AUDIOVISUAL DISCRIMINATION BETWEEN LAUGHTER AND SPEECH

Stavros Petridis[1] and Maja Pantic[1,2]
[1]Imperial College London, Computing, UK / [2]University of Twente, EEMCS, Netherlands
{stavros.petridis04; 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 leads to an improved reliability of audiovisual approach in comparison to single-modal approaches. We also investigated the level at which audiovisual information should be fused for the best performance. When tested on 96 audiovisual sequences depicting spontaneously displayed (as opposed to posed) laughter and speech episodes, the proposed audiovisual feature-level approach achieved a 86.9% recall rate with 76.7% precision.

Index Terms— Audiovisual data processing, laughter detection, data fusion, nonlinguistic information processing.

1. INTRODUCTION

Laughing, smiling, and talking are arguably our most prominent social signals [1]. Laughs are also very good predictors of affective states such as joy, humour, distress, and anxiety [2]. It is therefore not strange that laughter is reported to be the most often annotated paralinguistic event occurring in recorded natural speech [3].

Automatic laughter detection can be used as a tool for detecting the user's affective state and facilitating affect-sensitive human-computer interfaces [4]. It can be used to identify 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.

Little work has 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). 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 & Mueller [5], who used HMMs and spectral coefficients, the work of Cai et al. [6], who used HMMs with Mel-Frequency Cepstral Coefficients (MFCCs) and perceptual features, and the work of Campbell et al. [7], who used phonetic features and HMMs to detect four types of laughter. Another approach is that of Kennedy & Ellis [8], who trained Support Vector Machines (SVM) with MFCCs and delta MFCCs. The most extensive study in this area was made by Khiet & Leeuwen [3], 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, the only approach that uses audiovisual information is the work of Ito et al. [9]. They 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.

In this paper, we present an audiovisual approach to discriminating 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 [10]. At this point we would like to remark that we only use spontaneous (as opposed to posed) displays of laughter and speech episodes from the audiovisual recordings of the AMI meeting corpus [11]. We focus on person-independent recognition which makes the task of laughter detection even more challenging. We compare the performance of audio- and video-only laughter detection with that of audiovisual laughter detection. Finally, we also investigate two different kinds of multimodal data fusion: feature and decision level fusion. Independently of the type of data fusion, audiovisual laughter detection outperforms single-modal (audio/video only) laughter detection, attaining on average 84% recall.

2. DATASET

The AMI Meeting Corpus consists of 100 hours of meeting recordings. We only used the close-up video recordings of the subject's face (720 x 576 pixels, 25 frames per second) and the individual headset audio recordings (16 kHz). The language used in the meetings is English and the speakers are mostly non-native. 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.

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, i.e. only harmonic, acoustically symmetric laugh episodes were kept [2]. 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. DATA PROCESSING

3.1. Video Processing

To capture the facial expression dynamics, we track 20 facial points (Fig. 1) 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 Patras – Pantic particle filtering tracking scheme [12], 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 20 points tracked in K frames.

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) rather than deliberately displayed episodes of speech and laughter. Since we are interested in facial expression configuration (relevant to speech and laughter episodes) rather than in head movements, we need to distinguish changes in the location of facial points caused by changes in facial expression from those caused by rigid head movements. To do so we use Principal Component Analysis (PCA). PCA is defined as an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance by any projection 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

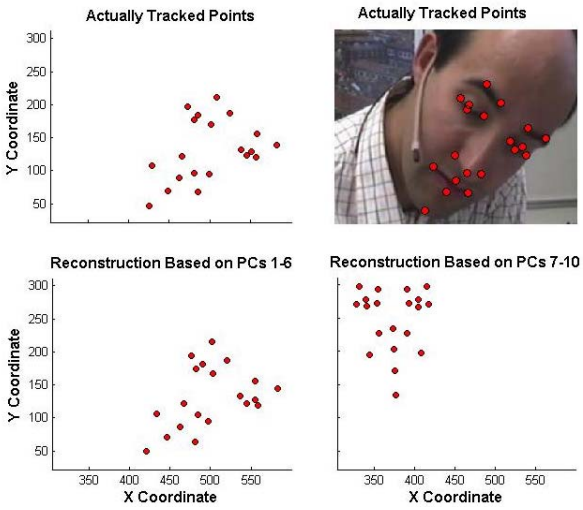

Fig. 1: PCA analysis of facial point tracking. Upper row: actually tracked facial points. Bottom row: (left) 20 facial points after they have been reconstructed using the first 6 principal components, (right) 20 facial points after they have been reconstructed using principal components 7 to 10.

using different combinations of the PCs. As can be seen from Fig. 1, it seems that indeed the lower-order PCss reflect rigid-movement aspects of the data, while the higher-order PCs reflect facial expression aspects of the data.

In our evaluation studies, we use PCs 7 to 10 to reconstruct the position of 20 facial points in each frame. Then we calculate all the distances between all these points. Hence, we end up with 190 distances per frame. We use only spatial features (rather than using temporal features as is often the case in automatic facial expression analysis) since the facial expression configuration during laughter episodes is significantly different than that occurring during speech episodes [1].

