Image and Vision Computing xxx (2012) xxx–xxx

Contents lists available at SciVerse ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

The MAHNOB Laughter database ☆

Stavros Petridis [a,*], Brais Martinez [a], Maja Pantic [a,b]

[a] Computing Department, Imperial College London, UK
[b] EEMCS, University of Twente, The Netherlands

ARTICLE INFO

Article history:
Received 25 October 2011
Received in revised form 2 August 2012
Accepted 23 August 2012
Available online xxxx

Keywords:
Laughter
Audiovisual
Thermal
Database
Audiovisual automatic laughter–speech discrimination

ABSTRACT

Laughter is clearly an audiovisual event, consisting of the laughter vocalization and of facial activity, mainly around the mouth and sometimes in the upper face. A major obstacle in studying the audiovisual aspects of laughter is the lack of suitable data. For this reason, the majority of past research on laughter classification/detection has focused on audio-only approaches. A few audiovisual studies exist which use audiovisual data from existing corpora of recorded meetings. The main problem with such data is that they usually contain large head movements which make audiovisual analysis very difficult. In this work, we present a new publicly available audiovisual database, the MAHNOB Laughter database, suitable for studying laughter. It contains 22 subjects who were recorded while watching stimulus material, using two microphones, a video camera and a thermal camera. The primary goal was to elicit laughter, but in addition, posed smiles, posed laughter, and speech were recorded as well. In total, 180 sessions are available with a total duration of 3 h and 49 min. There are 563 laughter episodes, 849 speech utterances, 51 posed laughs, 67 speech–laughs episodes and 167 other vocalizations annotated in the database. We also report baseline experiments for audio, visual and audiovisual approaches for laughter-vs-speech discrimination as well as further experiments on discrimination between voiced laughter, unvoiced laughter and speech. These results suggest that the combination of audio and visual information is beneficial in the presence of acoustic noise and helps discriminating between voiced laughter episodes and speech utterances. Finally, we report preliminary experiments on laughter-vs-speech discrimination based on thermal images.

1. Introduction

Laughter is an important social signal which plays a key role in social interactions and relationships. It is estimated to be about 7 million years old [1], and like other social signals, it is widely believed to have evolved before the development of speech [2,3]. It has been suggested that laughter evolved in order to facilitate the formation and maintenance of positive and cooperative relationships in social groups [4]. It usually expresses a state of positive emotion and induces positive reactions in the receiver and, contrary to speech, laughing at the same time with others is considered as positive feedback.

Laughter is a universal non-verbal vocalization, since there is evidence of a strong genetic basis in its development [5]. For example, babies have the ability to laugh before they can speak [2], children born both deaf and blind still have the ability to laugh [6], and the acoustic features of laughter produced by congenitally deaf and normally hearing students are similar [7]. Considering its prevalence and universality as a social signal, it is surprising that our knowledge about laughter is still incomplete and little empirical information is available [8].

From a technical perspective, automatic recognition of laughter can be useful for affect sensing [9] and affect-sensitive human–computer interfaces [10]. Laughter recognition can also be used as a useful cue for the detection of the users' conversational signals such as agreement [11] and can benefit the automatic analysis of multi-party meetings [12]. It is also useful in automatic speech recognition, to correctly identify laughter episodes as non-speech segments, as those usually degrade the performance of automatic recognizers. In addition, a laughter detector can be used for multimedia tagging and retrieval [13].

Previous works on laughter have focused on the use of audio information only, i.e., visual information carried by facial expressions of the observed person has been ignored. These works include studies analyzing the role of laughter in interpersonal communication [14,15] and study the acoustic properties of laughter [16–18], such as duration, formants and fundamental frequency. They further include works which aim to discriminate laughter from speech [19] (or from other vocalizations [20]), and those that aim to recognize laughter in a continuous audio stream [21–23]. More details about existing works in laughter recognition can be found in [24]. It should also be noted that a few works on smile recognition exist in the literature [25,26]. Although it has long been debated whether smile and

☆ This paper has been recommended for acceptance by Hatice Gunes and Bjoern Schuller.

* Corresponding author. Tel.: +44 207 594 8195.
E-mail addresses: stavros.petridis04@imperial.ac.uk (S. Petridis), b.martinez@imperial.ac.uk (B. Martinez), m.pantic@imperial.ac.uk (M. Pantic).

0262-8856/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.imavis.2012.08.014

laughter are the two extremes in the same continuum as suggested in [4], and it is very likely that they exhibit different characteristics in their visual appearance, the same tools could be used to extract information from the visual stream.

Laughter is clearly an audiovisual event consisting of a laughter vocalization and a facial expression. However, the lack of audiovisual data has prevented audiovisual research for a long time. Only recently, a few works have been published which combine visual information with audio. They reported improved performance over audio-only approaches. In these works the audio features were augmented with visual features, the most common being shape features encoding geometric relations between facial fiducial points [24,27–29]. Different types of visual cues have also been considered, like face and body actions [30]. More details about audiovisual approaches to laughter vs. speech discrimination can be found in [24].

A major challenge in studying laughter is the collection of suitable data. Since laughter usually occurs in social situations, it is not easy to obtain clear recordings of spontaneous and natural expressions of an individual. The existing work tackled this problem by following two different approaches. The first approach is to use existing data from multi-party meeting sessions during which the subjects sometimes laugh. In most occasions only microphones are used, restricting the laughter analysis to information coming from the audio channel only. Recently a few datasets of audiovisual meeting recordings have become available [31,32] as well. The main problem with these audiovisual data is that the audio signal is usually noisy due to the presence of several people. In addition, people tend to move their heads a lot and near-frontal views of the face are not always available, making the visual data difficult to process even with the state-of-the-art methods. The meeting recordings used in audiovisual laughter studies are described in Section 2.

As an alternative, some works use recordings of laughter elicited by showing funny videos to subjects. These data usually contain clean audio recordings and near-frontal view of the recorded subjects. Unfortunately, most of these data is not publicly available. To the best of our knowledge, only two publicly available audiovisual databases of elicited laughter provide suitable data for audiovisual laughter analysis, the AudioVisual Laughter Cycle database (AVLC) [33] and the MMI (part V) database [34]. The AVLC database contains recordings of 24 subjects showing 1066 laughter episodes while the MMI—part V contains recordings of 9 subjects showing 164 laughter episodes. In both databases, subjects were watching funny video clips and their reactions were recorded by a camera and a microphone. Although they are both useful databases, a number of drawbacks should be taken into account. These datasets do not contain speech utterances from the same subjects who produce laughter episodes, and analysis of differences between laughter and speech cannot be therefore conducted based on these data nor the training of a laughter-vs-speech discrimination system is possible. In addition, subjects in the AVLC database have markers on their faces, which hinders the automatic extraction of visual features in the majority of the state-of-the-art approaches. Both databases are described in more detail in Section 2.

It is clear that, unlike audiovisual speech recognition, where benchmark datasets exist, the lack of suitable audiovisual laughter data is a major obstacle for further research on laughter and its discrimination from speech.

In this paper, we present a new audiovisual laughter database, the MAHNOB Laughter database, which addresses the drawbacks of the existing relevant databases and aims to provide a benchmark for laughter classification. Laughter was elicited by showing short funny video clips to subjects and their reactions were recorded using two microphones (camera and lapel microphone), a video camera, and a thermal camera.

We included thermal image recordings as these provide physiological information related to the process of laughter. The inspiration comes from the work presented in [35] where the physiological reactions during the process of deceit were studied using thermal information. Since we wanted to include episodes of both spontaneous and acted (deliberately displayed) laughter, in order to enable studies in discriminating the two, we thought that inclusion of thermal imagery will enable a further insight — whether the physiological reactions are the same independently of whether we laugh spontaneously or deliberately. Furthermore, by including thermal recordings in the database a significant set of thermal recordings of naturalistic expressive facial videos becomes available representing one of the very few such resources.[1]

In total 22 subjects were recorded in 180 sessions. In addition to laughter, subjects were asked to produce posed smiles, posed laughs[2], and to speak for approximately 90 s in English and 90 s in their mother language. Multimodal recordings are synchronized and annotated in terms of laughter (spontaneous and posed) and speech. In total, there are 563 laughter episodes, 849 speech utterances, 51 posed laughter episodes, 67 speech–laughs episodes and 167 other vocalizations. The 2 audio streams, the video stream, the thermal camera stream and the annotation files can be viewed and downloaded (after an end user license agreement is signed) from http://mahnob-db.eu/laughter/. The database is described in detail in Section 3.

The paper presents further baseline methods for audio-only, video-only, and audiovisual laughter-vs-speech discrimination. The audiovisual approach uses standard audio and visual features combined using feature-level fusion. This method represents a version of the approach proposed in [24]. Results on the MAHNOB Laughter database show that the addition of visual information is beneficial when the audio signal is noisy (camera microphone), whereas in clean audio conditions (lapel microphone) the improvement is small.

We also present results for discrimination between voiced laughter, unvoiced laughter and speech. The distinction between voiced and unvoiced laughter is common and the differences between the two types have been studied by psychologists [37,14]. 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 the nose or mouth and the vocal folds are not involved in the production of laughter. This is an important distinction since it has been shown that the two kinds of laughter have different functions in social interactions. Grammer and Eibl-Eibesfeldt [37] found that male interest was partly predicted by the number of voiced laughs produced by female partners, while the opposite does not hold [14]. The latter study also demonstrated that voiced laughter usually elicited more positive evaluations than unvoiced laughter. It is also believed that voiced laughter is directly related to the experience of positive affect, whereas un-voiced laughter is used to negotiate social interactions [38]. Finally, in a previous preliminary study [39] we have shown that voiced laughter is more correlated with amusing multimedia content, whereas unvoiced laughter is more correlated with less amusing content.

Regarding the thermal images, we provide a baseline result in the task of discriminating between laughter and speech from still images. To this end, we do perform a fully automatic registration of the face following [40], define a region of interest heuristically to cover the whole mouth region, and compute a feature description for these image regions as in [41]. We also provide publicly available versions of the code (in the form of Matlab-compiled executables) used for the face registration through our website.

[1] The NVIE database [36] provides recordings of visible and infrared images of expressive faces, including both posed and natural expressions. No recordings of laughter nor audio have been included in this database. However, to the best of our knowledge the NVIE database is the only source of thermal facial videos made publicly available.

[2] Posed laugh is the laugh produced by subjects when they are asked to laugh on demand without the presence of humorous stimuli.

Please cite this article as: S. Petridis, et al., The MAHNOB Laughter database, Image Vis. Comput. (2012), http://dx.doi.org/10.1016/j.imavis.2012.08.014

2. Related databases

An overview of all existing audiovisual databases containing laughter is shown in Table 1. In what follows we describe briefly each of the databases listed on it. A more detailed summary of the databases can be found in [42].

SEMAINE [50]: In the SEMAINE database the users interact with 4 agents, which have different personalities. The agents are played by a human operator. The aim is to evoke emotionally colored reactions from the users, whose reactions are recorded by a camera and a microphone. All Elicited Laughter Meeting users interact with each agent for approximately 5 min. In total 150 users were recorded, most of them being native English speakers, and the total duration of the recordings comprising the database is 80 h.
Laughter annotations for some sequences have become available recently and therefore the total number of laughter episodes is not known. The entire database is available online from http://semaine-db.eu/.
Augmented Multi-party Interaction (AMI) corpus [31]: The AMI meeting corpus is a multi-modal database consisting of 100 h of meeting recordings where people show a huge variety of spontaneous expressions. In each meeting there are four participants who can move freely in the meeting room and interact with each other. All meetings are held in English, although most of the participants are non-native English speakers.
Laughter annotations are also provided but they are approximate: only one time stamp is used to indicate that laughter occurs, while its start and end times are not given. Furthermore, several smiles with no audible laughter sound are labeled as laughter. The database is available to download from http://corpus.amiproject.org/.
AudioVisual Interest Corpus (AVIC) [51]: The AVIC corpus is an audiovisual dataset containing scenario-based dyadic interactions. Each subject is recorded while interacting with an experimenter who plays the role of a product presenter and leads the subject through a commercial presentation. The subject's role is to listen to the presentation and interact with the experimenter depending on his/her interest on the product. The language used is English with most subjects being non-native speakers. In total 21 subjects were recorded (11 males, 10 females) and the total duration of the database recordings is 10 h and 22 min.
Annotations for some nonlinguistic vocalizations are available, including laughter, consent and hesitation. In total, 324 laughter episodes have been annotated but 57 are very short (less than 120 ms). The database is available upon request.
Free Talk (FT) database [32]: This is an audiovisual database of meeting recordings where people show a variety of spontaneous expressions similar to the AMI corpus. It consists of three multi-party conversations where the participants discuss without any constraint on the topic and they are allowed to move freely [30]. The language used is English with most participants being non-native speakers. In total 4 participants were recorded and the duration of each meeting is approximately 90 min.
Non-speech sounds, including laughter, were manually annotated. There are around 300 laughs with an average duration of 1.50 s and around 1000 speech samples with an average duration of 2 s. The database is available from http://www.speech-data.jp/corpora.html.
AudioVisual Laughter Cycle database (AVLC) [33]: The AVLC database was designed to elicit laughter from the participants, who were recorded while watching funny video clips for approximately 10 min. In total 24 subjects were recorded, 9 females and 15 males from various origins. The average age is 29, with a standard deviation of 7.3.
It should be noted that this database was mainly created for laughter synthesis purposes [55]. As a consequence, facial markers were

Table 1
Existing audiovisual databases containing laughter. Each table corresponds to one scenario: meeting, dyadic interaction and elicited laughter. Y: yes, N: no, C: camera, H: headset, L: lapel, F: far-field, ?: indicates that this information is not available.

