Multi-modal Laughter Recognition in Video Conversations

Sergio Escalera, Eloi Puertas, Petia Radeva, and Oriol Pujol
Universitat de Barcelona, Gran Via de les Corts, 585, 08007, Barcelona, Spain.
{sergio,eloi,petia,oriol}@maia.ub.es

Abstract

Laughter detection is an important area of interest in the Affective Computing and Human-computer Interaction fields. In this paper, we propose a multi-modal methodology based on the fusion of audio and visual cues to deal with the laughter recognition problem in face-to-face conversations. The audio features are extracted from the spectogram and the video features are obtained estimating the mouth movement degree and using a smile and laughter classifier. Finally, the multi-modal cues are included in a sequential classifier. Results over videos from the public discussion blog of the New York Times show that both types of features perform better when considered together by the classifier. Moreover, the sequential methodology shows to significantly outperform the results obtained by an Adaboost classifier.

1. Introduction

Nowadays, Human-Computer interaction topic has become a cutting edge field of research. Now, computers are more ubiquitous than never. It is usual to find them anywhere, but still the Human-Computer interaction (HCI) is only based on manipulation. Communication between computers and humans must evolve, thus everytime humans engage a transaction with computers, they should be able to interact in a more "human" way. Current researches on HCI are summing up the efforts from different fields involving Artificial Intelligence, Speech Recognition, or Computer Vision in order to make computers able to accordingly interact with their interlocutors. To achieve this goal, when computers are engaged in an interaction with humans, the acquisition of audiovisual feedback is a key point. This information can be provided in diferent ways. Speech is the most common one, but spoken words are highly person and context dependent [1]. Nonetheless, speech recognition and extraction of semantic information is a very active and challenging field. Thus, the information provided by this channel can be filtered and reinforced by means of a multi-modal fusion including other information channels such as the visual feedback. Facial expressions, gestures, and non-linguistic sounds are other ways which computers can acquire useful feedback from humans. Non-linguistic sounds such as laughter, caught, murmurs, or yawns provide very significant information. Humans have the ability of recognize emotions just by hearing such sounds [2]. For example, a person yawning is an indicator that he is bored or tired. Otherwise, if he is laughing, it is an indicator that he is happy.

Laughter is one of the most important non-linguistic sounds. In human-human conversations laughter appears frequently, not only when the interlocutors are in a more relaxed or informal situation, but also in meetings as an agreement signal or as a welcome response [3]. In HCI automatic, recognition of laughter can be used to detect the affective state of the interlocutor and also for detecting semantically meaningful events in meetings such agreements, topic changes, or jokes. Laughter detection can also be useful in another situations such as in automatic speech recognition or content-based video retrieval.

Previous works on automatic laughter detection have been mainly focused on audio information. Such works use spectral or phonetic features from the audio with a sequential learning approach as Hidden Markov Models [4, 5]. The most extensive study in this area was made by Truong and Leeuwen [6], who compared the performance of different utterance level features using different classifiers. They found that a fusion between classifiers based on Gaussian Mixture Models and classifiers based on Support Vector Machines increases discriminative power. They also found that a fusion between classifiers that use spectral features and classifiers that use prosodic information usually increases the performance for discrimination between laughter and speech. Fewer works have focused on the problem of automatic laughter detection using multi-modal audio and visual information. Reuderink [7] used visual features based on PCA and RASTA-PLP (Perceptual linear prediction) features for audio processing. They use Gaussian mixture models and SVM as classifiers, fusing them on decision level. Petridis and Pantic [8] used spectral features and prosodic features got from the audio channel plus visual

978-1-4244-3993-5/09/$25.00 ©2009 IEEE

features based semi-automated tracking of annotated visual face landmarks and head movement. To recognize whether an input audio or visual sample is an episode of speech or an episode of laughter a neural network is used. Then a fusion rule is used in order to combine both predictions.

