwen [24], who compared the performance of different auditory frame and utterance level features using different classifiers and different combinations thereof. To the best of our knowledge, four approaches have been proposed so far that are based on audiovisual information [8], [15], [16], [18]. Ito et al. [8] 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 two individual detectors are fused on decision level achieving 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 (PCA) 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%. Petridis and Pantic [15], [16] used spectral features and prosodic features together with visual features based on PCA as the audio and visual features respectively. Both decision- and feature-level fusion were used which outperformed single-modal detectors, achieving over 90% recall in a person-independent test.

In this paper, we present an audiovisual approach to discriminating laughter episodes from speech episodes based on temporal features, i.e. features which describe the evolution of static features over time. Our research on an audiovisual approach rather than an audio-only approach to laughter recognition is mainly driven by research on audiovisual speech and affect recognition that reported improved performance over audio-only speech/affect recognition [17], [28]. We should note that we use only spontaneous (as opposed to posed) displays of laughter and speech episodes from the audiovisual recordings of the AMI meeting corpus [12] in a person-independent way which makes the task of laughter detection even more challenging [28]. We compare the performance of different temporal features for both single-modal and audiovisual detectors. Our results show that each static feature is best described in time by the combination of several temporal features (which are different for each static feature) rather than a fixed set of temporal features applied to all static features. We used decision and feature level fusion and found that their performance is equivalent when temporal features are used. Our results also show that audiovisual laughter detection outperforms single-modal (audio / video only) laughter detection, attaining an F1 rate of over 89%.

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 than posed smiles [26]. It is also believed that spontaneous smiles exhibit the characteristics of automatic movement, i.e. the motor routines seem to be pre-programmed [3]. On the other hand, posed smiles are less likely to exhibit characteristics of pre-programmed motor routines, because they are mediated by greater cortical involvement [3]. It is reasonable to believe that this finding is also true for other expressions apart from smiles. In conclusion, it is widely believed that spontaneous expressions may significantly differ from posed expressions. Evidence supporting this hypothesis is provided by the significant degradation in performance of tools trained and tested on posed expressions when applied to spontaneous expressions. This is the reason we only used spontaneous expressions in this study.

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 (FPS)) 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-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.

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. For each channel the following two types of features are used:

Static: Static features are computed for each audio / video frame based on information provided by the current frame only and ignoring any past information.

Temporal: Temporal features are computed for each audio / video frame but information is extracted from a temporal window of duration T ending at the current frame. This is equivalent with using a sliding window of duration T for extracting the temporal features which moves forward one frame at a time. In this way, not only the current state of the frame is taken into account but also its history, i.e how it reached the current point. In this way we capture some temporal characteristics of displayed audiovisual behavioural cues which seem to be very important for interpretation of human behaviour, as argued in psychological literature (e.g., [19]). For both audio and video we compare 4 different temporal feature sets.

1. **PCA-based:** One of the most common approaches to extract features from temporal windows is to stack all the features together and then apply a dimensionality reduction technique, like PCA or LDA, to reduce the very high dimensionality. In this study, we extract the audio / visual features for each audio / video frame contained in the temporal window T and then we apply PCA. We keep the first n principal components which account for more than 90% of the variation. Therefore, the information of the temporal window is encoded in n temporal features.

2. **Mean and Standard Deviation:** Again, we extract the audio / visual features for each audio / video frame contained in the temporal window T, and then compute the mean and standard deviation of each feature. A similar approach was used by Kennedy and Ellis [9] but instead of considering a temporal window of length T, they considered the whole laughter segment. We prefer simple statistical features like mean and standard deviation following the findings presented in [22] and given their good performance in [9]. Using this approach the information of the temporal window is encoded in $2 * K$ temporal features, where K is the number of static features per frame.
3. **Polynomial Fitting:** The values that features take over the temporal window T, create a curve which can be approximately described by a p^{th} order polynomial. This approach was successfully adopted by [25] for facial action unit detection. We experimented with linear, quadratic and cubic polynomials. The best results were obtained with a quadratic polynomial. Only these results are presented in this paper. Therefore, the evolution of each feature f_k over the temporal window is described with 3 parameters, f_{k1}, f_{k2} and f_{k3} as shown in eq. 1.