	Elicited laughter				Meeting	
	MAHNOB	AVLC [33]	MMI-V [34]		FreeTalk [47]	AMI [31]
No. episodes	563	1066	164	No. episodes	~300	?
No. subjects	22	24	9	No. subjects	4	?
Video res.	720×576	640×480	640×480	Video res.	1040×1040[a]	720×576
FPS	25	25	25	FPS	60	25
Audio (kHz)	48/44.1	44.1	48	Audio (kHz)	16	16
Mic. type	C, L	H	C	Mic. type	F	H, L, F
Web-based	Y [43]	N[b][44]	Y [45]	Web-based	Y [47]	Y [48]
Searchable	Y	N	Y	Searchable	N[c]	N[c]
Used in		[46]	–	Used in	[30]	[24,29] [49,27]

	Dyadic interaction		
	SEMAINE [50]	AVIC [51]	DD [52]
No. episodes	?	324	238
No. subjects	150	21	41
Video res.	780×580	720×5760	640×480
FPS	50	25	30
Audio (kHz)	48	44.1	48
Mic. type	H, F	L, F	C
Web-based	Y[53]	N	N[d]
Searchable	Y	N	N
Used in	–	[51,54]	[49]

[a] This is the resolution of the 360° camera.
[b] Audio recordings are available for download online. Video recordings are available upon request to the authors.
[c] Browsing tools are provided but no search functionality is available.
[d] This database is not publicly available.

Please cite this article as: S. Petridis, et al., The MAHNOB Laughter database, Image Vis. Comput. (2012), http://dx.doi.org/10.1016/j.imavis.2012.08.014

placed on subjects' faces in order to enable facial motion tracking by infrared cameras, which are used to animate a laughter expression by an embodied conversational agent in [46]. This motivation also explains why there are few other vocalizations except laughter.

The database contains mostly laughter, but other vocalizations like breathing or short speech utterances are present as well. Manual annotation was performed by one annotator using a hierarchical protocol. The annotations rely mostly on the audio source, although the visual information was also taken into account for annotating the facial expression boundaries or detecting silent laughs. The database contains 1066 spontaneous laughs and 27 posed laughs. The duration of the annotated laughs ranges from 250 ms to 82 s. The audio recordings and the annotations file are available from http://tcts.fpms.ac.be/~urbain/, and the video recordings are available upon request from the authors.

MMI-database, part V [34]: The MMI database contains mainly videos of posed facial expressions. A recent addition, part V, is targeted at spontaneous expressions and includes a few recordings of elicited laughter. In total 9 subjects, 4 males and 5 females, were recorded, while watching funny videoclips.

The laughs have been annotated manually by one annotator and further divided into voiced and unvoiced laughter. Annotation was performed using both audio and video information, so some smiles which are not accompanied by an audible laughter sound have also been annotated as laughter. In total, 164 laughter episodes, 109 unvoiced and 55 voiced, have been annotated. The mean duration of the voiced and unvoiced laughter episodes is 3.94 and 1.97 s, respectively. The database is available from http://www.mmifacedb.com/.

3. The MAHNOB Laughter database

This section describes the MAHNOB Laughter database, together with the limitations of the currently available databases it aims to overcome. Its main advantages are, in the first place, that it contains a lot of speech from the same subjects that produce laughter.This makes it suitable for the analysis and discrimination of the audiovisual laughter and speech characteristics. In second place, the video recordings are more suited for automatic analysis. On the MMI—part V database the subjects were placed about 1.5 m away from the camera, resulting in a lower face resolution, whereas in the AVLC database a webcam was used to capture the video.

Furthermore, in the AVLC database the subjects have markers on their face, hindering the application of automatic visual feature extraction. Thirdly, the MAHNOB Laughter database contains cleaner audio recordings than MMI—part V, where the audio signal was captured using just the integrated camera microphone. Finally, thermal recordings are also available. To the best of our knowledge, only the NVIE database [36] offers publicly available recordings of facial expressions in thermal imagery. However, in this case, no audio recordings were made.

The MAHNOB Laughter database is freely available and can be downloaded from http://mahnob-db.eu/laughter/. The database is organized in sessions as described below. Each session is named to reflect the subject ID present and the session number it represents. The format used is SXXX-YYY where XXX is the subject ID (001 to 025) and YYY is the session number, e.g., S011-003.

3.1. Recording protocol

We used a large collection of funny video clips in order to elicit laughter from the subjects. This is a common way of eliciting laughter in psychological studies [16,17,56,7]. Two of the clips were used to elicit laughter in a previous psychological study [16], whereas the other clips were found on the Internet. The length of the clips ranged from just a few seconds to 2 min. Several sessions per subject were recorded, each of them using from 1 to 5 video clips depending on their duration.

In addition to the sessions for recording laughter, four more sessions were recorded for all subjects. In two sessions each subject was asked to speak in English and in his/her mother language, if different from English, for approximately 90 s in each case. They were also given the option to either select a topic and talk about it or to hold a conversation with a friend or an operator. The goal for recording in two languages was to create a multilingual speech corpus in order to investigate if language can affect the performance of a laughter-vs-speech discrimination system. To the best of our knowledge all previous works employ only English language, potentially causing a bias to the discrimination models.

Subjects were also asked to produce several posed laughter episodes. As suggested in [2] laughter on command is embarrassing and therefore most people cannot easily produce posed laughter. More than half of the subjects in our recordings, found it difficult to laugh on command with the most common reaction being spontaneous laughter while trying to produce a posed laughter. Therefore, sometimes this session had to be repeated several times. This agrees with [5], which reports that about half of the subjects could not laugh on command in a similar experiment.

Subjects were finally asked to produce several posed smiles, starting from neutral and going back to neutral. The vast majority of subjects were successful in this task.

In the laughter session the subjects were simply told that they should watch some video clips, without being informed about the content of the clips or the aim of the study (with the exception of the three authors, subjects S003, S004 and S009). In other words, no instructions were given on how they should react, and they were allowed to move their hands and heads freely. Sessions containing posed smile, posed laughter and speech were usually recorded after the laughter sessions, and subjects were explicitly told what they should do.

Finally, two people were present in the recording room in order to explain the procedure to the subjects, operate the systems, and interact with the subjects during the speech sessions if necessary. As shown in Figs. 2, 3 and 4 the background can differ significantly from one session to another. This was done on purpose in order to create more diverse background environments which resemble more realistic scenarios (compared to the artificial scenario of having the same background in all sessions). The background is usually the same for all sessions of the same subject but changes from one subject to another. In addition, a totally static background was not desired so some people sometimes appear in the background moving freely. This is especially important for the thermal recordings, where the difference between the skin temperature and the typical temperature of the rest of the background is enough to obtain a perfect foreground segmentation.

In total, 90 laughter sessions, 38 speech sessions, 29 posed smile sessions and 23 posed laughter sessions were recorded. It should be emphasized that spontaneous laughter episodes occur in all types of sessions. It is natural for people to laugh during the speech sessions or when trying to produce posed smiles or posed laughter.

3.2. Recording setup

Video recordings were made at 25 fps using a JVC GR-D23E Mini-DV video camera which has a resolution of 720×576 pixels. The video recordings were compressed using a H264 codec. Deinterlacing was performed using edge-directed interpolation.

The camera also records audio using its built-in microphone (2 channels, 48 kHz, 16 bits). Since the camera is positioned at some distance from the subject, the signal-to-noise ratio is low. In order to record clean audio, a lapel microphone was also used (1 channel, 44.1 kHz, 16 bits). The audio recordings were saved

Please cite this article as: S. Petridis, et al., The MAHNOB Laughter database, Image Vis. Comput. (2012), http://dx.doi.org/10.1016/j.imavis.2012.08.014

directly to a hard disk as uncompressed wav files using the EDIROL UA-25 audio interface.

Thermal recordings were obtained using a VarioCAM Head HiRes 384 camera. The thermal images are originally captured with a resolution of 384×288 pixels. With such resolution it was hard to distinguish some face structures, which hindered manual annotations, e.g., in terms of the facial component location. Therefore, they were subsequently transformed using the software provided by the manufacturer to 576×432 pixels. The frame rate is set to 25 fps. The temperature resolution of the camera is 0.1 °C. Since the sensor is not cooled, a non-uniformity correction (NUC) process is required to eliminate noise from the recordings. This stops the recordings for almost 2 s. To prevent any interruption of the recordings, this process was run prior to starting the sessions, which were kept shorter than 3 min. Even with this precaution, a noise increment may be observed for sequences longer than 2 min. In cases where the elicited reaction was longer than 3 min, the video and the audio recordings were kept until the subject returned to a neutral emotion state.

The video and the thermal cameras were placed next to each other and the distance between the lenses was about 10 cm. They were placed in front of the subjects at a variable distance between 0.5 and 1 m from them. A laptop was used to play the videos and placed directly under the cameras.

In order to allow the subjects to listen to the audio from the clips without interfering with the audio recordings, the subjects wore earphones whenever it was necessary. Finally, frontal or almost-frontal illumination was provided using halogen lamps placed behind the cameras, so the amount of shadows in the face and their variability is minimized.

3.3. Participants

In total 25 subjects were recorded, but 3 subjects were excluded due to equipment failure. Therefore, 22 subjects are included in the database, 12 males and 10 females, from 12 different countries and of different origins. The average age for males and females is 27 (standard deviation: 3) and 28 (standard deviation: 4), respectively. There is 1 male subject with a beard and 1 female subject with glasses (most subjects were asked to take off their glasses since they are opaque to thermal lighting). All subjects gave written consent to allow the use of their recordings for research purposes.

3.4. Synchronization of data streams

The audio from the camera microphone is already synchronized with the video stream since they are recorded by the same camera.[3] On the other hand, the audio from the lapel microphone needs to be synchronized with the camera audio and video streams since it was recorded by a different device. In order to automatically obtain this synchronization we use a cross-correlation measure, which is a widely used technique to measure similarity between two audio signals, and a detailed description can be found in [42].

The thermal video and the visible video are synchronized manually. We recorded at the beginning of each session a distinctive event, easily visible in both video streams. The most common was using a lighter, but a hand clapping was also used in some sequences. However, since the thermal video had to be directly recorded into a computer, some frames were dropped during the data transfer. To account for this, we use the timestamp information provided by the camera, that was accurate up to a second. Through this information, we can detect when a frame was dropped within a second, but we do not know exactly which frame is missing. The available synchronization information consists of one manually obtained synchronization point per sequence, and a list of synchronization points spaced 1 s in time, obtained using the timestamp information. Unfortunately, for some of the sequences the number of missing frames is too high to provide a good synchronization. These sequences were excluded, resulting in the removal of 48 out of 180 sessions recorded.

3.5. Annotation

It is relatively easy to define the start and end points of a speech utterance, but the same is not true for laughter. Little is said in the literature about when a laughter episode starts and ends [58] and even phoneticians do not agree on how laughter should be segmented. As a result there are no standard rules on how to label a laughter episode. In this work, we use the same terminology as in [2,58] where a laughter episode is defined as several bouts[4] separated by one or more inhalation phases. Furthermore, laughter episodes were annotated using the criterion suggested in [16]: "Laughter is defined as being any perceptibly audible expression that an ordinary person would characterize as laughter if heard under everyday circumstances". Annotation was performed using mainly the audio channel (lapel microphone audio) and the start and end points were defined as the start and end of the audible expression. 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 for all experiments conducted. In cases where it was not clear if a laughter should be considered as one episode or two consecutive episodes the video channel was used as well, similar to [33].