In this paper we propose a system for automatic laughter detection which extract information from both audio and video channels. The audio features considered are derived from spectrogram and complemented with accumulated power, spectral entropy, and fundamental frequency. On the other hand, the visual features correspond to the participant mouth movement and smile-laughter labels using a learnt classifier based on principal components. Then, we use a stacked sequential learning schema for laughter detection. We join audio and visual features and perform a two level stacked learning using Adaboost base classifier. First, an Adaboost is trained with all the multi-modal data. Then, an extended data set is created using a window of predictions for each sample. Finally, a second level classifier is learnt with this extended data set. Results over public video data shows better performance of the multi-modal features than using individual cues. Moreover, the sequential stacked approach significantly outperforms the results provided by an Adaboost classifier.

The layout of the paper is as follows: Section 2 describes the multi-modal features and the stacked learning methodology. Section 3 presents the experimental evaluation of our approach, and finally, Section 4 concludes the paper.

2. Laughter Recognition

In this section we describe in detail the spectrogram-based audio features, the visual features derived from mouth detection, and the stacked learning methodology used to learn the multi-modal features.

2.1. Audio features

In order to describe the laughter from audio we base our analysis on the work [9]. The feature vector uses a basic spectrogram $s(\omega)$ with an interleaved sliding window of size 256 samples with relative displacements of 128 samples. Moreover, the feature vector is enriched by three more general descriptive features, namely accumulated power ($\sum s(\omega)$), spectral entropy ($-\sum s(\omega)\log s(\omega)$), and fundamental frequency. The fundamental frequency is computed by finding the peak in the band between 20 - 500 Hz.

2.2. Visual features

The non-verbal cues used in this work consist of a frame-by-frame smile-laughter detection and a historial mouth movement. The use of these features will complement the audio features information in order to improve the laughter detection process.

On one hand, frame-to-frame smile-laughter detection increases the probability to find a laughter sequence. Since only using visual information smiles and laughters remain close similar, they are taken into account as the same visual category. One the other hand, sporadic smiling frames or mouth movement due to the speaking sequences can be confused with a false smile-laughter sequence. Because of this reason, the historial of the mouth movement is also included. The smiling and laughter sequences tend to have less mouth variation in time than the speaking ones. Thus, historial mouth movement can help the system to discriminate between smile-laughter and speaking sequences. Moreover, a post-filtering of the output features removes isolated smile-laughter detection and groups sequences with high smile-laughter densities. Next, we describe in detail the two visual cues used in this paper.

2.2.1 Visual smile-laughter detection

The procedure to learn a frame-by-frame smile-laughter classifier is shown on the top of Fig. 1. In order to define a groundtruth to learn the classifier, first we apply a face detection followed by a mouth extraction using face geometry.

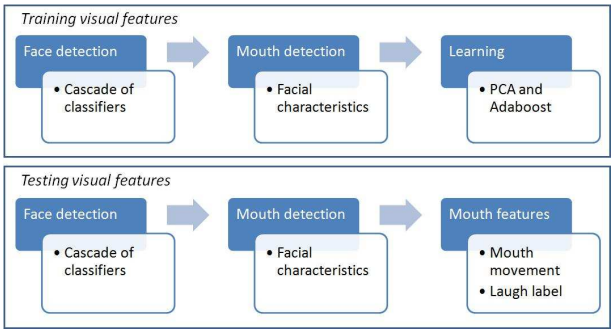
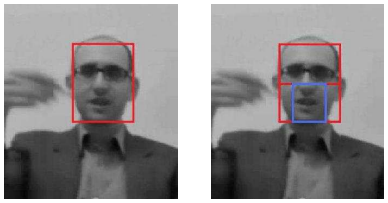


Figure 1. Learning and testing visual features.

