Smile and laughter detection for elderly people-robot interaction

Fan Yang[12], Mohamed A. Sehili[1], Claude Barras[12], and Laurence Devillers[13]

[1] Department of Human-Machine Communication, LIMSI-CNRS, Orsay, France
[2] Department of Computer Science, University Paris-Sud, Orsay, France
[3] University Paris-Sorbonne, Paris, France

Abstract. Affect bursts play an important role in non-verbal social interaction. Laughter and smile are some of the most important social markers in human-robot social interaction. Not only do they contain affective information, they also may reveal the user's communication strategy. In the context of human robot interaction, an automatic laughter and smile detection system may thus help the robot to adapt its behavior to a given user's profile by adopting a more relevant communication scheme. While many interesting works on laughter and smile detection have been done, only few of them focused on elderly people. Elderly people data are relatively rare and often carry a significant challenge to a laughter and smile detection system due to face wrinkles and an often lower voice quality. In this paper, we address laughter and smile detection in the ROMEO2 corpus, a multimodal (audio and video) corpus of elderly people-robot interaction. We show that, while a single modality yields a given performance, a fair improvement can be reached by combining the two modalities.

1 Introduction

Laughter and smile are considered as all-important human communication skills. They convey lots of information during human-human interaction such as emotional state, social communication strategy and personality. This kind of information also appears in human-robot interaction, especially with humanoid robots. This paper focuses on laughter and smile detection of elderly people who interact with a robot in a real-life situation. For this purpose, we use part of a social interaction corpus [17] recorded in two retirement homes in France. This multimodal corpus is collected under the ROMEO2 project[1] and features elderly people interacting with the humanoid robot Nao.

Audio and visual based smile and laughter detection has each its pros and cons. Actually, while audio is a suitable modality for laughter detection, particularly when a subject is not facing the video source, things seem to be less obvious when it comes to detecting a smile by merely using an audio signal. Visual detection however can be a good way for both smile and laughter as

[1] http://projetromeo.com

long as the subject faces the video camera and there is no obstacle between the two communicative parties. Another issue that visual smile detection faces is the similarity between a smiling mouth and a speaking mouth. The joint use of audio and video channels to improve the overall smile and laughter detection performance is investigated in this work.

Beside the afore mentioned issues, others are to be taken into account when dealing with smile and laughter detection for elderly people. Namely, the lack of relevant data of actual elderly people naturally interacting with a robot, voice quality and face wrinkles related difficulties are a supplementary challenge addressed in this work. We thusly consider the use of such realistic data with the proposed detection methods as the main contribution of this work.

The rest of this paper is organized as follows. Section 2 reviews some of the related work. Section 3 describes the data collection protocol, the corpus content, the annotation scheme, the questionnaires submitted to each subject after the experience, a statistical analysis and the data subset used in our experiments. In section 4 we describe our audio and video smile and laughter detection methods, as well as the fusion of both modalities. We then present our experiment protocols and the obtained results in section 5. We give our conclusion and perspectives in section 6.

2 Related Work

Due to the importance of smile and laughter in human-human interaction, many recent works have focused on smile and laughter research in the computer science area, especially in human-machine interaction. Many workshops dedicated to the topic have been organized such as the Interdisciplinary Workshop on Laughter and other Non-Verbal Vocalizations in Speech[2]. International projects like the ILHAIRE project[3] are also to be mentioned.

There are many acted or posed facial expression databases (e.g. the Cohn-Kanade database [6], the MMI Facial Expression database [13, 21], or the JAFFE database [10]), but only few realistic databases exist. They are even fewer when it comes to realistic data involving elderly people. Most researchers actually test and validate their methods on acted or posed corpora [1, 5, 8, 18]. In [22], however, the authors argue that spontaneous expressions are different from posed expressions both in appearance and in timing. This means that methods used for posed expressions recognition might not be suited to realistic expressions. Therefore, the detection methods proposed in this work and evaluated on our realistic social interaction corpus between elderly people and a robot need to be assessed on another acted corpus.

