Multimodal Laughter Detection in Natural Discourses

Stefan Scherer, Friedhelm Schwenker, Nick Campbell, and Günther Palm

Abstract. This work focuses on the detection of laughter in natural multiparty discourses. For the given task features of two different modalities are used from unobtrusive sources, namely a room microphone and a 360 degree camera. A relatively novel approach using Echo State Networks (ESN) is utilized to achieve the task at hand. Among others, a possible application is the online detection of laughter in human robot interaction in order to enable the robot to react appropriately in a timely fashion towards human communication, since laughter is an important communication utility.

1 Introduction

Paralinguistic dialog elements, such as laughter, moans and back channeling, are important factors of human to human interaction besides direct communication using speech. They are essential to convey information such as agreement or disagreement in an efficient and natural way. Furthermore, laughter is an indication for the positive perception of a discourse element by the laughing dialog partner, or an indication for uncertainty considering nervous or social laughters [1]. Overall laughter is a very communicative element of discourses that is necessary for "healthy" communication and it can be used to measure engagement in interaction [9, 16, 8, 10]. Lively discourses are not only important for face to face communication, but as we believe essential for the acceptance of artificial agents, such as robots or expert systems providing information using speech synthesis and other mainly human

Stefan Scherer · Friedhelm Schwenker · Günther Palm
Institute of Neural Information Processing, Ulm University
e-mail: {stefan.scherer,friedhelm.schwenker,
guenther.palm}@uni-ulm.de

Nick Campbell
Center for Language and Communication Studies, Trinity College Dublin
e-mail: nick@tcd.ie

H. Ritter et al. (Eds.): Human Centered Robot Systems, COSMOS 6, pp. 111–120.
springerlink.com

communication modalities, e.g. gestures etc., to communicate with human dialog partners [15]. Furthermore, laughter is acoustically highly variable and is expressed in many forms, such as giggles, exhaled or inhaled laughs, or even snort like laughters exist. Therefore, it is suspected, that laughter is difficult to model and to detect [1, 17].

However, modeling laughter and thereby detecting laughter in natural discourses has been the topic of related research: in [8] one second large segments of speech are considered. For each of these segments the decision, whether somebody of the speakers laughed or not, is being made using Mel Frequency Cepstral Coefficients (MFCC) and Support Vector Machines (SVM). The recognition accuracy of this approach reached 87%. One of the obvious disadvantages of this approach is that segments of one second in length are used and therefore no accurate on- and offsets of the laughs can be detected.

Truong and Leeuwen [16, 17] first recognized laughter in previously segmented speech data taken from a manually labeled meeting corpus containing data from close head mounted microphones. They used Gaussian Mixture Models (GMM) and pitch, modulation spectrum, perceptual linear prediction (PLP) and energy related features. They achieved the best results of 13.4% equal error rate (EER) using PLP features on pre-segmented audio samples of an average length of 2 seconds for laughter and 2.2 seconds for speech. In their second approach [17] they extracted PLP features from 32 ms windows every 16 ms using three GMMs for modeling laughter, speech, and silence. Silence was included since it was a major part of the meeting data. An EER on segmenting the whole meeting of 10.9% was achieved. In future work they want to use HMMs to further improve their results. Their second approach allows a very accurate detection of laughter on- and offsets every 16 ms. However, it does not consider the, in [9] mentioned, average length of a laughter segment of around 200 ms since only 32 ms of speech are considered for each decision.

In [9] the same data set as in [16, 17] was used for laughter detection. In a final approach after narrowing down the sample rate of their feature extractor to a frequency of 100 Hz (a frame every 10 ms) a standard Multi Layer Perceptron (MLP) with one hidden layer was used. The input to the MLP was updated every 10 ms, however the input feature vector considered 750 ms including the 37 preceding and following frames of the current frame for which the decision is computed by the MLP. The extracted features include MFCCs and PLPs since they are perceptually scaled and were chosen in previous work. Using this approach an EER of around 7.9% was achieved.