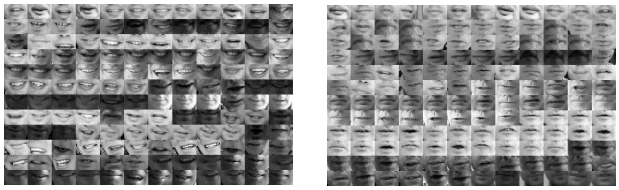
The face detection is performed by means of a cascade of classifiers using an Adaboost with decision stumps and the whole set of Haar-like features computed over the integral image [10]. The result of this step is shown in Fig. 2(a). Once a face $F_i \in \{0,..,255\}^{n\times m}$ is detected at frame i, the mouth region is defined as $M_i \in \{0,..,255\}^{n/2\times m/2}$, which corresponds to the center bottom half region of F_i. An example of a mouth localization is shown in Fig. 2(b).



(a) (b)

Figure 2. (a) Face and (b) mouth detection example.

With the extracted regions, we can define two categories in order to learn the smile-laughter classifier. At the beginning, we manually labeled the intervals of sequences corresponding to smiling or laughter. However, smiling sequences can contain frames corresponding to neutral or speaking mouth movements. In the same way, non-smiling sequences also contain isolated smiling frames. Thus, working at frame level we choose to manually define the positive and negative smile-laughter sets. Some examples from the positive and negative sets are shown in Fig.3(a) and (b), respectively. The dyadic sequences are obtained from the public New York Times opinion Bloggings data set [11]. In particular, the data set is composed by 600 positive and 2500 negative samples, respectively. All samples are resized to a resolution of 25× 40 pixels. Applying PCA and saving a 99% of principal components, 70 features per sample are obtained. This feature space is then learnt using 100 iterations of Adaboost with decision stumps. Then, to look for a smile-laughter frame in a video test sequence, the face detector, mouth localization, and learnt laughter classifier are applied to each test frame, as shown on the bottom of Fig. 1. Some examples of smile-laughter results with this learnt classifier are shown in Fig. 4(a) and (b), respectively. Finally, a post-filtering is applied over the output vector of labels from the dyadic sequences. In the filtering, the frames within a filter width of N frames are set to the label of the majority. An example is described in Fig. 4(c).



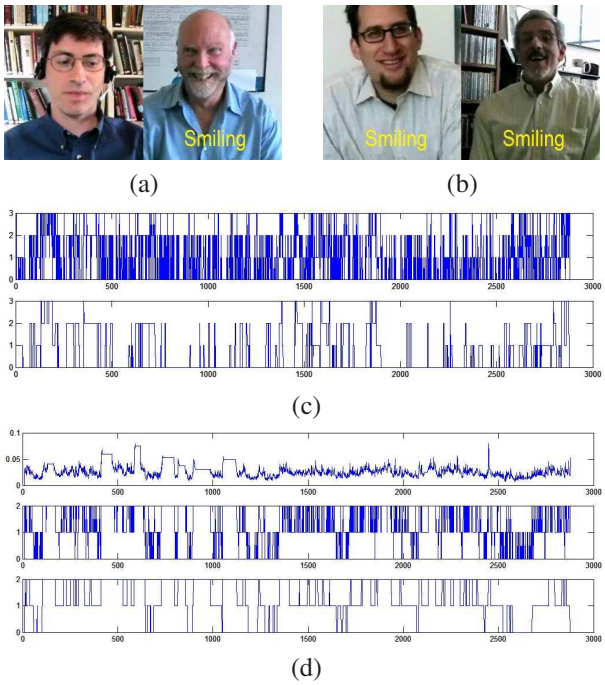
(a) (b)

Figure 3. (a) Positive and (b) negative laughter samples.

2.2.2 Mouth movement historial

In the final step of the testing scheme shown in Fig. 1 one can see that the smile-laughter labels are complemented with the information provided by the mouth movement. The main goal is that if we are able to compute the level of movement of the speaker mouth in a sequence, we will be able to split the smile-laughter behavior from the speaking one, since speaking sequences tend to have a higher movement degree. In order to deal with this problem, we compute the additive distance between the current localized mouth and the set of the previous mouth states.