$$f_k = f_{k1}t^2 + f_{k2}t + f_{k3}, k \in [1...K] \quad (1)$$

4. **AdaBoost-based:** When considering temporal features, which describe the evolution of static features over time T, it is common to apply the same set of functions to all static features. In other words, the assumption is made that the evolution of all static features in time can be described in the same way. However, this is not always true and it is reasonable to believe that the temporal evolution of (some) static features will be different. In order to capture those different characteristics we consider a pool of features, which contain all above-mentioned feature sets, together with the following statistical features: median, 3rd and 4th moments, 1st and 3rd quantiles, kurtosis, skewness, and weighted mean and standard deviations (samples which are further away in time from the current frame are weighted less than samples closer to the current time when the mean and standard deviation are computed). Then, we apply feature selection by means of AdaBoost. In this way, we can keep the temporal features that best describe the evolution of each static feature.

AdaBoost: AdaBoost is a machine learning technique and is one of the most popular ensemble learning methods. Training occurs in N rounds, by incrementally adding weak learners to a final strong learner. AdaBoost can be used either as a classification tool or as a feature selector [27]. In the latter case, if a pool of M features is available and the weak learner is restricted to use only one feature per round, then Adaboost finds the best single feature in each round. In other words, the feature used to train the best weak classifier in round N_i is the best feature for that round. In this way a feature set of the N best features will be available after N rounds.

Details on how the static audio and visual features are extracted are presented in sections 4 and 5. Once the static and temporal features are extracted for both modalities, then they are fused with the two commonly used fusion methods, decision level and feature level fusion.

Feature Level Fusion: The extracted audio and visual features are combined and then fed to a classifier. Processing of all features increases the dimensionality of the problem and makes the problem more complex since it requires a large amount of training data. In addition, an important issue that usually has to be addressed is that of the synchronization of features coming from different modalities. Once the features are synchronized, the most common approach to feature level fusion, which is also used in this paper, is their concatenation.

Decision Level Fusion: The most commonly used level of fusion is decision level fusion, which is based on the fusion of modalities on a higher level, i.e. each modality is processed independently and then the final outputs are fused using various integration rules. This approach does not require synchronization of features coming from different data streams but the correlation between the features across different sources is lost.

4. AUDIO MODULE

The audio module is responsible for the audio signals processing. It extracts features from the audio signal on a frame-by-frame basis which are then used by the classification algorithm. The features used in this study are the Perceptual Linear Prediction (PLP) coefficients . Spectral or cepstral features, such as PLP features [7], have been successfully used for speech recognition. Although they were designed for speech recognition applications they have been also successfully used for laughter detection as well [24], [15]. Petridis and Pantic [16], for example, reported a higher success rate in automatic laughter detection when using PLP features than other prosodic features. The same result is also reported by Truong and Leeuwen [24] who compared PLP with other non-spectral features. e.g. pitch and energy. By experimenting with the number of PLP coefficients we found that the use of 7 PLP coefficients [16], instead of 13 which is commonly used in speech recognition applications, leads to better performance for the task of laughter detection. This is also consistent with the finding of Kennedy and Ellis [9]. In [15], it was reported that a frame rate of 50 FPS with a 50% overlap has almost the same performance as when using higher frame rates. Therefore, the framerate of 50 FPS was selected, i.e., the window size is 40ms and the step-size is 20ms. In addition to the 7 PLP coefficients, their delta features were calculated as well. The delta features are calculated by a linear regression over a short neighborhood around the target spectral feature. The slope of the fitted line represents the derivative of the spectral feature and therefore captures some local temporal characteristics. In total, 14 auditory features are computed per frame.

5. VIDEO MODULE

The video module is responsible for the visual signals processing. The first step is to track some characteristic facial points, which will be used subsequently for feature extraction. Then a Point Distribution Model (PDM) is learnt with the aim of decoupling the head movement from the movement produced by the displayed facial expressions.

5.1 Tracking

To capture the facial expression dynamics, we track 20 facial points as shown in Fig. 1 in the video segments.

