# IS THIS JOKE REALLY FUNNY? JUDGING THE MIRTH BY AUDIOVISUAL LAUGHTER ANALYSIS

*S. Petridis*

sp104@doc.ic.ac.uk
Department of Computing
Imperial College London
London, UK

*M. Pantic*

m.pantic@imperial.ac.uk
Department of Computing
Imperial College London, UK
EEMCS, Univ. Twente, NL

## ABSTRACT

This paper presents the results of an empirical study suggesting that, while laughter is a very good indicator of amusement, the kind of laughter (unvoiced laughter vs.voiced laughter) is correlated with the mirth of laughter and could potential be used to judge the actual hilarity of the stimulus joke. For this study, an automated method for audiovisual analysis of laugher episodes exhibited while watching movie clips or observing the behaviour of a conversational agent has been developed. The audio and visual features, based on spectral properties of the acoustic signal and facial expressions respectively, have been integrated using feature level fusion, resulting in a multimodal approach to distinguishing voiced laughter from unvoiced laughter and speech. The classification accuracy of such a system tested on spontaneous laughter episodes is 74 %. Finally, preliminary results are presented which provide evidence that unvoiced laughter can be interpreted as less gleeful than voiced laughter and consequently the detection of those two types of laughter can be used to label multimedia content as little funny or very funny respectively.

***Index Terms***— Implicit content based indexing,audiovisual laughter detection

## 1. INTRODUCTION

Over the last decade the amount of multimedia data produced has dramatically increased, which requires the development of automatic methods for indexing and retrieval in order to benefit from it. Automatic content-based multimedia indexing has attracted much research interest recently with the goal of automatically labelling objects, scenes, and events in multimedia data [1]. A common approach is to use large amounts of annotated data and then to use machine learning methods, based on low-level audio and visual features, to learn the characteristics of each label in order to apply the trained systems on new unlabelled data. A different approach that has been recently proposed is implicit content based indexing [2]. In this approach the user's reaction is monitored and used to tag or to judge the accuracy of the provided tags of the multimedia content that is being watched. It is supposed that if the user shows agreement or confusion then the assigned tag can be considered as accurate or misleading respectively. Alternatively, the user's behavior itself while watching multimedia content can be used to assign new tags to the content. For example, laughing, crying, and disgust displays can be used as indicators of funny, sad and disgusting scenes respectively. Only very recently few related works have been presented which investigate the role of emotions in information seeking [3] and rank movie scenes based on affect-related physiological signals [4]. However, indexing of multimedia content based on the actual user's behavior, i.e., facial expressions and vocalisations, has not been attempted yet .

Within this framework of implicit tagging, we focus on laughter which is one of the most common and useful human social signals. It helps humans to express their emotions and intentions in social interactions and also provides useful feedback during human-human interaction. It is usually perceived as positive feedback, i.e., it shows joy, acceptance, agreement, but it can also be perceived as negative feedback, e.g., irony. We present results of a preliminary study in which users are asked to rate the mirth of a video clip by watching their own reaction (laughter) displayed while they watched that funny video clip. Initial results suggest that voiced laughters, i.e., laughters consisting mainly of voiced sounds (e.g. ha-ha-ha), are perceived as indicators of highly amusing scenes whereas unvoiced laughs (e.g. snorts) are perceived as indicators of less amusing scenes. We apply an adapted version of the audiovisual laughter detector proposed in [5] to discriminate between the two types of laughter and speech. The detector has been built using a dataset consisting of laughters occured in social interactions and laughters elicited by humourous stimuli, and it achieves a classification rate of 74 %.

## 2. DATASET

For the purposes of training our laugther detector we used two datasets: the AMI meeting corpus and our own dataset. The

AMI Meeting Corpus [6] consists of 100 hours of meetings recordings where people show a huge variety of spontaneous expressions. We only used the close-up video recordings of the subject's face (720 x 576 pixels, 25 frames per second) and the related individual headset audio recordings (16 kHz). The language used in the meetings is English and the speakers are mostly non-native speakers. For our experiments we used seven meetings and the relevant recordings of 10 participants (8 young males and 2 young females) with or without glasses and no facial hair. The laughter episodes contained in this corpus are the result of social interaction. Our own dataset contains laughter episodes elicited by humourous stimuli. We recorded 7 subjects while watching short funny video clips. We annotated and extracted the laughter episodes of 2 subjects, one male and one female and added them to the laughter episodes from the AMI meeting corpus. The combined dataset has been used to train our laughter detector.