This is performed in order to avoid the bias that can appear due to the translation of mouth detection among consecutive frames. Thus, to compute the mouth movement MM_{il} at frame i, we estimate an accumulated subtraction



(a) (b) (c) (d)

Figure 4. (a)(b) Smiling detection results using visual frame-to-frame classification. (c) Post-filtering results: On the top, the initial frame-by-frame output is shown in a four minutes conversation at 12 FPS. Zero label means no-one smile-laughter, one label means left speaker smile-laughter, two label means right speaker smile-laughter, and three label means both speakers smile-laughter. On the bottom, the result after filtering with a filter size of ten frames is shown. (d) Similar to previous procedure using the mouth movement cue. At the bottom, the output obtained using eq. (1) is shown. In the middle, a discretization using three movement levels is obtained. Finally, at the bottom, the result after filtering with a filter size of ten frames is shown.

of l mouth regions previous to the mouth at frame i. From the face region $F_i \in \{0,..,255\}^{n\times m}$ detected at frame i and the mouth region $M_i \in \{0,..,255\}^{n/2\times m/2}$, the mouth movement feature MM_{il} is computed as follows:

$$MM_{il} = \frac{1}{n \cdot m/4} \sum_{j=i-l}^{i-1} \sum_{k} |M_{i,k} - M_{j,k}| \qquad (1)$$

where $M_{i,k}$ is the kth pixel in a mouth region M_i, $k \in \{1,..,n\cdot m/4\}$, and $n\cdot m/4$ is a normalizing factor. The accumulated subtraction avoids false positive mouth activity detection due to noisy data and translation artifacts of the mouth region.

The top vectors in Fig. 4(d) correspond to the vectors of global movement features MM_{il} in a sequence of 2880 frames at 12 FPS. At the post-processing step, first, we filter the vectors in order to obtain a 3-value quantification. For this task, all vectors from all speakers for each movement feature are considered together to compute the corresponding feature histogram (i.e. histogram of global

mouth movement h_{MM}), which is normalized to unit in order to estimate the probability density function (i.e. pdf of global mouth movement P_{MM}). Then, two thresholds are computed in order to define the three values of movement, corresponding to low, medium, and high mouth movement quantifications:

$$t_1 : \int_0^{t_1} P_{MM} = \frac{1}{3}, \quad t_2 : \int_0^{t_2} P_{MM} = \frac{2}{3} \qquad (2)$$

The result of this step is shown in the middle vectors of Fig. 4(d) for the input vectors on the top of Fig. 4(d). Finally, in order to avoid abrupt changes in short sequences of frames, we apply a sliding window filtering of size q using a majority voting rule, as in the previous laughter label vectors. The result of this step is shown in the bottom vector of Fig. 4(d).

2.3. Stacked Sequential Learning

The laughter signal displays a clear pattern alternating voice and non-voice segments. Thus, discriminant information does not only come from each single signal example but from how they are presented along the time axis. For this reason, a successful system dealing with this problem must be aware of the non-independence of samples and their temporal coherence. Sequential learning is the discipline of machine learning that deals with dependent data such that neighboring examples exhibit some kind of relationship. In literature, sequential learning has been addressed from different perspectives: from the point of view of meta-learning by means of sliding window techniques, recurrent sliding windows [12], or stacked sequential learning [13, 14]. From the point of view of graphical models, Hidden Markov Models and Conditional Random Fields are used to infer the joint or conditional probability of the sequence. Graph Transformer Networks considers the input and output as a graph and looks for the transformation than minimizes a loss function of the training data using a Neural Network. Recently, Cohen et al. [14] shown that stacked sequential learning (SSL from now on) performed better than CRF and HMM on a subset of problems called "sequential partitioning problems". These problems are characterized by long runs of identical labels. Moreover, SSL is computationally very efficient since it only needs to train two classifiers a constant number of times. The basic idea of