For the visual detection system, we use Support Vector Machines (SVM) with a Radial Basis Function (RBF) kernel for classification. We use Local Binary Patterns (LBP) [4, 24] for feature extraction. There exist a variety of feature extraction methods in the literature (e.g. local Gabor binary patterns [11], local

[2] https://laughterworkshop2015.wordpress.com/

[3] http://www.ilhaire.eu/

phase quantization, histogram of oriented gradients [2], Haar filters [9] and FACS coding system action unit detection [3]).

Laughter detection in audio is also addressed in the literature. One can distinguish two types of laughter recognition: recognition with prior segmentation or segmentation by recognition. In the first approach, short audio segments representing acoustic events (anything that is not a silence) are either manually or automatically extracted from a continuous audio stream before they are classified. As for the second approach, segmentation by recognition, there is no prior knowledge about where an acoustic event starts and where it ends. The whole audio stream is analyzed to tell apart the classes of interest (e.g. speech, laughter, silence, other human or environmental sounds, etc.). In [7], MFCC and Modulation Spectrum features are used with SVM for laughter detection in meeting rooms. The main focus of the authors was the detection of laughter events where more than one person simultaneously laugh. Spacial cues were therefore calculated by cross-correlating the audio signals acquired by two tabletop microphones. The goals of this cross correlation is to better distinguish one-participant and multi-participant laughters. In [20] many sets of acoustic features (Perceptual Linear Prediction Coding features, Energy, Pitch and Modulation Spectrum) and classification algorithms (Gaussian Mixture Models, Hidden Markov Models and Multi Layer Perceptrons) are investigated for laughter-speech classification. The best baseline performance is obtained PLP features with GMM. Improvement could be observed by combining the PLP with GMM system with a system based on Pitch related features with SVM. [16] address a 5-class classification problem. Each audio segment is classified into four human classes (breathing, consent, hesitation and laughter) or a garbage class used to model background noise. The best reported performance was obtained with HMMs and PLP feature.

These methods address all the problem of classification after segmentation. Many other works focus on laughter detection using segmentation by classification scheme. In [19] a stream is segmented into laughter, speech and silence intervals using PLP features and GMM. A 3-state Viterbi decoder is first used to find the most likely sequence of states given a stream. The sequence of states is seen as a preliminary segmentation. The log likelihood of each segment given each of the GMM models is calculated to determine the final class of the segment. In [15] MFCC and HMMs are used to label a stream with a set of classes containing laughter, filler, silence and speech. A higher level model, a bigram language model, is used to explicitly model the order in which the labels appear in training data. [14] show that adding visual information (head pose and facial expression) slightly improve the performance of the audio-based system.

3 ROMEO2 corpus

The main motivation of this data collection was to build an elderly people-robot interaction corpus within the ROMEO2 project [17]. The collected corpus is made up of audio and video streams of the whole interaction for each subject as well as two questionnaires (a satisfaction questionnaire and a personality questionnaire) and detailed logs of the robot actions (time stamped utterances, sounds, gestures, played songs, etc.). A high definition webcam set up behind the sitting robot was used for frontal video capture, alongside with another camera used to capture both interlocutors from a profile perspective. The number of the participating subjects is 27 (3 men and 24 women), with an average age of 85.

Audio and video tracks were both annotated for this experiment. For the audio part, speech (start time and end time of each utterance), affect bursts (including laughter) and emotion in voice (one tag among happiness, sadness, anger, doubt, surprise or neutral) were annotated. In the visual part, head pose, head gesture, certain mouth movements, eyebrow movements and body movements and perceived emotion of face were also annotated.

In order to analyze the connection between a subject's behavior and their profile, we calculate the correlation between answers to 3 questions from the satisfaction questionnaire and the number of smiles and laughters within the interaction. As shown in table 1, the more relaxed and enjoyed a subject was, the more they expressed smile and laughter during the interaction.

Table 1. Correlation between experience enjoyment and the number of smiles and laughter during an interaction. +1 and -1 are the numerical translations of the answer used for correlation computation.

