AUDIOVISUAL DETECTION OF LAUGHTER IN HUMAN-MACHINE INTERACTION

Stavros Petridis
Department of Computing
Imperial College London
sp104@imperial.ac.uk

Maelle Leveque
Department of Computing
Imperial College London
maelle.leveque11@ic.ac.uk

Maja Pantic
Imperial College London
EEMCS, University of Twente
m.pantic@imperial.ac.uk

***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. However, past research on laughter recognition has mainly focused on the information available in the audio channel only, mainly due to the lack of suitable audiovisual data. Only recently few works have been published which combine audio and visual information and most of them deal with the problem of discriminating laughter from speech or other nonlinguistic vocalisations using presegmented data. There are very few works on audiovisual laughter detection from unsegmented audiovisual streams and have either been tested on small datasets or use coarse visual features. As a consequence, results are mixed and it is not clear to what extent the addition of visual information to audio is beneficial for laughter detection. In this work, we attempt to overcome the limitation of previous studies and investigate the performance of audiovisual fusion for laughter detection using audiovisual continuous streams from the SEMAINE database. Our results suggest that there is indeed an improvement in laughter detection with the addition of visual information which is dependent on the performance of the voice activity detector.**

***Keywords*-Laughter Detection, Nonlinguistic Vocalisations, Audiovisual Fusion**

I. INTRODUCTION

Nonlinguistic vocalizations (or nonverbal vocalizations) are very brief, discrete, nonverbal expressions of affect in both face and voice [1]. Although information related to human emotions is conveyed by these vocalizations research on their automatic recognition is limited compared to facial expressions or speech recognition. One of the most important nonlinguistic vocalizations is laughter, which is the most frequently annotated acoustic nonverbal behaviour in meeting corpora.

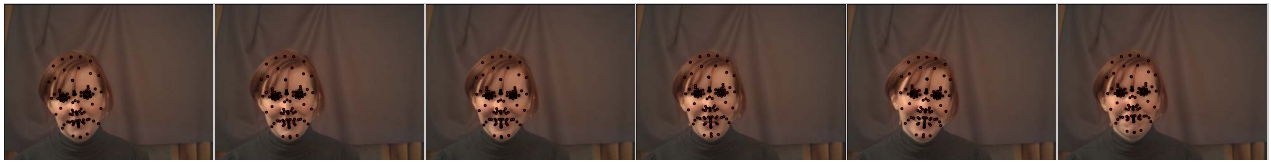
Therefore it is not surprising that previous work on nonlinguistic vocalisations has mainly focused on the detection/classification of laughter. The vast majority of these works fall into two categories: 1) audio-only laughter detection [2], [3], [4], [5], where the aim is to segment an audio stream into laughter and nonlaughter segments, and 2) laughter-versus-speech classification/discrimination, [6], [7], [8], [9], where the aim is to correctly classify presegmented episodes of laughter and speech or different types of laughter. There have been also a few attempts to discriminate between different non-linguistic vocalisations, like laughter, hesitation and consent [10] based on audio only.

The common characteristic of all these works is that they are based on audio information only, i.e., information carried by the facial expressions has been ignored. It has been shown that speech and laughter become more intelligible, especially under noisy conditions, for humans when visual information is present [11], [12]. This finding has inspired research on audiovisual speech and emotion recognition. However, the lack of suitable audiovisual data has prevented research on audiovisual laughter recognition for a long time.

Only recently, few works on audiovisual discrimination between speech and laughter [13], [14], [15], [16], [17] or between laughter, hesitation and consent [18] based on presegmented episodes have been published. However, work on audiovisual detection of laughter or other nonlinguistic vocalisations using continuous audiovisual streams is limited. To the best of our knowledge, there are only three works in this area. The first one is by Ito et al. [19] who used basic geometric and appearance features like lip lengths and cheek mean intensities in combination with MFCCs to detect laughter. However the dataset is very small, just 3 short dialogues (4-8 min each) were used, so the results should be treated as preliminary. In a more recent work Escalera et al. [20] used features based on mouth movements together with pitch and spectral entropy to detect laughter in 18 videos of dyadic interaction (4 min each). This is a larger dataset but the results are not conclusive whether the combination of audio and visual information is beneficial. In the last work [21], a much larger dataset has been used, 2 videos of 90 min each containing 4 subjects, but the visual features are coarse, based only on head and body movements. The results on laughter recognition are rather inconclusive whether the addition of visual information improves the performance or not. Probably this happens due to the coarse visual features that had to be used due to the nature of the data.