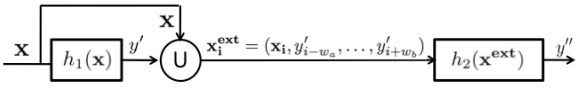


Figure 5. Stacked Sequential Learning scheme.

stacked sequential learning is to create an extended data set that joins the original training data features with the predicted labels considering a neighborhood around the example. Figure 5 shows a block diagram of the SSL method. The basic SSL method uses a five-fold cross-validation on the training set to obtain the predicted set Y' and considers a sliding window of length w with origin in the prediction of the current example to extend its features. That is, for each example in the training set $x_i \in \mathbf{X}$, the predicted values $y'_i \in Y'$ are obtained and joined creating an extended example $x_i^{ext} = (x_i, y'_{i-w_a}, \ldots, y'_{i+w_b}) \in \mathbf{X^{ext}}$, where the number of added features is $w = w_a + w_b + 1$. The extended training set is used to train a second classifier that is expected to capture the sequentiality of the data.

3. Results

In order to present the results, first we discuss the data, methods, validation protocol, and experiments.

• *Data*: The data used for the experiments consists in dyadic video sequences from the public New York Times opinion video library [11]. In each conversation, two speakers with different points of view discuss about a direct question (i.e. In the fight against terrorism, is an American victory in sight?). From this data set, 18 videos has been selected. These videos are divided into two mosaics of nine videos. The two mosaics are shown in Fig. 6. Each video has a frame rate of 12 FPS and a duration of four minutes, which corresponds to 2880 frames per video sequence.

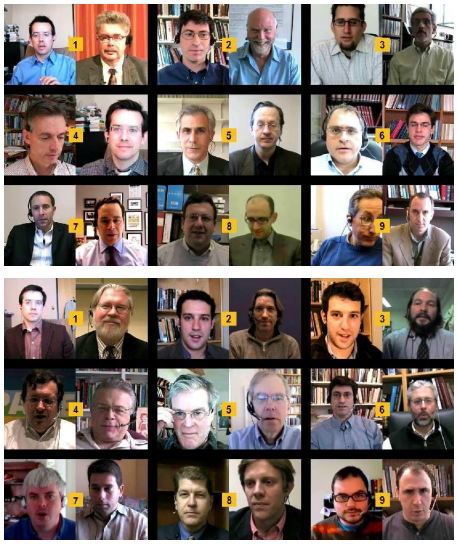


Figure 6. Two video sets.

• *Methods*: We learn the multi-modal cues described on the previous sections with the standard Gentle Adaboost classifier [15] and the previous sequential stacked methodology. Adaboost is run with 50 iterations of decision stumps, and the windows for the sequential strategy is of size 11. The details of the feature vectors are described next.

	Video 1	Video 2	Video 3	Video 4	Video 5	Video 6	Video 7	Video 8	Video 9
Mosaic 1	5.4(1.0)	5.3(0.8)	4.3(0.9)	3.3(0.6)	**2.7(0.6)**	6.7(0.8)	6.4(1.0)	3.1(1.0)	**7.9(0.6)**
Laugh period	8	6	20	14	**39**	3	3	15	**0**
Mosaic 2	3.4(0.9)	4.3(0.8)	4.8(0.9)	**7.2(1.0)**	4.2(1.2)	5.9(1.0)	4.2(1.0)	6.8(0.8)	4.3(0.9)
Laugh period	3	0	0	**0**	33	11	4	4	2

Table 1. Ranking positions and confidence interval of dyadic interactions.

• *Validation protocol*: In order to validate our results, we perform a ten-fold evaluation and compute the accuracy, sensitivity, and specificity measures.

• *Experiments*: First, we analyze the laughter characteristic as a feature to predict observers interest in dyadic video sequences. After, laughter detection is performed using the multimodal audio-video cues.

