Audiovisual Discrimination Between Speech and Laughter: Why and When Visual Information Might Help

Stavros Petridis, *Member, IEEE*, and Maja Pantic, *Senior Member, IEEE*

***Abstract*—Past research on automatic laughter classification/detection has focused mainly on audio-based approaches. Here we present an audiovisual approach to distinguishing laughter from speech, and we show that integrating the information from audio and video channels may lead to improved performance over single-modal approaches. Both audio and visual channels consist of two streams (cues), facial expressions and head pose for video and cepstral and prosodic features for audio. Two types of experiments were performed: 1) subject-independent cross-validation on the AMI dataset and 2) cross-database experiments on the AMI and SAL datasets. We experimented with different combinations of cues with the most informative being the combination of facial expressions, cepstral, and prosodic features. Our results suggest that the performance of the audiovisual approach is better on average than single-modal approaches. The addition of visual information produces better results when it comes to female subjects. When the training conditions are less diverse in terms of head movements than the testing conditions (training on the SAL dataset, testing on the AMI dataset), then no improvement was observed with the addition of visual information. On the other hand, when the training conditions are similar (cross validation on the AMI dataset), or more diverse (training on the AMI dataset, testing on the SAL dataset), in terms of head movements than is the case in the testing conditions, an absolute increase of about 3% in the F1 rate for laughter is reported when visual information is added to audio information.**

***Index Terms*—Human behavior analysis, laughter-versus-speech discrimination, neural networks, principal components analysis (PCA).**

I. Introduction

In human–human interaction, communication is 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 dependent [18], making the speech recognition and extraction of semantic information about the underlying intent a very challenging task for machines [68]. Other channels which provide useful feedback in human–human interactions include facial expressions, head and hand gestures, and nonlinguistic vocalizations. While there are numerous works on automatic recognition of facial expressions and head and hand gestures, automatic recognition of nonlinguistic vocalizations has attracted less attention [37], [68]. Scherer [54] defines nonlinguistic vocalizations (or nonverbal 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 [55], which suggests that information related to human emotions is conveyed 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 nonlinguistic vocalizations is laughter, which is reported to be the most frequently annotated acoustic nonverbal behavior in meeting corpora [30]. In the same work, it is reported that 8.6% of the time when a person vocalizes in a meeting is spent on laughing and an additional 0.8% is spent on laughing while talking. Laughter is a powerful affective and social signal since people very often express their emotion and regulate conversations by laughing. It has been reported that people frequently laugh after their own utterances and it has been suggested that this provides a mechanism to change the meaning of the utterance [63].

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 [6]. This information can be used by affect-sensitive human–computer interfaces [37] to make the interaction between humans and machines more natural and user-friendly. Another area of application is computer-aided psychotherapy where a computer not only saves the patient's answers while performing some computer-based exercises but also monitors his/her reactions, which is of particular interest to psychologists [34]. Also, semantically meaningful events in meetings such as topic change or jokes can be identified with the help of a laughter detector [64]. Provine [46] has shown that people tend to laugh at places where punctuation would be placed in a transcript of a conversation. Hence, such a detector can be used in automatic speech recognition for speech segmentation and for the recognition of nonspeech segments. Finally, a laughter detector can be a useful tool for multimedia tagging and retrieval. A user can watch a

Manuscript received December 09, 2009; revised July 26, 2010 and December 02, 2010; accepted December 02, 2010. Date of publication December 23, 2010; date of current version March 18, 2011. This work was supported by the European Research Council under the ERC Starting Grant agreement no. ERC-2007-StG-203143 (MAHNOB). The work of S. Petridis was supported in part by the European Community's 7th Framework Programme [FP7/2007-2013] under grant agreement no. 211486 (SEMAINE). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Shrikanth Narayanan.

S. Petridis is with the Department of Computing, Imperial College London, London, U.K. (e-mail: stavros.petridis04@imperial.ac.uk).

M. Pantic is with the Department of Computing, Imperial College London, London, U.K., and also with EEMCS, University of Twente, Enschede, The Netherlands (e-mail: m.pantic@imperial.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2010.2101586

video and then a tag can be automatically generated based on his or her reaction classifying the video as funny or not (implicit tagging), or he/she can search a multimedia database based on the specific content (content-based video retrieval) [38].

It is clear that laughter is an audiovisual event. It consists of an audio component, the laughter vocalization and a visual component which involves facial activity around the mouth, the cheeks, and often the upper face. Changes in the upper face appearance may not be present in ironic (not genuine) smiles [12], but they are usually present in laughter (a typically genuine expression of amusement) and especially in intense laughter episodes [53]. Therefore, it seems logical that the additional information carried by the visual modality would be beneficial for solving the problem of laughter detection/classification in an automatic way. Audiovisual approaches have been successfully applied to speech recognition [15], [45] and affect recognition [68]. In general, the main contribution of the visual information is the addition of complementary and redundant information which cannot be corrupted by acoustic noise in the environment and therefore may improve the performance of a recognition system.

In this paper, we present our research on audiovisual discrimination of laughter from speech. We extend our previous works [41]–[43], by using two cues per channel, spectral and prosodic cues for audio and head pose and facial expressions for video, with features extracted either per frame (video) or in a sliding window (audio) and fused using feature-level fusion. Our research on an audiovisual approach rather than an audio-only approach to laughter classification is mainly driven by research on audiovisual speech recognition that reported improved performance over audio-only speech recognition [15], [45]. Given that the previous research in the field has been focused on laughter classification/detection from the audio signal only (see Section II for overview of the past research), our objective in this study is to investigate if the addition of visual features helps the discrimination between laughter and speech.

We only use spontaneous (as opposed to posed) displays of laughter and speech episodes from the audiovisual recordings of the AMI meeting corpus [35] and the SAL database [14]. We focus on person-independent classification, which makes the task of laughter-versus-speech discrimination even more challenging. The paper is further organized as follows. Section II provides an overview of the past research on laughter classification/detection. Section III details the utilized datasets. Sections IV and V explain audio and video signals processing, respectively. Section VI describes the experimental setup and Section VII presents the experimental results. We compare the performance of several approaches to audiovisual laughter-versus-speech discrimination where different combinations of audio and visual cues are used in the process. The best individual cues were found to be facial expressions and spectral features for discriminating laughter from speech. The combination of facial expressions, cepstral features, and prosody resulted in a relatively small but statistically significant improvement over single-modal approaches. This improvement was found to be person-dependent and more pronounced when it comes to female subjects than male subjects. More specifically, for the majority of female subjects, a statistically significant improvement in the performance of the method has been attained with the addition of visual information, while only in the case of 25% of male subjects did the method benefit from adding the visual information.

When tested on 278 audiovisual sequences from the AMI corpus in a cross validation manner, the absolute increase for the classification rate and F1 rate for laughter is 2.4% and 3%, respectively. When training on the AMI dataset and testing on the SAL dataset, which is a less diverse dataset than AMI, an absolute increase of 2.3% and 3.4% is achieved for the classification rate and F1 rate for laughter, respectively. However, when a system is trained on the SAL dataset and tested on the AMI dataset, no improvement was observed with the integration of audio and visual information. These experiments also reveal that the AMI dataset is a more challenging dataset since systems trained on it and tested on the SAL dataset achieve much better performance than systems trained on the SAL dataset and tested on the AMI dataset.

II. Past Research on Laughter-versus-Speech Discrimination and Laughter Detection

A. Research in Psychology

Laughter is one of the most common and useful human social signals [64]. It helps humans to express their emotions and intentions in social interactions and provides useful feedback during interpersonal interactions. It is usually perceived as positive feedback, i.e., it shows joy, acceptance, and agreement, but it can also be used as negative feedback, e.g., irony. Campbell [10] presented results from the telephone conversations between Japanese speakers, showing that the speakers varied their laughing styles according to the sex and nationality of the partner. Provine [47] found that in the absence of stimulating media, e.g., television, people are about 30 times more likely to laugh, whereas they are only four times more likely to talk, when they are in company than when they are alone. Vettin and Todt [63] found that laughter is much more frequent in conversations than what had been previously reported in self-report studies. Babies have the ability to laugh before they can speak [53] and children who were born both deaf and blind still have the ability to laugh [16]. This suggests that at least some features of laughter can be developed without the experience of hearing/seeing laughter which is the evidence of a strong genetic basis [47]. These facts illustrate the significance of laughter and explain why it is considered one of the most important universal nonverbal vocalizations. However, it is surprising that our knowledge about laughter is still incomplete and little empirical information is available [28].

Since laughter has attracted interest by researchers from many disciplines, the terminology is sometimes confusing. Ruch and Ekman [53] point out that laughter is not a term used consistently nor is it precisely defined in research. In addition, Trouvain [59] points out that terms related to laughter are either not clearly defined or they are used in different ways in different studies.

Several classifications have been proposed in the literature regarding different types of laughter. The most commonly accepted one is the discrimination of laughter into two types: voiced and unvoiced [3], [21]. Voiced laughter is a harmonically rich, vowel-like sound with a measurable periodicity in vocal fold vibration, whereas unvoiced laughter is a noisy exhalation through nose or mouth and the vocal folds are not involved in the production of laughter. These two broad categories are characterized by significant variability. Especially the unvoiced class can contain different unvoiced variants such as grunts, pants, cackles, and snort-like sounds. Another classification has been proposed by Campbell *et al.* [11], which does not label an entire laughter episode but assumes that each laughter is composed of different combinations of four laughter segments: voiced, chuckle, breathy, and nasal.

It has been demonstrated that different types of laughter have different functions in social interactions. Grammer and Eibl-Eibesfeldt [21] found that male interest was partly predicted by the number of voiced laughs produced by female partners. The opposite does not hold and this result has also been confirmed by Bachorowski and Owren [3]. The latter study also demonstrated that voiced laughter always elicited more positive evaluations than unvoiced laughter. It is also believed that voiced laughter is directly related to the experience of positive affect, whereas unvoiced laughter is used to negotiate social interactions [24]. Except judging social signals like interest, the distinction between voiced and unvoiced laughter could be useful for judging the mirth of the laughter. This could be used for assessing the hilarity of observed material like movies and tagging the material in question accordingly (see [44] for a preliminary study).

Regarding the acoustics of laughter, two main streams can be distinguished in the literature. One suggests that the acoustic features of laughter are stereotyped [48], whereas the other suggests that its acoustics are variable and complex so laughter can be considered as a repertoire of sounds [4], [28]. Although not all studies agree on the findings regarding acoustic parameters of laughter, the majority of them agree on some general principles. Perhaps the most studied parameter in this area is the fundamental frequency F_0 and almost all recent studies agree that mean F_0 is higher in both male and female laughter than it is in speech [4], [52], [60]. The average duration of a laughter episode varies from less than 1 s [4], [52], to approximately 2 s [60]. It is also common to consider laughter as a series of successive elements whose parameters are not constant but changing between or even within elements [28]. Another characteristic of laughter is the alternation of voiced and unvoiced segments with the proportion of unvoiced segments being higher in laughter than in speech [60]. Finally, it has also been reported that the intensity of laughter goes down over time [53].

B. Automatic Laughter Classification/Detection

Relatively few works exist in the literature on automatic laughter classification/detection. These are summarized in Table I. As can be seen from Table I, there is a lack of a benchmark dataset based on which different methods could be compared. The use of different datasets in combination with the use of different performance measures makes the comparison of different approaches almost impossible. Further, as can be seen from Table I, both static and dynamic modeling approaches have been attempted. For dynamic modeling, hidden Markov models (HMMs) are commonly used just as is the case in automatic speech recognition. This is mainly due to suitability of HMMs to represent temporal characteristics of the phenomenon. For static modeling, support vector machines (SVMs) and neural networks (NNs) are the most commonly used tools in this field. Unlike automatic speech recognition where HMMs usually outperform static approaches, initial results on presegmented episodes using static models were very promising, and that explains why these methods are still commonly used. This is also confirmed by Schuller *et al.* [57], who have recently shown that the performance of SVMs is comparable to that of HMMs for the classification of nonlinguistic vocalizations. Another recent study [40] comparing NNs and coupled HMMs for discrimination of laughter-versus speech and posed-versus-spontaneous-smiles has come to a similar conclusion.

Regarding the audio features, several different features have been used with the most popular being the standard features used in automatic speech recognition, mel-frequency cepstral coefficients (MFCCs) and perceptual linear predictive (PLP) coefficients. Pitch and energy, which have been used in emotion recognition from speech [68], are commonly used as well.

From Table I, it can also be seen that the vast majority of the attempts towards automatic laughter classification/detection used only audio information, i.e., visual information carried by facial expressions of the observed person is ignored. Recently, few works on audiovisual laughter detection have been reported, which use information from both the audio and visual channel (see Table I and the end of this section).

1) Audio-Only Laughter Classification/Detection: In this category, works can be divided into two groups: those which focus on the detection of laughter in unsegmented audio stream or on the discrimination between several nonlinguistic vocalizations in presegmented audio episodes (where each episode contains exactly one of the target nonlinguistic vocalizations) and those which perform audio segmentation/classification into several audio categories, which are usually not nonlinguistic vocalizations, but one class is laughter. In the first group, there are usually two approaches:

1) laughter detection/segmentation, e.g., [27], [29], [31], where the aim is to segment an unsegmented audio stream into laughter and nonlaughter episodes;
2) laughter-versus-speech classification/discrimination, e.g., [33], [57], [60], where the aim is to correctly classify presegmented episodes of laughter and speech.