In this work we attempt to overcome some of the limitations of the previous work by using a large dataset from the SEMAINE database, 42 sessions with a total duration of 188

978-0-7695-5048-0/13 $26.00 © 2013 IEEE
DOI 10.1109/ACII.2013.28



(a) Frame 150 (b) Frame 200 (c) Frame 250 (d) Frame 350 (e) Frame 450 (f) Frame 550

Figure 1: Example of Laughter from the SEMAINE database, session 35. The 113 tracked points can also be seen.

Table I: Training, Validation and Test Sets

	Training Set	**Validation Set**	**Test Set**
Sessions	19-22, 29-31, 40-43, 70-73, 76, 79	46-49, 77, 78, 82-85	34-37, 60, 94, 95, 106-109, 112-115
Total Duration	77.3 min	60.2 min	51.3 min
No Speech Instances	2972	2174	2081
No Laughter Instances	45	77	54

min, and more refined visual features based on 113 facial points. Our aim is to perform audiovisual laughter detection using unsegmented audiovisual streams and investigate if the addition of visual information helps. We use 10 subjects and contrary to previous works only one subject appears both in the training and test set. So our experiments are close to but not completely subject independent. We combine audio and visual features on decision level and we show that audiovisual fusion leads to improved performance, which however depends on the performance of the voice activity detector.

II. Database

The SEMAINE database [22] is used for the experiments in this study since it contains audiovisual recordings of several subjects which make it suitable for detection of non-linguistic vocalisations.

Each subject interacts with 4 agents, which have different personalities, for approximately 5 min. The agents are played by human operators who have become thoroughly familiar with the agent's personality. The aim is to evoke emotionally coloured reactions from the users whose reactions are recorded by a camera with resolution 780 x 580 at 50 frames per second (fps) and a headset microphone at 48 kHz. It should be noted that although each user has his own microphone the voice of the agent is also recorded by it. So even when the user is silent some background speech is present.

The recordings have been annotated in terms of the six basic emotions and continuous emotional dimensions. Recently word level annotations including a few non-linguistic vocalisations like laughter, breath and sigh have been added to the database. An example is shown in Fig. 1.

III. Features

A. Audio Features

Cepstral features, such as MFCCs, have been widely used in speech recognition and have also been successfully used for laughter detection [2], [3] and laughter-vs-speech discrimination [13], [15]. Therefore, we use 13 MFCCs in this study as well, which are computed every 10ms over a window of 40ms, i.e. the frame rate is 100 fps. The MFCCs are augmented with the addition of the ΔMFCCs, which capture some local temporal characteristics, and this leads to an audio feature vector with 26 dimensions. The MFCCs and ΔMFCCs are computed using the functions provided in [23].

B. Visual Features

Changes in facial expression are captured by tracking 113 facial points using the facial point tracker described in [24]. This is a 3D tracker so the output is automatically head posed normalised, i.e., rotation, translation and scale of the face have been removed. The features we use are facial animation parameters (FAPS) which are automatically computed by the tracker. FAPS are defined in the ISO MPEG-4 standard.

IV. Experimental Protocol

A. Data Preparation

We use the subset of publicly available sessions which contain word level annotations. Sessions that cannot be tracked because parts of the face are not visible have been excluded. The remaining 42 sessions, which have a total duration of over 3 hours, are divided into three sets, training, validation and test sets as shown in Table I. Only one subject appears in all three sets and this is because it has been recorded in two different sessions (34 to 37 and 76 to 79). All other subjects appear only in one set in an effort to make the experiments as subject independent as possible. Finally,

breath and sighs have been excluded due to the very small number of instances available.

Normalisation: All the audio and visual features on the training set are normalised to a zero mean and unity standard deviation. The features on the validation and test set are normalised by subtracting the mean μ_{tr} and dividing by the standard deviation σ_{tr} which are computed on the training set only. It should be noted that the values for μ_{tr}, σ_{tr} are computed only for the parts of the training set that the subjects vocalises, i.e. either speaks or laughs.

Synhronisation: As described in section III the audio and visual features are extracted at different frame rates, 100 and 50 fps, respectively. In order to synhronise them, we upsample the visual features by linear interpolation as in [25].

