Audiovisual Discrimination Between Laughter and Speech

Stavros Petridis [a] Maja Pantic [a b]

[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 and we show that the integration of audio and visual information leads to improved performance over single-modal approaches. We consider two cases, one that we discriminate between laughter and speech and one that we discriminate between voiced laughter, unvoiced laughter and speech. When tested on 207 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 90% and a classification rate of over 80%.

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. The next step is to divide the laughter episodes into 2 widely accepted laughter categories, voiced and unvoiced, with the aim to investigate if it is possible to discriminate these two types of laughter from speech using again audiovisual features.

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. The visual channel is divided into 2 streams (cues): face and head movements as shown and the audio channel is divided into 2 streams as well: spectral and prosodic features.

The visual features used are 10 shape parameters computed in each video frame. 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 first 6 shape parameters used correspond to head movements and the remaining 4 correspond to facial expressions. The audio spectral features extracted are the mean and standard deviation of 6 Mel Frequency Cepstral Coefficients (MFCCs) and Δ MFFCs over a temporal window of 320ms which slides forward 160 at a time. The

Cues	F1 Voiced Laughter	F1 Unvoiced Laughter	F1 Speech	CR 3 classes	F1 (2 classes)	CR 2 classes
Face	34.60	61.06	82.37	66.22	83.73	84.37
Head	24.93	29.74	53.68	39.22	54.70	55.15
MFCC	57.97	69.14	85.68	74.83	86.16	86.69
Prosody	57.24	59.65	74.39	66.20	72.69	74.43
Face + MFCC + Prosody	65.05	76.76	93.33	81.93	93.39	93.63

Table 1: Columns 2 to 4: F1 measure per class, Column 5: Classification Rate of the 3 class problem, Column 6: F1 measure for the laughter vs speech detector, Column 7: Classification rate for the 2 class problem

prosodic features used are the mean and standard deviation of pitch and energy together with the unvoiced ratio over the same window.

Once the audio and visual 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. In the case of the 3 class problem, i.e. voiced laughter vs unvoiced laughter vs speech, 3 one-vs-all classifiers are trained which are combined to achieve the final result.

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 67 audio-visual voiced laughter segments, 48 unvoiced laugther segments and 92 audio-visual speech segments from 10 subjects.

4 Results

Table 1 shows the performance of each cue separately and the best combination of cues when using audiovisual fusion on feature level. We see that face and MFCCs are the best performing cues in each modality and their combination with prosodic features on feature level results in the best audiovisual performance. The last columns show the results of the laughter vs speech detector which has a very good performance. Columns 2 to 5 show the results of the 3 class detector (voiced laughter / unvoiced laugther / speech). The F1 measure is computer per class, whereas the 5th column shows the overall classification rate. As expected the performance is worse when compared to the 2 class problem, however it is possible to discriminate the 2 types of laughter satisfactorily. Regarding the level at which data fusion should be performed, initial experiments suggest that feature level fusion performs slightly better decision level fusion . Our results also show that audiovisual laughter detection outperforms single-modal (audio / video only) laughter detection, attaining an classification rate of over 80% for the 3 class problem and over 90% for the 2 class problem (see Table 1).

5 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] 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.