One of the first works on laughter detection is that of Kennedy and Ellis [27], who trained SVMs with MFCCs, spatial cues, and modulation spectrum features (MSFs) to detect group laughter, i.e., when more than a certain percentage of participants are laughing. They used the ICSI corpus achieving true positive and false positive rates of 87% and 13%, respectively. However, inconsistent results were obtained when the system was tested on unseen datasets from NIST RT-04 [2]. Truong and van Leeuwen [61] used cepstral features (PLP) for laughter segmentation in meetings. GMMs were trained for speech, laughter, and silence, and the system was evaluated on the ICSI corpus achieving an EER of 10.9%. Laskowski and Schultz [31] present a system for the detection of laughter and its attribution to specific participants in multi-channel recordings. Each participant can be in one of the three states (silence, speech, laughter) and the

TABLE I

Previous Works on Audio-Only and Audiovisual Laughter Classification/Detection. A: Audio, V: Video, L: Laughter, NL: Nonlaughter, S: Speech, NT: Neutral, Subj: Number of Subjects, Y: Yes, N: No, CV: Cross Validation, SI: Subject Independent, CR: Classification Rate, TP: True Positive Rate, FP: False Positive Rate, EER: Equal Error Rate, R: Recall, PR: PRecision, ER: Error Rate. Window: Duration of the Window Used for Classification/Segmentation. The Label "Sequence" Means That Classification is Performed Directly on the Entire Presegmented Episode. When no Information is Provided in a Study, Then This is Denoted by ?

Study	A/V	Classifier	Features	Window	Dataset	Size	Testing	SI	Classes	Perfomance
Classification										
Truong & van Leeuwen (2005) [60]	A	SVM, GMM	PLP, Pitch&Energy, Pitch&Voicing, MSF	Sequence	ICSI (Bmr,Bed), CGN [36]	L: 3264, S: 3574	Train: L:2422, S:2680 Test: L:894, S:842	Y(Bed, CGN)	Laughter / Speech	EER: Bmr - 2.6% Bed - 2.9% CGN - 7.5%
Campbell (2005) [11]	A	HMM	?	Sequence	ESP [9]	L: 3000	?	?	4 Types / Segments of Laughter	CR: 81% Segments, 75% Laughter
Schuller et al. (2008) [57]	A	SVM, HMM, HCRF	PLP, MFCC	Sequence	AVIC [58]	2901 examples L: 261, Subj: 21	3-fold Stratified CV	Y	5 classes	CR: 80.7% R: 87.7 PR: 75.1%
Lockerd & Mueller (2002) [33]	A	HMM	Spectral Coefficients	Sequence	Own	L: 40, S: 210 Subj: 1	Train: 70% Test: 30%	N	Laughter / Speech	CR: 88%
Reuderink et al. (2008) [50]	AV	A: GMMs, HMMs V: SVMs	A:RASTA-PLP, V: Shape Parameters	Sequence	AMI [35]	L: 60, S: 120 Subj: 10	2 x 15-fold CV	N	Laughter / Speech	EER: 14.2% AUC: 0.93
Petridis & Pantic (2008) [41]	AV	NNs	A: PLP, V: Facial Points Distances	Audio (20ms) / Video (40ms) Frame	AMI	L: 40 (58.4 sec) S: 56 (118.08 sec) Subj: 8	Leave-one-subject-out CV	Y	Laughter / Speech	R: 86.9% PR: 76.7%
Petridis & Pantic (2008) [42]	AV	NNs	A: PLP, V: Shape Parameters	320ms	AMI	L: 40 (58.4 sec) S: 56 (118.08 sec) Subj: 8	Leave-one-subject-out CV	Y	Laughter / Speech	F1: 89.31%
Detection / Segmentation										
Kennedy & Ellis (2004) [27]	A	SVM	MFCC, Mod. Spectrum, Spatial Cues	1 sec	ICSI(Bmr) [26], NIST RT-04 [2]	L: 1926 (ICSI) 44 (NIST), Subj: 8	CV (ICSI) Train: 26 meetings Test: 3 meetings	Y (NIST)	Laughter / Non-Laughter	ICSI TP: 87% FP: 13%
Truong & van Leeuwen (2007) [61]	A	GMMs	PLP	16 ms	ICSI(Bmr)	L: 91min, S: 93min Subj: 10	Train: 26 meetings Test: 3 meetings	N	Laughter / Speech / Silence	EER: 10.9%
Laskowski & Schultz (2008) [31]	A	HMM	MFCC, Energy	100ms	ICSI (Bmr, Bro, Bed)	NT: 716.2min S: 94.4min L: 16.6min Subj: 23	Train: 26 meetings Bmr Test: 3 meetings Bmr, Bro, Bed	N	Laughter / Speech / Neutral	F1: 34.5%
Knox et al. (2008) [29]	A	NNs	MFCC, Pitch, Energy, Phones, Prosodics, MSF	Audio frame (10ms)	ICSI (Bmr)	L: 6641 sec NL: 98848 sec	Train: 26 meetings Test: 3meetings	N	Laughter / Non-Laughter	EER: 5.4%
Ito et al. (2005) [25]	AV	A: GMMs, V: LDFs	A: MFCC, V: Lip angles, lengths, Cheek mean intensities	Audio / Video Frame	Own	3 dialogues, 4 - 8 min each Subj: 3	5-fold CV	N	Laughter / Non-Laughter	R: 71% PR: 74%

aim is to decode the vocal activity of all participants simultaneously. HMMs are used with MFCCs and energy features. The system is tested on the ICSI meeting corpus. To reduce the amount of states that a multi-party conversation can have, they apply minimum duration constraints for each vocalization and overlap constrains which assume that no more than a specific number of participants speak or laugh at the same time. The F1 rate achieved is 34.5%. When tested on unseen datasets, the F1 is less than 20%, but the system does not rely on manual presegmentation. Knox *et al.* [29] used MFCCs, pitch, energy, phones, prosodics, and MSFs with neural networks in order to segment laughter by classifying audio frames as laughter or nonlaughter. A window of 1010 ms (101 frames) was used as input to the neural network and the output was the label of the center audio frame (10 ms). The ICSI corpus was used and an equal error rate of 5.4% was achieved.

The most extensive study in laughter-versus-speech discrimination was made by Truong and Leeuwen [60], who compared the performance of different audio-frame-level features (PLP, Pitch and Energy) and utterance-level features (Pitch and Voicing, Modulation Spectrum) using SVMs and Gaussian mixture models (GMMs). They used the ICSI corpus [26] and CGN corpus [36] achieving an equal error rate of 2.6% and 7.5% in subject-dependent and subject-independent experiments, respectively. Campbell *et al.* [11] first divided laughter into four classes: hearty, amused, satirical, and social and decomposed each laughter into four laughter segments: voiced, chuckle, breathy, and nasal. They used HMMs to recognize these four laughter segments and the four classes of entire laugh episodes from the ESP corpus [9] resulting in classification rates of 81% and 75%, respectively. Schuller *et al.* [57] used the AudioVisual Interest Corpus (AVIC) [58] to classify five types of nonlinguistic vocalizations: laughter, breathing, hesitation, consent, and other vocalizations including speech. They used HMMs and hidden conditional random fields (HCRF) with PLP, MFCC and energy features and SVMs with several statistical features, e.g., mean, standard deviation, etc., which describe the variation over time of other low level descriptors, e.g., pitch, energy, zero-crossing rate, etc. Using a 3-fold stratified cross validation, they reported an overall classification rate of 80.7%. From the confusion matrix provided in [57], the recall and precision of laughter can be computed which are 87.7% and 75.1%, respectively. Lockerd and Mueller [33] used spectral coefficients and HMMs with the aim to detect when the operator of a video camera laughs. The system was trained using data of a single subject achieving a classification rate of 88%.

In the second group of approaches, there are usually several classes which correspond to different sounds, e.g., laughter, applause, music, scream, etc. Because of the nature of this problem, the features used are more diverse. That includes zero crossing rate (ZCR), brightness (BRT), bandwidth (BW), total spectrum power (TSP) and subband powers (SBP) and short time energy (STE) in addition to the standard features mentioned above. SVMs [22] and HMMs [8] have been used and the results of these methods can be seen in Table I. Since these works are not focused on laughter detection/classification, they are not described in this paper in further detail.

2) Audiovisual Laughter Classification/Detection: To the best of our knowledge, there is only one work on audiovisual laughter detection/segmentation and just a few works on audiovisual laughter-versus-speech discrimination and, as a consequence, the approaches followed are less diverse. In all works, the aim is to discriminate laughter from nonlaughter (speech [41]–[43], or speech and silence [25], [50]). The only study on audiovisual laughter detection was conducted by Ito *et al.* [25], who built an image-based laughter detector based on geometric features (lip lengths and angles), mean intensities in the cheek areas (grayscale images were used), and an audio-based laughter detector based on MFCC features. Linear discriminant functions (LDFs) and GMMs were used for the image-based and audio-based detectors, respectively, and the output of the two detectors were combined with an AND operator to yield the final classification for an input sample. They attained 71% recall rate and 74% precision rate using three sequences of three subjects in a person-dependent way. Reuderink *et al.* [50] used visual features based on principal components analysis (PCA) and RASTA-PLP features for audio processing. GMMs and HMMs were used for the audio classifier, whereas SVMs were used for the video classifier. The outputs of the classifiers were fused on decision level, by weighted combination of the audio and video modalities, obtaining an equal error rate of 14.2% in a subject-dependent way on 60 episodes of laughter and 120 episodes of speech from the AMI corpus.

In our previous works on audiovisual laughter-versus-speech discrimination, we used either PLP features alone [41], [42] or together with pitch and energy [43]. As visual features, we used either displacements of the tracked facial points [41], or visual features based on PCA [42], [43]. Neural networks were used as the single-modal classifiers (for audio and video separately), which were fused on the decision and feature level achieving an F1 rate of 89% in a subject-independent test for 40 presegmented laughter episodes and 56 presegmented speech episodes from the AMI corpus. The work presented in this paper represents a continuation and enhancement of this earlier work.

III. Dataset

Posed (acted) 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 [62]. Also it seems that spontaneous smiles exhibit characteristics of automatic movement, i.e., the motor routines seem to be preprogrammed [12]. On the other hand, posed smiles are less likely to exhibit characteristics of preprogrammed motor routines, because they are mediated by greater cortical involvement [12]. In the case of laughter, Ruch and Ekman [53] point out that laughing on command may be embarrassing, and thus, the results obtained from voluntary laughter are of limited value for describing spontaneous laughter. In conclusion, spontaneous expressions may significantly differ from posed expressions. An additional evidence supporting this hypothesis is apparent from the significant degradation in performance of tools trained and tested on posed expressions and applied to spontaneous expressions [68]. For this reason, we only used spontaneous expressions in this study.

Another challenge in studying laughter is the lack of data. Since laughter usually occurs in social situations, when people are in groups, it is not easy to obtain clear recording of individual spontaneous expressions of laughter. Consequently, meeting corpora are commonly used where laughter often occurs as described below. For the purpose of this study we used two datasets, one containing social interactions between four subjects (AMI dataset) and the other one containing interaction between subjects and an artificial agent (SAL dataset).

1) AMI Dataset: The AMI Meeting Corpus [35] is an ideal dataset for our study since it consists of 100 h of meeting recordings where people show a huge variety of spontaneous expressions. We only used close-up video recordings of the subjects' faces (720×576 pixels, 25 frames per second), and the related individual headset audio recordings (16 kHz). The language used in the meetings is English, with speakers being mostly nonnative speakers. For our experiments, we used seven meetings (IB4001 to IB4005 and IB4010, IB4011) and the relevant recordings of ten participants, eight young males (subjects: s01, s04, s05, s06, s07, s08, s09, s10) and two young females (subjects: s02, s03), with or without glasses and no facial hair. Nine participants are of Caucasian origin and one is of Asian origin.

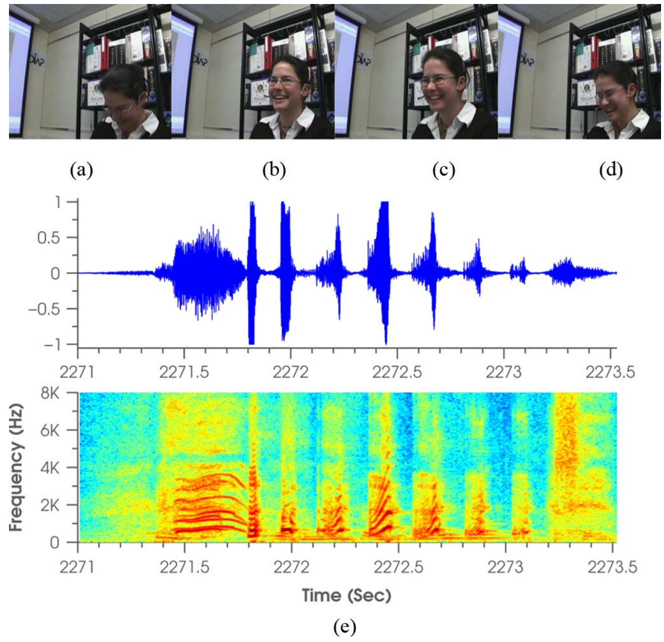


Fig. 1. Example of voiced laughter from subject s02, AMI dataset, Meeting ID:IB4011_1. (a) Frame 1. (b) Frame 21. (c) Frame 41. (d) Frame 63. (e) Top row: Audio signal, Bottom Row: Spectrogram.