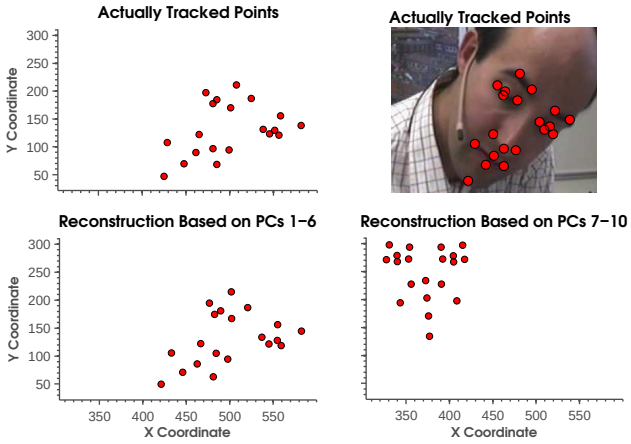

Figure 1: PCA analysis of facial point tracking. Upper row: actually tracked facial points. Bottom row: (left) 20 facial points after they have been reconstructed using the first 6 principal components, (right) 20 facial points after they have been reconstructed using principal components 7 to 10.

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 [14]. 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 these 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. As shown in [26] large head movements typify spontaneous rather than acted behaviour. Since we are interested in learning facial expression configurations typical for speech and laughter episodes, 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 need to decouple head movements from facial expressions. To do so we use a similar approach to that of Gonzalez-Jimenez and Alba-Castro [6], 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 [4].

First, we concatenate the (x, y) coordinates of the 20 tracked 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, b is a N-dimensional vector, $\overline{x}$ is the mean shape and x is the input tracked points. With the help of eq. 2 we can compute the shape parameters b and then the face can be reconstructed using eq. 3. As can be seen from Fig. 1, it seems that indeed the lower-order PCs (1 to 6) reflect rigid-movement aspects of the data, while the higher-order PCs (7 - 10) reflect facial expression aspects of the data. The same has been reported by Gonzalez-Jimenez and Alba-Castro [6]. Ideally, we would like that the first 6 PCs contain only rigid head motion information whereas the other PCs (7-10) contain only non-rigid facial motion. However, this claim cannot be made in a general case as it depends on the training data used to built the PDM. As shown in [16] the head pose does not contain useful information therefore only the PCs 7 to 10 were used.

6. EXPERIMENTAL STUDIES

In order to investigate which temporal features are most informative for the task at hand, we conducted several experimental studies by using different temporal features 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. Then the results obtained in each fold are averaged in order to get the final results. In this way it is guaranteed that the obtained results are subject independent. Since neural networks are used as classifiers, each time a cross-validation loop runs, the obtained results are slightly different due to different initialization conditions. The results presented in Fig. 2 and Tables 1, 2 and 3 are the average results of running each cross-validation loop 25 times. In order to assess if the performance of two classifiers is statistically significant t-tests have been conducted. In each cross validation fold, all 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 set. The F1 measure is used as the performance measure.

F1 Measure: Recall and precision are two commonly used rates for measuring the performance of binary classifiers. Recall 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 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:

$$F_\alpha = \frac{(1+\alpha) \times precision \times recall}{\alpha \times precision + recall} \tag{4}$$

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 $\alpha = 1$.

6.1 Single-modal laughter detection

In this set of experiments, the laughter vs speech detector uses information extracted from only one modality, video or audio. The static audio-based detector uses a neural network trained using 14 PLP-related features (described in section 4) extracted in each frame at 50 FPS. Similarly, the static video-based detector uses a neural network trained using the 4 shape parameters (PCs 7 - 10), which describe the facial expressions as shown in Section 5, extracted in each frame at 25 FPS. Classification is done per frame. In order to investigate the performance of the temporal features, we trained four classifiers using one temporal feature set (see section 3) at a time in addition to the static features. The shortest laughter segment in the dataset is 360 ms and the longest temporal window considered was 320 ms. The results for video and audio are shown in Tables 1 and 2 respectively.

