Audiovisual Laughter Detection Based on Temporal Features

Stavros Petridis [a] Maja Pantic [ab]

[a] *Dept. Computing, Imperial College London,180 Queen's Gate, SW7 2AZ, London, UK*
[b] *EEMCS, Univ. Twente, Drienerlolaan 5, 7522 NB, Enschede, NL*

Abstract

Previous research on automatic laughter detection has mainly been focused on audio-based detection. In this study we present an audiovisual approach to distinguishing laughter from speech based on temporal features and we show that the integration of audio and visual information 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. 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%.

1 Introduction

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. Laughter is a powerful affective and social signal since people very often express their emotion and regulate conversations by laughing. In human - computer interaction (HCI), automatic detection of laughter can be used as a useful cue for detecting the user's affective state and, in turn, facilitate affect-sensitive human-computer interfaces. 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 the majority of these studies is that only audio information is used, i.e., visual information carried by facial expressions of the observed person is ignored. Here we present an audiovisual approach in which audio and visual features are extracted from the audio and video channels respectively and fused on decision- or feature-level fusion. The aim of this approach is to discriminate laughter episodes from speech episodes based on temporal features, i.e. features which describe the evolution of static features over time.

2 System Overview

As an audiovisual approach to laughter detection is investigated in this study, information is extracted simultaneously from the audio and visual channels. For each channel two types of features are computed: static and temporal. The static features used are the PLP coefficients for audio and 4 shape parameters for video computed in each audio/video frame respectively. The shape parameters are computed by a point distribution model, learnt from the dataset at hand, with the aim of decoupling the head movement from the movement produced by the displayed facial expressions [1]. The 4 shape parameters used are those which correspond to the facial expressions. The temporal features considered are simple statistical features, e.g. mean, standard deviation, etc, calculated over a window T together with the coefficients of a quadratic polynomial fitted in the same window T. When considering temporal features, which describe the evolution of static features over time T (size of the used temporal window), 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

Type of Fusion	Audio features	Visual Features	F1
Static Features			
Audio only	PLP + ΔPLP	-	68.18
Video Only	-	4 Shape Param.	83.49
Decision Level	PLP + ΔPLP	4 Shape Param.	86.53
Feature Level	PLP + ΔPLP	4 Shape Param.	83.72
Static Features + Temporal Features			
Decision Level	PLP + ΔPLP + AdaBoost	4 Shape Param. + Quadratic Fitting	89.31
Feature Level	PLP + ΔPLP + AdaBoost	4 Shape Param. + Quadratic Fitting	89.08

Table 1: F1 measure for the two different types of audiovisual fusion, decision and feature level fusion

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 feature(s) will be different. In order to capture those different characteristics we consider a pool of features, which contains all the temporal features. Then AdaBoost is applied (as a feature selector) to select the temporal features that best describe the evolution of each static feature.

Once the static and temporal features are extracted for both modalities, then they are fused with the two commonly used fusion methods, decision- and feature- level fusion. Neural networks are used as classifiers for both types of fusion.

3 Dataset

Posed expressions may differ in visual appearance, audio profile, and timing from spontaneously occurring behavior. 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 use only spontaneous (as opposed to posed) displays of laughter and speech episodes from the audiovisual recordings of the AMI meeting corpus [2] in a person-independent way which makes the task of laughter detection even more challenging.In total, we used 40 audio-visual laughter segments, 5 per person, and 56 audio-visual speech segments.

4 Results

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. It has been also demonstrated that the additional information provided by the temporal features is beneficial for this task. 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. Our results also show that audiovisual laughter detection outperforms single-modal (audio / video only) laughter detection, attaining an F1 rate of over 89% (see Table 1).

Acknowledgements

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

References

[1] 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.

[2] 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.