Question	**Events**	**Correlation**	**P-value**
(Q7) Would you like it to address you using the familiar form (+1) or using the formal form (-1)?	Laughter + Smile	0.424	0.04395
(Q10) Would you prefer a robot that looks like a robot (+1) or a human (-1)?	Smile with open mouth	-0.465	0.02528
(Q11) Do you consider the robot as a machine (+1) or as a friend or a (human) companion (-1)?	Laughter + Smile	-0.429	0.04134

In this work, since nearly 90% of audio annotated laughters have intersection with the visual annotated laughter and smile events in our corpus, we use visual annotation as the annotation reference for our experiment, and only obvious events such as laughter and smile with open mouth are studied. This results in about 575 events from the 27 subjects subjects. Since the repartition of the events was not balanced among the subjects, all male participants (3 subjects) and all the subjects with less than 10 events of laughter and open-mouth smile were discarded. As a result, our experimental data subset consists in 15 female subjects and contains 168 laughter events and 218 open-mouth smiling events (386 smile related events in total).

4 Proposed smile and laughter detection methods

4.1 Audio-visual fusion of smile and laughter detection

Smile detection from images has been found to be efficient in many acted emotion corpora [1, 5, 8, 18]. However, this has not yet been studied on realistic data of elderly people-robot interaction. Besides the face wrinkles related issue inherent to elderly people, the distinction between an open mouth smile and a speaking mouth raises another important challenge for visual smile detection. To overcome this, speech detection from audio is used in this work to back up the visual detection system and let the visual system only focus on smile detection while the subject is not talking.

4.2 Visual smile and laughter detection

For visual laughter and smile detection, we use uniform-Local Binary Patterns [12] with a SVM classifier using a RBF kernel. This method was tested on the GENKI-4K corpus and performed with an accuracy of over 83%.

Video was recorded at 30 frames per second with a resolution of 1280x720 pixels. For each frame, a frontal facial image of the subject was extracted using the Viola and Jones face detector [23]. After histogram equalization, facial images were reshaped to 64x64 pixels and processed using LBF with a 10x10 grid to get one feature vector for each frame. The performance of the SVM with a RBF kernel classifier was optimized, by varying the values of the c and γ parameters of SVM between 0.01 and 100 on a logarithmic grid with a multiplying step of 3. The evaluation is run at a frame-level and segment-level. For the segment-level evaluation, we consider that a segment contains smile or laughter if the majority of frames it is composed of are classified as smile or laughter.

4.3 Laughter detection from audio

As an audio stream of a typical human-robot interaction is globally made up of speech, laughter, robot prompts and silence, we used a 4-class classification scheme. Audio classification was done on frame-level using 13 MFCC coefficients and four GMM models to represent the 4 classes. Each frame is 20 ms long and has an overlap of 10 ms with the previous one. Audio frames within a visual laughter annotation are classified, which yields a sequence of symbols belonging a 4-letter alphabet. From our annotation experience we noticed that, visually, a laughter lasts longer than its respective audio signal. This respective perceived audio laughter can actually be very brief in comparison to what the subject's face or body depicts, partially or completely masked by the a robot's utterance or merged with speech and/or silence. Therefore, for audio segment classification illustrated in figure 1, we used an aggregation strategy that applies on short sequences of consecutive frames (a sliding window of 800 ms for instance) instead the whole audio slice aligned with a visual annotation. the goal of this is to detect the presence of brief audio laughter events within a long "humanly" perceived