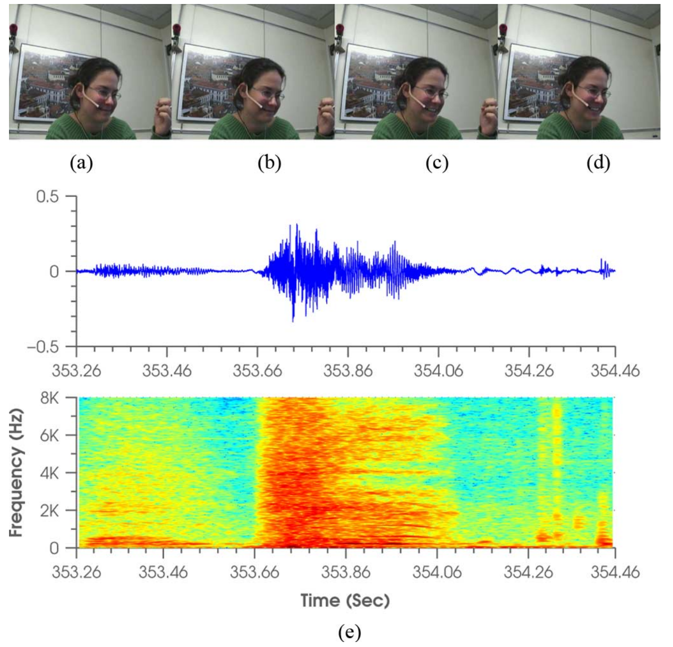


Fig. 2. Example of unvoiced laughter from subject s02, AMI dataset, Meeting ID:IB4002_2. (a) Frame 1. (b) Frame 11. (c) Frame 21. (d) Frame 31. (e) Top row: Audio signal, Bottom Row: Spectrogram.

2) SAL Dataset: The Sensitive Artificial Listener (SAL) technique is described in [14] as “a specific type of induction technique that focuses on conversation between a human and an agent, that either is or appears to be a machine and it is designed to capture a broad spectrum of emotional states”. The subjects interact with four different agents that have different personalities and the audiovisual response of the subjects is recorded. For our experiments, we used 15 subjects in total, eight males (subjects: s03, s05, s06, s07, s08, s09, s10, s13) and seven females (subjects: s01, s02, s04, s11, s12, s14, s15). We used the close-up video recordings of the subject’s face (720 × 576 pixels for 12 subjects and 352 × 288 for subjects s04, s07, and s13) and the related audio recording (48 kHz for 12 subjects and 44.1 kHz for subjects s04, s07, and s13). The language used in the human–agent interaction is English, with all speakers being native.

All laughter and speech episodes used in this study were presegmented based on audio. This means that the start and end point of a laughter episode is defined for the audio signal and then the corresponding video frames are extracted. All methods presented in this study used such audiovisual data for training and testing. So, instead of using information only from the audio modality, as by audio-only approaches, our approach uses the visual information co-occurring with laughter as well, representing an audiovisual approach to laughter classification.

Initially, laughter episodes were selected based on the annotations provided with the AMI Corpus. After examining these episodes, we only kept those that do not co-occur with speech, do not contain profile views of the face (i.e., all facial components are still visible), and satisfy the criterion as suggested in [4]: “Laughter is defined as being any perceptibly audible expression that an ordinary person would characterize as laughter if heard under everyday circumstances”. For the SAL dataset, we manually annotated laughter episodes according to these rules.

In total, there are 633 laughter episodes annotated in the subset of the AMI corpus we use. However, the majority of them violates the aforementioned rules. Specifically, many episodes are with subjects in a profile view to the camera, subjects laughing altogether so the individual subject’s laughter is not perceivable and subjects smiling rather than laughing, i.e., there is no audible laughter expression. For our study, we randomly selected 124 laughter episodes that do not violate the above-mentioned rules. The reason why 124 episodes were selected is that the number of episodes per subject varies a lot. For s01, only seven episodes do not violate the above-mentioned rules. Therefore, in order not to have an extremely unbalanced dataset, the number of laughter episodes per subject was randomly set to be between 7 and either 21 (which is three times the minimum number of episodes) or the maximum number of episodes that do not violate the above rules if this number is less than 21. So in total, 124 laughter episodes were selected, with s01 and s03 being the subjects with the minimum and maximum number of laughter episodes, 7 and 17, respectively. It is important to note that laughter episodes included in our dataset were selected such as to meet the above criteria, no matter how similar they are to a prototypical “ha-ha-ha” laughter expression. Therefore, there are both prototypical and nonprototypical laughter episodes (i.e., unvoiced laughter like snorts and cackles) included in the datasets.

Examples of voiced laughter and unvoiced laughter included in the AMI dataset are shown in Figs. 1 and 2, respectively. Examples of voiced and unvoiced laughter included in the SAL dataset are shown in Figs. 13–15. It is common that audible inhalations are present either at the beginning or at the end of episodes or both and phoneticians argue whether or not they

TABLE II
DESCRIPTION OF THE TWO DATASETS USED IN THIS STUDY

Type	No Episodes (No Subjects)	Total Duration (sec)	Mean / Std (sec)
AMI			
Laughter	124 (10)	145.36	1.17 / 0.73
Speech	154 (10)	285.92	1.86 / 1.12
SAL			
Laughter	94 (15)	136.96	1.45 / 0.78
Speech	177 (15)	377.32	2.13 / 0.80

TABLE III
AVERAGE NUMBER OF VOICED AND UNVOICED LAUGHTER EPISODES PRODUCED BY MALES AND FEMALES

	AMI		SAL	
Gender	Unvoiced Laughter	Voiced Laughter	Unvoiced Laughter	Voiced Laughter
Female	2.0	14.5	2.0	4.4
Male	2.8	8.6	3.3	2.9

should be considered as belonging to an instance of laughter [59]. In this study, inhalations were considered a part of laughter episodes if they were present exactly after or before the laughter.

For the AMI dataset, speech segments were determined by the annotations provided with the AMI Corpus. For the SAL dataset, speech segments were manually annotated and segmented. Finally, speech segments were randomly selected such that they do not contain long pauses between two consecutive words. As a result, only few speech segments are adjacent to laughter segments. Details of the two datasets used in this study are given in Tables II and III. As can be seen from Table II, female subjects produce more voiced laughter than male subjects, and this is consistent with findings in psychology [4]. In this study, a laughter episode is labeled as voiced if at least 20% of its frames are voiced. The annotated episodes used from both corpora can be found in [1].

IV. AUDIO MODULE

The audio module extracts features from the audio part of the input episode, which are then used by the classification algorithm to classify the episode as a speech or laughter episode. Two different kinds of features are used in this study: 1) cepstral features and 2) prosodic features.

A. *Cepstral Features*

Cepstral features, such as MFCC or PLP coefficients, have been widely used in audiovisual speech recognition [15], [45] and language identification [67]. From Table I, it can be concluded that they are very often used for laughter classification/detection as well. On average, MFCC and PLP show very similar performance [56], [67]. This is also confirmed by experiments in our study. We have chosen to use MFCC but the use of PLP results in an equally good performance. The MFCCs were computed using the MATLAB functions provided in [17].

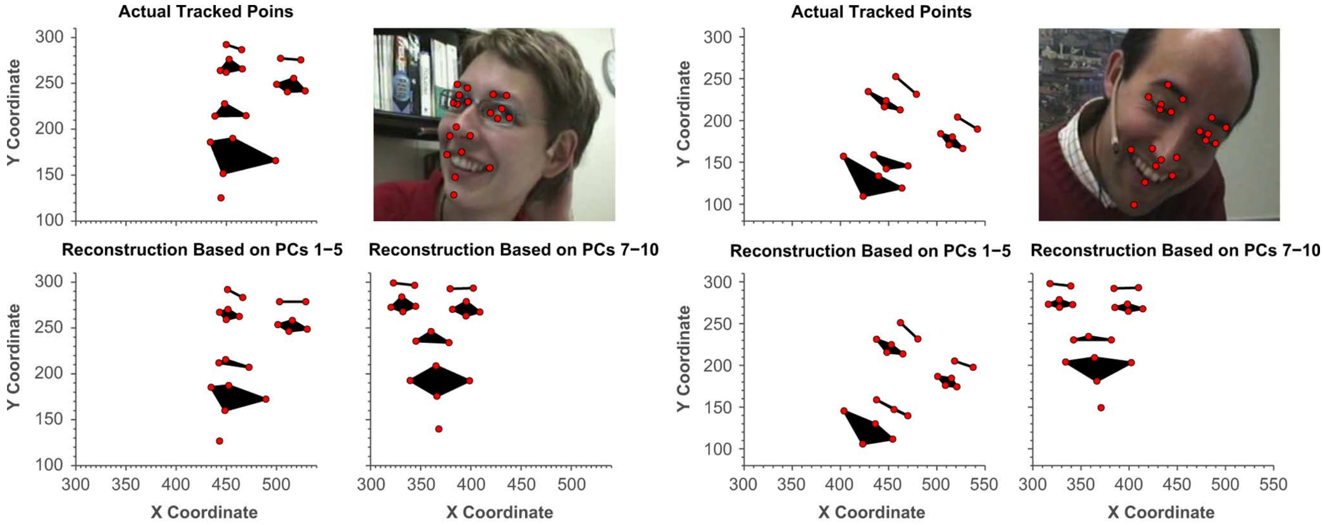
An important issue when using cepstral features is the number of coefficients to be used. The use of 12 or 13 MFCC/PLP coefficients is common in speech recognition. However, using 6 or 7 MFCC/PLP coefficients have been reported to lead to either the same or an improved performance in laughter detection [27], [43] and language identification [67]. Hence, we use 6 coefficients based on the finding of Kennedy and Ellis [27], who reported that using 6 MFCCs results in the same performance as using 13 MFCCs. In addition to the 6 MFCCs, their delta features ($\Delta\mathrm{MFCC}$) were calculated as well. The delta features 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. So in total, 12 features are computed every 10 ms over a 40 ms long frame.

Since not much information is carried by a single frame, it is beneficial to compute features over longer temporal windows as shown in [42]. In order to do that, we compute the mean and standard deviation of each MFCC and $\Delta\mathrm{MFCC}$ over temporal windows of 160 ms. A similar approach was used by Kennedy and Ellis [27], who used windows of 1 s. As shown in [42], choosing a longer temporal window is beneficial in the case of presegmented data as is the case with our data. However, note that this could degrade the performance in a real-world scenario where segmentation is not available. For computational efficiency, there is no overlap between consecutive windows, since the improvement is marginal as shown in [42]. Using this approach, the information contained in each temporal window is encoded in terms of $2 * 12 = 24$ features.

B. *Prosodic Features*

The two most commonly used prosodic features in studies on vocal affect recognition are pitch and energy [68]. Both of them have also been used in previous works on audio-only laughter-versus-speech discrimination (see Table I). The addition of prosodic features to MFCCs, both on decision level and feature level, has been proven to be beneficial for deceptive speech detection [20] and for language identification [67]. In addition, Bachorowski *et al.* [4] found that the mean pitch in both male and female laughter was higher than in modal speech. Hence, in this study, we use both pitch and energy for discriminating laughter from speech episodes.

Pitch (P) was computed using a MATLAB implementation of the Praat pitch estimator described in [5]. The pitch ceiling for the algorithm was set to 1000 Hz. For each frame, the harmonics-to-noise ratio is computed, and if it is lower than 0.45, then the frame is labeled as unvoiced and pitch is not defined. Energy (E) of a signal is simply the sum of squares of the signal's raw values. The root mean square energy is used in this study as the energy feature. Both pitch and energy are computed every 10 ms over a window of 40 ms. Again, we compute statistics of pitch and energy features over a 160 ms long window with no overlap. We compute the same statistics as in [7], i.e., mean, standard deviation, range, median, interquartile range, lower quartile, upper quartile, minimum, and maximum. The statistics for pitch are computed only from the voiced frames. In addition, the unvoiced ratio is computed as well, i.e., the proportion of unvoiced frames contained in the window.



(a) (b)

Fig. 3. PCA analysis of facial point tracking using PCs computed from the AMI dataset. Upper row: actually tracked facial points. Bottom row: (left) 20 facial points after they have been reconstructed using the first 5 principal components, (right) 20 facial points after they have been reconstructed using principal components 7 to 10. (a) Subject s03. (b) Subject s08.

In addition to these prosodic features, the zero crossing rate (ZCR) was computed, too. The reason for using ZCR is its sensitivity to the difference between voiced and unvoiced sections. High zero crossing rates usually indicate noise and low rates usually indicate periodicity [49]. So ZCR is likely to be beneficial in laughter-versus-speech discrimination, since laughter contains more unvoiced frames than speech [60]. The same statistics as for pitch and energy were computed for the ZCR.

In order to avoid the need for synchronization between cepstral and prosodic features, the same window length was used for both. In other words, all the above-mentioned statistical features were computed over a 160 ms long window with no overlap between two consecutive windows.

V. Video Module

The video module is responsible for processing the visual part of an input episode. The first step is to track a number of characteristic facial points. Then, a Point Distribution Model (PDM) is learnt with the aim of decoupling rigid from nonrigid face movements. Both are used in this study: 1) features which correspond to rigid head movements and 2) features which correspond to facial expressions.

A. Tracking

To capture face movements in an input video, we track 20 facial points, as shown in Fig. 3. 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 [39], applied to 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 episode. Hence, for each episode containing K video frames, we obtain a set of K vectors containing 2-D coordinates of the 20 points.

B. Decoupling of Rigid and Nonrigid Face Movements

While speaking and especially while laughing, people may 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 separating 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 rigid head movements from nonrigid head movements (facial expressions) so that we can investigate the effect of each cue separately. To do so, we use a similar approach to that by Gonzalez-Jimenez and Alba-Castro [19], in which PCA is used for decoupling. Our approach is based on PDMs [13] and has also been used in [42], [43], and [50].