All laughter and speech segments were pre-segmented based on audio. We only kept those segments that do not co-occur with speech, do not contain profile views of the face, are longer than 450ms, and satisfy the criterion as suggested in [7]: "Laughter is defined as being any perceptibly audible that an ordinary person would characterize as a laugh if heard under everyday circumstances". The laughter episodes were further divided into 2 groups: voiced and non-voiced laughter. This was done following the same procedure as in [7], i.e., those who contained primarily voiced sounds were labelled as voiced, and those which contained primarily unvoiced sounds were labelled as unvoiced. For speech segments we selected those that do not contain long pauses between two consecutive words. In total, we used 82 audio-visual voiced laughter segments, with a total duration of 118.23 seconds, 51 unvoiced laughter segments with a total duration of 62.50 seconds, and 109 audio-visual speech segments with a total duration of 200 seconds.

## 3. FEATURE EXTRACTION

In [8], it was shown that the most informative cues for discriminating laughter (this includes both voiced and unvoiced) from speech were cepstral features, pitch and energy, and features derived from facial expressions. The same features are used in this study as well.

### 3.1. Spectral Features

Spectral or cepstral features, such as Mel Frequency Cepstral Coefficients (MFCCs), have been widely used in speech recognition. We use 6 MFCCs since it has been reported [9] that these achieve the same performance in laughter detection application as when using 13 MFCCs, which are commonly used in speech recognition applications. In addition to the 6 MFCCs, their delta features were calculated, in order to capture some local temporal characteristics. So in total 12 features are computed per audio frame, where the duration of each frame is 40 ms and the overlap between successive frames is 20ms. Since not much information is carried by a single frame, it is beneficial to compute features over longer temporal windows as proposed in [5]. In order to do that, we compute the mean and standard deviation of each MFCC and $\Delta$MFCC over a 320ms temporal window. We use simple statistical features, like mean and standard deviation, since they were found to achieve very good performance [5]. Using this approach the information of the temporal window is encoded in $2 * 12 = 24$ features.

### 3.2. Prosodic features

The two most commonly used prosodic features in studies of emotion detection are pitch and energy [10]. Pitch is the perceived fundamental frequency of a sound. While the actual fundamental frequency can be precisely determined through physical measurement, it may differ from the perceived pitch. Bachorowski et al. [7] found that the mean pitch in both male and female laughter was higher than in modal speech. Energy of a signal is simply the sum of squares of the signal's raw values. We compute pitch and energy every 20ms over a window of 40ms. The energy features used are the mean and standard deviation over a temporal window of length 320ms. The pitch features used in this study are the mean and standard deviation of the voiced frames over a window of 320ms. In addition, the unvoiced ratio is computed, i.e., the proportion of unvoiced frames in the same window of 320ms. When pitch can not be estimated then a zero value is assigned.

### 3.3. Visual Features

To capture the facial expression dynamics, we track 20 facial points, as shown in Fig. 1, in the video segments. These points are the corners / extremities of the eyebrows (2 points), the eyes (4 points), the nose (3 points), the mouth (4 points) and the chin (1 point). To track these facial points we used the Patras - Pantic particle filtering tracking scheme [11]. The points were manually annotated in the first frame of an input video and tracked for the rest of the sequence. Hence, for each video segment containing K frames, we obtain a set of K vectors containing 2D coordinates of the 20 points.

The next step is the decoupling of head movements from facial expressions. To do so we use a similar approach to that proposed in [12] in which Principal Component Analysis (PCA) is used for decoupling, skipping the alignment of the shapes in order to capture the head movement as well. This approach has also been used in [13],[8],[5].

First, we concatenate the (x, y) coordinates of the 20 tracking points in a 40-dimensional vector. Then we use PCA to extract 40 principal components (PCs) for all the frames in the dataset. PCA is defined as an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance of the data comes to lie on the 1st PC, the 2nd greatest variance on the 2nd PC, and so on. Given that in our dataset head movements account for most