3.1. Correlation between smile-laughter and observers' interest

One of our hypotheses is that laughter detection can be an indicator of interest in dyadic conversations. In order to analyze this hypothesis, we ranked the two mosaics shown in Fig. 6. For this task, 40 people from 10 different nationalities categorized from one (maximum interest) to 9 (lowest interest) the videos of both mosaics separately. In each mosaic, the nine conversations are displayed simultaneously during the four minutes analyzed by our methodology. Table 1 shows the mean rank and confidence interval of each dialog considering the observers' interest ranks. The ranks are obtained estimating each particular rank r_i^j for each observer i and each video j, and then, computing the mean rank R for each video as $R_j = \frac{1}{P}\sum_i r_i^j$, where P is the number of observers. At bottom of each rank, the number of laughter seconds of the video sequence is shown. Note that the first and last interest choice from the first mosaic matches with the highest and lest number of laughter seconds labeled in the videos. On the second mosaic, the last interest choice also matches with the less number of laughter seconds. This suggests laughter as a discriminative feature for the observers' interest rating problem.

3.2. Multi-modal Laughter Detection Performance

In our scheme, the verbal and non-verbal cues are combined in order to be learnt by Adaboost and the Sequential Stacked Learning procedures. The verbal cue is obtained by sampling at 8000 Hz per second. Our video sequences have a fixed length of four minutes, and a windowing procedure of width size of 256 with intersection of 128 is applied among iterations in order to compute the FFT transform. Thus, a final audio sequences of 15000 positions and 134 features is obtained.

In order to combine the previous audio information with the non-verbal cues, the 2880 vectors data from four minute videos at 12 FPS of the smile-laughter labels and mouth movement level are projected into a 15000 value vectors by expanding each value in a range of $\simeq$5.2 positions, matching with the audio data length. Before applying this step, the smile-laughter labels are re-labeled so that values of left speaker laughter, right speaker laughter, and both speakers laughter are set to one, while the non-laughter sequences are set to zero. It has been re-labeled since in the combination with the audio data, the source of useful information is just if laughter exists.

With the previous feature vectors obtained from the 18 videos shown in Fig. 6, we labeled the sequences within the four minute videos at which laughter appears, having a total of 270000 samples of 136 features, where only 2% of the samples have been manually labeled as laughter. Then, a ten-fold procedure is applied with Adaboost learner and audio features, Adaboost learner and multi-modal features, Sequential stacked with audio features, and Sequential stacked with multi-modal features. Given the confusion matrix M obtained at each iteration of the ten-fold procedures $M = \begin{pmatrix} TN & FP \\ FN & TP \end{pmatrix}$, where TN, FP, FN, and TP stand for the true negatives, false positives, false negatives, and true positives, respectively, we compute the accuracy ACC as:

$$ACC = \frac{TN + TP}{TN + FP + FN + TP} \tag{3}$$

the sensitivity SE as:

$$SE = \frac{TP}{TP + FN} \tag{4}$$

and the specificity SP as:

$$SP = \frac{TN}{TN + FP} \tag{5}$$

Computing the mean ACC, SE, and SP measures of the ten runs of the evaluation with the different features sets and classifiers we obtain the results shown in Table 2. The results for Adaboost using the multi-modal cues and only audio features remains close similar. In fact, the only difference is an small increment on the sensitivity when including the visual features. On the other hand, when using the sequential methodology, the results show significant improvements on the three measures. Moreover, one can see that the video cues also increase the sensitivity. It can be

understood by the fact that spontaneous smiling can be rejected by the sequential methodology meanwhile more continuous smiling can help the laughter classifier to generalize better. On the other hand, the reduction of the global accuracy and specificity occurs because of the unbalanced influence of the samples of the no-laughter category and the need of increasing the sensitivity of the laughter class.

	ACC	SE	SP
Adaboost Audio	0.70	0.51	0.70
Adaboost Audio-Video	0.70	0.52	0.70
Sequential Audio	0.81	0.61	0.81
Sequential Audio-Video	0.77	0.65	0.77

Table 2. Laughter recognition results.