Laughter is often followed by an audible inhalation and it is not clear if this belongs to laughter or not [58]. In the AVLC database [33] it was considered as part of the laughter whereas in [16] it was not. Therefore, we introduce two different sets of labels, one that contains the inhalation and one which does not contain it as described below.

Annotation was performed by one annotator using the ELAN annotation tool [59,60] using the following 9 labels: *Laughter*: Used for laughter as defined above. *Speech*: Used for speech. *Speech–Laugh*: Used to label segments where speech and laughter occur simultaneously. It is still unclear whether speech laughter is closer to laughter or speech. In some studies it has been assigned to speech [22,61]. On the other hand, it has been proposed that speech laughter should represent an independent category [58] since it is not simply laughter superimposed on articulation but a more complex vocalization. In other databases [33] and studies [62,63] it has been annotated and studied as an independent category, and we follow the same approach. *Posed Smile*: Used to label posed smiles. *Posed Laughter*: Used to label posed laughter. *Other*: Used to label other human sounds, e.g., coughing, throat clearing, etc. *Laughter + Inhalation*: Used to label laughter but also contains the terminal inhalation whenever it is present. *Speech Laugh + Inhalation*: Used to label speech laughter but also contains the terminal inhalation whenever it is present. *Posed Laughter + Inhalation*: Used to label posed laughter but also contains the terminal inhalation whenever it is present.

Fig. 1 shows a screenshot of the annotation window. It shows a series of vocalizations: *Laughter–Speech Laugh–Laughter–Speech Laugh*, *Laughter*, *Speech*, and *Laughter*. The annotation of the inhalation at the end of laughter episodes is also visible in label *Laughter + Inh*. It can be seen that both the duration and the time interval between the end of the audible laughter and the inhalation vary a lot.

In some studies [61,16,37,62,38] laughter has been divided into voiced and unvoiced categories, so we have used a second level of annotation for all the episodes of laughter, i.e., those that fall into the category *Laughter*. In the literature, these two types of laughter can be distinguished either manually by human annotators [62] or automatically using the pitch contour, e.g., in [38] "Laughs were coded

[3] In [57] it is reported that this camera synchronizes audio and video with 38 ms offset (constant).

[4] Bout is a sequence of laughter syllables in one exhalation phase [58].

Please cite this article as: S. Petridis, et al., The MAHNOB Laughter database, Image Vis. Comput. (2012), http://dx.doi.org/10.1016/j.imavis.2012.08.014

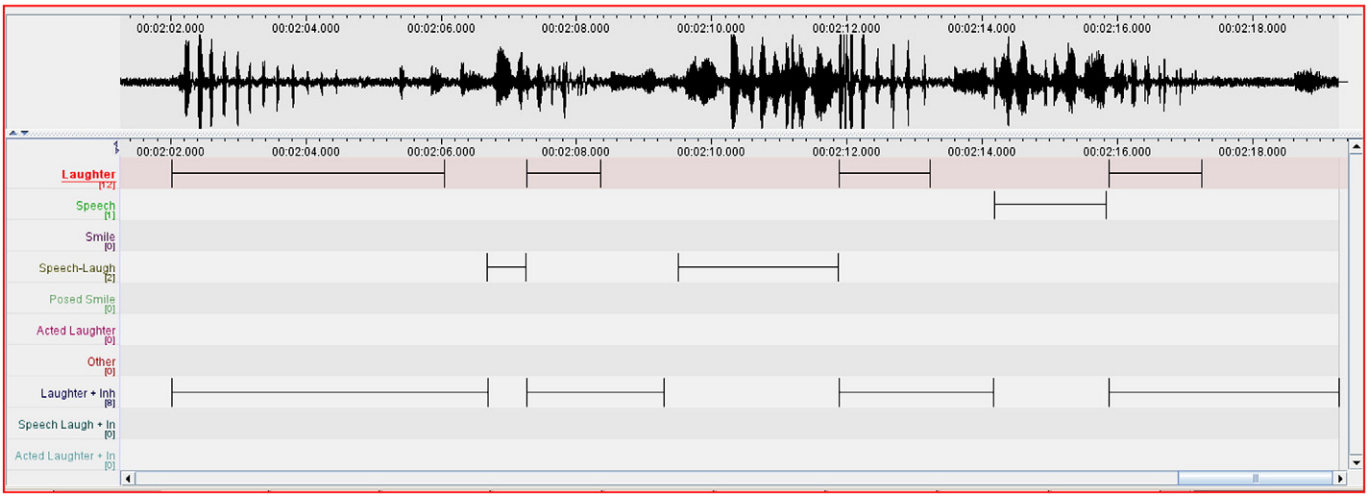

Fig. 1. Screenshot from the ELAN annotation tool. The labels together with the annotations are shown for a segment from session S023-003.

as voiced if there were stereotyped episodes of multiple vowel-like sounds with evident F0 modulation in 50% or more of the sound".

In this study we used a combination of both approaches. First, two human annotators labeled all the laughter episodes. Then PRAAT software [64] was used to compute the number of unvoiced frames in each episode, which was then divided by the total number of frames, resulting in the unvoiced ratio of the episode. If the unvoiced ratio was higher than 85% then an unvoiced label was given, otherwise the episode was labeled as voiced. In addition, two human annotators manually labeled each episode. Finally, each episode was labeled based on majority voting of the 3 votes (2 human votes and 1 based on PRAAT). The agreement between the two annotators was 91% and the agreement between the human annotators and the PRAAT-based annotation was 79% and 83%, respectively. That set of annotation cannot be found on the ELAN files, since it was a postprocessing stage, so a separate file is provided with the labels (voiced/unvoiced) on http://mahnob-db.eu/laughter/media/site/documents/voicedUnvoicedLaughter_Annotations.xls.

Examples of voiced laughter is shown in Fig. 2, an example of unvoiced laughter in Fig. 3, and posed laughter is shown in Fig. 4. The corresponding audio signals and spectrograms from the lapel microphone are shown in Fig. 5a to c, with the exception of posed smiles which do not produce any vocalization. Fig. A.17a to c in the Appendix show the same audio signals and spectrograms from the camera microphone.

3.6. Database statistics

In this section we provide relevant statistics of our database, as the total number of episodes, their duration and pitch for all subjects, and male and female subjects separately (summarized in Table 2). Furthermore, we provide these statistics for the cases of voiced and unvoiced laughter. To this end, we considered the 563 laughter episodes and 849 speech utterances annotated, while two subjects were not able to produce posed laughter. During the speech sessions 35 laughter episodes occur with 27 labeled as voiced laughter. Results are shown for all subjects, but also for each gender separately.

3.6.1. Duration

The mean duration of the laughter episodes is 1.65 s. This is similar to the mean duration reported in [19] and [65] of 1.80 and 1.615 s, respectively, being both studies based on multi-party meeting recordings. In other studies using elicited laughter the mean duration reported varies a lot, from less than 1 s [17,14] to 3.5 s [33]. In these cases the differences steam mainly from either the subjects showing very moderate laughing response to the video clips [17] or laughter being annotated using a different criterion. Fig. 6a shows the histogram of the laughter episodes duration. It can be seen that the majority of laughter episodes are relatively short, as 63.7% of them have a duration of up to 1.25 s, and laughs with a duration of up to 3.75 s account for 90.2% of all episodes. In our recordings posed laughter episodes are almost twice as long as spontaneous laughter episodes, being consistent with the results presented in [33]. However, we define the boundaries of the laughter episodes differently, so the total duration of the episodes is not directly comparable. In our study their average duration is 3.13 s for posed laughter and 1.65 s for spontaneous, while in their study the average duration is 7.7 s and 3.5 s, respectively.

3.6.2. Pitch

The mean pitch for laughter, 477 Hz, is very similar to the mean pitch of 475 Hz reported in [19]. The range reported in the literature [56,15–17,66,7,63] for male and female mean pitch of laughter is 126–424 Hz and 266–502 Hz, respectively. The male mean pitch, 400 Hz, falls in this range but the female pitch, 535 Hz, is slightly higher than what has been previously reported. In agreement with all the previous studies, the mean laughter pitch for both males and females is higher than the mean pitch of speech. The variability of

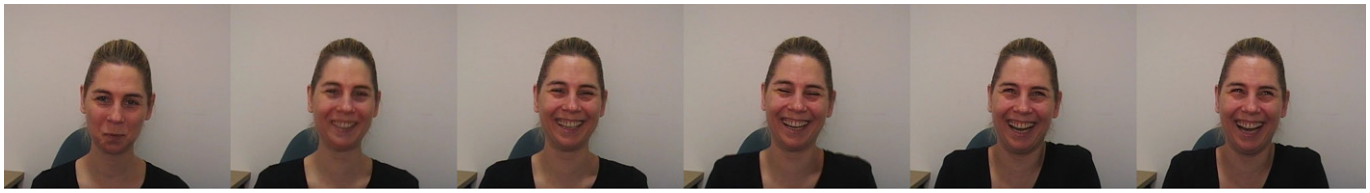

Fig. 2. Example of voiced laughter displayed by subject S009, Session S009-004.

Please cite this article as: S. Petridis, et al., The MAHNOB Laughter database, Image Vis. Comput. (2012), http://dx.doi.org/10.1016/j.imavis.2012.08.014

Fig. 3. Example of unvoiced laughter displayed by subject S005, Session S005-008.

pitch in female subjects is also much higher than for male subjects [16] (standard deviation of 169 and 96 respectively). The mean pitch for posed laughter is 372 Hz, lower than for spontaneous laughter, which is 477 Hz, being the former less variable than the latter. In the particular case of speech–laughter, the mean pitch is even lower than for posed laughter but higher than for speech.

3.6.3. Voiced/unvoiced laughter

Statistics for voiced and unvoiced laughter are shown in Table 3. Consistent with the results presented in [16], voiced laughs are longer than unvoiced laughs for both genders. Voiced laughs with a duration of up to 1.25 s account for 51% of all the voiced episodes, whereas unvoiced laughs account for 80% of all the unvoiced episodes. It can also be seen that female subjects produce more voiced laughter episodes than male subjects, 20.3 and 10.45 respectively. This also agrees with the results in [16]. On the other hand, there is no difference in the average number of unvoiced laughs produced by males and females, 11.2 and 11.1, respectively. This is in contrast to the finding that males produce more unvoiced laughs than females [16]. Fig. 6b and c shows the histograms of the duration of voiced and unvoiced laughs.

4. Experimental setup

4.1. Visual features

To capture face movements in an input video, we track the 20 facial points shown in Fig. 7. These points are the corners/extremities of the eyebrows (4 points), the eyes (8 points), the nose (3 points), the mouth (4 points) and the chin (1 point). To track these facial points we used the particle filtering algorithm proposed by Patras and Pantic [67], applied to tracking color-based templates centered around the facial points. The features are computed using the Point Distribution Model (PDM) built in [24] from the AMI dataset [68]. As suggested in [69] the facial expression movements are encoded by the projection of the tracking points coordinates to the N principal components (PCs) of the PDM which correspond to facial expressions. As shown in [24] PCs 7–10 were found to correspond to facial expressions (confirmed through visual inspection) so our shape features (shape parameters) are the projection of the 20 points to those 4 PCs. Further details of the feature extraction procedure can be found in [24].

4.2. Audio features

We use a set of 6 Mel Frequency Cepstral Coefficients (MFCCs). MFCCs have been widely used in speech recognition and have also been successfully used for laughter detection [23]. Although it is common to use 12 MFCCs for speech recognition we only use the first 6 MFCCs, given the findings in [23], where 6 and 12 MFCCs resulted in the same performance for laughter detection. These 6 audio features are computed every 10 ms over a window of 40 ms, i.e., the frame rate is 100 fps. In addition to MFCCs, the zero crossing rate (ZCR) has been used for the experiment on discrimination between voiced laughter, unvoiced laughter and speech. The reason for using ZCR is its sensitivity to the difference between voiced and unvoiced sections. High ZCR usually indicate noise, and low rates usually indicate periodicity [70].

4.3. Evaluation procedure

All the experiments described here follow a leave-one-subject-out cross-validation methodology. Due to some random components on the training procedure, each time we run a cross validation experiment we get slightly different results. In order to assess the stability of the experiments, each cross validation experiment is executed 10 times, and the mean and standard deviation are reported.