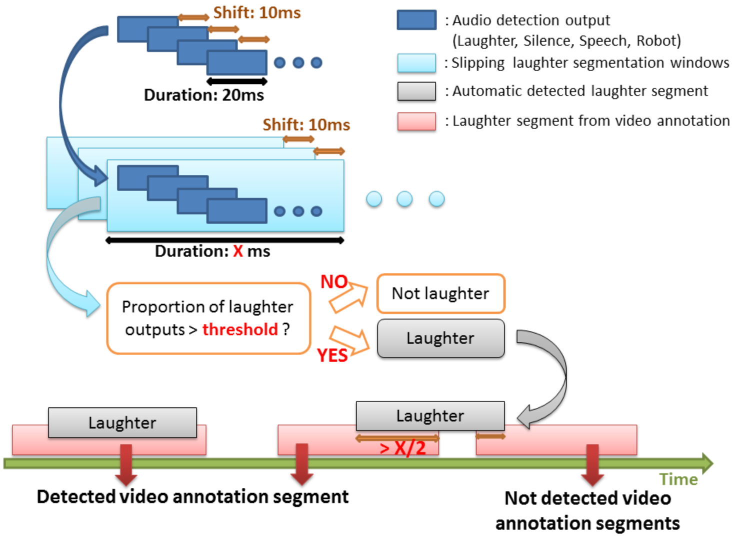
Shift: 10ms
Duration: 20ms
Shift: 10ms
Duration: X ms
: Audio detection output (Laughter, Silence, Speech, Robot)
: Slipping laughter segmentation windows
: Automatic detected laughter segment
: Laughter segment from video annotation
Proportion of laughter outputs > threshold ?
NO
YES
Not laughter
Laughter
Laughter
Laughter
> X/2
Time
Detected video annotation segment
Not detected video annotation segments

Fig. 1. Audio-level detection scheme

laughter. A window is recognized as belonging to class C if C boasts the highest number of frames within that window.

5 Experiments

5.1 Visual smile and laughter detection results

Three different protocols are used in our experiments:

- **Mono-subject:** for each subject, half the data is used for training and the remaining half for test.
- **Multi-subject:** for each subject, half the data of the subject plus all data from all the other subjects are used for training. The remaining subject's half is used for test.
- **Leave-one-out:** for each subject, training is performed using data from all the other subject's whereas test is done on all the subject's data

A total of 386 visual annotated smile and laughter events are used in our experiments. We also extracted 378 random segments outside laughter and smile annotations for the test. The visual system faces mainly two kinds of issues, missed frames due to a face detection problem (often caused by subjects' head-turning) and the apparent resemblance between a talking mouth and open-mouth smiles or laughter. This resemblance is accentuated by face wrinkles. To better assess the performance of the visual detection system, we suppose the use of an speech activity detection system with perfect outputs. Hence, all segments that contain speech are considered as non laughter and non smile events.

The obtained results with the three test protocols are illustrated in table 2. From this table we can observe that the best performance is reached using the speaker dependent protocol (i.e. mono-subject) with no extra data from other speakers for training. When we add data from other speakers to a given speaker's training data, we notice a performance decrease. This may be explained by the fact that, for elderly people, as for younger people, smile and laughter sensibly differ from one person to another. Finally, the leave-one-out protocol (no data from the test speaker were used to train the model) gave the lowest performance among the three protocols.

These results are obtained using all laughter and smile events (368). When using only 289 segments (the ones without a face detection problem), the obtained recall is 68.5%, 58.5% and 41.2% for the three evaluation protocols respectively.

Table 2. Visual detection evaluation. Total frame accuracy refers to the global system's accuracy (for laughter/smile against non-laughter/non-smile annotations). The two most-right columns are the recall and accuracy of laughter and smile annotations at frame-level and segment-level respectively. A. stands for **accuracy**, R. for **recall**, P. for **precision** and BER. for **Balanced Error Rate**.

Protocol	Frame-level	Segment-level
Mono-subject	A.85.7% R.80.6% BER.23.7%	R.51.3% P.95.2%
Multi-subject	A.73.9% R.62.4% BER.31.4%	R.43.8% P.95.5%
Leave-one-out	A.72.8% R.40.8% BER.46.6%	R.30.8% P.73.9%

5.2 Audio laughter detection results

For laughter detection, we also used 386 annotated smile and laughter events as well as 378 segments randomly cut from regions that do not contain a smile nor a laughter. An actual laughter event is considered as correctly classified by the system if at least one analysis window is recognized as a laughter by the system. A non laughter segment is however considered as correctly classified if none of the analysis windows it contains is recognized as laughter. Moreover, as each analysis window represents a sequence of frames, two frame aggregation strategies are used, a majority voting and an "at least half the frames" strategies. For the first strategy, the window is given the label of the most represented class. As for the second strategy, at least half the frames of the window have to be belong to one single class so that the window is labeled with this class.