In the current work Echo State Networks (ESN) are used to recognize laughter in a meeting corpus [2], comprising audio and video data for multimodal detection experiments. The approach utilizing ESNs, is making use of the sequential characteristics of the modulation spectrum features extracted from the audio data [7]. Furthermore, the features are extracted every 20 ms and comprise data of 200 ms in order to be able to give accurate on- and offset positions of laughter, but also to comprise around a whole “laughter syllable” in one frame [9]. In a second approach the video data containing primary features such as head and body movement are

incorporated into the ESN approach for a multimodal laughter detection method. One of the main goals of this work is to provide a classification approach that is only relying on unobtrusive recording gear, such as a centrally placed microphone and a 360 degrees camera, since it is particularly important to provide a natural environment for unbiased communication. However, this constraint only allows the extraction and analysis of basic features.

The remainder of this paper is organized as follows: Section 2 gives an introduction on the used data and explains the recording situation in detail, Section 3 describes the utilized features and the extraction algorithm, Section 4 comprises detailed descriptions of the approaches for classifying the data. Section 5 reports the obtained results of the single and multimodal approaches. Finally, Section 6 concludes the paper and summarizes the work.

2 Utilized Data

The data for this study consists of three 90 minutes multi-party conversations in which the participants originating from four different countries each speaking a different native language. However, the conversation was held in English. The conversations were not constrained by any plot or goal and the participants were allowed to move freely, which renders this data set very natural and therefore it is believed that the ecological validity of the data is very high. The meetings were recorded by using centrally positioned, unobtrusive audio and video recording devices. The audio stream was directly used as input for the feature extraction algorithm at a sample rate of 16 kHz. A 360 degree video capturing device was used for video recording and the standard Viola Jones algorithm was used to detect and track the faces of the participants throughout the 90 minutes. The resulting data has a sample rate of 10 Hz and comprises head and body activity [2]. In Figure 1 one of the 360 degree camera frames including face detection margins is seen. It is clear that only the heads and upper parts of the body are visible since the participants are seated around a table. However, hand gestures, head and body movements are recorded without any obstruction. Separate analysis has revealed high correlations of activity between body and head movement of active speakers, as well as it is expected that the coordinates of the head positions should move on a horizontal axis while the speaker produces a laughter burst, which could be a sequence learnt by the ESN.

The little constraints on the conversation provide very natural data including laughters, and other essential paralinguistic contents. The data was annotated manually and non-speech sounds, such as laughters or coughs were labeled using symbols indicating their type. Laughter including speech, such as a laughter at the end of an utterance, was labeled accordingly, but was not used as training data for the classifiers in order not to bias the models by the included speech. However, all the data was used in testing. For training a set of around 300 laughters containing no speech of an average length of 1.5 seconds and around 1000 speech samples of an average length of 2 seconds are used. A tenth of this pool of data was excluded from training

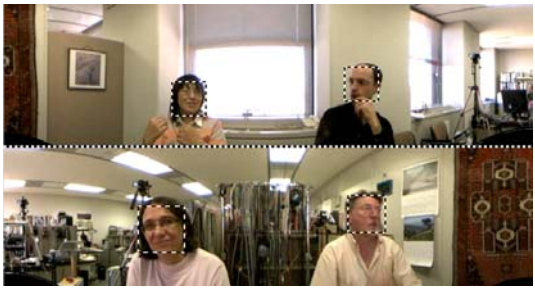

Fig. 1 A frame of the 360 degree camera positioned in the center of the conference table. The image includes the boundaries of the detected faces using the Viola Jones approach.

in each fold of the 10-fold cross validations, except in the experiments in which all the dialog data is presented, including all the laughs including speech in the labeled segment. Overall, laughter is present in about 5-8% of the data. This variance is due to the laughters including speech[1].

3 Features

As mentioned before the available data for the experiments is comprised of two different modalities. The conversations are available as audio and video files. Therefore, suitable features for the laughter detection task were extracted. In the following two paragraphs the extraction procedures are explained in some detail, for further information refer to the cited articles.