Audio: From Table 1 it can be seen that as the length of the temporal window increases, so does the F1 measure for the temporal features based on mean and standard deviation and Adaboost. On the other hand, the PCA-based and polynomial-fitting features result in lower and almost steady performance respectively as the temporal window increases. It is worth pointing out that the addition of the simplest features based on mean and standard deviation leads to a significant performance increase from 68.18% to 74.68%. Similarly, the use of AdaBoost-based features results in an even higher increase in performance obtaining an F1 measure rate of 76.77%. Therefore, we only consider the Adaboost features with the longest possible temporal window (320ms) for audiovisual fusion. When using AdaBoost as a feature selector we need a stopping criterion, i.e., we need a way to define the number of rounds, N, it will run directly influencing the number of most informative features it will select. In order to do that, we add a large number of features and then select N as the value that gives a good compromise between the number of features (we want as few features as possible) and error over the training set. Using this approach, AdaBoost stops after 23 rounds. The total number of temporal features considered, which includes the features described in section 3, is 196.

Video: As can be seen from Table 2, when it comes to video modality, only the PCA-based and the polynomial-fitting features result in an increasing performance as the temporal window increases. However, the addition of the PCA-based feature degrades the performance of the video classifier since the obtained F1 measures are lower than the F1 measures for the static-video-features based classifier. On the other hand, the addition of the polynomial-fitting features is beneficial since an F1 rate of 85% is achieved for the longest temporal window (in correspondence to F1 of 83.49% achieved when only static features are used). This increase is statistically significant at a 95% confidence interval (p-value = 0.0001). The performance of the other 2 temporal feature sets results in a peak either for a 160ms or a 240ms temporal window. The improvement when adding the mean and standard deviation is statistically significant at a 95% confidence interval (p-value = $8 * 10^{-4}$) but this is not the case with AdaBoost features since the p-value is 0.0617. The same stopping criterion is used for AdaBoost as in the case of audio features, which results in the selection of 13 features. The total number of temporal features considered (see section 3) is 56. Although AdaBoost was successfully used to select the most informative temporal features in audio the same is not true for video. A possible explanation is that the we use a fixed-length window. Most of the audio feature sets reach the maximum performance at the same window length. For example, mean and standard deviation (Table 1) and the other statistical features considered in AdaBoost reach the maximum performance in 320ms. This is why AdaBoost reaches its best performance for the window length of 320ms. This does not hold for video, since different temporal features achieve the maximum performance in different window lengths. So, it is expected that feature selection by AdaBoost will benefit with the inclusion of temporal features calculated over various window lengths.

We also notice that the performance gain from the inclusion of the temporal features is not as high as in the audio modality. This is not surprising since the information contained in a single video frame is much richer compared to the information contained in a single audio frame. This is evident from the successful development of several frame-based computer vision applications, for example [23], [27], and it is also supported by our results. Therefore, the extra information made available by means of adding more video frames is less significant than that made available by means of adding of more audio frames. The main conclusions drawn from the above experiments can be summarized as follows:

1. The best temporal features for audio-based laughter detection are the statistical features selected by Adaboost whereas for video-based laughter detection the best temporal features are the quadratic-fitting features.
2. For audio-based laughter detection, the longer the window the better the performance calculated in terms of the F1 measure. This is not true for video, since the peak performance is achieved for window lengths of 160s and 320ms.
3. The use of temporal features is more beneficial for audio than it is for video-based laughter detection.
4. It is beneficial for audio-based laughter detection to consider a combination of different temporal features, which describe the evolution of static features over time in a sliding window, rather than considering a fixed set of predefined temporal features, e.g. mean and standard deviation only.
5. We see that even simple statistical temporal features like the mean and standard deviation can lead to a significant improvement in performance. In both cases, audio- and video-based laughter detection, they are the second best feature set. Particularly for video, the difference between the best result of mean and standard deviation (84.70%) and the best result of quadratic-fitting (85%) is not statistically significant at a 95% confidence interval (p -value = 0.0209).
6. AdaBoost achieves very good results in audio even with short windows. For example, the temporal features extracted over a 160ms window is better than any other feature set no matter how long the window is.