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 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-face-movement aspects of the data, while higher-order PCs are expected to retain nonrigid-face-movement (facial expression) aspects of the data. To test this assumption, we computed the PCs for the whole AMI 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:

$$b = (x - \bar{x})P \tag{1}$$

$$x \approx \hat{x} = \bar{x} + bP^T \tag{2}$$

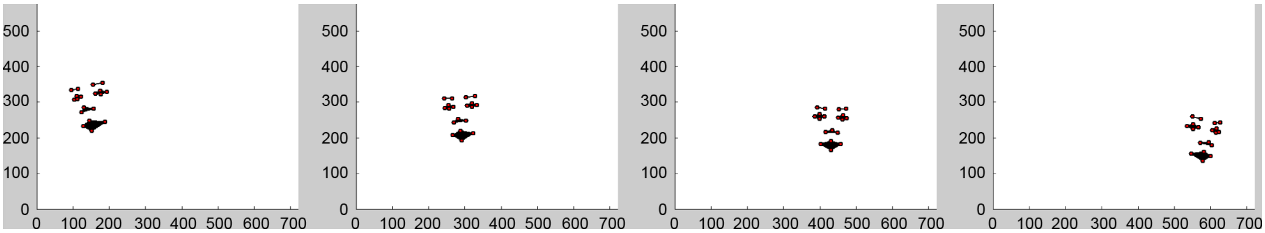


Fig. 4. Principal component analysis, AMI dataset—Mode 1: Effect of varying b_1.

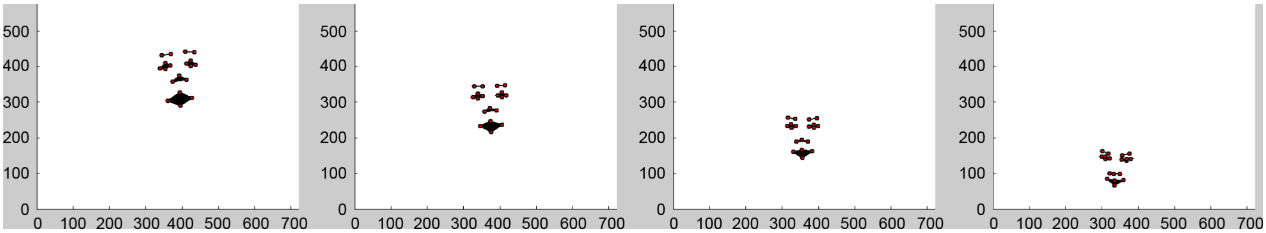


Fig. 5. Principal component analysis, AMI dataset—Mode 2: Effect of varying b_2.

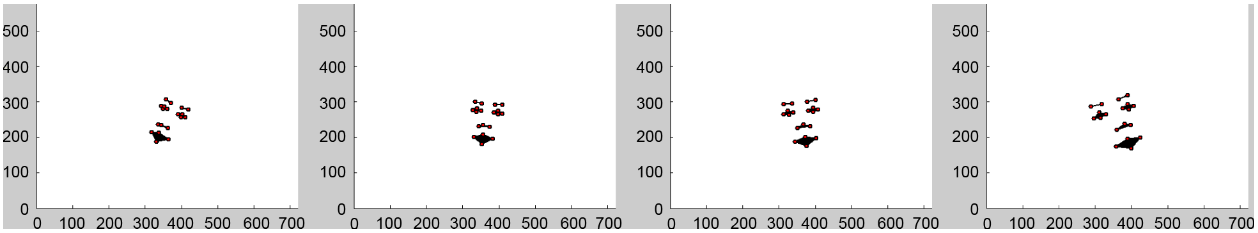


Fig. 6. Principal component analysis, AMI dataset—Mode 3: Effect of varying b_3.

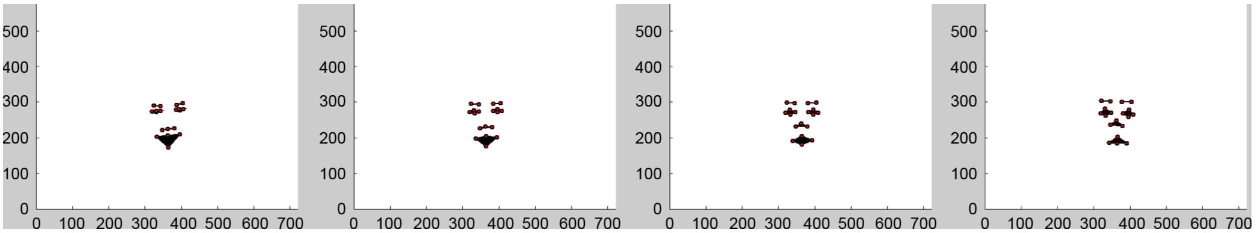


Fig. 7. Principal component analysis, AMI dataset—Mode 7: Effect of varying b_7.

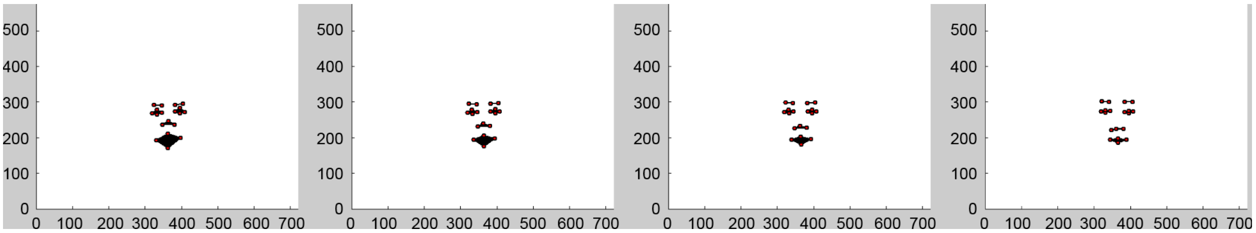


Fig. 8. Principal component analysis, AMI dataset—Mode 8: Effect of varying b_8.

where x is a 40-dimensional vector containing the (x, y) coordinates of the 20 tracked points (1×40), $\bar{x}$ is the mean shape (1×40), P contains N out of the 40 eigenvectors $(40 \times N)$, and b is an N-dimensional vector $(1 \times N)$. With the help of equation (1), we can compute the shape parameters b and then the face can be reconstructed using equation (2).

As can be seen from Figs. 3–8, it seems that indeed the lower-order PCs reflect rigid-face-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 [19]. To further investigate what is captured by each PC, shape parameters b can be varied one at a time. In other words, we can vary shape $\hat{x}$ using equation (2). The variance of the ith 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 ith parameter, b_i, is called the ith mode of the model [13]. By visual inspection, we can identify which PCs reveal information about the facial expression information (nonrigid face motion) and which the information about the head pose (rigid face motion).

1) Head Pose: Modes 1, 2, and 3 are shown in Figs. 4–6. We see that the 1st, 2nd, and 3rd modes correspond to horizontal head movement, vertical head movement, and head rotation, respectively. Similarly, the 4th and 5th modes, which are not shown, correspond to changes in scale and head yaw, respectively. In other words, the first five PCs contain head movement information. Shape parameters [equation (1)] are computed in each frame and since no temporal information is used they correspond to head pose information. Therefore, we use the first five shape parameters, i.e., b_1 to b_5, as the head pose descriptive features.

2) Facial Expressions: Modes 7 and 8 are shown in Figs. 7 and 8. As can be seen, the 7th and 8th modes correspond to mouth movements (mouth closing). Modes 9 and 10, which are not shown, correspond to other facial expressions. Therefore, we use shape parameters 7 to 10, i.e., b_7 to b_{10}, as the facial expressions features. Mode 6 and modes > 10 do not account for any visible or clearly distinguishable change in head pose or facial expression, and therefore, they are not used in further processing.

Ideally, we would like the first five shape parameters to contain only head pose information whereas the other shape parameters (7–10) to contain only nonrigid facial motion. However, this largely depends on the training data used to built the PDM. This is shown, for example, in Fig. 4, where apart from the horizontal head movement, subtle facial expressions are also present. The same problem is reported in [19]. We should note here that this analysis holds when PCs are computed from AMI. In case of the SAL dataset, which is less diverse, the first four modes correspond to head pose and modes 5 to 7 correspond to facial expressions.

As shown in [42], a video frame already contains a lot of information for discriminating laughter from speech. Hence, computing statistical features, like mean and standard deviation, over longer temporal windows, i.e., over several video frames, is not very beneficial [42]. An absolute increase in the F1 measure of 1.21% (from 83.49% to 84.70%) is reported in [42], when using the mean and standard deviation of shape parameters b_7 to b_{10} over a 240 ms window with the maximum overlap between consecutive windows. This result is also confirmed in this study, with an increase of 1.2% in the F1 measure when statistical features of $b_7 - b_{10}$ were considered as well for a 240 ms window with 50% overlap. Since this increase in the classification accuracy is relatively small and in order to make synchronization between audio and visual features easier (since a 160 ms audio window contains exactly four video frames), we use the shape parameters $b_1 - b_5$ and $b_7 - b_{10}$ per frame as the sole visual features.

VI. Classification Methodology and Experimental Setup

Neural networks were used as classifiers in this study since they are able to learn nonlinear functions from examples. As already mentioned in Section II-B, some recent works [40], [57] have shown that the performance of static classifiers like NNs and SVMs is comparable to that of HMMs and coupled HMMs for the classification of nonlinguistic vocalizations. Feedforward neural networks with one hidden layer are used as classifiers in this study and the resilient backpropagation training algorithm [51] is used for training. The learning rate is set to 0.05 and the training is stopped when either the maximum number of epochs is reached (500 in our case) or the magnitude of the gradient is less than 0.04. The number of hidden neurons is defined by means of a 2-fold cross validation in the following way. The subjects used for training are randomly divided into two groups. Then, several networks are trained, with different numbers of hidden neurons, using only subjects from one group and tested on the other group and vice versa. The number of hidden neurons leading to the best performance in terms of the F1 measure is chosen for training a network on the entire training set.

As explained in Section V-B and as shown in Figs. 4–8, modes 1 to 5 and 7 to 10 (based on the AMI dataset) represent head pose and facial expression information. Therefore, when training on the AMI dataset (Sections VII-A and VII-BI), PCs 1 to 5 and 7 to 10 are used which are computed from the AMI dataset. When training on the SAL dataset (Section VII-BII), PCs 5 to 7 are used (since only facial expression information is used), computed from the SAL dataset.

Both audio and visual features are z-normalized per subject, to a zero mean and unity standard deviation, by subtracting the features' means and dividing the result by their standard deviation. Subject normalization has been useful in emotion detection from speech [65] and helps removing subject and recording variability [7].

In the first set of experiments (Section VII-A), we performed leave-one-subject-out cross validation on the AMI dataset, using in each cross validation cycle all samples of one subject as the test data and all other samples from the remaining nine subjects as the training data. In this way, the obtained results are subject independent. The performance of the cross validation overall is the average of the performances in each cross validation cycle (fold). In the second set of experiments (Section VII-B), we performed cross-database experiments on the AMI and SAL datasets, by training a classifier on one dataset and testing it on the other and vice versa. For both types of experiments, ROC curves, area under the ROC curve (AUC), F1 measure, and classification rate are used as the performance measures.

The training and testing of the classifiers is performed on a frame-level basis for video (40 ms) and on window-level basis (160 ms) for audio, as described in Sections IV and V, respectively. Since the audio and visual features are extracted at different frame rates, they must be synchronized for audiovisual fusion. This is achieved by upsampling the audio features, which are extracted at a lower frame rate than the visual features, by simply copying each feature so as to match the rate at which the visual features are extracted [15]. After synchronization, the audio and visual features are concatenated in a single vector for feature-level fusion. In previous work [42], we have shown that the performance of decision-level and feature-level fusion is comparable in case of the laughter-versus-speech discrimination problem. In case of other problems, like audiovisual speech recognition [15], [45], feature-level fusion is still the most commonly used type of fusion, although this solution is suboptimal given the asynchronous nature between audio and video. Motivated by this past research, we use feature-level fusion in this work. Classification is performed by applying either the single-modal or the bimodal classifiers to all individual frames/windows of the given episode resulting in a series of "speech" and "laughter" labels. The majority of these labels is assigned as the label of the entire episode in question.

As shown in Table III, the total duration of speech episodes is higher than the total duration of laughter episodes. Consequently, there are many more speech than laughter frames/windows, which means that the training set can become unbalanced with the speech class containing more than two times more examples than the laughter class. Such an unbalanced set tends to degrade the performance of the classifier [32]. In order to avoid this problem, the speech class is created by randomly sampling examples, such that it contains not more than two times more examples than the laughter class.

Due to the random initialization of the weights of the neural networks and the random sampling, as explained above, each time we run a cross validation or a cross database experiment, we get slightly different results. In order to assess the replicability of the experiments, each cross validation and cross database experiment is executed ten times and the mean and standard deviation are reported.

In order to compare the performance of different combination of audiovisual cues with audio or visual cues, a paired T-test is used. Given the small number of instances, the assumptions of the paired T-test may be violated, so the results should be interpreted carefully. The paired T-test is applied on the average performance measure, i.e., over all subjects, of each cross-validation/cross database experiment. Since each experiment is conducted ten times, we end up with ten paired differences. In this way, we compare the overall performance of the cues in question. In order to get a more detailed view of the comparison, we also apply the paired T-test for each subject separately. The significance level used was set to 5%.