The first experiment, laughter-vs-speech, is a 2-class discrimination problem. Therefore, we train a binary classifier. The second problem consists of a 3-class classification problem, i.e., discrimination between voiced laughter, unvoiced laughter and speech. In this case we train 3 one-vs-all binary classifiers and assign the class based on the classifier with the maximum output. The classifier used is a feedforward neural network with one hidden layer and the resilient backpropagation training algorithm [71] is used for training. The number of hidden neurons, which ranges from 6 to 32, is optimized by an inner 2-fold cross-validation loop over the training subjects.

The performance is measured in terms of the F1 measure (or F1 score) and the Classification Rate (CR):

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{1}$$

$$CR = (TP + TN)/(TP + TN + FP + FN) \tag{2}$$

where $TP/TN/FP/FN$ stands for true positive, true negative, false positive and false negative respectively.

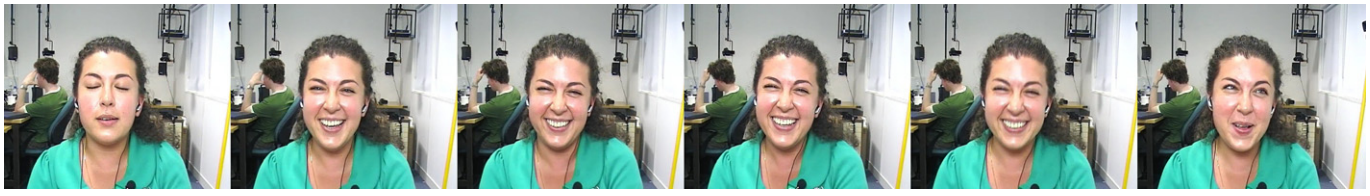

Fig. 4. Example of posed laughter displayed by subject S023, Session S023-008.

Please cite this article as: S. Petridis, et al., The MAHNOB Laughter database, Image Vis. Comput. (2012), http://dx.doi.org/10.1016/j.imavis.2012.08.014

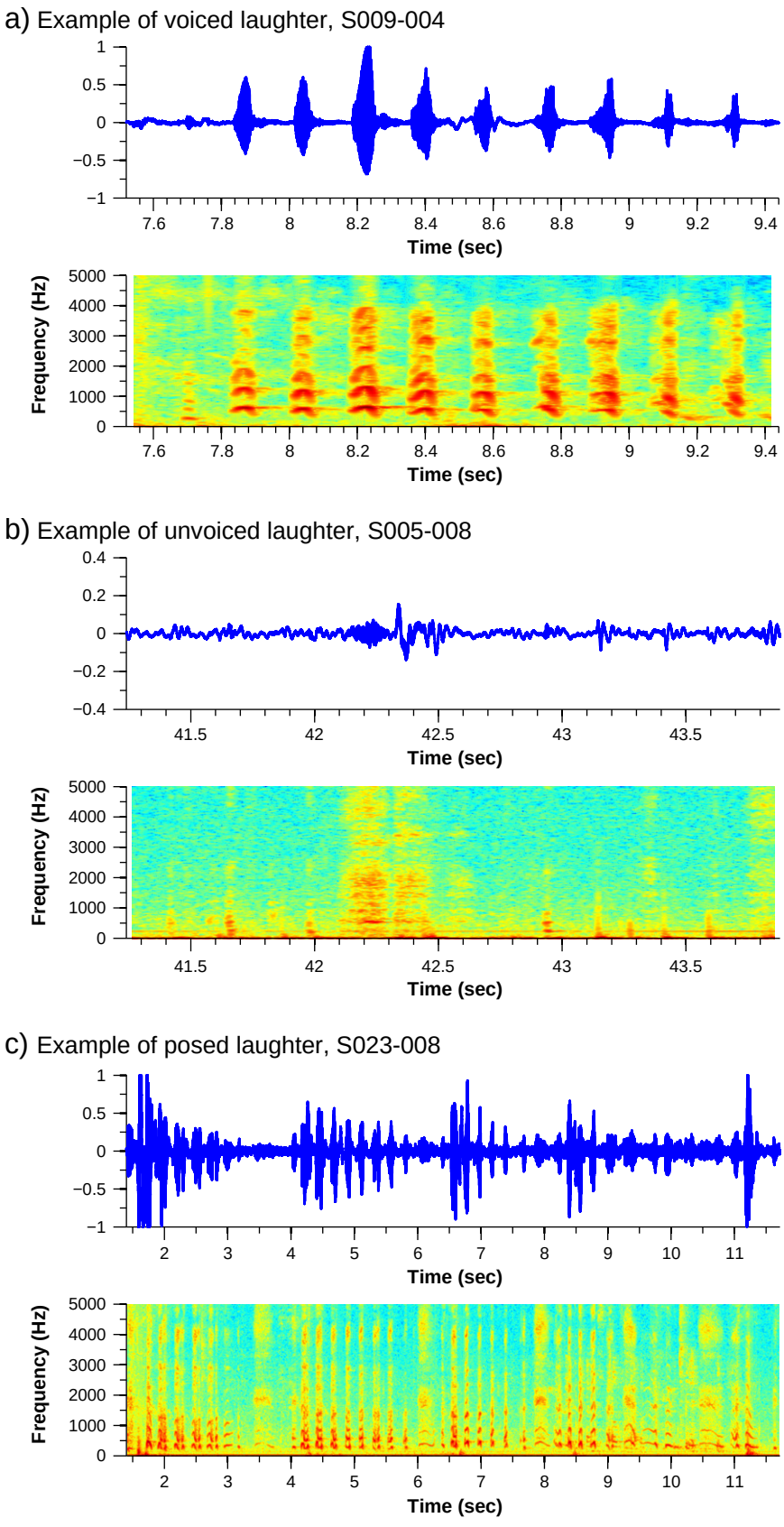

Fig. 5. Top row: audio signal (lapel microphone), bottom row: spectrogram.

We also consider the per class F1 measure. It results from considering one class as positive while the other as negative and vice versa. In particular, the measure used in the inner cross-validation loop of the parameter optimization uses the combined F1 measure, defined as $0.5 \times \text{F1 Positive Class} + 0.5 \times \text{F1 Negative Class}$, as a performance measure.

Please cite this article as: S. Petridis, et al., The MAHNOB Laughter database, Image Vis. Comput. (2012), http://dx.doi.org/10.1016/j.imavis.2012.08.014

Table 2
Database statistics for all subjects. Statistics are also shown for male and female subjects separately.

Type	No episodes/no subjects	Total duration (s)	Duration (s) Mean/Std	Pitch (Hz) Mean/Std
All subjects				
Laughter	563/22	930.72	1.65/2.32	477/157
Speech	849/22	2474.11	2.91/2.28	185/64
Posed laughter	51/20	159.79	3.13/4.45	372/120
Speech laughter	67/17	94.05	1.40/0.82	332/103
Male subjects				
Laughter	249/12	385.48	1.55/1.93	400/96
Speech	532/12	1586.85	2.98/2.19	146/35
Posed laughter	31/12	93.41	3.01/4.40	342/103
Speech laughter	23/9	33.00	1.44/0.88	293/86
Female subjects				
Laughter	314/10	545.25	1.74/2.60	535/169
Speech	317/10	887.26	2.80/2.41	249/46
Posed laughter	20/8	66.37	3.32/4.62	420/131
Speech laughter	44/8	61.04	1.39/0.79	353/107

Table 3
Statistics for voiced and unvoiced laughter episodes.

Subjects	No episodes/no subjects	Total duration (s)	Duration (s) Mean/Std	Average no laughs per subject
Voiced laughter				
All	318/21	651.81	2.05/2.64	15.14
Males	115/11	241.87	2.10/2.39	10.45
Females	203/10	409.94	2.02/2.78	20.30
Unvoiced laughter				
All	245/22	278.92	1.14/1.71	11.14
Males	134/12	143.61	1.07/1.24	11.17
Females	111/10	135.31	1.22/2.14	11.10

Both audio and visual features are z-normalized to zero mean and unity standard deviation. The means and standard deviations are computed on the training set only, and then applied to the test set. In addition, the audio and visual features are synchronized for audiovisual fusion, since they are extracted at different frame rates. This is achieved by upsampling the visual features by linear interpolation as in [72]. For the experiments using audiovisual information, the audio and visual features are concatenated in a single vector for feature-level fusion.

The training and testing of the classifiers is performed on a frame-level basis. Classification is performed by applying either the single-modal or the bimodal classifiers to all frames of the given episode resulting in a series of scores between −1 and 1 for each frame. These scores are summed across all frames and the entire episode is labeled based on the classifier that produced the maximum score over the entire sequence.

As shown in Table 2, the total duration of speech episodes is higher than the total duration of laughter episodes. Consequently, there are many more speech frames than laughter ones. This means that the training set can become unbalanced. This is even worse in the 3-class problem, where the laughter examples are further divided into 2 classes. Such an unbalanced set tends to degrade the performance of the classifier [73]. Consequently, the negative classes are constructed by randomly sampling negative examples, so they contain no more than twice the number of examples of the positive class.

A randomization test [74] is used to compare the performance of audio-only classification with the audiovisual fusion. As discussed in [75], the randomization test performs similarly to the commonly used T-test when the normality assumption is met, but outperforms the T-test when the normality assumption over the sampled population is not met. The randomization test is applied on the average performance measure, i.e., averaging over all subjects of a cross-validation experiment. Since each experiment is conducted 10 times, we end up with 10 paired differences. The significance level used was set to 5%. In order to get a more detailed view of the comparison, we also apply the randomization test for each subject separately.

5. Experimental results

In this section we report audio-only, video-only and audiovisual fusion results on two sets of experiments: 1) discrimination between laughter and speech and 2) discrimination between voiced laughter, unvoiced laughter and speech. Both types of speech, English and the subjects' mother language, are merged and considered as one class. For a preliminary analysis of language effects interested readers are referred to [42]. We also present results on laughter-vs-speech discrimination based on thermal images. Please note that the results presented here should be considered as baseline results.

A total of 13 examples (9 laughter, 4 speech) have been excluded due to either facial occlusions while laughing/speaking, or the face the field of view of the camera. Therefore, no useful visual information could be extracted.

In total, 554 laughter episodes (311 voiced and 243 unvoiced) and 845 speech utterances were used. The complete list of annotations can be found in [43].

5.1. Laughter-vs-speech discrimination results

The aim in this set of experiments is to discriminate laughter from speech using the audio channel, the visual channel and the combination of both. Two different scenarios are considered, one using the lapel microphone, where the audio noise levels are low, and one using the

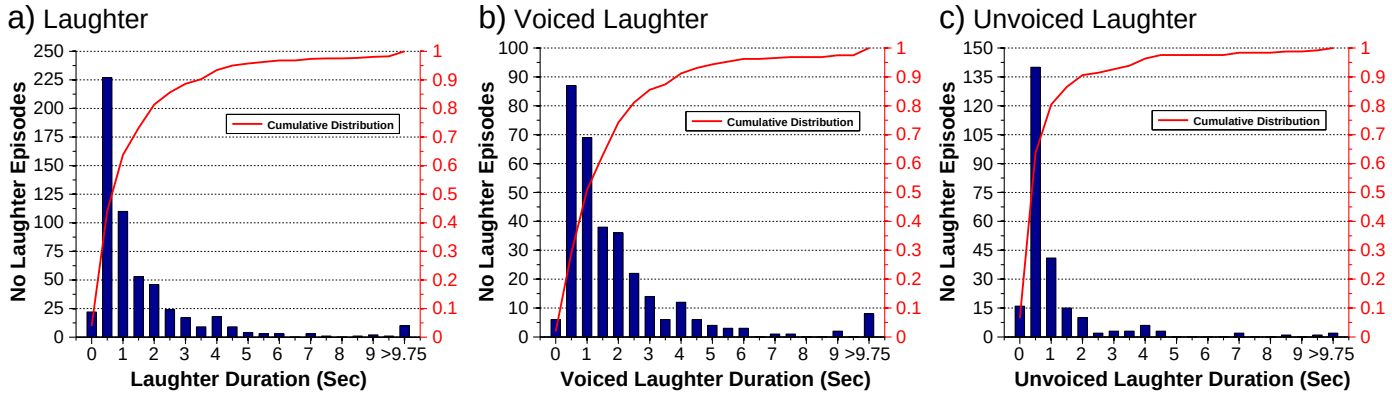

Fig. 6. Histograms and cumulative distribution of all laughter, voiced laughter and unvoiced laughter duration. Each bin corresponds to an interval of 0.5 s, except for the first, with an interval of 0.25 s, and the last bin (>9.75 s). The y-axis for the cumulative distribution is shown on the right.