Feature Sets	Dim	T = 80ms	T = 160ms	T = 240ms	T = 320ms
Static Features + Delta Features					
PLP + Δ PLP	14	68.18	68.18	68.18	68.18
Static Features + Temporal Features					
PCA-based	41	68.17	68.01	67.42	65.73
Mean + Std	42	71.95	72.87	73.71	74.68
AdaBoost-based	37	72.32	75.24	75.50	76.77
Quadratic Fitting	56	70.06	69.96	70.04	69.38

Table 1: Mean F1 rate over 25 experiments for various feature sets/window lengths for the audio-only detector

Feature Sets	Dim	T = 80ms	T = 160ms	T = 240ms	T = 320ms
Static + Delta Features					
$b_7 - b_{10}$	4	83.49	83.49	83.49	83.49
Static + Delta + Temporal Features					
PCA-based	11	82.58	82.97	83.09	83.21
Mean + Std	12	84.15	84.66	84.70	83.63
AdaBoost-based	17	82.81	84.01	83.63	83.58
Quadratic Fitting	16	83.74	84.56	84.66	85

Table 2: Mean F1 rate over 25 experiments for various feature sets/window lengths for the video-only detector

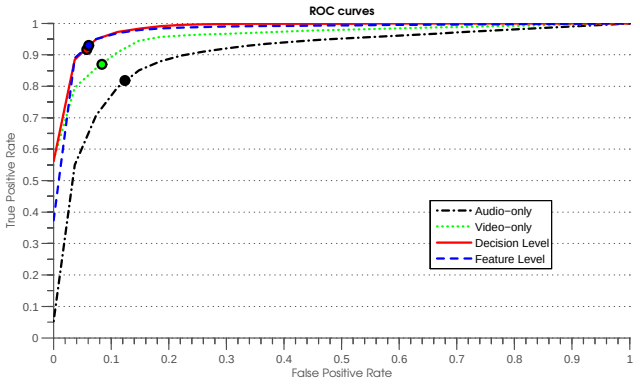

Figure 2: ROC curves for audio-, video-only, decision and feature level fusion. The markers indicate the operating point of each detector which corresponds to the F1 rates presented in Tables 1, 2, 3

6.2 Audiovisual laughter detection

In this set of experiments, we use information extracted from both modalities, video and audio. As described in section 3, multimodal data is fused on decision and feature level. The SUM rule is used as the integration function for the decision level fusion. In feature level fusion, the video features are upsampled to 50 FPS and then concatenated with the audio features.

Table 3 shows the F1 measure for the two different types of fusion and the different types of feature sets used for fusion. We used the best audio and visual features for combination, i.e. AdaBoost-based for audio with a window length of 320ms and quadratic-fitting features for video with a window length of 320ms. In addition, we also consider one more case that is worth investigating and is shown in the last row of Table 3. In the case in question, we first extract all temporal features for audio and video and concatenate them (resulting in a feature vector with $196 + 56 = 252$ dimensions). Then we apply AdaBoost to select the best 36 features. This choice is based on the fact that the best performance of AdaBoost is achieved with 23 and 13 features for audio and video respectively.

Fig. 2 shows the ROC curves for the best audio-only, video-only and audiovisual detectors. The markers indicate the operating point of each detector which corresponds to the F1 rate presented in Tables 1 (7th row), 2 (8th row) and 3 (6th row). These results, together with the results presented in Table 3 clearly indicate that integrating the temporal information from audio and video leads to an improved performance over single-modal and static-features-only approaches. When comparing the two different types of fusion, we see that decision and feature level fusion are almost identical when the temporal feature sets are used. However, decision-level clearly outperforms feature-level fusion when static features are used. In other words, we see that feature-level fusion benefits more from the addition of the temporal features. This is not surprising since feature-level fusion is expected to take advantage of the correlation between the two modalities which is stronger (more apparent) when longer temporal windows are considered. The correlation between the synchronized audio and video frames is weaker (less apparent) when only one frame is used.

It is also interesting to note that the second type of feature level fusion (concatenation of audio and video features followed by the AdaBoost-based selection of a smaller set of the most informative featrues) performs equally well with the first type of feature fusion, i.e. when the best features sets from each modality are used. The main conclusions drawn from the above experiments can be summarized as follows:

1. For both types of multimodal data fusion, the inclusion of the temporal features results in improved performance over the static feature set performance.
2. Audiovisual laughter detector outperforms single-modal detectors when temporal features are used. When static features are used, only the decision-level leads to significant improvement over single-modal laughter detection.