Table 3 shows the obtained results. We can see that a short window results in a better recall but a relatively low precision whereas a long window leads to the opposite result. Moreover, a more rigorous aggregation strategy (at least 50%) improves the precision at the expense of the recall.

Table 3. Audio laughter detection evaluation using various analysis window durations on 386 laughter and smile events and 378 non laughter segments. Maj. voting: a window is considered as laughter if the most represented class within it is laughter. ≧ 50%: a window must contain at least 50% of laughter frames to be considered as a laughter.

Analysis window duration	Maj. voting	≧ 50%
600ms	R.80.1% P.54.0%	R.64.5% P.56.7%
800ms	R.74.9% P.54.3%	R.56.5% P.58.9%
1s	R.71.0% P.55.7%	R.49.7% P.59.6%

Note that out of the 386 visually annotated laughter and smile events, only 168 are laughter. Therefore, when we run the audio system on these 168 events, using an analysis window of 800 ms, we obtain a recall of 85.7% and 69.6% for the majority voting and the "at least 50%" aggregation strategies respectively.

5.3 Video-audio smile and laughter detection

In our detection system based on the fusion of audio and video decision, we consider a laughter or a smile detection if either of the two modalities decides a positive detection. To make a trade-off between precision and recall in the audio system, we use an analysis window of 800 ms with the two audio frames aggregation strategies mentioned in sub-section 5.2. Table 4 shows the obtained results. The fusion resulted in a fairly good recall improvement with a precision decrease. The recall of the fusion system is better than either one-modality based systems, regardless of the audio aggregation strategy. The precision however lays between that of the two systems and is mostly better than the audio system's.

Table 4. Audio-visual detection evaluation on 386 laughter and smile events. For audio detection, an analysis window of 800 ms is used.

		Video/audio fusion	
Protocol	Video only	≧ 50%	Maj. voting
Mono-subject	R.51.3% P.95.2%	R.78.8% P.65.1%	R.86.8% P.57.4%
Multi-subject	R.43.8% P.95.5%	R.75.4% P.64.7%	R.86.8% P.57.6%
Leave-one-out	R.30.8% P.73.9%	R.66.8% P.58.4%	R.80.6% P.54.3%

6 Conclusion and future work

This paper presents an audio-visual smile and laughter detection system in the context of elderly people-robot social interaction. The audio-video fusion system performs in two steps: first, each one-modality system is separately run to obtain its individual decision given a segment of aligned audio and video signals. The final decision is positive (there is a laughter or a smile) if at least one of the two systems outputs a positive decision.

The video system has a very good segment-level precision in subject dependent evaluation and a rather good precision in subject independent evaluation. It has however a lower recall in comparison to the audio system. This is probably due to the fact that many subjects laughed whilst turning their head as well as to mis-detections of the face detection system. This said, the audio system runs continuously which results in a high recall rate even when a long analysis window is used. We believe that the recall of the video system can be improved by dealing with the head turning issue whereas the audio system can achieve a better precision by using more data for training and using classification methods of a more discrimination power.

The fusion of the two system seems to lead to good compromise between precision and recall. Improving each of the one-modality systems will indeed result in an overall improvement of the fusion system performance.

In future work, we will consider other state-of-the-art video feature extractors in order to improve the performance of our system and to compare the performance of detection in posed database and realistic databases of elderly people-robot social interaction. We also plan to experiment other fusion strategies (e.g. frame-level of two modalities fusion or feature fusion).