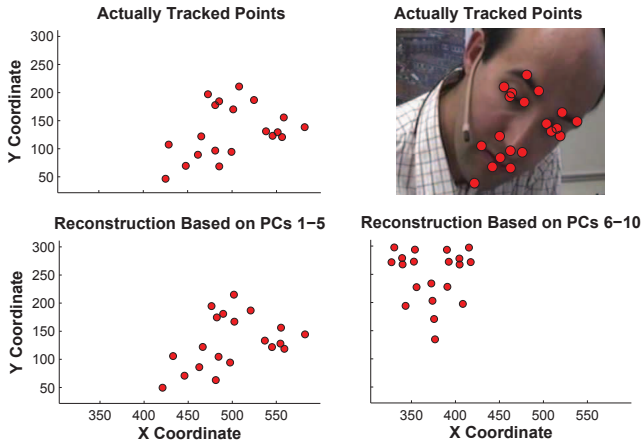


**Fig. 1**. PCA analysis of facial point tracking. Upper row: actually tracked facial points. Bottom row: (left) 20 facial points after they have been reconstructed using the first 5 principal components, (right) 20 facial points after they have been reconstructed using principal components 6 to 10.

of the variation in the data, lower-order PCs are expected to reflect rigid-movement aspects of the data while higher-order PCs are expected to retain non-rigid-movement (facial expression) aspects of the data. As can be seen from Fig. 1, it seems that indeed the lower-order PCs reflect rigid-movement aspects of the data, while the higher-order PCs reflect facial expression aspects of the data. In this study we found that head movements and facial expressions are encoded in PCs 1 to 5 and 6 to 10 respectively. So the visual features used are the projections of the coordinates of an input frame to PCs 6 to 10 as explained in [5].

## 4. AUTOMATIC AUDIOVISUAL DISCRIMINATION BETWEEN VOICED AND UNVOICED LAUGHTER AND SPEECH

In order to investigate if the automatic discrimination between the two types of laughter and speech is possible we use a slightly modified version of the audiovisual system described in [5] with the features described in section 3. The extracted audio and visual features are concatenated, i.e. fusion is performed at feature level, and fed to a neural network classifier. To test the detector, we performed leave-one-subject-out cross validation, using in every validation fold all samples of one subject as test data and all other samples as training data. Then the results obtained in each fold are averaged in order to get the final results. In this way the obtained results are subject independent. In addition, since in each fold the proportion of examples for each class can vary significantly, which can affect the classifier's performance, the training set is balanced by randomly selecting almost equal number of examples for each class prior to training. In each cross validation fold, all features used for training are z-normalized to a mean

| Type of Detector | F1 | Classification Rate |
|---|---|---|
| Laughter - Speech | 87.9 | 88.2 |
| Unvoiced Laughter - | 66.6 | |
| Voiced Laughter - | 70.4 | 74.7 |
| Speech | 87.8 | |

**Table 1**. Performance of the audiovisual laughter detector for the 2-class and 3-class problem

$\mu = 0$ and standard deviation $\sigma = 1$. Then, the obtained $\mu$ and $\sigma$ are used to z-normalize the features in the test set. The performance measures used are the classification rate and F1 measure which is a weighted combination of recall and precision.

The performance of our audiovisual detector is presented in Table 1. The second row presents the results when both types of laughter are merged into one class, and the third row presents the results for the 3-class problem. In the latter case, the F1 measure is reported per class. As can be seen from Table 1 it is a relatively easy problem to discriminate laughter from speech achieving high accuracy, whereas it is much harder to discriminate the three classes although the accuracy is still relatively high, 74.72%.

## 5. USER STUDY

The aim of this preliminary study is to investigate how the two different types of laughter (voiced / unvoced) are related to the video content which is presented to the user. Bachorowski and Owren [14] have shown that voiced laughter always elicited more positive evaluations than unvoiced laughter in social interactions. Driven by that result we wanted to investigate whether voiced and unvoiced laughters produced while watching funny video clips correlate with the perceived mirth of the clip.

In order to perform this study we used the recordings of 7 subjects, 4 males and 3 females, as described in section 2. We extracted 4 to 5 laugh segments per subject, containing both voiced and unvoiced laughters, and we asked the subjects to rate the mirth of video clips based on a scale from 1 to 3. The rating was the answer to the following question "Based on your reaction what do you think you were watching? A scene that was a little funny (1), just funny (2), or very funny (3)?". The results are shown in Fig. 2. Each group of bars shows the percentage of voiced and unvoiced laughters that were assigned to this category by the subjects. From Fig. 2 we can see that the vast majority of laughters assigned to the "little funny" category were unvoiced. The results for the second category are more balanced with 43 % of laughters being voiced and 57 % unvoiced. Finally, for the "very funny" category 63 % of the laughters were voiced and 37 % unvoiced. These preliminary results provide evidence that unvoiced laughters are perceived by human observers as