3.2. Audio Processing

Spectral or cepstral features, such as Perceptual Linear Prediction coding features (PLP) [13], have been successfully used for speech recognition. They have been successfully used for laughter detection as well. Truong & Leeuwen [3], for example, reported a higher success rate in automatic laughter detection when using PLP features than when using prosodic features like pitch and energy. Hence, we adopt this approach as well and compute, for each window, 13 PLP features and their temporal derivatives resulting in 26 features per window. All the features are z-normalized to a mean $\mu = 0$ and standard deviation $\sigma = 1$.

We also experimented with different frame rates in order to find an optimum for this application. Fig. 2 shows the Receiver Operating Characteristic (ROC) curves for four different frame rates, 25, 50, 75, and 100 frames per second (FPS) respectively. These correspond to a step-size of 40, 20, 13.3, and 10 ms respectively. The length of the window

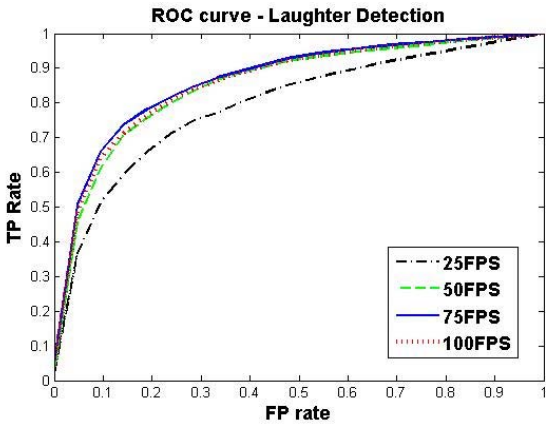

Fig. 2: ROC curves for audio-only classification using different frame rates

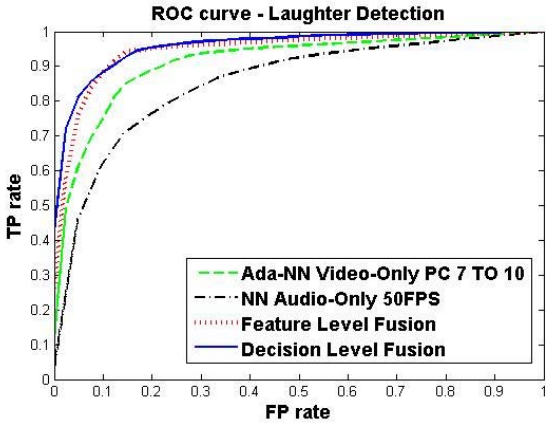

Fig. 3: ROC curves for audio-, video-only, feature level and decision level fusion

is the double of the step-size. As can be seen from Fig. 2, the worst results in audio-only classification between laughter and speech are obtained when the frame rate is 25FPs. However, there are no significant differences in results when the other three frame rates are used. In turn, and in order to keep the number of feature vectors computed for each segment relatively small, we set the frame rate to 50 FPS.

4. LAUGHTER DETECTION

To recognize whether an input audio and / or visual sample is an episode of speech or an episode of laughter we use either a neural network classifier or a combination of AdaBoost and neural networks, where AdaBoost is used as the feature selector rather than a classifier.

Neural Networks were used as the classifier since they are able to learn non-linear function from examples but any learning algorithm which can learn complex functions from examples such as SVM is expected to perform equally well. We used feed-forward neural networks where the number of training epochs varied from 50 to 100 depending on the amount of data that were fed to the network.

AdaBoost creates an ensemble of T weak classifiers, by resampling the data in T rounds, which are combined using a combination rule with the goal of improving the accuracy of any given learning algorithm. AdaBoost can be used either as a classification method or as a feature selector (a feature set of the T best features will be available after T rounds, [14]). In this study, AdaBoost is used as a feature selector. The video channel results in a large amount of features which has the potential to significantly degrade the performance of a learning algorithm unless a large amount of training data is available. Hence, we use AdaBoost to reduce the number of visual features by selecting the most discriminative features, which are then fed to a neural network. We found that good results can be obtained by keeping just the first 18 visual features selected by AdaBoost (from a total of 190 features computed for each frame, see section 3.1). It is interesting to note that in all the experiments described in section 5, AdaBoost always selected the distance between the lower lip and the right lip corner as the most informative feature. The next most commonly selected informative feature is the distance between the lower lip and the lower left eyebrow. While the first feature is related to the extent to which the mouth corner is raised (which is larger for smiles than for speech), both features are directly related to the extent to which the mouth is open (which is much larger for laughter than for speech, [1]).

5. EXPERIMENTAL STUDIES

In order to investigate: (i) whether integrating audio and visual information on laughter/ speech episodes leads to an improved classification performance, and (ii) on which level the fusion of audiovisual information should be carried out for the best performance, we conducted several experimental studies including feature- and decision-level-based audiovisual laughter detection, audio-only-based, 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 Table 1 and Fig. 3 represent an average of the results obtained for each fold. In this way it is guaranteed that the obtained results are subject independent. The training and testing of the classifiers is performed on a video / audio frame-level basis. ROC curves, recall and precision rates are used as the performance measures.