Fig. 7. Example of the 20 tracking points used. Session S007-004, frame 1099.

camera microphone, where the audio signal is noisy. Results for both scenarios are shown in Table 4. Video-only classification results in both cases in the worst performance. Note that in both cases the results for video-only classification are the same since they differ only in the audio channel used. This is not surprising since the audio channel carries most of the information whereas the video channel carries some extra complementary information. In addition, laughter is usually accompanied by a facial expression but there are some cases that these expressions are subtle so the additional visual information is negligible. This is in agreement with previous studies in audiovisual laughter-vs-speech discrimination, e.g., [24], but also on audiovisual speech [72,76] and affect recognition [77].

In the case of lapel microphone audio, it can be seen from Table 4 that the audiovisual fusion does not result in an improved performance over the audio-only performance. The F1 measure for speech and the CR are practically the same for both approaches. Only the F1 measure for laughter benefits a little, with a small but statistically significant absolute increase of 1.1%. The results are different in the case of the camera audio. As a consequence of the lower signal to noise ratio, the audio-only performance decreases but it is still better than the video-only performance. The audiovisual fusion significantly outperforms the audio-only performance resulting in a 7.2%, 2.5% and 3.9% absolute increase in F1 laughter, F1 speech and CR, respectively.

Previous findings state that the combination of audio and visual information was beneficial for speech-vs-laughter discrimination [24,78,28,30], in contradiction with the experiments with low audio noise levels. However, in all previous works a meeting scenario was considered and, as a consequence, the audio signal was noisy. The results from both experiments are in agreement with [72,76], which report that in audiovisual speech recognition the benefits of fusion become apparent when the noise level increases.

Fig. 8a shows the classification rates per subject (male subjects only) for the audio-only, video-only and audiovisual approaches when the camera audio is considered. In the single-modal experiments the classification performance per subject varies a lot. It ranges from 65% for subject S05 to 100% for subject S16 for video-only, and from 73% for subject S03 to 100% for subject S21 for audio-only. The audiovisual approach is less subject-dependent with all subjects achieving an accuracy of over 86%, except for subject S05 that yielded a CR of 76%. The difference between the audio-only and audiovisual classifiers is statistically significant for all male subjects except for S06, S07, S21 and S24, with fusion being worst for S05, S17 and better for the rest.

For the case of female subjects (Fig. 8b) similar conclusions can be drawn regarding the variability of the performance. In this case the difference between the audio-only and audiovisual classifiers is statistically significant for all subjects except for S09, S11 and S22, with the audiovisual fusion being beneficial for S02, S14, S20 and S23 and worse for 3 subjects, S08, S13 and S25.

It can be seen in Fig. 8, that for some subjects the video-only performance is much lower than the audio-only performance and vice versa. In case of the video-only approach this is likely the result of bad tracking of the facial points. In particular, subjects S17 and S25 produce large head movements which significantly degrade the quality of tracking and subject S22 wears glasses which also affects the performance of the tracker. On the other hand, subjects like S05 and S07 only produce subtle facial expressions and mainly produce unvoiced laughs (3 voiced and 19 unvoiced laughter episodes for subject S05 and 6 voiced and 21 unvoiced laughter episodes for subject S07) which are harder to detect from video-only as explained in Section 5.2. Regarding the degraded audio-only performance for subjects S04, S19, S20 and S23 this could be the result of different background noises, since the recording conditions are not the same for all sessions. In addition, the fact that subjects have different accents, since almost all of them are non-native English speakers, should also be taken into account.

Table A.8 shows the normalized confusion matrices for the audio, video and audiovisual classifiers. It can be seen that the main source of confusion for video is laughter episodes being classified as speech (31%), whereas much fewer speech utterances are classified as laughter (6.4%). The same is true for audio, where very few speech utterances are confused with laughter (0.9%).

When combining audio and video, the number of speech utterances confused with laughter increases (compared to audio-only), from 0.9% to 3.4%, but the number of laughter episodes classified as speech significantly decreases from 33.4% to 19.9%.

In order to illustrate when the addition of visual information helps, examples of speech and laughter are shown in Figs. 9, 10 and 11. Fig. 9 shows a laughter episode produced by a male subject which is confused with speech by audio, but since it is accompanied by a pronounced smile the audiovisual classifier is able to and correctly classify it as laughter. The previously shown example in Fig. 2 also corresponds to this same situation. Fig. 10 shows another example of laughter misclassified by the audio classifier. In this case, the laughter is accompanied by a very weak smile. Therefore, the addition of visual information does not help so the example is misclassified as speech also by the audiovisual classifier. Finally, Fig. 11 shows a speech example which is correctly classified by the audio classifier but since it is produced with a smile it is confused with laughter by the audiovisual classifier.

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

1. Audiovisual fusion is beneficial when the audio signal is noisy, but does not help when the noise level is low.

Table 4
Mean (and standard deviation) of the F1 and classification rates (CR) for laughter-vs-speech discrimination of 22-fold cross validation conducted 10 times. When the difference between the audio and the audiovisual classifier is statistically significant then this is denoted by † next to the best value.

Lapel microphone audio				Camera microphone audio		
Modalities	F1 laughter	F1 speech	CR	F1 laughter	F1 speech	CR
Video	77.2 (0.7)	87.5 (0.3)	83.9 (0.4)	77.2 (0.7)	87.5 (0.3)	83.9 (0.4)
Audio	84.7 (0.5)	92.0 (0.2)	89.5 (0.3)	79.3 (1.1)	89.7 (0.4)	86.2 (0.6)
Audiovisual	85.8 (0.5)†	91.8 (0.2)	89.6 (0.3)	86.5 (0.6)†	92.2 (0.3)†	90.1 (0.4)†

Please cite this article as: S. Petridis, et al., The MAHNOB Laughter database, Image Vis. Comput. (2012), http://dx.doi.org/10.1016/j.imavis.2012.08.014

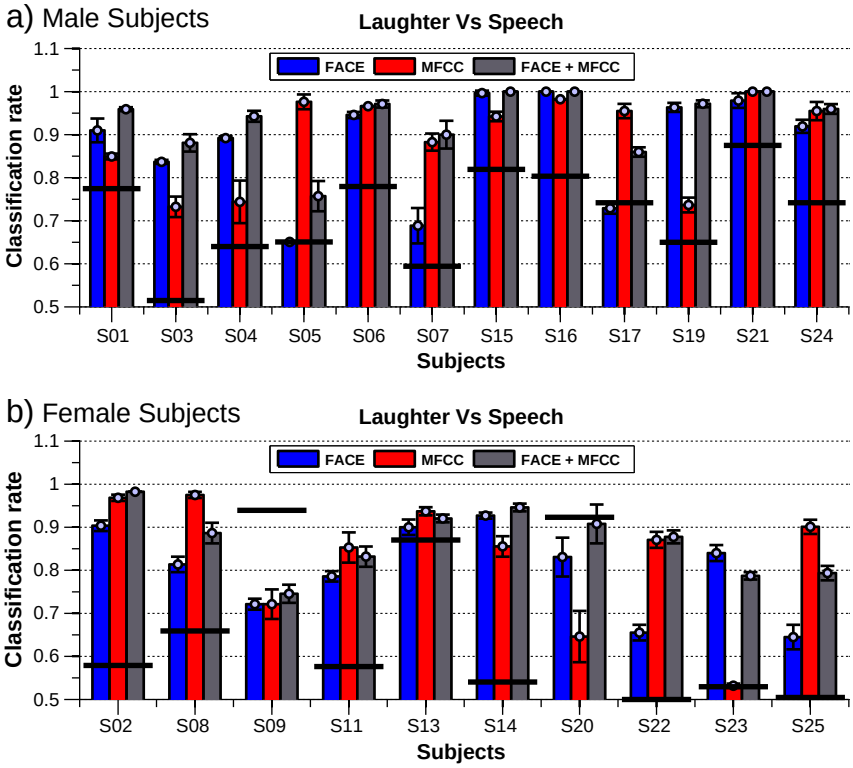

Fig. 8. CR per subject for audio-based, video-based, and audiovisual fusion approaches to laughter-vs-speech discrimination. The results presented are the mean and standard deviation for each subject averaged over 10 runs. Horizontal lines indicate the CR attained by a system that always predicts the most common class on the test set for a given subject. No horizontal line means the majority guessing CR is less than the lower limit of the plot (50%).

2. Most errors are laughter episodes confused with speech. Relatively few speech utterances are confused with laughter.
3. The addition of visual information to audio slightly increases the number the speech utterances misclassified but significantly reduces the number of misclassified laughter episodes.

5.2. Voiced laughter–unvoiced laughter–speech discrimination results

In this set of experiments laughter is divided into 2 classes, voiced and unvoiced, as described in Section 3.5. We have a 3-class problem and we use three one-vs-all classifiers.

Results are shown in Table 5. It is obvious that the performance is worse than laughter-vs-speech discrimination for both types of audio used. The video-only classifier performs better than the audio-only classifier for voiced laughter, but it performs much worse for unvoiced laughter for both types of audio.

In the case of the lapel microphone audio, the audiovisual fusion leads to an improvement over the audio-only classification of voiced laughter and speech, with an absolute increase in F1 for laughter of 10.2% and 2.3%, respectively. On the other hand it is harmful for unvoiced laughter, resulting in an 11.9% absolute decrease in the F1 measure for laughter. This is due to the poor performance of the video classifier on unvoiced laughter. Overall, a 0.8% improvement in the CR is reported, which is a statistical significant difference.

In the case of the camera microphone audio, the performance of the audio classifiers is further degraded due to noise. The same conclusion as above can be drawn. The audiovisual fusion outperforms the audio-based classification for voiced laughter and speech, resulting in an absolute increase of the F1 measure of 17.5% and 4.6%, respectively. Similarly, the addition of visual information to audio leads to a 6.3% absolute decrease in the F1 measure for the case of unvoiced laughter. Overall, a 4.1% increase in the CR is reported.

Fig. 12a shows the classification rates per subject (male subjects only) for the audio-only, video-only and audiovisual approaches when the camera audio is considered. In the single-modal experiments the classification performance per subject varies a lot. It ranges from 62% for subject S03 to 96% for subject S15 for video-only, and from 54% for subject S04 to 100% for subject S21 for audio-only. The audiovisual approach is less subject-dependent with all subjects achieving an accuracy between 73% for S05 and 98% for S21. The difference between the audio-only and audiovisual classifiers is statistically significant for all male subjects, being better for S01, S03, S04, S06, S07, S15, S16, S19 and S24 and worse for the other three.

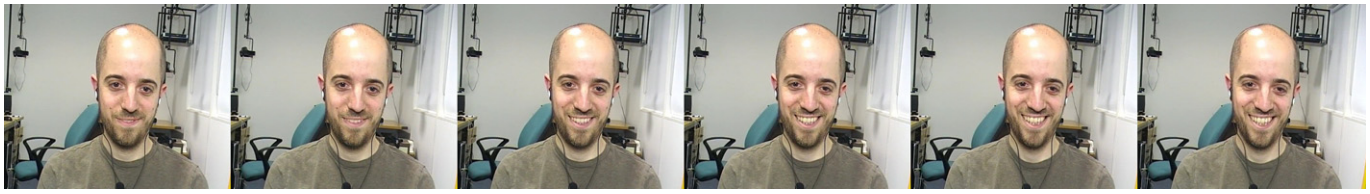

Fig. 9. Example of voiced laughter displayed by subject S019, Session S019-002. This example is misclassified by the audio classifier but correctly classified by the audiovisual classifier.

Please cite this article as: S. Petridis, et al., The MAHNOB Laughter database, Image Vis. Comput. (2012), http://dx.doi.org/10.1016/j.imavis.2012.08.014

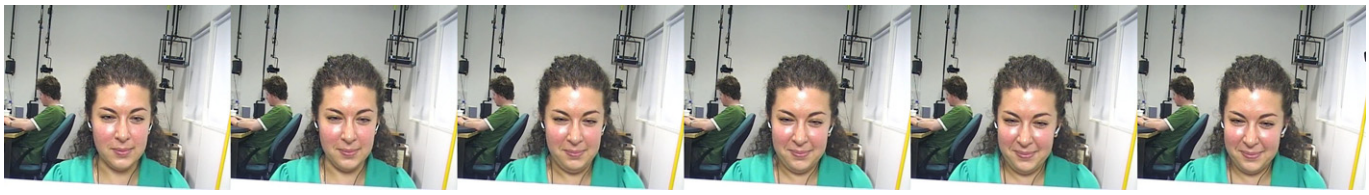

Fig. 10. Example of voiced laughter displayed by subject S023, Session S023-004. This example is misclassified by both the audio and audiovisual classifiers.