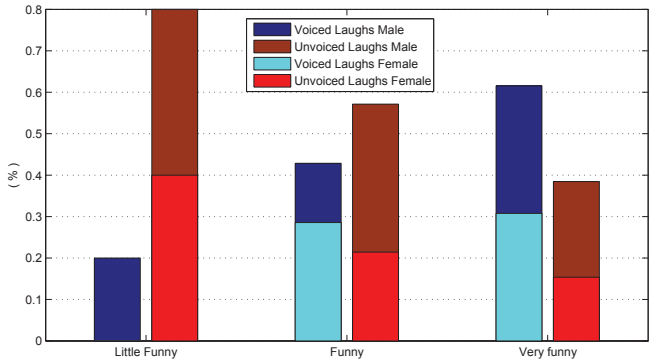


**Fig. 2**. Distribution of subjects laugh ratings for each category (best seen in color)

an indication of low amusement (in our case of a less funny scene) whereas voiced laughters are perceived as an indication of high amusement, i.e., a "very funny" scene. It is also interesting that females rated all voiced laughters as corresponding either to "funny" or "very funny" scenes whereas males assigned them in all categories with more votes going to the "very funny" category. Regarding unvoiced laughters, we notice that males are more likely to associate them with low amusement scenes than females.

Overall, we see that both sexes perceive voiced laughters as an indication of more amusing stimuli material than unvoiced ones, which are usually perceived as an indication of less amusing stimuli material. We also noticed that this distinction is more clear for females than for males.

## 6. CONLUSIONS

In this work, a preliminary study was conducted with the aim of associating different types of laughter with the perceived hilarity of the multimedia content being watched. Initial results suggest unvoiced laughter is correlated with not so amusing multimedia content and voiced laughter is correlated with highly amusing multimedia content. However, a more thorough study is needed to confirm the findings reported here.

## 7. ACKNOWLEDGEMENTS

The research presented in this paper has been funded in part by the EC's 7th Framework Programme [FP7 / 2007-2013] under grant agreement no 211486 (SEMAINE) and in part by the European Research Council under the ERC Starting Grant agreement no. ERC-2007-StG-203143 (MAHNOB).

## 8. REFERENCES

[1] WH Adams, G. Iyengar, C.Y. Lin, M.R. Naphade, C. Neti, H.J. Nock, and J.R. Smith, "Semantic Indexing of Multimedia Content Using Visual, Audio, and Text Cues," *EURASIP Journal On Applied Signal Processing*, vol. 2, pp. 170–185, 2003.

[2] "http://www.doc.ic.ac.uk/ maja/hct.html," .

[3] I. Arapakis, J.M. Jose, and P.D. Gray, "Affective feedback: an investigation into the role of emotions in the information seeking process," in *Proc. of the 31st Intern. ACM SIGIR conference on Research and development in information retrieval*. ACM New York, NY, USA, 2008, pp. 395–402.

[4] M. Soleymani, G. Chanel, J.M. Kierkels, and T. Pun, "Affective ranking of movie scenes using physiological signals and content analysis," in *MS '08: Proc. of the 2nd ACM workshop on Multimedia semantics*, New York, NY, USA, 2008, pp. 32–39, ACM.

[5] S. Petridis and M. Pantic, "Audiovisual laughter detection based on temporal features," in *ACM ICMI*, 2008, pp. 37–44.

[6] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, and V. Karaiskos, "The ami meeting corpus," in *Int'l. Conf. on Methods and Techniques in Behavioral Research*, 2005, pp. 137–140.

[7] J. A. Bachorowski, M. J. Smoski, and M. J. Owren, "The acoustic features of human laughter," *Journal-Acoustical Society of America*, vol. 110, no. 1, pp. 1581–1597, 2001.

[8] S. Petridis and M. Pantic, "Fusion of audio and visual cues for laughter detection," in *ACM CIVR*, 2008, pp. 329–337.

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

[10] Z. Zeng, M. Pantic, G.I. Roisman, and T.S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 1, pp. 39–58, Jan. 2009.

[11] I. Patras and 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.

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

[13] B. Reuderink, M. Poel, K. Truong, R. Poppe, and M. Pantic, "Decision-level fusion for audio-visual laughter detection," in *MLMI*, 2008.

[14] J.A. Bachorowski and M.J. Owren, "Not All Laughs Are Alike: Voiced but Not Unvoiced Laughter Readily Elicits Positive Affect," *Psychological Science*, vol. 12, no. 3, pp. 252–257, 2001.