VII. Experimental Studies

As explained in the previous sections, we extract information simultaneously from the audio and the visual channel. The extracted visual information concerns two cues: facial expressions and head pose. Similarly, the extracted audio information channel concerns two cues as well: cepstral (MFCC) and prosodic features (including ZCR). In order to investigate which cues carry useful information for the task in question (i.e., speech versus laughter discrimination), we conducted several experimental studies combining different audio and visual cues for audiovisual, audio-based, and video-based laughter-versus-speech discrimination.

A. *Cross Validation Experiments on the AMI Dataset*

1) Single-Modal Approach: In this set of experiments, the laughter-versus-speech classifier uses information extracted only from one modality, either video or audio. The results for each cue separately are shown in Table IV. As can be seen, the best performing single cue for video are the nonrigid facial movements (facial expressions) and for audio are the spectral features, achieving a CR of 83.9% and 92.3%, respectively. Cepstral features have been already found informative for discriminating laughter from nonlaughter in earlier studies [27], [60]. Face is the channel that carries most information in interpersonal communications [37]. Speech and laughter are arguably the two most common ways of communicating, changing the facial appearance in two very different ways. It is, therefore, only logical that facial expression is informative for the task in question.

TABLE IV
F1 Rates and Classification Rate (CR) for the Audio-Based and Video-Based Discrimination Between Laughter and Speech. The Results Presented are the Mean (and Standard Deviations) of 10-Fold Cross Validation Conducted Ten Times Using the AMI Dataset. The Highest Mean Performances in Each Column are in Bold. FF: Feature-Level Fusion

Cues	Features Used	F1 Laughter	F1 Speech	CR
		Video Only		
Face	b_7 to b_{10}	**80.4 (0.8)**	**86.3 (0.5)**	**83.9 (0.6)**
Head	b_1 to b_5	35.8 (3.9)	63.1 (1.5)	53.2 (1.5)
Face + Head (FF)	b_1 to b_5 + b_7 to b_{10}	78.9 (1.2)	84.7 (0.7)	82.3 (0.9)
		Audio Only		
MFCC	Mean+Std of 6 MFCC + 6 ΔMFCC	89.9 (1.5)	92.8 (0.9)	91.6 (1.1)
P & E + ZCR (FF)	Statistics of Pitch (10), Energy (9), ZCR (9)	81.9 (1.7)	87.8 (1.1)	85.4 (1.4)
MFCC + P & E ZCR (FF)	24 Spectral + 19 Prosodic 9 ZCR	**90.9 (2)**	**93.4 (1.3)**	**92.3 (1.6)**

The classification based on head pose is much worse (CR of 53.2%), indicating that head pose, as used in this study, is not very informative for laughter-versus-speech discrimination when used alone. Even if delta features are added, which can be considered to capture local head movements, the performance remains the same. Head pose/movements have been found to be useful for the discrimination between posed and spontaneous smiles [62]. In addition, there is evidence in the literature that head movements are linked to prosody, pitch in particular [66], during speech and that intense laughter results in significant head movement [53]. However, there is no evidence that head pose/movements are different in laughter than in speech, especially when both of them are spontaneous.

Prosodic features perform quite well achieving a similar performance to facial expression features but then perform worse than cepstral features. The addition of prosodic features to cepstral features slightly improves the performance of the audio-only classifier, but this improvement is not statistically significant.

2) Audiovisual Approach: In this set of experiments, we use information extracted from both modalities, video and audio.

TABLE V
F1 RATES AND CLASSIFICATION RATE (CR) FOR THE AUDIOVISUAL DISCRIMINATION BETWEEN LAUGHTER AND SPEECH. THE RESULTS PRESENTED ARE THE MEAN (AND STANDARD DEVIATIONS) OF 10-FOLD CROSS VALIDATION CONDUCTED TEN TIMES USING THE AMI DATASET. THE TWO HIGHEST MEAN PERFORMANCES IN EACH COLUMN ARE IN BOLD

Feature-level Fusion			
Cues	F1 Laughter	F1 Speech	CR
Face + MFCC	**93.5 (1.0)**	**95.1 (0.7)**	**94.4 (0.8)**
Face + P & E + ZCR	88.3 (1.1)	91.6 (0.7)	90.2 (0.8)
Face + MFCC + P & E + ZCR	**93.9 (0.6)**	**95.3 (0.4)**	**94.7 (0.5)**
Head + MFCC	91.2 (1.7)	93.6 (1.1)	92.6 (1.3)
Head + P & E + ZCR	78.8 (2.3)	85.8 (1.3)	83.0 (1.7)
Head + MFCC + P & E + ZCR	91.7 (2.2)	93.9 (1.5)	93.0 (1.8)
Face + Head + MFCC	90.1 (1.6)	92.9 (1.1)	91.7 (1.3)
Face + Head + P & E + ZCR	85.4 (1.5)	89.5 (1.0)	87.8 (1.2)
Face + Head + MFCC + P & E + ZCR	92.3 (0.8)	94.3 (0.6)	93.4 (0.7)

The results for different combinations of audio and visual cues are shown in Table V. The best two results in each column are shown in bold. As can be seen, this is achieved when combining facial expression features with cepstral or cepstral and prosodic features, resulting in a CR of 94.4% and 94.7%, respectively. We should point out that the improvement of 2.8% and 2.4%, respectively, is attained when using only four visual features, i.e., the projection of the coordinates of the 20 tracked points to PCs 7–10 as explained in Section V [equation (1)] in addition to 52 audio features.

The difference in CR and F1 rates between the best two audiovisual approaches, FACE + MFCC + P & E + ZCR and FACE + MFCC and the best two single-modal approaches, MFCC + P & E + ZCR and MFCC, is statistically significant. But the difference between the two audiovisual approaches is not statistically significant. The confusion matrices for the audio-only, video-only, and audiovisual approaches are shown in the Appendix, Tables VII–IX.

The addition of head pose features is not beneficial in general. This shows that the information conveyed by the head pose, as used in this study, is rather uninformative for laughter-versus-speech discrimination. The addition of prosodic features on the other hand results in an improved performance, but it is not always statistically significant.

The main conclusions drawn from the above experiments can be summarized as follows.

1) Facial expression and cepstral features are the most informative visual and audio cues for discrimination between laughter and speech episodes.
2) Head pose features seem to be uninformative for laughter-versus-speech discrimination in spontaneous data.
3) Prosodic features when combined with other audio and visual cues lead to a slight improvement. However, this improvement is not always statistically significant.
4) The best audiovisual approach to discrimination of laughter from speech episodes is to combine facial expression with cepstral features or to combine with cepstral and prosodic features.

B. Cross Databases Experiments on AMI and SAL Datasets

In all the above experiments, we reported results on the AMI dataset based on a subject-independent cross validation. Although each time the trained system was tested on the data of a subject that has not been used for training, the recording conditions for both the training and the test data were the same. In order to evaluate the generalization performance of the proposed classifiers, we used the SAL dataset, which was recorded under very different conditions as explained in Section III.

We conducted two experiments. In the first one, the entire AMI dataset was used for training a laughter-versus-speech classifier, which was then tested on the entire SAL dataset. In the second experiment, the same systems were trained on the SAL dataset and then tested on the AMI dataset. Only the best audio cues, i.e., cepstral features and combination of cepstral and prosodic features, the best video classifier, i.e., facial-expression-based classifier and the best audiovisual classifiers combining facial expression and cepstral or cepstral and prosodic features, found by means of the experiments described in Section VII, were used. The results are shown in Table VI.

1) Training on the AMI Dataset and Testing on the SAL Dataset: When classifiers are trained on the AMI dataset and tested on the SAL dataset, similar conclusions to those obtained for subject-independent cross validation on the AMI dataset can be drawn as shown in Table VI. More specifically, the two audiovisual approaches (combining facial expression features with cepstral or cepstral and prosodic features) result in a statistically significant improvement over the two best single-modal approaches (based on cepstral features or cepstral features and prosodic features), in terms of the average CR and F1 rates. For example, the addition of facial features to cepstral and prosodic features leads to an absolute increase of 2.3%, 1.8%, and 3.4% for CR, F1 for speech, and F1 for laughter, respectively. Similarly to the approach presented in Section VII-AII, since the training was performed on AMI, the same four visual features were used. It is also worth emphasizing the fact that, although the systems were tested on a very different database, the performance is still very good. The confusion matrices for the audio-only, video-only, and audiovisual approaches are shown in the Appendix, Tables X–XII, respectively.

Fig. 9 shows the ROC curves for audio-based, video-based, and audiovisual classifiers achieving the highest CR. As can be seen, the video-only classifier has the worst ROC curve from the plotted cases with AUC 0.896. The audio-only classifier has the second best ROC curve with AUC 0.947. From Fig. 9, it is obvious that the audiovisual classifier achieves the highest AUC 0.979. This result agrees with the results listed in Table VI.

Fig. 11 shows the classification rates per subject for the audio-only, video-only, and audiovisual approaches. The video-only classifier, based on the facial expression features, is usually the worst performing approach. Its classification performance depends heavily on the subjects, ranging from 70% for subject s11

TABLE VI

F1 RATES AND CR FOR CROSS DATABASE EXPERIMENTS. THE RESULTS PRESENTED ARE THE MEAN (AND STANDARD DEVIATIONS) OF TEN EXPERIMENTS. THE TWO HIGHEST MEAN PERFORMANCES IN EACH COLUMN ARE GIVEN IN BOLD. FOR THE TRAIN AMI → TEST SAL (TRAIN SAL → TEST AMI) EXPERIMENT, PCS COMPUTED FROM AMI (SAL) ARE USED. IN ORDER TO INVESTIGATE HOW THE LIMITED DIVERSITY OF THE SAL DATASET AFFECTS THE GENERALIZATION PERFORMANCE OF CLASSIFIERS TRAINED ON SAL, EXPERIMENTS WITH PCS COMPUTED FROM AMI WERE ALSO PERFORMED

	Feature-level Fusion					
	Train AMI → Test SAL			**Train SAL → Test AMI**		
Cues	F1 Laughter	F1 Speech	CR	F1 Laughter	F1 Speech	CR
Face (PCs AMI)	88.1 (0.8)	93.6 (0.5)	91.7 (0.7)	67.6 (1.8)	83.1 (0.6)	77.8 (0.9)
Face (PCs SAL)	-	-	-	65.2 (1.8)	80.0 (0.6)	74.6 (0.9)
MFCC	92.5 (1.8)	96.1 (0.8)	94.9 (1.1)	70.6 (4.1)	84.2 (1.6)	79.4 (2.4)
MFCC + P & E + ZCR	93.2 (1.8)	96.4 (0.9)	95.3 (1.2)	70.7 (5.6)	84.4 (2.1)	79.7 (3.1)
Face (PCs AMI) + MFCC	**96.4 (0.5)**	**98.1 (0.3)**	**97.5 (0.3)**	82.8 (2.0)	89.4 (0.9)	86.9 (1.3)
Face (PCs SAL) + MFCC	-	-	-	**72.7 (3.4)**	**84.9 (1.3)**	**80.5 (1.9)**
Face (PCs AMI) + MFCC + P & E + ZCR	**96.6 (1.0)**	**98.2 (0.6)**	**97.6 (0.7)**	79.4 (3.0)	87.9 (1.3)	84.7 (1.9)
Face (PCs SAL) + MFCC + P & E + ZCR	-	-	-	**71.2 (4.0)**	**84.6 (1.3)**	**80.0 (2.0)**

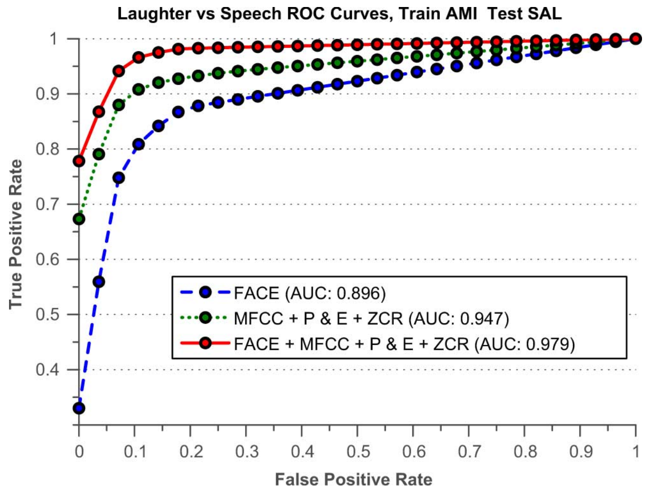


Fig. 9. ROC curves for audio-, video-only, and audiovisual feature-level-fusion approaches to laughter-versus-speech discrimination, when a classifier is trained on the AMI dataset and tested on the SAL dataset.

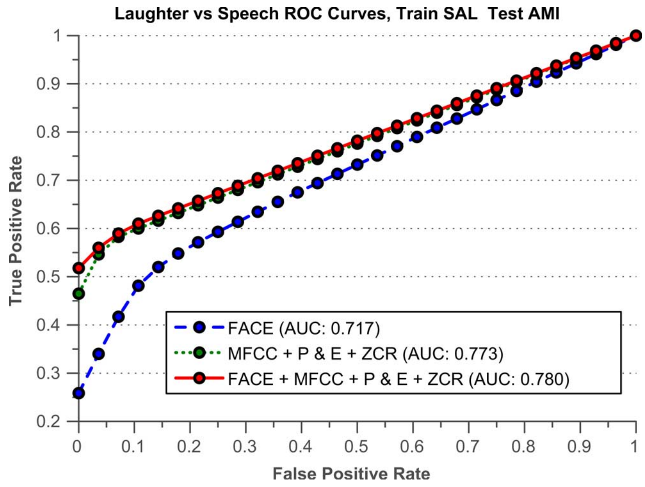


Fig. 10. ROC curves for audio-, video-only, and audiovisual feature-level-fusion approaches to laughter-versus-speech discrimination, when a classifier is trained on the SAL dataset and tested on the AMI dataset.

to 100% for subjects s01, s06, and s10. The audio-only classifier usually performs better than the video-only classifier, and its performance is less subject-dependent, ranging from 88.3% for subject s11 to 100% for subjects s01 and s09. The audiovisual approach is even less subject-dependent than the single-modal approaches, with the classification accuracy ranging from 91.7% for subject s14 to 100% for subjects s01, s02, s06, s09, s10, s12, and s15. The difference between the audio-only and audiovisual classifiers are statistically significant for subjects s02, s06, s07, s11, s12, s14, and s15. By looking at the performances in Fig. 11, this means that the audiovisual approach outperforms audio-only classification for six subjects, is worse in the case of one subject, and results in similar performance in the case of eight subjects.