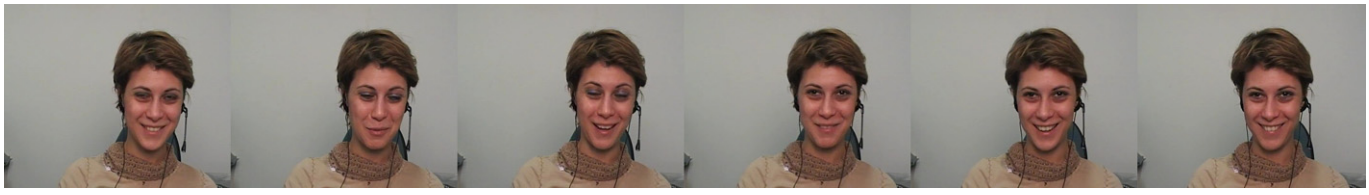

Fig. 11. Example of speech displayed by subject S020, Session S020-002. This example is classified correctly by the audio classifier but misclassified by the audiovisual classifier.

For the case of female subjects (Fig. 12b) similar conclusions can be drawn regarding the variability of the performance. In this case the difference between the audio-only and audiovisual classifiers is statistically significant for all subjects except for S09, S11 and S13, with the audiovisual fusion being beneficial for S02, S14, S20 and S23 and worse for 3 subjects, S08, S22 and S25.

By comparing Fig. 12a and b we see that male subjects tend to have higher performance. The mean CR for male subjects is 80.8%, 82.9% and 88.1% for video-only, audio-only and audiovisual classification, respectively whereas the mean CR for female subjects are 64.9%, 75% and 77.2%.

Table A.9 shows the normalized confusion matrices for the audio, visual and audiovisual classification, respectively. The audio classifier mostly confuses voiced laughter with speech. This is not surprising since voiced laughter is usually described as vowel-like in literature [16].

The video classifier mostly confuses unvoiced laughter with speech, but there is also significant confusion between voiced and unvoiced laughter. The visual information yields bad results for unvoiced laughter discrimination against voiced laughter and speech, whereas it discriminates efficiently between the other two classes (see Table A.9). We believe visual information can efficiently code the differences between a fully expressive open-mouth smile typical of voiced laughter and the more closed-mouth configurations typical for speech. In contrast, unvoiced laughter visual appearance is much more evenly distributed among the closed and open mouth configurations. This can be illustrated by Fig. 13, where we show a histogram of values of the first visual feature. This feature mainly controls the opening (lower values) and closing (higher values) of the mouth [24]. It is obvious that while the histograms for male speech and voiced laughter are concentrated over more extreme values, the unvoiced laughter histogram is more balanced throughout the range of values.

Furthermore, female voiced laughter is more prone to be confused with speech than male voiced laughter (see Fig. 12), since in our database females produce voiced laughter with closed mouth more often, as shown in Fig. 15. An example can be seen in Fig. 10.

Table 6 shows the symmetric Kullback–Leibler divergence [79] for the histograms presented in Figs. 13–15. It is also obvious from this table that the distributions of the first visual feature for female speech and voiced laughter are much more similar than the corresponding distributions for male subjects. This also explains why female voiced laughter is confused more with speech than male voiced laughter.

Finally, the audiovisual classifier significantly reduces the number of voiced and unvoiced laughs confused with speech, compared to the audio classifier. This comes at the expense of a small increase in the confusions between the two types of laughter.

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

1. Audiovisual fusion is beneficial for the recognition of voiced laughter and speech but not for unvoiced laughter.
2. The audio classifier mainly misclassifies voiced laughs as speech, whereas the video classifier mainly misclassifies unvoiced laughs as speech. In addition, the video classifier systematically confuses voiced and unvoiced laughter due to similar facial expressions.
3. The addition of visual information to audio significantly decreases the confusion of the two types of laughter with speech, compared to audio, but slightly increases the confusion between them.
4. Unvoiced laughter episodes tend to be accompanied by more subtle facial expressions than voiced laughter, but there is no strong correlation between the mouth configuration and the type of laughter.
5. In this database, female subjects also produce a significant number of voiced laughter with closed mouth which confuses the video classifiers, and therefore leads to a lower classification rate compared to male subjects.

Table 5
Mean (and standard deviation) of the F1 and classification rates (CR) for voiced laughter, unvoiced laughter and speech discrimination of 22-fold cross validation conducted 10 times. When the difference between the audio and the audiovisual classifier is statistically significant then this is denoted by † next to the best value.

Cues	F1 voiced laughter	F1 unvoiced laughter	F1 speech	CR
Lapel microphone audio				
Video	60.8 (0.9)	34.0 (2.7)	87.6 (0.4)	74.5 (0.6)
Audio	57.1 (1.8)	73.6 (1.3)†	89.7 (0.5)	81.5 (0.6)
Audiovisual	67.3 (0.8)†	61.7 (1.6)	92.0 (0.3)†	82.3 (0.4)†
Camera microphone audio				
Video	60.8 (0.9)	34.0 (2.7)	87.6 (0.4)	74.5 (0.6)
Audio	52.3 (1.6)	68.3 (1.6)†	88.2 (0.3)	79.1 (0.4)
Audiovisual	69.8 (0.7)†	62.0 (1.2)	92.6 (0.3)†	83.2 (0.2)†

5.3. Results using thermal images

We aim to discriminate speech and laughter using static images on the thermal infrared spectrum. We followed a fully automatic system, described in the following. First the face is detected using a Viola and Jones-like algorithm. The location of both eyes and both nostrils

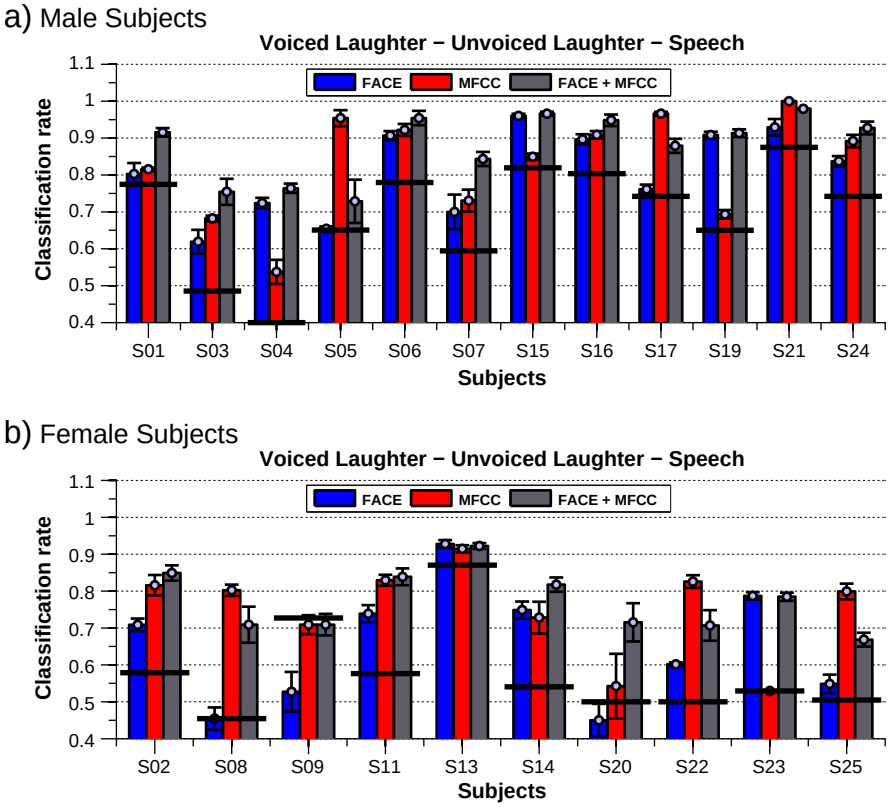

Fig. 12. CR per subject for audio-based, video-based, and audiovisual feature-level-fusion approaches for voiced laughter-vs-unvoiced laughter-vs-speech discrimination. The results presented are the mean and standard deviation averaged over 10 runs. Horizontal lines indicate the CR attained by a system that always predicts the most common class on the test set for a given subject. No horizontal line means the majority guessing CR is less than the lower limit of the plot (40%).

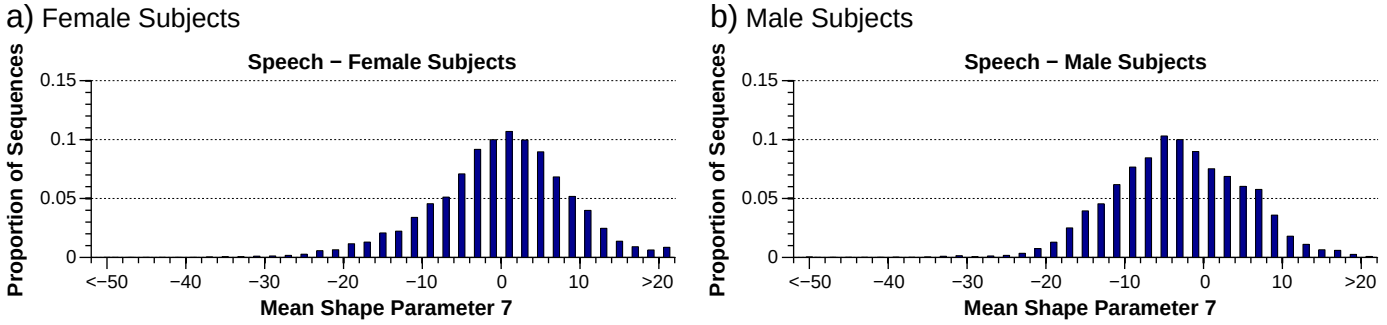

Fig. 13. Histogram of the first shape feature for female (a) and male (b) speech utterances.

is subsequently detected using the algorithm presented in [40].[5] Using these 4 points, the face is registered to a default template through a non-reflective affine similarity to eliminate the effects of translation and scaling, which is not assume isotropic.

A default region of interest (ROI) is selected over the registered face, which is specified to cover the whole mouth region even for the cases of wide mouth opening due to laughter. The set of features described in [41] is then computed over the ROI per image. Finally, exactly the same methodology as in Section 4.3 is used for classification. We opted for using the features in [41] rather than the ones in [36] since for the latter, the images are registered using manual annotations. The precision of the registration has a great influence on the PCA, since without such precise alignment, the PCA will encode differences produced by the misalignment rather than the information produced by facial expressions. In our case we aimed at performing an automatic alignment, so we discarded this approach.

In those cases where the face detection or the registration failed, the image was excluded from the dataset. However, in the cases where the registration was not accurate but did not fail, we consider the image, since such imprecision is typical from an automated system.

An example of the obtained automatic facial component localization and the ROI from where the features are extracted is shown in Fig. 16. The results obtained for the speech–laughter discrimination task are

[5] An executable for detecting the face and the facial components in thermal images is available from the group's website (http://ibug.doc.ic.ac.uk).

Table 6
Symmetric Kullback–Leibler divergence for the histograms presented in Figs. 13–15. S: speech, L_{UNV}: unvoiced laughter, and L_V: voiced laughter.

Compared distributions	Female subjects	Male subjects
$S-L_{UNV}$	12.64	14.92
$S-L_V$	6.99	18.58
$L_{UNV}-L_V$	0.79	1.45

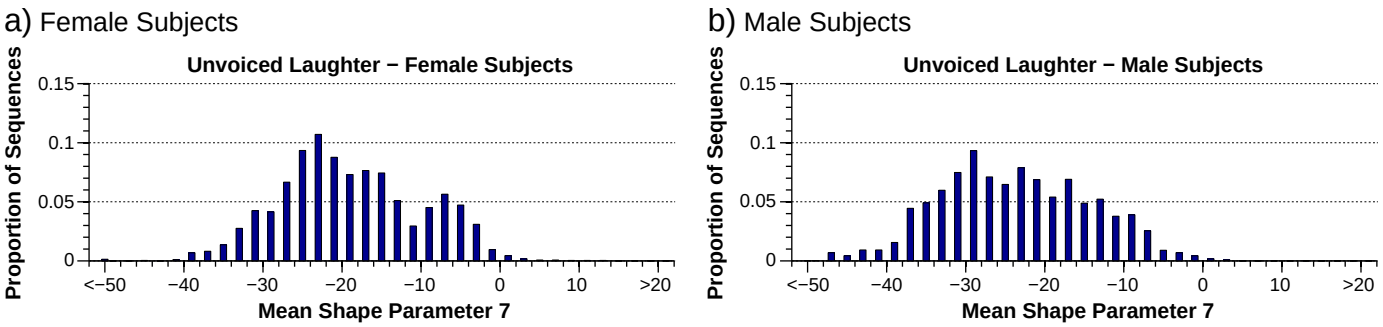

Fig. 14. Histogram of the first shape feature for female (a) and male (b) unvoiced laughter episodes.