B. Training

Based on the annotations provided we segment the training sessions into sequences that contain only one class, i.e., speech or laughter sequences. Each sequence is used as a training example to train time delay neural networks (TDNNs) with one hidden layer and two outputs, one for laughter and one for speech. The resilient backpropagation algorithm [26] is used to train the networks for 1000 epochs. We should point out that other learning algorithms like Hidden Markov Models (HMMs) could have been used but the goal of this work is to investigate if the combination of audio and visual information is beneficial for laughter detection. As shown in [21], the performance of HMMs and dynamic neural networks is comparable for laughter detection.

Two networks are trained, one using the audio features and one using the visual features and fusion is performed on the decision level. We also tried feature-level fusion but it performed slightly worse on the validation set, so we only report results for decision-level fusion in this study.

The main problem is that the vast majority of training examples belong to the speech class, given the nature of the interaction between the user and the agent. It is well known that highly imbalanced sets can degrade the performance of classifiers, therefore we randomly downsample the speech examples used for training to 100. This means that we throw away a significant number of training speech examples, so we repeat training 10 times. Each time we randomly select a different subset of size 100 for speech, and we report the mean and standard deviation of recall, precision and F1 measure, which are the performance measures used in this study. We should emphasise that we report the performance per frame and not per laughter/speech instance.

C. Parameter Optimisation

For each network the number of hidden neurons and the inputs delays need to be optimised on the validation set. The optimal number of hidden neurons found for the audio and visual networks are 30 and 5, respectively, and the optimal number of input delays for both networks is 4. In addition, the decision fusion weights need to be optimised on the validation set for each class. The optimal audio weights for the two classes (speech, laughter) are the following: 0.9 and 0.4. Therefore the visual weights are 0.1 and 0.6.

D. Voice Activity Detection

In order to deal with the silent segments a voice activity detector (VAD) is needed. It detects the non-silent segments where the audio-only, visual-only and audiovisual detectors are applied. The segments identified as silence are not used further. In this study we use the VAD described in [27]. We also we use an ideal VAD, which is based simply on the silence annotations provided. This is definitely not a realistic scenario, since we make the assumption that we have a perfect detector that does not make any mistakes. However, we would like to investigate the influence of the voice activity detector in segmentation. This is important, since as described in section II, crosstalk is present and although the subject may be silent the voice of the agent can be clearly heard.

V. Results

The performance of the audiovisual segmentation is measured on the 15 test sessions. Initially, a VAD is applied to identify the segments where the subject vocalises. Then, the trained TDNNs are fed with the audio/visual features of the voice activity segments and they label each frame as speech or laughter.

Results when the ideal VAD is used are shown in Table II. It is obvious that the combination of audio and visual information is beneficial for speech and laughter. The recall rate for speech increases by 3.1%, whereas the precision rate remains almost the same and this leads to an increase in the F1 measure of 1.6%. Similarly, the F1 measure for laughter increases by 8.4 %, due to a 7.9% increase in the precision rate despite a decrease in the recall rate.

Results when the non-ideal VAD is used are shown in Table III. The benefits of combining visual and audio features follow the same pattern as above. The recall rate for speech goes up by 2.6% and precision remains almost the same and this results in a 1.1% increase in the F1 rate for speech. The recall rate for laughter goes down by 3.4% but at the same time the precision rate goes up by 2.7% which leads to an increase in the F1 rate for laughter of 3.5%.

It should also be noted that silence achieves a high precision rate of 90.6% but a much lower recall rate of 66.5%. In other words, this means that those segments labelled as silence are indeed silence most of the time, however there are several silent segments mislabeled as non-silence. This is expected due to the voice of the agent being audible when the subject does not vocalise, which is recognised as voice activity by the detector. As a consequence, the precision rate

Table II: Segmentation performance of the audio, video and audiovisual classifiers on the test sessions, using an ideal voice activity detector. The results presented are the mean and (st. dev.) of the 10 experiments conducted. DF: Decision Fusion, R: Recall, PR: Precision, F1: F1 measure.

	R	PR	F1
Silence	100.0 (0.0)	100.0 (0.0)	100.0 (0.0)
Audio			
Speech	92.7 (1.6)	99.3 (0.0)	95.9 (0.8)
Laughter	46.1 (2.9)	10.6 (2.4)	17.1 (3.2)
Video			
Speech	81.8 (6.9)	98.9 (0.1)	89.4 (4.2)
Laughter	31.3 (9.0)	5.9 (1.2)	9.8 (1.6)
Audiovisual (DF)			
Speech	95.8 (1.2)	99.2 (0.0)	97.5 (0.6)
Laughter	43.2 (3.3)	18.5 (4.4)	25.5 (4.2)

Table III: Segmentation performance of the audio, video and audiovisual classifiers on the test sessions, using the voice activity detector from [27]. The results presented are the mean and (st. dev.) of the 10 experiments conducted. DF: Decision Fusion, R: Recall, PR: Precision, F1: F1 measure.