Type of Fusion	Audio features	Visual Features	F1
Static Features			
Decision Level	PLP + ΔPLP	$b_7 - b_{10}$	86.53
Feature Level	PLP + ΔPLP	$b_7 - b_{10}$	83.72
Static Features + Temporal Features			
Decision Level	PLP + ΔPLP + AdaBoost	$b_7 - b_{10}$ + Quadratic Fitting	89.31
Feature Level	PLP + ΔPLP + AdaBoost	$b_7 - b_{10}$ + Quadratic Fitting	89.08
Feature Level	PLP + ΔPLP + $b_7 - b_{10}$ + AdaBoost		89.23

Table 3: F1 measure for the two different types of audiovisual fusion, decision and feature level fusion

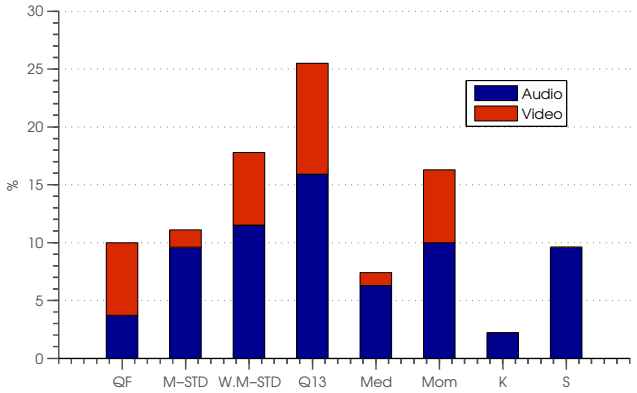

Figure 3: Distribution of the selected audio and visual features. QF: quadratic fitting, M-STD: mean & standard deviation, W.M-STD: weighted M-STD, Q13: 1st/3rd quantiles, Med: Median, Mom: 3rd/4th moments, K:Kurtosis, S:Skewness

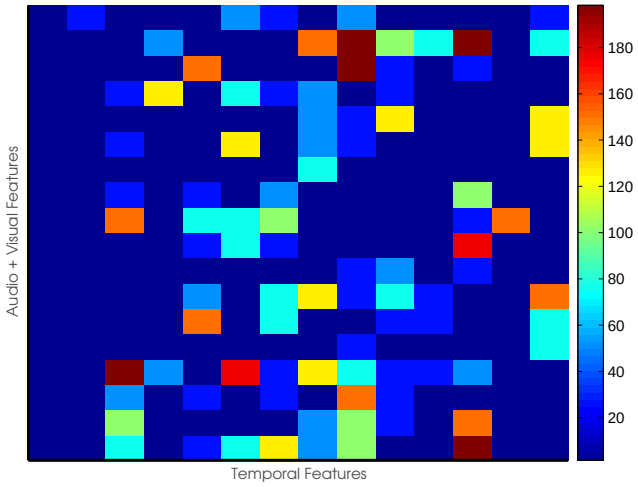

Figure 4: Map showing the total number of times a temporal feature (horizontal axis) was selected to describe a static audio/visual feature(vertical axis)

3. Performances of laughter detectors based on feature-level fusion and decision-level fusion are equivalent when temporal features are used. This can be interpreted as an indication that the correlations between the used audio and visual features are weak.
4. Decision-level outperforms feature-level-fusion-based laughter detectors in the case of static features.

6.3 Feature Analysis

As mentioned in section 6.1, the concatenation of all temporal audio and video features followed by feature selection by AdaBoost performs equally well with the first type of feature fusion, i.e. when the best features sets from each modality are used. Therefore, it is worthy investigating which audio and video features are selected. Fig. 3 shows the distribution of the AdaBoost-selected audio and visual features from the various temporal feature sets. The features that have been most frequently selected by the AdaBoost from either audio or video features are those belonging to the 1st/3rd quantiles. The audio features being least frequently selected are those from the kurtosis feature set. The video features that have not been selected at all are those from the kurtosis and skewness feature sets. We can also see that the audio temporal features are more often selected than the visual temporal features with the only exception being the quadratic-fitting feature set from which more visual than audio temporal features are selected. In total, the ratio between the selected audio and visual temporal features is 68.9:31.1. It is also interesting to point out that features from a single feature set were never selected for more than 15.9% for audio (quantiles) and 9.6% for video (quantiles). In other words, the fact that there is no dominant feature set suggests that indeed the time evolution of each static feature is different and various temporal features should be used. This is better illustrated in Fig. 4 which shows the temporal features selected by AdaBoost for each static audio and visual features in all 25 experiments. It is obvious that the selected temporal features vary a lot between different static features. Since in each experiment an 8-fold cross-validation is run each temporal feature cannot be selected more than 200 times.