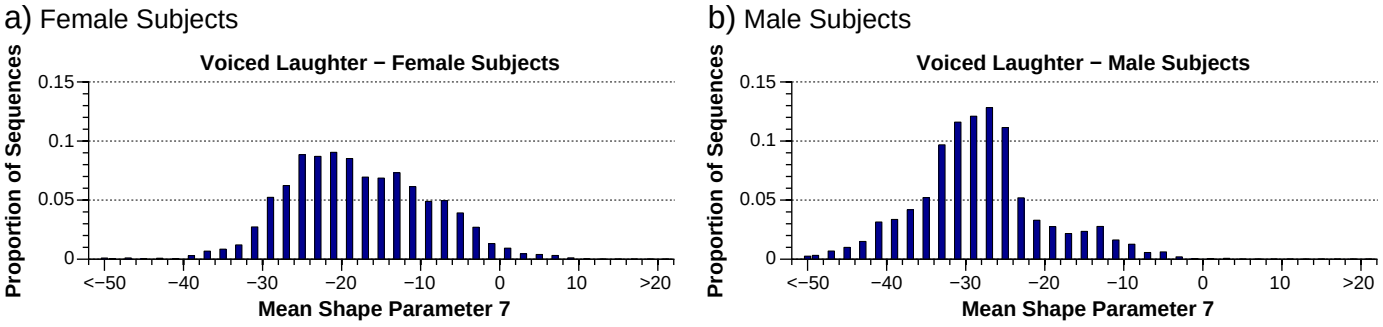

Fig. 15. Histogram of the first shape feature for female (a) and male (b) voiced laughter episodes.

shown in Table 7. It can be seen that similarly to the video-only classification speech can be much better recognized than laughter. It is also clear that the performance is much lower than the audio-only or video-only classification. However, it should be noted that both in audio and video processing there are several good performing features sets, whereas research in thermal videos processing is at an early stage and there are no standard feature sets. We simply provide a baseline performance in this study, but it could be possible to further improve the performance by using more sophisticated features, like features from difference thermal images.

6. Conclusions

In this paper we presented a new database focused on laughter, offering a lot of advantages with respect to previously existing databases. Firstly, a variety of modalities are included, namely, two sources of audio of different quality, a camera video stream and a thermal video stream. Secondly, the focus of the database on audiovisual recording of laughter means that the recording conditions facilitated that visual facial information is almost always available, and that the amount of laughter sessions and the number of subjects producing them is large enough. The database can be used both for audiovisual laughter-vs-speech classification and audiovisual laughter detection experiments using the speech sessions. Thirdly, we include laughter and speech recordings for almost all subjects, whereas speech and laughter from the same subject is usually not included in most laughter-oriented databases. Finally, the database is broadly annotated and highly accessible, since it is publicly available and searchable through the internet. Along with the presented database, we provide extensive baseline experiments for automatic discrimination of laughter and speech, and between voiced laughter, unvoiced laughter and speech. Experimental results are detailed in terms of video-based, audio-based and audiovisual discrimination, and also depending on gender, both for noisy and clean audio.

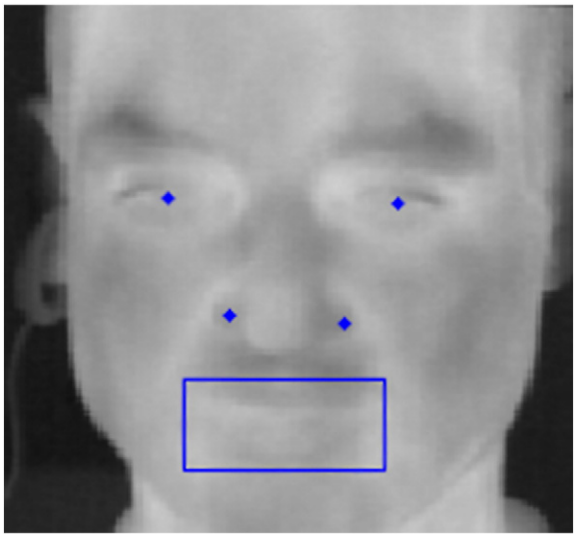

Fig. 16. Automatic detection of both eyes and both nostrils. The face is registered using these points, and a region of interest that encloses the mouth is defined over the registered face.

Acknowledgements

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 and B. Martinez is also supported in part by EPSRC grant EP/H016988/1: Pain rehabilitation: E/Motion-based automated coaching and EC's 7th Framework Programme [FP7/2007-2013] under the grant agreement no 231287 (SSPNet).

Table 7
Average error in discrimination between laughter and speech using thermal imagery (left), and corresponding confusion matrix (right).

F1-laugh	F1-sp	CR		Pred. laugh.	Pred. speech
46 (2.9)	74 (1.8)	65 (2.0)	Laugh.	49.9	50.1
			Speech	28.5	71.5

Appendix A

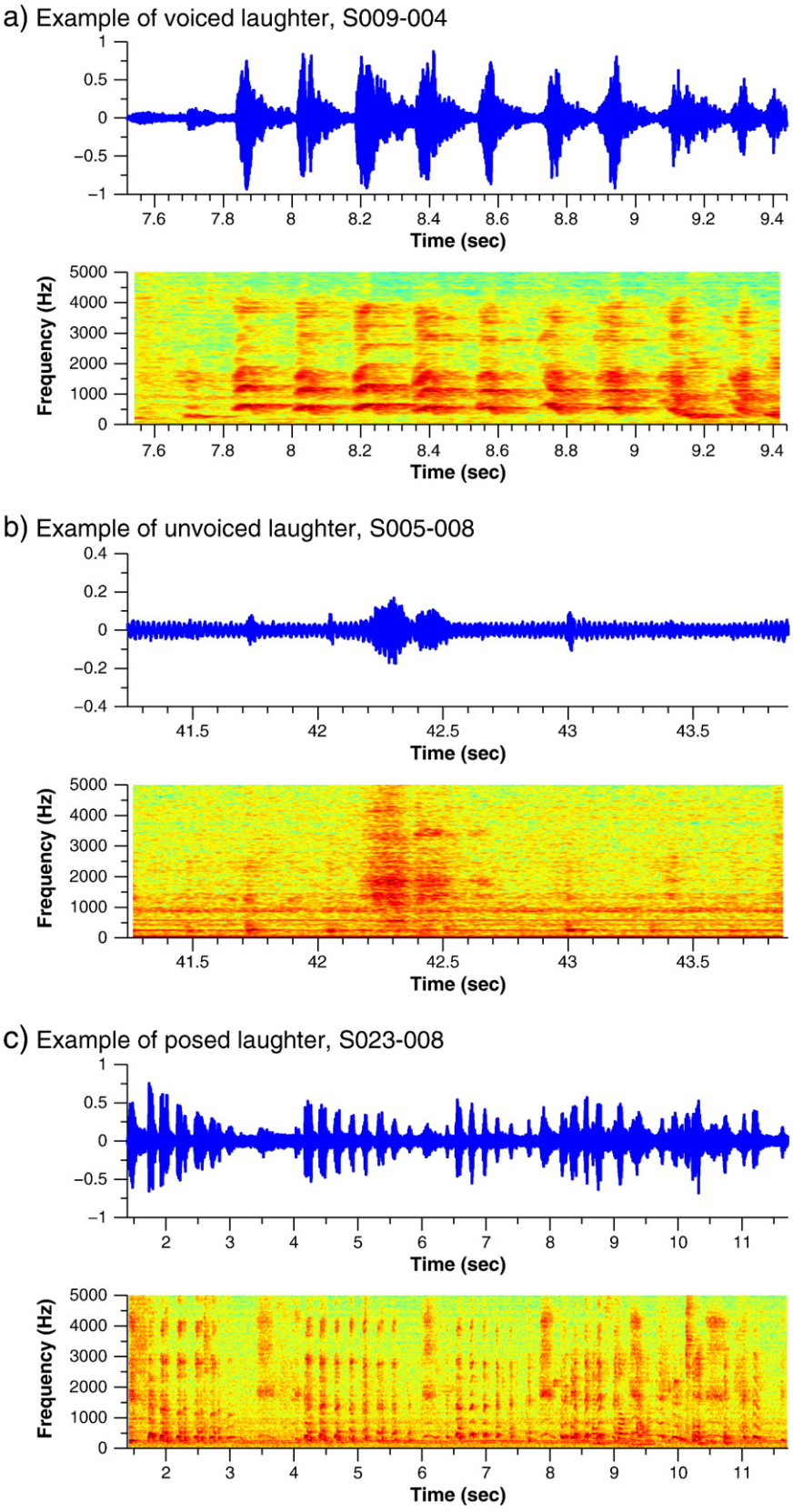

Fig. A.17. Top row: audio signal (camera microphone), bottom row: spectrogram.

Please cite this article as: S. Petridis, et al., The MAHNOB Laughter database, Image Vis. Comput. (2012), http://dx.doi.org/10.1016/j.imavis.2012.08.014

Table A.8
Confusion matrices for laughter-vs-speech discrimination.

	Actual laughter	Actual speech	Actual laughter	Actual speech	Actual laughter	Actual speech
	Video		Audio		Audiovisual	
Predicted laughter	69.0	6.4	66.6	0.9	80.1	3.4
Predicted speech	31.0	93.6	33.4	99.1	19.9	96.6

Table A.9
Confusion matrices for voiced laughter–unvoiced laughter-vs-speech discrimination.

	Actual			Actual			Actual		
	LV	*LU*	*S*	*LV*	*LU*	*S*	*LV*	*LU*	*S*
	Video			Audio			Audiovisual		
Predicted *LV*	57.3	30.5	2.7	39.9	15.6	0.2	66.9	24.5	2.0
Predicted *LU*	18.2	27.3	2.8	9.2	58.4	0.3	14.1	53.5	0.3
Predicted *S*	24.5	42.2	94.5	50.9	26.1	99.5	19.0	22.0	97.7