It is interesting to point out that subjects s02, s11, s12, s14, and s15 are females. Consequently, this means that the addition of the visual information to the audio information is beneficial in the case of 4 out of 7 female subjects, and it is not beneficial in the case of one subject. On the other hand, it is only beneficial in the case of two out of eight male subjects. Also, as mentioned in Section III, subjects s04, s07, and s13 of the SAL dataset have lower video resolution than subjects of the AMI dataset or the remaining subjects of the SAL dataset, but nonetheless a statistically significant improvement is reported for subject s04.

2) Training on the SAL Dataset and Testing on the AMI Dataset: For this experiment, the results are quite different, as shown in Table VI. Generally, the overall performance of the classifiers is much lower than is the case in the previous experiment, indicating that the AMI dataset is a more challenging dataset than the SAL dataset. This is due to the specific recording conditions under which the SAL dataset has been recorded. More specifically, in the SAL dataset, subjects always look at the camera, retaining such a frontal view throughout the recording, whereas in the AMI dataset, subjects are rarely in a frontal view since they participate in a meeting and tend to move their head a lot; see, for example, Figs. 1 and 2. Also the audio conditions are different with a lot of noise present in the

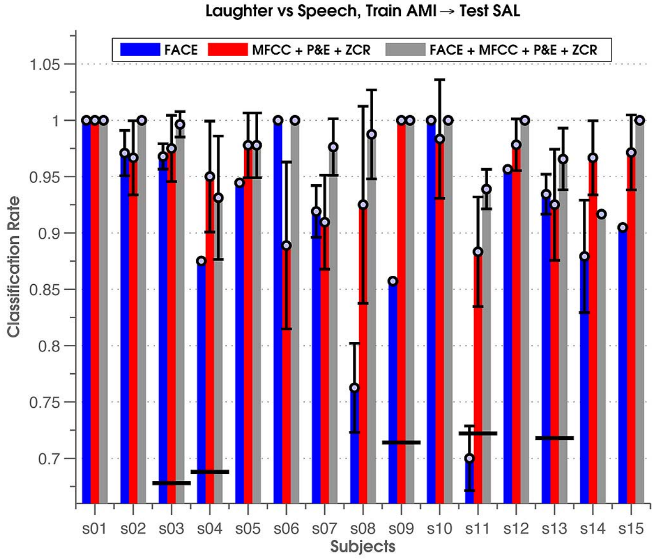


Fig. 11. Classification rate per subject for audio-based, video-based, and audiovisual feature-level-fusion approaches to laughter-versus-speech discrimination. The results presented are the mean and standard deviations of the classification rates achieved for each subject over ten experiments, when a classifier is trained on the AMI dataset and tested on the SAL dataset. The horizontal lines indicate the CR if the majority class is always guessed for each subject. In case there is no horizontal line for a subject, then this means that the majority guessing CR is less than 66%, which is the lower limit of the plot.

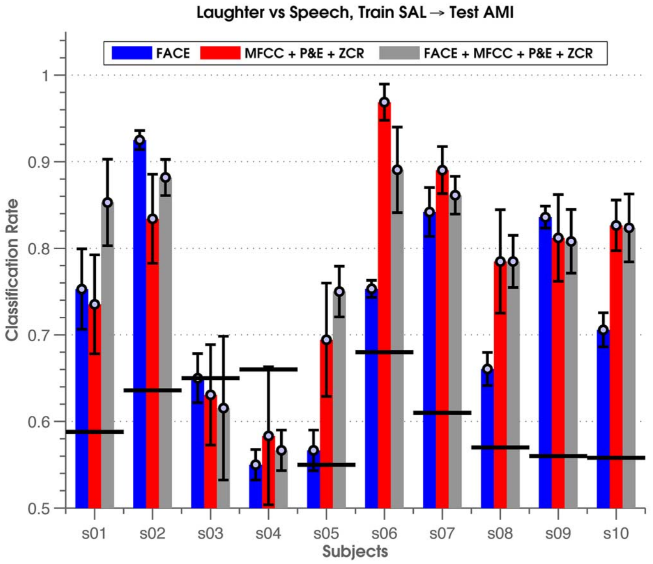


Fig. 12. Classification rate per subject for audio-based, video-based, and audiovisual feature-level-fusion approaches to laughter-versus-speech discrimination. The results presented are the mean and standard deviations of the classification rates achieved for each subject over ten experiments, when a classifier is trained on the SAL dataset and tested on the AMI dataset. The horizontal lines indicate the CR if the majority class is always guessed for each subject.

AMI dataset due to multiple subjects being recorded at the same time. To wit, the AMI recordings are of four subjects, where each subject is recorded by a separate camera, whereas the SAL recordings are of only one subject at the time interacting with an agent. Therefore, a system trained on the SAL dataset fails to generalize well on the AMI dataset.

In this experiment, the audiovisual approaches also achieve higher CR and F1 rates than the corresponding audio-only approaches. However, this difference is not statistically significant. As explained in Section VI, three visual features are used in this case, i.e., the projection of the coordinates of the 20 tracked points to PCs 5, 6, and 7 computed based on the SAL dataset [equation (1)]. In order to confirm the hypothesis, that the limited diversity of the SAL dataset in terms of head movements affects the generalization performance of the systems when tested on the AMI dataset, we used the PCs 7–10 computed based on the AMI dataset, in order to train a classifier on the SAL dataset and then test it on the AMI dataset. The results are shown in the corresponding rows of Table VI. Indeed, as can be seen, a significant improvement in the performance, up to an absolute increase of 10.1% for the F1 rate for laughter, is achieved. In this case, the differences between the audiovisual classifiers and the audio classifiers are all statistically significant. The confusion matrices for the audio-only, video-only, and audiovisual approaches are shown in the Appendix, Tables XIII–XV, respectively.

Fig. 10 shows the ROC curves for audio-based, video-based, and audiovisual classifiers achieving the highest CR. As can be seen, the video-only classifier has the worst ROC curve from the plotted cases with AUC 0.717. The audio-only and audiovisual classifiers have very similar ROC curves with an AUC of 0.773 and 0.780, respectively.

Fig. 12 shows the CR for each subject of the AMI dataset. As can be seen in all cases, the performance varies a lot depending on the subject. The difference between the audio-only and the audiovisual classifiers is statistically significant for subjects s01, s02, s05, s06, and s07. In other words, the audiovisual approach performs better than audio-only classification for three subjects, it performs worse for two subjects, and performs the same in the case of the remaining five subjects. It is interesting to point out that there are only two female subjects in the AMI dataset, s02 and s03. So the addition of the visual information to the audio information is beneficial for one out of two female subjects. On the other hand, it is beneficial only for two out of eight male subjects and it is not beneficial for the other two male subjects.

3) Discussion: From the experiments explained above, we see that a system trained on the AMI dataset can generalize very well on the SAL dataset and the addition of visual information to the audio information results in a small but statistically significant improvement. On the other hand, a system trained on the SAL dataset, which is a less challenging dataset than the AMI dataset, does not generalize well on the AMI dataset and the addition of the visual information is not beneficial. Taking into account the results of the evaluation studies conducted on the AMI dataset, it can be concluded that when test data are similar to (cross validation on AMI) or less diverse than training data (training on the AMI dataset and testing on the SAL dataset) in terms of head movements, then the combination of audio and visual cues leads on average to improved classification rates. It should also be emphasized that although on average over all subjects the performance improves, there are subjects in whose cases the addition of the visual information does not lead to an improved performance. On the other hand, when test data are more diverse than training data (as when training on the SAL dataset and testing on the AMI dataset), then the addition of the

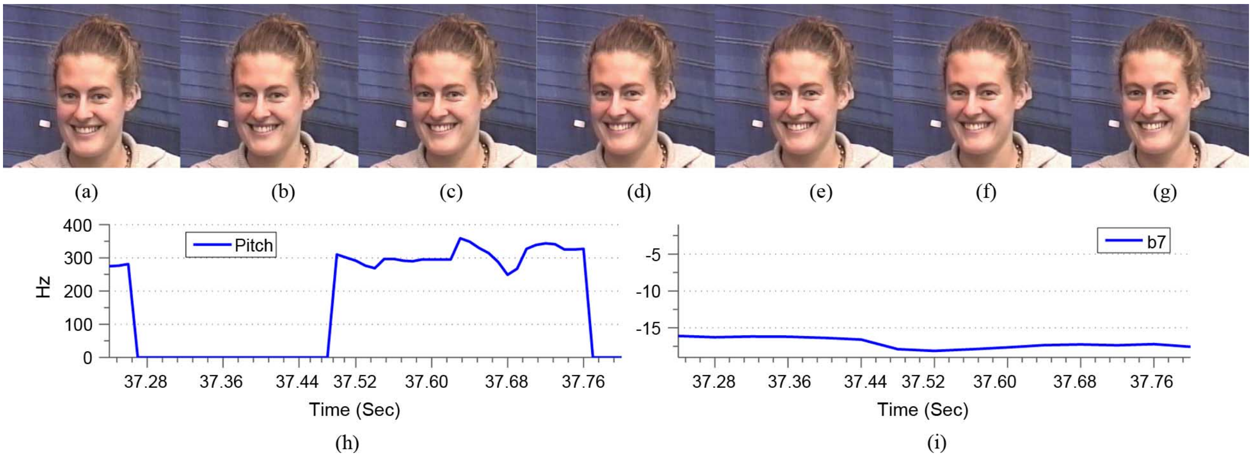


Fig. 13. Example of voiced laughter displayed by subject s02 (Alis), SAL dataset. (a) Frame 1. (b) Frame 3. (c) Frame 5. (d) Frame 8. (e) Frame 10. (f) Frame 12. (g) Frame 15. (h) Pitch. (i) Evolution of shape parameter 7 (using PCs-AMI) over time.

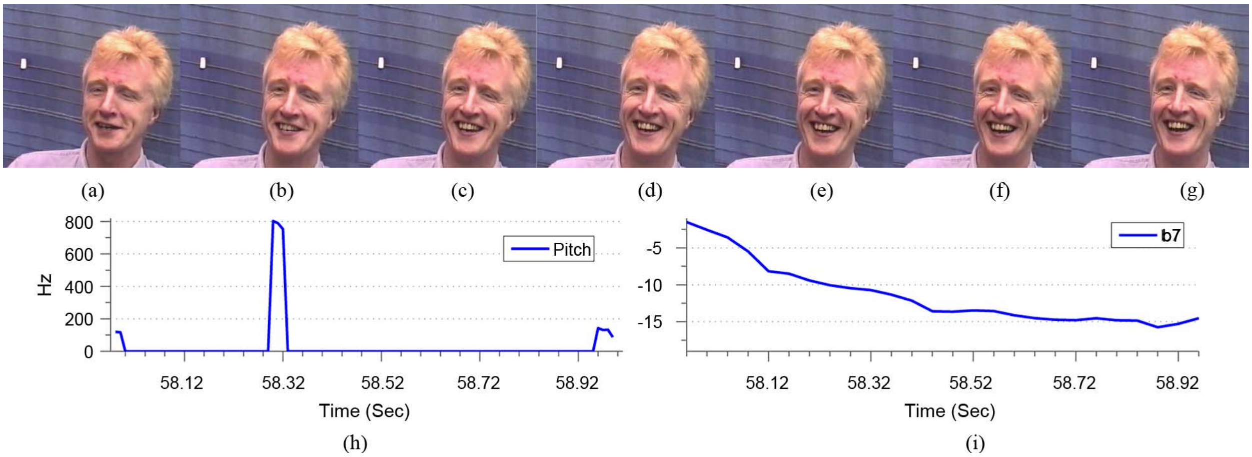


Fig. 14. Example of unvoiced laughter displayed by subject s06 (GHillSect3), SAL dataset. (a) Frame 1. (b) Frame 5. (c) Frame 9. (d) Frame 13. (e) Frame 18. (f) Frame 22. (g) Frame 26. (h) Pitch. (i) Evolution of shape parameter 7 (using PCs-AMI) over time.

visual information is not expected to improve the performance of the system.

Based on the above-described experiments, there is also evidence to suggest that in the case of female subjects, the addition of the visual information leads to a better performance more often than is the case with male subjects. To wit, in the case of five out of nine female subjects, adding the visual information resulted in a better performance and only in the case of one female subject did it result in a degraded performance. This result is significant when compared to that attained for male subjects, where adding the visual information resulted in an improved performance in the case of four out of 16 male subjects and in a degraded performance in the case of two out of 16. This is also consistent with findings in psychology [4] which suggest that females produce voiced laughter more often than unvoiced laughter, which is typically accompanied with more pronounced smile and opened mouth than is the case with the unvoiced laughter produced commonly by males.