Single-modal laughter detection: As explained in previous sections, audio-based detector of laughter vs. speech utilizes a neural network trained using 26 PLP features per window at 50FPS. Video-only-based detector uses a neural network trained using 18 visual features (distances between the facial points) selected by AdaBoost in each frame of an input video. Fig. 3 and Table 1 summarize the classification results attained by these detectors. These results clearly

indicate that laughter detection based on visual information significantly outperforms the one based on audio information only. Given that the results obtained for audio-only-based laughter detection are comparable to those obtained by other researchers in the field (e.g., [3], [7]), we can conclude that visual information is very important 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 (e.g., [2], [3], [8]).

Audiovisual laughter detection: While all agree that multi-sensory fusion including audiovisual data fusion would be highly beneficial for machine analysis of human behavior (e.g., analysis of affective states), it remains unclear how the fusion should be accomplished [4]. Studies in neurology on fusion of sensory neurons are supportive of early data fusion (i.e., feature-level fusion) rather than late data fusion (i.e., decision-level fusion). However, it is an open issue how to construct suitable joint feature vector composed of features from different modalities with different time scales, different metric levels and different time scales. We approach the problem in a very simple (arguably oversimplified) manner. To achieve decision-level fusion, the input coming from each modality (audio and video) is modeled independently by a neural network, and these single-modal recognition results are combined at the end using the SUM function. To achieve feature-level fusion, we concatenate audio and video features into a single feature vector. To do so, we up-sample the video modality so that the video frame rate (originally 25 FPS) equals the audio frame rate of 50 FPS. As suggested by research in audiovisual speech recognition [10], up-sampling is done by copying each frame. The resulting feature vector is then used to train the target classifier. Fig. 3 and Table 1 summarize the classification results attained for feature- and decision-level laughter vs. speech detection. These results clearly indicate that integrating the information from audio and video leads to an improved reliability of audiovisual approach in comparison to single-modal approaches. Yet, the results are inconclusive when it comes to the level at which the two data streams should be integrated. Although the feature-level fusion attains significantly higher recall rates, the precision decreases. More experimental results using more data are needed if this issue is to be properly investigated.

Table 1: Recall and precision for audio-, video-only, feature level and decision level fusion

	Recall	Precision
Audio-Only	66.04%	61.54%
Video-Only	82.55%	78.16%
Decision Level Fusion	81.66%	82.28%
Feature Level Fusion	86.87%	76.67%

6. 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 both (i) whether integrating audio and visual information on laughter/ speech episodes leads to an improved classification performance, and (ii) on which level audiovisual information should be fused for 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 indicate that integrating the information from audio and video leads to an improved reliability of audiovisual approach in comparison to single-modal approaches. However, the results are inconclusive when it comes to the level at which the two data streams should be integrated. Further research using more data samples is needed on this topic.

ACKNOWLEDGEMENTS

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).

REFERENCES

[1] R.R Provine. Yawns, laughs, smiles, tickles, and talking. In: The Psychology of Facial Expression, J.A. Russell & J.M. Fernandez-Dols, (Eds.), 158-175, 1997.

[2] M. Schroeder. Experimental study of affect bursts. Speech Communication, 40, 99-116, 2003.

[3] K.P. Truong, D.A. van Leeuwen. Automatic discrimination between laughter and speech. Speech Comm., 49, 144-158, 2007

[4] M. Pantic, A. Pentland, A. Nijholt, T. Huang. Human computing and machine understanding of human behaviour: A survey. LNAI, 4451, 47–71, 2007.

[5] A. Lockerd, F. Mueller. LAFCam – leveraging affective feedback camcorder. Proc. CHI'02, 574–575, 2002.

[6] R. Cai, L. Lie, H-J. Zhang, L-H. Cai. 2003. Highlight sound effects detection in audio stream. Proc. ICME'03, 37–40, 2003.

[7] N. Campbell, H. Kashioka, R. Ohara. No Laughing Matter. Proc. Interspeech, 465–468, 2005.

[8] L.S Kennedy, D.P.W Ellis. Laughter detection in meetings. Proc. ICASSP Meeting Recognition Workshop, 2004.

[9] A. Ito, X. Wang, M. Suzuki, S. Makino. Smile and laughter recognition using speech processing and face recognition from conversation video. Proc. Int'l Conf. Cyberworlds, 2005.

[10] S. Dupont, J. Luettin. Audio-Visual Speech Modeling for Continuous Speech Recognition. IEEE Trans. Mult, 141-151, 2000

[11] I. McCowan, et al. The AMI Meeting Corpus. Proc. Measuring Behavior, 2005.

[12] I. Patras, M. Pantic. Particle filtering with factorized likelihoods for tracking facial features. Proc. FG, 97–102, 2004

[13] H. Hermansky. Perceptual linear predictive (PLP) analysis of speech. J. Acoustical Society of America, 87, 1738–1752, 1990.

[14] P. Viola, M. Jones. Robust real-time face detection. Proc. ICCV, 2001