7. CONCLUSIONS

In this paper we proposed a (semi-)automated audiovisual system for distinguishing laughter from speech episodes. We investigated the use of different temporal features in order to describe the time evolution of the static features. It has been shown that the additional information provided by the temporal features is beneficial for this task. It has been also demonstrated by means of experimental evaluation that it is better to use a (different) combination of temporal features for each static feature rather than applying the same set of temporal features. Regarding the level at which multimodal data fusion should be performed, both decision- and feature-level fusion approaches resulted in equivalent performances when temporal features were used. However, when static features were used, decision-level fusion outperformed

feature-level fusion. The results also suggest that integrating the information from audio and video leads to improved reliability over single-modal approaches when temporal features are used. Future work includes investigations of different temporal features per static feature computed in windows of different lengths, which is expected to be beneficial for video.

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] R. Cai, L. Lu, H.-J. Zhang, and L.-H. Cai. Highlight sound effects detection in audio stream. In *ICME*, volume 3, pages 37–40, 2003.

[2] N. Campbell, H. Kashioka, and R. Ohara. No laughing matter. In *European Conf. on Speech Comm. and Technology*, pages 465–468, 2005.

[3] J. F. Cohn and K. L. Schmidt. The timing of facial motion in posed and spontaneous smiles. *Intern. Journal of Wavelets Multiresolution and Information Processing*, 2:121–132, 2005.

[4] 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.

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

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

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

[8] A. Ito, W. Xinyue, M. Suzuki, and S. Makino. Smile and laughter recognition using speech processing and face recognition from conversation video. In *Intern. Conf. on Cyberworlds, 2005*, pages 8–15, 2005.

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

[10] K. Laskowski and S. Burger. Analysis of the occurrence of laughter in meetings. In *10th ISCA Intern. Conf. on Spoken Language Processing, INTERSPEECH*, pages 1258–1261, 2007.

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

[12] 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 *Int'l. Conf. on Methods and Techniques in Behavioral Research*, pages 137–140, 2005.

[13] M. Pantic, A. Pentland, A. Nijholt, and T. Huang. Human computing and machine understanding of human behavior: A survey. In *ICMI*, pages 239–248, 2006.

[14] I. Patras and M. Pantic. Particle filtering with factorized likelihoods for tracking facial features. In *Int'l Conf. on Automatic Face and Gesture Recognition*, pages 97–104, 2004.

[15] S. Petridis and M. Pantic. Audiovisual discrimination between laughter and speech. In *ICASSP*, pages 5117–5120, 2008.

[16] S. Petridis and M. Pantic. Fusion of audio and visual cues for laughter detection. In *ACM Intern. Conf. on Image and Video Retrieval*, pages 329–337, 2008.

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

[18] B. Reuderink, M. Poel, K. Truong, R. Poppe, and M. Pantic. Decision-level fusion for audio-visual laughter detection. In *Joint Workshop on Machine Learning and Multimodal Interaction*, 2008.

[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 *Speech prosody*, pages 1–4, 2006.

[22] B. Schuller, A. Batliner, D. Seppi, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and V. Aharonson. The relevance of feature type for the automatic classification of emotional user states: Low level descriptors and functionals. In *INTERSPEECH*, pages 2253–2256, 2007.

[23] M. Slaney and M. Covell. FaceSync: A Linear Operator for Measuring Synchronization of Video Facial Images and Audio Tracks. *NIPS*, pages 814–820, 2000.

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

[25] M. Valstar and M. Pantic. Combined support vector machines and hidden markov models for modeling facial action temporal dynamics. In *IEEE Int'l Workshop on HCI, LNCS*, pages 118–127, 2007.

[26] M. F. Valstar, H. Gunes, and M. Pantic. How to distinguish posed from spontaneous smiles using geometric features. In *ACM ICMI*, pages 38–45, 2007.

[27] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. *CVPR*, 1:511–518, 2001.

[28] Z. Zeng, M. Pantic, G. Roisman, and T. Huang. A survey of affect recognition methods: Audio, visual and spontaneous expressions. *IEEE PAMI, accepted for publication*, 2008.