Fig. 13 shows a voiced laughter episode which is confused by the audio classifier (FACE + MFCC + P & E +ZCR) for a speech episode. It is accompanied by a smile which is picked up by the visual module, helping the audiovisual classifier to correctly label the episode as laughter. Fig. 14 shows an example of an unvoiced laughter episode which is again confused by the audio classifier for a speech episode. The smile produced by the subject helps the audiovisual classifier to classify this episode correctly. Fig. 15 shows an example where an unvoiced laughter is produced with closed mouth and, as a result, the visual information is not helpful in this case. This example is misclassified by the audio classifier, but since the smile is barely visible, the video module cannot pick it up and, as a result, the audiovisual classifier also confuses this laughter episode for a speech episode. Note the higher values that the 7th shape parameter takes in Fig. 15 compared to those depicted in Figs. 13 and 14. Usually, lower values of the 7th shape parameter correspond to a more open mouth (for example, see how the value of b_7 decreases as the mouth opens in Fig. 14). Generally, visual cues do not carry much discriminative information when laughter episodes are accompanied by subtle facial expressions, and in such cases, the visual information is not beneficial. As mentioned above, subtle facial expressions occur more often with unvoiced laughter episodes than with voiced ones, making the visual information less beneficial in case of unvoiced laughter that is more often displayed by males than by females.

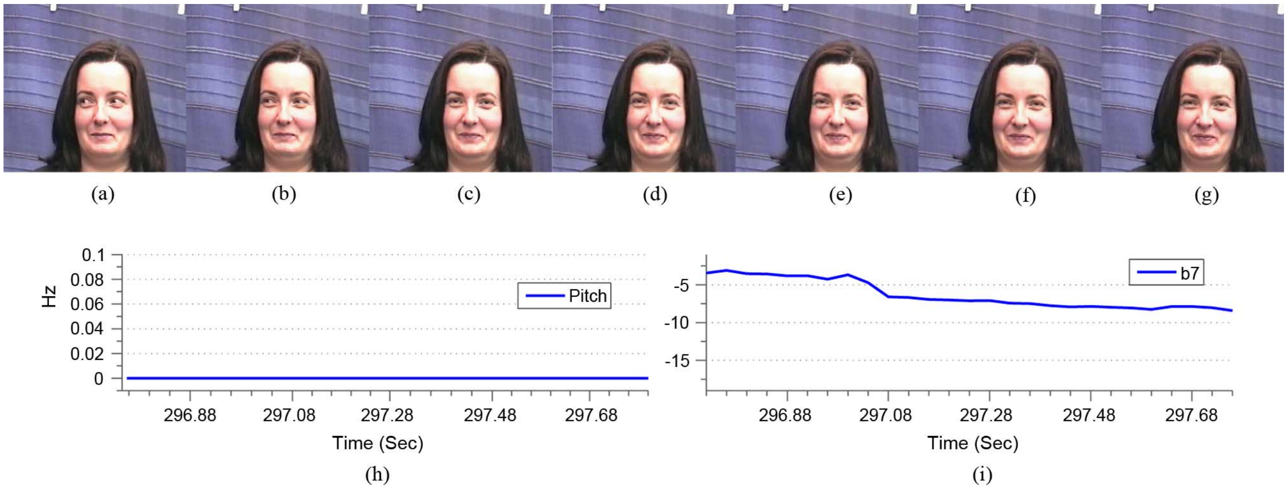


Fig. 15. Example of unvoiced laughter displayed by subject s14 (RuthSect1), SAL dataset. (a) Frame 1. (b) Frame 5. (c) Frame 9. (d) Frame 13. (e) Frame 18. (f) Frame 22. (g) Frame 27. (h) Pitch. (i) Evolution of shape parameter 7 (using PCs-AMI) over time.

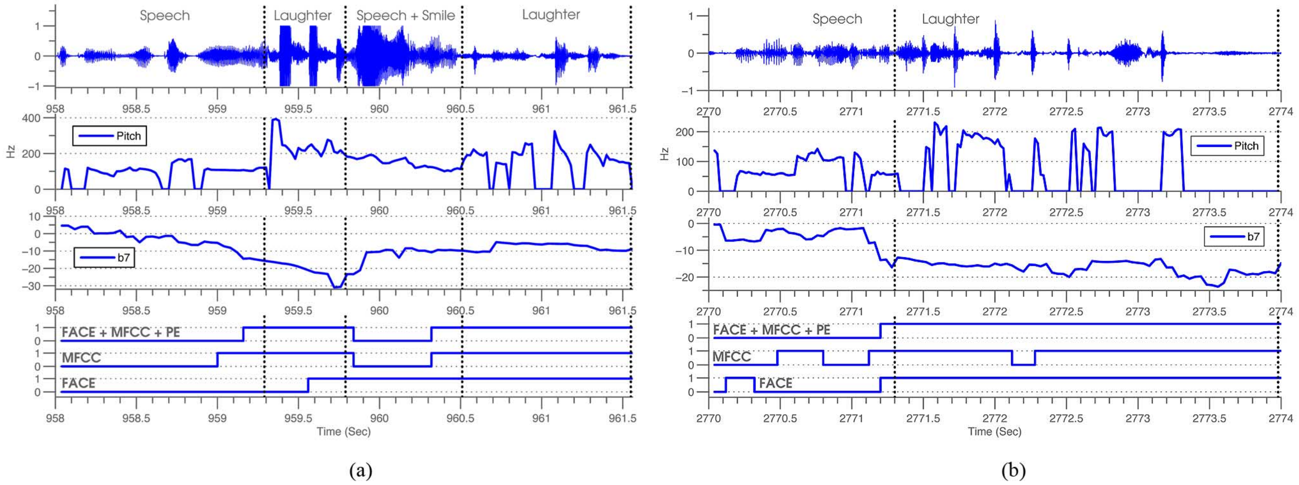


Fig. 16. 1st row: Audio Signal. 2nd row: Pitch. 3rd row: Values of shape parameter 7 (b7). 4th row: Outputs of the audio, video, and audiovisual systems, 1 is laughter, 0 is speech. Ground truth is indicated on the top of the graphs (a) Subject 10, AMI dataset. Meeting ID: IB4010_3. (b) Subject 8, AMI dataset. Meeting ID: IB4010_2.

The main conclusions drawn from the above experiments can be summarized as follows.

1) The combination of visual and audio information leads to improved performance over single modal approaches, but it is not beneficial for all subjects.
2) The addition of visual information to audio information seems to affect more the performance of female than male subjects. For the majority of female subjects, a statistically significant improvement has been attained, while only 25% of male subjects benefited by adding the visual information.
3) The audiovisual approach is beneficial when training conditions are similar or less diverse in terms of head movements than training conditions.
4) When testing conditions are more diverse than training conditions in terms of head movements, the addition of the visual information to the audio information does not seem to help.

C. Segmentation Example

The methods described in this study work with presegmented sequences, i.e., episodes. In a real-life scenario, where presegmentation is not available, the methods could work with a fixed window length. Alternatively they could work on the frame level directly by labeling each frame independently of the others followed by a smoothing step.

Two examples of how the best performing audiovisual approach works by labeling directly each frame is shown in Fig. 16. In both cases, the method was trained on the AMI dataset coming from the nine subjects and was tested on data containing both speech and laughter coming from the subject who was left out. As can be noticed, the pitch is higher in laughter than in speech, as described in the psychology literature as well. The 7th facial shape parameter (b7) takes lower values for laughter, indicating that mouth shapes differ in laughter and in speech. In Fig. 16(a), it can be seen that the video-only approach misclassifies the first half of the first laughter episode. The person smiles while speaking

TABLE VII
CROSS VALIDATION ON AMI—FACIAL EXPRESSIONS FEATURES

	Predicted Laughter	Predicted Speech
Actual Laughter	91.8 (1.7)	32.2 (1.7)
Actual Speech	12.6 (1.1)	141.4 (1.1)

between the two laughter episodes and the video-based detector labels all these frames as a laughter instance. The audiovisual approach performs slightly better than the audio-only approach, classifying just a few frames belonging to speech episodes as laughter. As can be seen from Fig. 16(b), the audio-based and video-based approaches misclassify some frames. However, these false detections occur at different times, and the audiovisual approach is able to successfully combine the audiovisual information to eliminate these misclassifications.

VIII. CONCLUSION

In this paper, we presented an automated audiovisual approach to distinguishing laughter from speech episodes. Very high performance measures were reported for the laughter-versus-speech discrimination problem when training data were similar or more diverse than test data (in terms of head movements). In these cases, adding the visual information to the audio information leads to an improved classification performance on average. This is especially so in the case of female subjects, who produce voiced laughter more often than unvoiced laughter, which is characterized by distinct changes in the facial expression (a wider smile and more open mouth), resulting in significant improvements of performance of the method when the visual information is added to the audio information. On the other hand, when training data are less diverse than test data (in terms of head movements), then adding visual information does not seem to help. We also investigated which cues are informative for the target discrimination. Facial expression and cepstral features play a very important role in discriminating laughter from speech, and prosodic features may help as well. Future research includes the investigation of the performance of the audiovisual approach in the presence of acoustic noise when the addition of the visual information is expected to be particularly beneficial (as is the case in audiovisual speech recognition). Finally, it is also interesting to investigate the use of more sophisticated classification and fusion tools that can take into account the asynchronous nature of the audio and visual streams as well as the contextual information, like asynchronous HMMs and long short-term memory networks [23], having the potential to outperform the standard multimodal data fusion approaches.

APPENDIX
CONFUSION MATRICES—MEAN AND (ST. DEV.) OF NUMBER OF INSTANCES OVER TEN EXPERIMENTS

Tables VII–IX show the cross validation on AMI—facial expressions features, cepstral + prosodic features, and facial expressions + cepstral + prosodic features, respectively. Tables X–XII show the train AMI, test SAL—facial expressions features, cepstral + prosodic features, and facial expressions + cepstral + prosodic features, respectively. Tables XIII–XV show the train SAL, test AMI—facial expressions features, cepstral + prosodic features, and facial expressions + cepstral + prosodic features, respectively.

TABLE VIII
CROSS VALIDATION ON AMI—CEPSTRAL + PROSODIC FEATURES

	Predicted Laughter	Predicted Speech
Actual Laughter	106.5 (3.5)	17.5 (3.5)
Actual Speech	3.8 (1.5)	150.2 (1.5)

TABLE IX
CROSS VALIDATION ON AMI—FACIAL EXPRESSIONS + CEPSTRAL + PROSODIC FEATURES

	Predicted Laughter	Predicted Speech
Actual Laughter	113.1 (1.0)	10.9 (1.0)
Actual Speech	3.9 (0.6)	150.1 (1.5)

TABLE X
TRAIN AMI, TEST SAL—FACIAL EXPRESSIONS FEATURES

	Predicted Laughter	Predicted Speech
Actual Laughter	84.4 (0.7)	10 (0.7)
Actual Speech	12.6 (1.7)	164.4 (1.7)

TABLE XI
TRAIN AMI, TEST SAL—CEPSTRAL + PROSODIC FEATURES

	Predicted Laughter	Predicted Speech
Actual Laughter	86.6 (2.5)	7.4 (2.5)
Actual Speech	5.3 (2.8)	171.7 (2.8)

TABLE XII
TRAIN AMI, TEST SAL—FACIAL EXPRESSIONS + CEPSTRAL + PROSODIC FEATURES

	Predicted Laughter	Predicted Speech
Actual Laughter	91.5 (0.7)	2.5 (0.7)
Actual Speech	3.9 (2.0)	173.1 (2.0)

TABLE XIII
TRAIN SAL, TEST AMI—FACIAL EXPRESSIONS FEATURES

	Predicted Laughter	Predicted Speech
Actual Laughter	66.1 (2.9)	57.9 (2.9)
Actual Speech	12.7 (1.3)	141.3 (1.3)

TABLE XIV
Train SAL, Test AMI—Cepstral + Prosodic Features

	Predicted Laughter	Predicted Speech
Actual Laughter	69 (8.6)	55 (8.6)
Actual Speech	1.5 (0.5)	152.5 (0.5)

TABLE XV
Train SAL, Test AMI—Facial Expressions + Cepstral + Prosodic Features

	Predicted Laughter	Predicted Speech
Actual Laughter	69.2 (6.2)	54.8 (6.2)
Actual Speech	0.9 (1.3)	153.1 (1.3)

References

[1] [Online]. Available: http://www.doc.ic.ac.uk/~maja/AMI-SAL-annotations.xls.

[2] Nist (2004), Rich Transcription 2004 Spring Meeting Recognition Evaluation, Documentation. [Online]. Available: http://www.nist.gov/speech/tests/rt/rt2004/spring/.

[3] J. Bachorowski and M. Owren, “Not all laughs are alike: Voiced but not unvoiced laughter readily elicits positive affect,” *Psychol. Sci.*, vol. 12, no. 3, pp. 252–257, 2001.

[4] J. A. Bachorowski, M. J. Smoski, and M. J. Owren, “The acoustic features of human laughter,” *J. Acoust. Soc. Amer.*, vol. 110, no. 1, pp. 1581–1597, 2001.

[5] P. Boersma, “Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound,” *Proc. Inst. Phonet. Sci.*, vol. 17, pp. 97–110, 1993.

[6] K. Bousmalis, M. Mehu, and M. Pantic, “Spotting agreement and disagreement: A survey of nonverbal audiovisual cues and tools,” in *Proc. IEEE Int. Conf. Affective Computing and Intelligent Interfaces*, 2009, vol. 2.

[7] C. Busso, S. Lee, and S. Narayanan, “Analysis of emotionally salient aspects of fundamental frequency for emotion detection,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 4, pp. 582–596, 2009.