The results reported in this section can be considered significantly good since the classification is performed frame-by-frame. In our current research we are combining the visual knowledge of smile and laughter as a post-processing filtering step of audio outputs considering that laughter can not appear on isolated samples.

Acknowledgements

This work has been partially supported by the projects TIN2006-15694-C02-02 and CONSOLIDER-INGENIO 2010 (CSD2007-00018).

4. Conclusions

In this paper we presented a multi-modal solution to the laughter detection problem. We defined a set of audio and visual cues which are learnt integrating an Adaboost classifier in a sequential classification methodology. The audio features are derived from the spectogram and the visual cues are based on the participant mouth movement and a smile-laughter classifier. We used the public discussion blog of the New York Times to validate our methodology, showing high sensitivity and better performance of the sequential classifier in a frame-by-frame evaluation.

References

[1] G. W. Furnas, T. K. Landauer, L. M. Gomez, S. T. Dumais, The vocabulary problem in human-system communication, Commun. ACM 30 (11) (1987) 964–971. 1

[2] M. Schröder, D. K. J. Heylen, I. Poggi, Perception of non-verbal emotional listener feedback, in: R. Hoffmann, H. Mixdorff (Eds.), Speech Prosody 2006, Dresden, Vol. 40 of Studientexte zur Sprachkommunication, TUDpress, Dresden, 2006, pp. 43–46. 1

[3] M. Pantic, A. Pentland, A. Nijholt, T. Huang, Human computing and machine understanding of human behavior: a survey, in: ICMI '06: Proceedings of the 8th international conference on Multimodal interfaces, ACM, New York, NY, USA, 2006, pp. 239–248. 1

[4] R. Cai, L. Lu, H.-J. Zhang, L.-H. Cai, Highlight sound effects detection in audio stream, in: ICME '03: Proceedings of the 2003 International Conference on Multimedia and Expo - Volume 3 (ICME '03), IEEE Computer Society, Washington, DC, USA, 2003, pp. 37–40. 1

[5] N. Campbell, H. Kashioka, R. Ohara, "no laughing matter", in INTERSPEECH-2005, 465-468 (2005). 1

[6] K. P. Truong, D. A. van Leeuwen, Automatic discrimination between laughter and speech, Speech Commun. 49 (2) (2007) 144–158. 1

[7] B. Reuderink, M. Poel, K. Truong, R. Poppe, M. Pantic, Decision-level fusion for audio-visual laughter detection, in: MLMI '08: Proceedings of the 5th international workshop on Machine Learning for Multimodal Interaction, Springer-Verlag, Berlin, Heidelberg, 2008, pp. 137–148. 1

[8] S. Petridis, M. Pantic, Fusion of audio and visual cues for laughter detection, in: CIVR '08: Proceedings of the 2008 international conference on Content-based image and video retrieval, ACM, New York, NY, USA, 2008, pp. 329–338. 1

[9] R. Caneel, Social signaling in decision making, in: Master Thesis, 2005. 2

[10] M. Jones, P. Viola, Robust real-time face detection, in: International Journal of Computer Vision, Vol. 57, 2004, pp. 137–154. 2

[11] http://video.nytimes.com/. 3, 4

[12] T. G. Dietterich, Machine learning for sequential data: A review, in: Proc. on Joint IAPR International Workshop on Structural, Syntactic, and Statistical Pattern Recognition, 2002, pp. 15–30. 4

[13] D. H. Wolpert, Stacked generalization, Neural Networks 5 (2) (1992) 241–259. 4

[14] W. W. Cohen, V. R. de Carvalho, Stacked sequential learning, in: Proc. of IJCAI 2005, 2005, pp. 671–676. 4

[15] J. Friedman, T. Hastie, R. Tibshirani, Additive logistic regression: a statistical view of boosting, in: The annals of statistics, Vol. 38, 1998, pp. 337–374. 4