References

1. Oscar Déniz, M Castrillon, J Lorenzo, L Anton, and Gloria Bueno. Smile detection for user interfaces. In *Advances in Visual Computing*, pages 602–611. Springer, 2008.
2. Abhinav Dhall, Akshay Asthana, Roland Goecke, and Tom Gedeon. Emotion recognition using phog and lpq features. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 878–883. IEEE, 2011.
3. P Ekman, WV Friesen, and JC Hager. Facs manual. *A Human Face*, 2002.
4. Xiaoyi Feng, Matti Pietikäinen, and Abdenour Hadid. Facial expression recognition based on local binary patterns. *Pattern Recognition and Image Analysis*, 17(4):592–598, 2007.
5. Akinori Ito, Xinyue Wang, Motoyuki Suzuki, and Shozo Makino. Smile and laughter recognition using speech processing and face recognition from conversation video. In *Cyberworlds, 2005. International Conference on*, pages 8–pp. IEEE, 2005.
6. Takeo Kanade, Jeffrey F Cohn, and Yingli Tian. Comprehensive database for facial expression analysis. In *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, pages 46–53. IEEE, 2000.
7. Lyndon S Kennedy and Daniel PW Ellis. Laughter detection in meetings. In *NIST ICASSP 2004 Meeting Recognition Workshop, Montreal*, pages 118–121. National Institute of Standards and Technology, 2004.
8. Uwe Kowalik, Terumasa Aoki, and Hiroshi Yasuda. Broaference–a next generation multimedia terminal providing direct feedback on audiences satisfaction level. In *Human-Computer Interaction-INTERACT 2005*, pages 974–977. Springer, 2005.

9. Gwen Littlewort, Jacob Whitehill, Tingfan Wu, Ian Fasel, Mark Frank, Javier Movellan, and Marian Bartlett. The computer expression recognition toolbox (cert). In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 298–305. IEEE, 2011.
10. Michael Lyons, Shota Akamatsu, Miyuki Kamachi, and Jiro Gyoba. Coding facial expressions with gabor wavelets. In *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, pages 200–205. IEEE, 1998.
11. S Moore and R Bowden. Local binary patterns for multi-view facial expression recognition. *Computer Vision and Image Understanding*, 115(4):541–558, 2011.
12. Timo Ojala, Matti Pietikäinen, and Topi Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(7):971–987, 2002.
13. Maja Pantic, Michel Valstar, Ron Rademaker, and Ludo Maat. Web-based database for facial expression analysis. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pages 5–pp. IEEE, 2005.
14. Stavros Petridis and Maja Pantic. Audiovisual discrimination between speech and laughter: Why and when visual information might help. *Multimedia, IEEE Transactions on*, 13(2):216–234, 2011.
15. Hugues Salamin, Anna Polychroniou, and Alessandro Vinciarelli. Automatic detection of laughter and fillers in spontaneous mobile phone conversations. In *Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on*, pages 4282–4287. IEEE, 2013.
16. Björn Schuller, Florian Eyben, and Gerhard Rigoll. Static and dynamic modelling for the recognition of non-verbal vocalisations in conversational speech. In *Perception in multimodal dialogue systems*, pages 99–110. Springer, 2008.
17. Mohamed Sehili, Fan Yang, V Leynaert, and L Devillers. A corpus of social interaction between nao and elderly people. In *5th International Workshop on Emotion, Social Signals, Sentiment & Linked Open Data (ES3LOD2014)*. LREC, 2014.
18. Yusuke Shinohara and Nobuyuki Otsu. Facial expression recognition using fisher weight maps. In *Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on*, pages 499–504. IEEE, 2004.
19. K Truong and D Van Leeuwen. Evaluating automatic laughter segmentation in meetings using acoustic and acoustics-phonetic features. In *Proc Proc Workshop on the Phonetics of Laughter at the 16th International Congress of Phonetic Sciences (ICPhS)*, pages 49–53, 2007.
20. Khiet P Truong and David A Van Leeuwen. Automatic discrimination between laughter and speech. *Speech Communication*, 49(2):144–158, 2007.
21. Michel Valstar and Maja Pantic. Induced disgust, happiness and surprise: an addition to the mmi facial expression database. In *Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect*, page 65, 2010.
22. Michel F Valstar, Maja Pantic, Zara Ambadar, and Jeffrey F Cohn. Spontaneous vs. posed facial behavior: automatic analysis of brow actions. In *Proceedings of the 8th international conference on Multimodal interfaces*, pages 162–170. ACM, 2006.
23. Paul Viola and Michael Jones. Robust real-time object detection. *International Journal of Computer Vision*, 4:51–52, 2001.
24. Guoying Zhao and Matti Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(6):915–928, 2007.