[8] R. Cai, L. Lu, H.-J. Zhang, and L.-H. Cai, “Highlight sound effects detection in audio stream,” in *Proc. Int. Conf. Multimedia and Expo*, 2003, vol. 3, pp. 37–40.

[9] N. Campbell, “Recording techniques for capturing natural everyday speech,” in *Proc. Language Resources and Evaluation Conf.*, 2002.

[10] N. Campbell, “Whom we laugh with affects how we laugh,” in *Proc. Workshop Phonetics of Laughter*, 2007, pp. 61–65.

[11] N. Campbell, H. Kashioka, and R. Ohara, “No laughing matter,” in *Proc. Eur. Conf. Speech Communication and Technology*, 2005, pp. 465–468.

[12] J. F. Cohn and K. L. Schmidt, “The timing of facial motion in posed and spontaneous smiles,” *Int. J. Wavelets Multires. Inf. Process.*, vol. 2, pp. 121–132, 2005.

[13] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, “Active shape models—their training and application,” *Comput. Vis. Image Understand.*, vol. 61, no. 1, p. 38, 1995.

[14] E. Douglas-Cowie, R. Cowie, C. Cox, N. Amir, and D. Heylen, “The sensitive artificial listener: An induction technique for generating emotionally coloured conversation,” in *Proc. Workshopn Corpora for Research on Emotion and Affect*, pp. 1–4.

[15] S. Dupont and J. Luettin, “Audio-visual speech modeling for continuous speech recognition,” *IEEE Trans. Multimedia*, vol. 2, no. 3, pp. 141–151, 2000.

[16] I. Eibl-Eibesfeldt, “The expressive behavior of the deaf-and-blind born,” in *Social Communication and Movement: Studies of Interaction and Expression in Man and Chimpanzee*. New York: Academic, 1973, pp. 163–193.

[17] D. P. W. Ellis, PLP and RASTA (and MFCC and Inversion) in Matlab, 2005. [Online]. Available: http://www.ee.columbia.edu/dpwe/resources/matlab/rastamat.

[18] G. Furnas, T. Landauer, L. Gomez, and S. Dumais, “The vocabulary problem in human-system communication,” *Commun. ACM*, vol. 30, no. 11, pp. 964–972, 1987.

[19] D. Gonzalez-Jimenez and J. L. Alba-Castro, “Toward pose-invariant 2-D face recognition through point distribution models and facial symmetry,” *IEEE Trans. Inf. Forensics Security*, vol. 2, no. 3, pp. 413–429, 2007.

[20] M. Graciarena, E. Shriberg, A. Stolcke, F. Enos, J. Hirschberg, and S. Kajarekar, “Combining prosodic lexical and cepstral systems for deceptive speech detection,” in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2006, vol. 1, pp. 1033–1036.

[21] K. Grammer and I. Eibl-Eibesfeldt, “The ritualisation of laughter,” in *Die Natürlichkeit der Sprache und der Kultur*. Bochum, Germany: Brockmeyer, 1990, pp. 192–214.

[22] G. Guo and S. Li, “Content-based audio classification and retrieval by support vector machines,” *IEEE Trans. Neural Netw.*, vol. 14, no. 1, pp. 209–215, 2003.

[23] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computat.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[24] W. Hudenko, W. Stone, and J. Bachorowski, “Laughter differs in children with autism: An acoustic analysis of laughs produced by children with and without the disorder,” *J. Autism Develop. Disorders*, vol. 39, no. 10, pp. 1392–1400, 2009.

[25] A. Ito, W. Xinyue, M. Suzuki, and S. Makino, “Smile and laughter recognition using speech processing and face recognition from conversation video,” in *Proc. Int. Conf. Cyberworlds*, 2005, pp. 8–15.

[26] A. Janin *et al.*, “The ICSI meeting corpus,” in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2003, vol. 1, pp. 364–367.

[27] L. Kennedy and D. Ellis, “Laughter detection in meetings,” in *Proc. NIST Meeting Recognition Workshop*, 2004.

[28] S. Kipper and D. Todt, “The role of rhythm and pitch in the evaluation of human laughter,” *J. Nonverb. Behav.*, vol. 27, no. 4, pp. 255–272, 2003.

[29] M. Knox, N. Morgan, and N. Mirghafori, “Getting the last laugh: Automatic laughter segmentation in meetings,” in *Proc. INTERSPEECH*, 2008, pp. 797–800.

[30] K. Laskowski and S. Burger, “Analysis of the occurrence of laughter in meetings,” in *Proc. INTERSPEECH*, 2007, pp. 1258–1261.

[31] K. Laskowski and T. Schultz, “Detection of laughter-in-Interaction in multichannel close-talk microphone recordings of meetings,” *Lecture Notes in Computer Science*, vol. 5237, pp. 149–160, 2008.

[32] Y. Liu, N. Chawla, M. Harper, E. Shriberg, and A. Stolcke, “A study in machine learning from imbalanced data for sentence boundary detection in speech,” *Comput. Speech Lang.*, vol. 20, no. 4, pp. 468–494, 2006.

[33] A. Lockerd and F. Mueller, “LAFCAM: Leveraging affective feedback camcorder,” in *Proc. CHI, Human Factors in Computing Systems*, 2002, pp. 574–575.

[34] I. Marks, K. Cavanagh, and L. Gega*, Hands-on Help: Computer-Aided Psychotherapy*. East Sussex, U.K.: Psychology Press Hove, 2007.

[35] 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 *Proc. Int. Conf. Methods and Techniques in Behavioral Research*, 2005, pp. 137–140.

[36] N. Oostdijk, “The spoken Dutch corpus: Overview and first evaluation,” in *Proc. Int. Conf. Language Resources and Evaluation*, 2000, pp. 887–894.

[37] M. Pantic, A. Pentland, A. Nijholt, and T. Huang, “Human computing and machine understanding of human behavior: A survey,” *Lecture Notes in Computer Science*, vol. 4451, pp. 47–71, 2007.

[38] M. Pantic and A. Vinciarelli, “Implicit human-centered tagging,” *IEEE Signal Process. Mag.*, vol. 26, no. 6, pp. 173–180, 2009.

[39] I. Patras and M. Pantic, “Particle filtering with factorized likelihoods for tracking facial features,” in *Proc. Int. Conf. Automatic Face and Gesture Recognition*, 2004, pp. 97–104.

[40] S. Petridis, H. Gunes, S. Kaltwang, and M. Pantic, “Static vs. dynamic modeling of human nonverbal behavior from multiple cues and modalities,” in *Proc. ICMI*, 2009, pp. 23–30.

[41] S. Petridis and M. Pantic, “Audiovisual discrimination between laughter and speech,” in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, 2008, pp. 5117–5120.

[42] S. Petridis and M. Pantic, “Audiovisual laughter detection based on temporal features,” in *Proc. ACM Int. Conf. Multimodal Interfaces*, 2008, pp. 37–44.

[43] S. Petridis and M. Pantic, “Fusion of audio and visual cues for laughter detection,” in *Proc. ACM Int. Conf. Image and Video Retrieval*, 2008, pp. 329–337.

[44] S. Petridis and M. Pantic, "Is this joke really funny? Judging the mirth by audiovisual laughter analysis," in *Proc. IEEE Int. Conf. Multimedia & Expo*, 2009, pp. 1444–1447.

[45] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior, "Recent advances in the automatic recognition of audiovisual speech," *Proc. IEEE*, vol. 91, no. 9, pp. 1306–1326, 2003.

[46] R. Provine, "Laughter punctuates speech: Linguistic, social and gender contexts of laughter," *Ethology*, vol. 95, no. 4, pp. 291–298, 1993.

[47] R. Provine*, Laughter: A Scientific Investigation*. New York: Viking, 2000.

[48] R. Provine and Y. Yong, "Laughter: A stereotyped human vocalization," *Ethology*, vol. 89, no. 2, pp. 115–124, 1991.

[49] L. Rabiner and R. Schafer*, Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1978.

[50] B. Reuderink, M. Poel, K. Truong, R. Poppe, and M. Pantic, "Decision-level fusion for audio-visual laughter detection," *Lecture Notes in Computer Science*, vol. 5237, pp. 137–148, 2008.

[51] M. Riedmiller and H. Braun, "A direct adaptive method for faster backpropagation learning: The RPROP algorithm," in *Proc. IEEE Int. Conf. Neural Networks*, 1993, vol. 1, pp. 586–591.

[52] H. Rothgänger, G. Hauser, A. Cappellini, and A. Guidotti, "Analysis of laughter and speech sounds in Italian and German students," *Naturwissenschaften*, vol. 85, no. 8, pp. 394–402, 1998.

[53] W. Ruch and P. Ekman, "The expressive pattern of laughter," in *Emotions, Qualia and Consciousness*. Singapore: World Scientific, 2001, pp. 426–443.

[54] K. Scherer, "Affect bursts," in *Emotions: Essays on Emotion Theory*, S. van Goozen, N. van de Poll, and J. Sergeant, Eds. Hillsdale, NJ: Erlbaum, 1994, pp. 161–193.

[55] M. Schroeder, D. Heylen, and I. Poggi, "Perception of non-verbal emotional listener feedback," in *Proc. Speech Prosody*, Dresden, Germany, 2006, pp. 1–4.

[56] B. Schueller, 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 *Proc. INTERSPEECH*, 2007, pp. 2253–2256.

[57] B. Schueller, F. Eyben, and G. Rigoll, "Static and dynamic modelling for the recognition of non-verbal vocalisations in conversational speech," *Lecture Notes in Computer Science*, vol. 5078, pp. 99–110, 2008.

[58] B. Schueller, R. Mueller, B. Hoernler, A. Hoethker, H. Konosu, and G. Rigoll, "Audiovisual recognition of spontaneous interest within conversations," in *Proc. ACM Int. Conf. Multimodal Interfaces*, 2007, pp. 30–37.

[59] J. Trouvain, "Segmenting phonetic units in laughter," in *Proc. Int. Conf. Phonetic Sciences*, 2003, pp. 2793–2796.

[60] K. P. Truong and D. A. van Leeuwen, "Automatic discrimination between laughter and speech," *Speech Commun.*, vol. 49, no. 2, pp. 144–158, 2007.

[61] K. P. Truong and D. A. van Leeuwen, "Evaluating laughter segmentation in meetings with acoustic and acoustic-phonetic features," in *Proc. Workshop Phonetics of Laughter*, 2007.

[62] M. F. Valstar, H. Gunes, and M. Pantic, "How to distinguish posed from spontaneous smiles using geometric features," in *Proc. ACM Int. Conf. Multimodal Interfaces*, 2007, pp. 38–45.

[63] J. Vettin and D. Todt, "Laughter in conversation: Features of occurrence and acoustic structure," *J. Nonverb. Behav.*, vol. 28, no. 2, pp. 93–115, 2004.

[64] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image Vis. Comput.*, vol. 27, no. 12, pp. 1743–1759, 2009.

[65] B. Vlasenko, B. Schueller, A. Wendemuth, and G. Rigoll, "Frame vs. turn-level: Emotion recognition from speech considering static and dynamic processing," in *Proc. ACII*, 2007, pp. 139–147.

[66] H. Yehia, T. Kuratate, and E. Vatikiotis-Bateson, "Linking facial animation, head motion and speech acoustics," *J. Phonet.*, vol. 30, no. 3, pp. 555–568, 2002.

[67] B. Yin, E. Ambikairajah, and F. Chen, "Combining cepstral and prosodic features in language identification," in *Proc. Int. Conf. Pattern Recognition*, 2006, vol. 4, pp. 254–257.

[68] Z. Zeng, M. Pantic, G. Roisman, and T. Huang, "A survey of affect recognition methods: Audio, visual and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, 2009.

Stavros Petridis (S'07–M'11) received the B.Sc. degree in electrical and computer engineering from the Aristotle University of Thessaloniki, Thessaloniki, Greece, in 2004 and the M.Sc. degree in advanced computing from Imperial College London, London, U.K., in 2005. Since 2007, he has been pursuing the Ph.D. degree in the Department of Computing, Imperial College London, working on audiovisual recognition of nonlinguistic vocalizations and particularly laughter.

He was a research intern in the Image Processing Group at University College London in 2003 and in the Field Robotics Centre, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, in 2006. His research interests lie in the areas of pattern recognition and machine learning and their application to multimodal recognition of human nonverbal behavior and nonlinguistic vocalizations. He is currently working on audiovisual fusion approaches inspired by recent findings in neuroscience.

Maja Pantic (SM'06) is a Professor in affective and behavioral computing in the Department of Computing at Imperial College London, London, U.K. (http://ibug.doc.ic.ac.uk/), and at the Department of Computer Science, University of Twente, Enschede, The Netherlands.

She is one of the world's leading experts in the research on machine understanding of human behavior, including vision-based detection, tracking, and analysis of human behavioral cues like facial expressions and body gestures and multimodal human affect/mental-state understanding. She has published more than 100 technical papers in these areas of research. In 2008, for her research on Machine Analysis of Human Naturalistic Behavior (MAHNOB), she received European Research Council Starting Grant as one of 2% best young scientists in any research field in Europe. She is also a partner in several FP7 European projects, including the currently ongoing FP7 SSPNet NoE, for which she is the scientific coordinator.

Prof. Pantic currently serves as the Editor-in-Chief of the *Image and Vision Computing Journal* and as an Associate Editor for the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS, PART B: CYBERNETICS (TSMC-B). She has also served as the General Chair for several conferences and symposia, including the IEEE FG 2008 and the IEEE ACII 2009.