References

[1] C. Niemitz, Visuelle zeichen, sprache und gehirn in der evolution des menscheneine entgegnung auf McFarland (Visual signs, language and the brain in the evolution of humans — a reply to McFarland, Z. Sem. 12 (1990) 323–336.
[2] W. Ruch, P. Ekman, The expressive pattern of laughter, in: Emotions, Qualia, and Consciousness, 2001, pp. 426–443.
[3] A. Pentland, S. Pentland, Honest Signals: How They Shape Our World, The MIT Press, 2008.
[4] M.J. Owren, J.-A. Bachorowski, The evolution of emotional expression: a selfish-gene account of smiling and laughter in early hominids and humans, in: T.J. Mayne, G.A. Bonanno (Eds.), Emotion: Current Issues and Future Directions, Guilford, New York, 2001, pp. 152–191.
[5] R. Provine, Laughter: A Scientific Investigation, Viking, New York, 2000.
[6] 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, 1973, pp. 163–193.
[7] M. Makagon, E. Funayama, M. Owren, An acoustic analysis of laughter produced by congenitally deaf and normally hearing college students, J. Acoust. Soc. Am. 124 (2008) 472–483.
[8] S. Kipper, D. Todt, The role of rhythm and pitch in the evaluation of human laughter, J. Nonverbal Behav. 27 (4) (2003) 255–272.
[9] J.A. Russell, J.A. Bachorowski, J.M. Fernandez-Dols, Facial and vocal expressions of emotion, Annu. Rev. Psychol. 54 (2003) 329–349.
[10] M. Pantic, A. Pentland, A. Nijholt, T. Huang, Human computing and machine understanding of human behavior: a survey, Lect. Notes Comput. Sci. 4451 (2007) 47–71.
[11] K. Bousmalis, M. Mehu, M. Pantic, Spotting agreement and disagreement: a survey of nonverbal audiovisual cues and tools, in: IEEE Intl Conf. Affective Computing and Intelligent Interfaces, Vol. 2, 2009.
[12] D. Gatica-Perez, Automatic nonverbal analysis of social interaction in small groups: a review, Image Vision Comput. 27 (12) (2009) 1775–1787.
[13] M. Pantic, A. Vinciarelli, Implicit human-centered tagging, IEEE Signal Process. Mag. 26 (6) (2009) 173–180, http://dx.doi.org/10.1109/MSP.2009.934186.
[14] J. Bachorowski, M. Owren, Not all laughs are alike: voiced but not unvoiced laughter readily elicits positive affect, Psychol. Sci. 12 (3) (2001) 252–257.
[15] J. Vettin, D. Todt, Laughter in conversation: features of occurrence and acoustic structure, J. Nonverbal Behav. 28 (2) (2004) 93–115.
[16] J.A. Bachorowski, M.J. Smoski, M.J. Owren, The acoustic features of human laughter, J. Acoust. Soc. Am. 110 (1) (2001) 1581–1597.
[17] H. Rothg¨anger, G. Hauser, A. Cappellini, A. Guidotti, Analysis of laughter and speech sounds in Italian and German students, Naturwissenschaften 85 (8) (1998) 394–402.
[18] D. Szameitat, K. Alter, A. Szameitat, D. Wildgruber, A. Sterr, C. Dar- win, Acoustic profiles of distinct emotional expressions in laughter, J. Acoust. Soc. Am. 126 (2009) 354–366.
[19] K.P. Truong, D.A. van Leeuwen, Automatic discrimination between laughter and speech, Speech Commun. 49 (2) (2007) 144–158.
[20] B. Schueller, F. Eyben, G. Rigoll, Static and dynamic modelling for the recognition of non-verbal vocalisations in conversational speech, Lect. Notes Comput. Sci. 5078 (2008) 99–110.
[21] M. Knox, N. Morgan, N. Mirghafori, Getting the last laugh: automatic laughter segmentation in meetings, in: INTERSPEECH, 2008, pp. 797–800.
[22] K. Laskowski, T. Schultz, Detection of laughter-in-interaction in multi-channel close-talk microphone recordings of meetings, Lect. Notes Comput. Sci. 5237 (2008) 149–160.
[23] L. Kennedy, D. Ellis, Laughter detection in meetings, in: NIST Meeting Recognition Workshop, 2004.
[24] S. Petridis, M. Pantic, Audiovisual discrimination between speech and laughter: why and when visual information might help, IEEE Trans. Multimedia 13 (2) (2011) 216–234.
[25] J. Whitehill, G. Littlewort, I. Fasel, M. Bartlett, J. Movellan, Toward practical smile detection, IEEE Trans. Pattern Anal. Mach. Intell. 31 (11) (2009) 2106–2111.
[26] M.F. Valstar, H. Gunes, M. Pantic, How to distinguish posed from spontaneous smiles using geometric features, in: Proc. ACM Int'l Conf. Multimodal Interfaces, 2007, pp. 38–45.
[27] S. Petridis, A. Asghar, M. Pantic, Classifying laughter and speech using audio-visual feature prediction, in: IEEE Int'l Conf. on Acoustics Speech and Signal Processing, 2010, pp. 5254–5257.
[28] A. Ito, W. Xinyue, M. Suzuki, S. Makino, Smile and laughter recognition using speech processing and face recognition from conversation video, in: Intern. Conf. on Cyberworlds, 2005, 2005, pp. 8–15.
[29] B. Reuderink, M. Poel, K. Truong, R. Poppe, M. Pantic, Decision-level fusion for audio-visual laughter detection, Lect. Notes Comput. Sci. 5237 (2008) 137–148.
[30] S. Scherer, F. Schwenker, N. Campbell, G. Palm, Multimodal laughter detection in natural discourses, in: Human Centered Robot Systems, 2009, pp. 111–120.
[31] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, et al., The AMI meeting corpus: a pre-announcement, in: Machine Learning for Multimodal Interaction, 2006, pp. 28–39.
[32] N. Campbell, Tools and resources for visualising conversational-speech interaction, in: Multimodal Corpora, Lecture Notes in Computer Science, 5509, 2009, pp. 176–188.
[33] J. Urbain, E. Bevacqua, T. Dutoit, A. Moinet, R. Niewiadomski, C. Pelachaud, B. Picart, J. Tilmanne, J. Wagner, The AVLaughterCycle database, in: Int'l Conf. on Language Resources and Evaluation, 2010.
[34] M. Valstar, M. Pantic, Induced disgust, happiness and surprise: an addition to the MMI facial expression database, in: Proceedings of Int'l Conf. Language Resources and Evaluation, Workshop on EMOTION, 2010, pp. 65–70.
[35] P. Tsiamyrtzis, J. Dowdall, D. Shastri, I. Pavlidis, M.G. Frank, P. Ekman, Imaging facial physiology for the detection of deceit, Int. J. Comput. Vis. 71 (2) (2007) 197–214.
[36] S. Wang, Z. Liu, S. Lv, Y. Lv, G. Wu, P. Peng, F. Chen, X. Wang, A natural visible and infrared facial expression database for expression recognition and emotion inference, IEEE Trans. Multimedia 12 (7) (2010) 682–691.
[37] K. Grammer, I. Eibl-Eibesfeldt, The ritualisation of laughter, in: Die Natu¨rlichkeit der Sprache und der Kultur, Brockmeyer, Bochum, 1990, pp. 192–214.
[38] W. Hudenko, W. Stone, J. Bachorowski, Laughter differs in children with autism: an acoustic analysis of laughs produced by children with and without the disorder, J. Autism Dev. Disord. 39 (10) (2009) 1392–1400.
[39] S. Petridis, M. Pantic, Is this joke really funny? Judging the mirth by audiovisual laughter analysis, in: Proc. IEEE Intl Conf. Multimedia & Expo, 2009, pp. 1444–1447.
[40] B. Martinez, X. Binefa, M. Pantic, Facial component detection in thermal imagery, in: IEEE Int'l Conf. Computer Vision and Pattern Recognition Workshops, Vol. 3, 2010, pp. 48–54.
[41] B. Hern'andez, G. Olague, R. Hammoud, L. Trujillo, E. Romero, Visual learning of texture descriptors for facial expression recognition in thermal imagery, Comput. Vision Image Understanding 106 (2–3) (2007) 258–269.
[42] S. Petridis, Audiovisual discrimination between laughter and speech, Ph.D. thesis, Imperial College London (2011).
[43] http://mahnob-db.eu/laughter/.
[44] http://tcts.fpms.ac.be/~urbain/.
[45] http://www.mmifacedb.com/.
[46] J. Urbain, R. Niewiadomski, E. Bevacqua, T. Dutoit, A. Moinet, C. Pelachaud, B. Picart, J. Tilmanne, J. Wagner, AVLaughterCycle, J. Multimodal User Interfaces (2010) 1–12.
[47] http://www.speech-data.jp/corpora.html.
[48] http://corpus.amiproject.org/.
[49] S. Petridis, M. Pantic, J.F. Cohn, Prediction-based classification for audiovisual discrimination between laughter and speech, in: IEEE Int'l Conf. on Automatic Face and Gesture Recognition, 2011.

Please cite this article as: S. Petridis, et al., The MAHNOB Laughter database, Image Vis. Comput. (2012), http://dx.doi.org/10.1016/j.imavis.2012.08.014

[50] G. Mckeown, M.F. Valstar, R. Cowie, M. Pantic, M. Schroeder, *IEEE Transactions on Affective Computing*. 3 (1) (2012) pp. 5–17.
[51] B. Schueller, R. Mueller, F. Eyben, J. Gast, B. Hoernler, M. Woellmer, G. Rigoll, A. Hoethker, H. Konosu, Being bored? Recognising natural interest by extensive audiovisual integration for real-life application, Image Vision Comput. 27 (12) (2009) 1760–1774.
[52] J. Cohn, T. Kruez, I. Matthews, Y. Yang, M. Nguyen, M. Padilla, F. Zhou, F. De la Torre, Detecting depression from facial actions and vocal prosody, in: Intern. Conf. on ACII 2009, IEEE, 2009, pp. 1–7.
[53] http://semaine-db.eu/.
[54] F. Eyben, S. Petridis, B. Schueller, G. Tzimiropoulos, S. Zafiriou, M. Pantic, Audiovisual classification of vocal outburst in human conversation using long–short-term memory networks, in: IEEE Int'l Conf. on Acoustics Speech and Signal Processing, 2011, pp. 5844–5847.
[55] S. Sundaram, S. Narayanan, Automatic acoustic synthesis of human-like laughter, J. Acoust. Soc. Am. 121 (1) (2007) 527–535.
[56] D. Mowrer, L. LaPointe, J. Case, Analysis of five acoustic correlates of laughter, J. Nonverbal Behav. 11 (3) (1987) 191–199.
[57] J. Lichtenauer, J. Shen, M.F. Valstar, M. Pantic, Image and Vision Computing. 29 (2011) pp. 666–680.
[58] J. Trouvain, Segmenting phonetic units in laughter, in: Proc. Int'l Conf. Phonetic Sciences, 2003, pp. 2793–2796.
[59] Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands, http://www.lat-mpi.eu/tools/elan/
[60] H. Brugman, A. Russel, Annotating multimedia/multi-modal resources with ELAN, in: Int'l Conf. on Language Resources and Evaluation, 2004, pp. 2065–2068.
[61] K. Laskowski, Contrasting emotion-bearing laughter types in multiparticipant vocal activity detection for meetings, in: IEEE Int'l Conf. on Acoustics Speech and Signal Processing, 2009, pp. 4765–4768.
[62] A. Batliner, S. Steidl, F. Eyben, B. Schuller, On laughter and speech laugh, based on observations of child–robot interaction, The Phonetics of Laughing (2011).
[63] E. Nwokah, H. Hsu, P. Davies, A. Fogel, The integration of laughter and speech in vocal communication: a dynamic systems perspective, J. Speech Lang. Hear. Res. 42 (4) (1999) 880.
[64] P. Boersma, D. Weenink, Praat: doing phonetics by computer (version 4.3.01) (www.praat.org)), in: Tech. Rep., 2005.
[65] M. Knox, N. Mirghafori, Automatic laughter detection using neural networks, in: INTERSPEECH, 2007, pp. 2973–2976.
[66] C. Bickley, S. Hunnicutt, Acoustic analysis of laughter, in: Second Int'l Conf. on Spoken Language Processing, 1992.
[67] I. Patras, M. Pantic, Particle filtering with factorized likelihoods for tracking facial features, in: Int'l Conf. on Automatic Face and Gesture Recognition, 2004, pp. 97–104.
[68] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, The AMI meeting corpus, in: Int'l. Conf. on Methods and Techniques in Behavioral Research, 2005, pp. 137–140.
[69] D. Gonzalez-Jimenez, J.L. Alba-Castro, Toward pose-invariant 2-d face recognition through point distribution models and facial symmetry, IEEE Trans. Inf. Forensics Secur. 2 (3) (2007) 413–429.
[70] L. Rabiner, R. Schafer, Digital Processing of Speech Signals, Prentice Hall, 1978.
[71] M. Riedmiller, H. Braun, A direct adaptive method for faster backpropagation learning: the RPROP algorithm, in: Proc. IEEE Int'l Conf. on Neural Networks, Vol. 1, 1993, pp. 586–591.
[72] G. Potamianos, C. Neti, G. Gravier, A. Garg, A.W. Senior, Recent advances in the automatic recognition of audiovisual speech, Proc. IEEE 91 (9) (2003) 1306–1326.
[73] Y. Liu, N. Chawla, M. Harper, E. Shriberg, A. Stolcke, A study in machine learning from imbalanced data for sentence boundary detection in speech, Comput. Speech Lang. 20 (4) (2006) 468–494.
[74] R. Fisher, The Design of Experiments, Oliver & Boyd, Edinburgh and London, 1935.
[75] M. Smucker, J. Allan, B. Carterette, A comparison of statistical significance tests for information retrieval evaluation, in: ACM Conf. Information and Knowledge Management, 2007, pp. 623–632.
[76] S. Dupont, J. Luettin, Audiovisual speech modeling for continuous speech recognition, IEEE Trans. Multimedia 2 (3) (2000) 141–151.
[77] Z. Zeng, J. Tu, B. Pianfetti, T. Huang, Audio-visual affective expression recognition through multistream fused HMM, IEEE Trans. Multimedia 10 (4) (2008) 570–577.
[78] S. Petridis, M. Pantic, Audiovisual laughter detection based on temporal features, in: Proc. ACM Int'l Conf. on Multimodal interfaces, 2008, pp. 37–44.
[79] D. Johnson, S. Sinanovic, Symmetrizing the kullback–leibler distance, Technical Report, Rice University (2011), http://www.ece.rice.edu/~dhj/resistor.pdf.

Please cite this article as: S. Petridis, et al., The MAHNOB Laughter database, Image Vis. Comput. (2012), http://dx.doi.org/10.1016/j.imavis.2012.08.014