	R	PR	F1
Silence	66.5 (0.0)	90.6 (0.0)	76.7 (0.0)
Audio			
Speech	87.6 (1.4)	76.6 (0.3)	81.7 (0.7)
Laughter	45.3 (2.3)	7.6 (1.5)	12.9 (2.2)
Video			
Speech	76.6 (6.6)	74.9 (1.5)	75.6 (3.7)
Laughter	29.2 (8.5)	3.7 (0.7)	6.5 (1.0)
Audiovisual (DF)			
Speech	90.2 (1.1)	76.5 (0.3)	82.8 (0.5)
Laughter	41.9 (2.9)	10.3 (1.7)	16.4 (2.2)

for all classes decreases since several new segments, which correspond to silence, are misclassified as speech or laughter.

It is also interesting to point out that the improvement in speech detection is the result of increased recall, whereas the improvement in laughter detection is the result of increased precision. This means that visual information helps in recognising speech frames, which had been labelled as laughter by the audio-only classifier, more accurately. As a consequence, the recall rate of speech increases and the precision rate of laughter increases as well, as fewer speech frames are confused with laughter.

We also see that the precision rate for laughter is low. This is a consequence of the highly imbalanced test set. Even though a small fraction of the speech examples is misclassified as laughter, e.g., 4.2% when an audiovisual detector is used with an ideal VAD, the actual number of those misclassified examples is higher than the total number of laughter examples.

An example of how the audiovisual detector works can be seen in Fig. 2. The beginning of laughter is misclassified as speech but the rest is correctly detected. We also notice that there is a short false laughter and silence detection after the end of the laughter episode. In addition, it is obvious that the last silent segment is detected with a delay.

Overall, we see that audiovisual fusion improves the recognition of speech and laughter, but it is affected by the performance of the voice activity detector since it is more pronounced in the case of the ideal voice activity detector. We should also point out that the performance of laughter detection is far from the performance of laughter-vs-speech classification based on presegmented episodes where F1 measures close to 90% are achieved [13], [15]. It is also difficult to compare with other approaches, given the completely different datasets and features used. For example, an F1 measure of 63% is reported in [21] using Echo State Neural Networks, but a laughter is considered as detected as long as only a few frames are detected within the episode. On the other hand, here we report the frame accuracy, so by applying the same rule and then reporting performance per episode will definitely improve the results. In an audio-only study [3] where accuracy is reported per frame an F1 measure of less than 20% is reported on unseen subjects which is comparable to the performance reported here.

VI. Conclusions

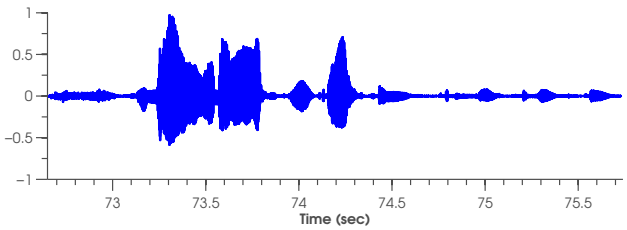
We presented an audiovisual approach for laughter detection. We combined audio and visual information on the decision level and showed that this combination leads to improved performance over the audio-only detection. The laughter detection performance on unsegmented streams is significantly lower than laughter-vs-speech discrimination using presegmented episodes which means there is a lot of room for improvement. The need for more audiovisual data is also evident, given the relatively low number of laughter episodes present in the SEMAINE database, which is expected to improve further the performance.

Acknowledgment

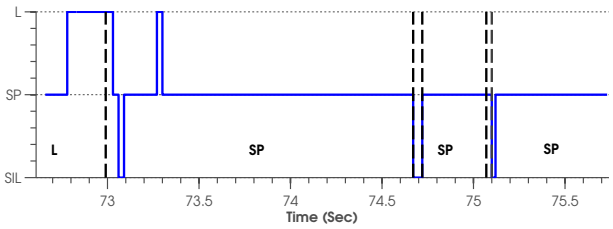
This work has been supported by the European Community's 7th Framework Programme [FP7/20072013] under grant agreement no. 231287 (SSPNet). The work of Maja Pantic was supported by the European Research Council under the ERC Starting Grant agreement no. ERC-2007-StG-203143 (MAHNOB).

References

[1] K. Scherer, "Affect bursts," in *Emotions: Essays on emotion theory*, S. van Goozen, N. van de Poll, and J. Sergeant, Eds., pp. 161–193. 1994.



(a) Audio Signal



(b) Detection Result

Figure 2: Example of how the audiovisual detector works using the non-ideal VAD on a segment from session 35. The black dashed lines denote the boundaries of the ground truth labels. The two short segments at the end of the segment correspond to silence. The blue solid line corresponds to the audiovisual detector's output where SIL, SP, L correspond to silence, speech and laughter, respectively.

[2] L. Kennedy and D. Ellis, "Laughter detection in meetings," in *NIST Meeting Recognition Workshop*, 2004.

[3] 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.

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

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

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

[7] A. Lockerd and F. Mueller, "Lafcam: Leveraging affective feedback camcorder," in *CHI, Human factors in computing systems*, 2002, pp. 574–575.

[8] N. Campbell, H. Kashioka, and R. Ohara, "No laughing matter," in *Europ. Conf. on Speech Comm. and Technology*, 2005, pp. 465–468.

[9] A. Batliner, S. Steidl, F. Eyben, and B. Schuller, "On laughter and speech laugh, based on observations of child-robot interaction," *The Phonetics of Laughing*, 2010.

[10] B. Schuller, 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.

[11] WH Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *Journal of the Acoustical Society of America*, vol. 26, no. 2, pp. 212–215, 1954.

[12] T.R. Jordan and L. Abedipour, "The importance of laughing in your face: Influences of visual laughter on auditory laughter perception," *Perception*, vol. 39, no. 9, pp. 1283–1285, 2010.

[13] S. Petridis and M. Pantic, "Audiovisual discrimination between speech and laughter: Why and when visual information might help," *IEEE Trans. on Multimedia*, vol. 13, no. 2, pp. 216 –234, April 2011.

[14] 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.

[15] S. Petridis, B. Martinez, and M. Pantic, "The MAHNOB laughter database," *Image and Vision Computing Journal*, vol. 31, no. 2, pp. 186–202, February 2013.

[16] S. Petridis, M. Pantic, and Cohn. J. F., "Prediction-based classification for audiovisual discrimination between laughter and speech," in *IEEE FG*, 2011, pp. 619–626.

[17] S. Petridis, S. Bilakhia, and M. Pantic, "Comparison of prediction-based fusion and feature-level fusion across different learning models," in *ACM Multimedia 2012*, Nara, Japan, November 2012, pp. 813–816.

[18] F. Eyben, S. Petridis, B. Schuller, and M. Pantic, "Audiovisual vocal outburst classification in noisy acoustic conditions," in *IEEE ICASSP*, Kyoto, Japan, March 2012, pp. 5097–5100.

[19] A. Ito, W. Xinyue, M. Suzuki, and S. Makino, "Smile and laughter recognition using speech processing and face recognition from conversation video," in *Intern. Conf. on Cyberworlds*, 2005, pp. 8–15.

[20] S. Escalera, E. Puertas, P. Radeva, and O. Pujol, "Multi-modal laughter recognition in video conversations," in *IEEE CVPR Workshops 2009*, pp. 110–115.

[21] Stefan Scherer, Michael Glodek, Friedhelm Schwenker, Nick Campbell, and Günther Palm, "Spotting laughter in natural multiparty conversations: A comparison of automatic online and offline approaches using audiovisual data," *ACM Trans. Interact. Intell. Syst.*, vol. 2, no. 1, pp. 4:1–4:31, Mar. 2012.

[22] G. Mckeown, M. F. Valstar, R. Cowie, M. Pantic, and M. Schroeder, "The semaine database: Annotated multimodal records of emotionally coloured conversations between a person and a limited agent," *IEEE Transactions on Affective Computing*, vol. 3, pp. 5–17, April 2012, Issue 1.

[23] Daniel P. W. Ellis, "PLP and RASTA (and MFCC, and inversion) in Matlab," 2005, http://www.ee.columbia.edu/ dpwe/resources/matlab/rastamat.

[24] F.J. Orozco, F.A. García, L. Arcos, and J. Gonzàlez, "Spatio-temporal reasoning for reliable facial expression interpretation," in *International Conference on Computer Vision Systems (ICVS)*. Bielefeld University, 2007.

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

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

[27] J. Sohn and W. Sung, "A voice activity detector employing soft decision based noise spectrum adaptation," in *IEEE ICASSP*, 1998, vol. 1, pp. 365–368.