For the detection of laughter we extracted modulation spectrum features from the audio source [5]. These features have been used in previous experiments and tasks such as emotion recognition and laughter recognition [13, 16, 11]. The features are extracted using standard methods like Mel filtering and Fast Fourier Transformations (FFT). In short they represent the rate of change of frequency, since they are based on a two level Fourier transformation. Furthermore, they are biologically inspired since the data is split up into perceptually scaled bands. In our experiment we used 8 bands in the Mel filtering regarding frequencies up to 8 kHz. Since the audio is sampled at a rate of 16 kHz this satisfies the Nyquist theorem [5]. These slow temporal modulations of speech emulate the perception ability of the human auditory system. Earlier studies reported that the modulation frequency components from the range between 2 and 16 Hz, with dominant component at around 4 Hz, contain important linguistic information [4, 3]. Dominant components represent strong rate of change of the vocal tract shape.

[1] The data, and annotations are freely available on the web, but access requires a password and user name, which can be obtained from the authors on request, subject to conditions of confidentiality.

As mentioned before the video features were extracted from the 360 degree recordings of a centrally placed camera. The sample rate of the face tracking using the Viola Jones approach was 10 Hz and the provided data comprised coordinates of the faces at each frame composed by the exact spot of the top left corner and the bottom right corner of the surrounding box of the face as seen in Figure 1. However, these coordinates are highly dependent on the distance of the person to the camera and therefore relative movement data of the face and body were taken as input to the classifiers. The coordinates were normalized to movement data of a mean value of 0 and a standard deviation of 1. Therefore, the movement ranged from -1 to 1 for each tracked face and body individually.

4 Echo State Network Approach

For the experiments a relatively novel kind of recurrent neural networks (RNN) is used, the so called Echo state network (ESN) [7]. Among the advantages of an ESN over common RNNs are the stability towards noisy inputs [14] and the efficient method to adapt the weights of the network [6]. Using the direct pseudo inverse adaptation method the ESN is trained in a very efficient way. With regard to these advantages and considering the targeted application area of the network the ESN is a fitting candidate. In contrast to for example SVMs used in [8] for the detection of laughter the ESN incorporates previous features and states for the decision whether or not a laughter is present, rendering it an ideal approach for online detection tasks.

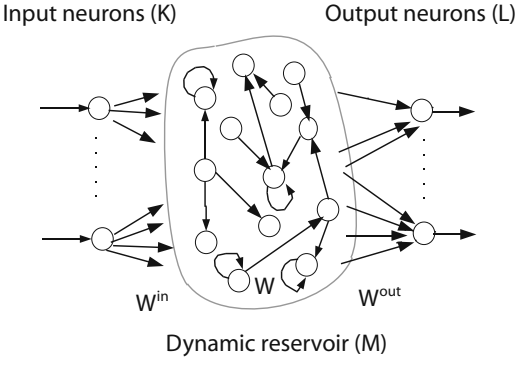

Fig. 2 Scheme of an Echo state network

In Figure 2 the scheme of an ESN is shown. The input layer K is fully connected to the dynamic reservoir M, and M is fully connected to the output layer L. ESNs are characterized by their dynamic memory that is realized by a sparsely interconnected reservoir M that is initialized randomly. The connection matrix is normalized to a so called spectral width parameter α guaranteeing that the activity within the dynamic reservoir is kept at a certain level. In order to train an ESN it is only necessary to

adapt the output weights W^{out} using the direct pseudo inverse method computing the optimal values for the weights from M to L by solving the linear equation system $W^{out} = M^{+}T$. The method minimizes the distance between the predicted output of the ESN and the target signal T. For the detailed algorithm refer to [6].

After training the ESN output is being post processed in order to avoid rapid shifts between states. Stability is being achieved by several smoothing steps as follows: First the output is smoothed using a fourth grade Butterworth filter with a cut off frequency rate of 0.3. If the output exceeds a threshold $\theta = 0.4$ it is counted as a hit and the output is set to 1, values below are set to 0. After generating this binary vector a Gaussian filter of a length of 25 is applied in order to get the outputs shown in the figures in the following section.

5 Experiments and Results

The basis architecture used are the ESNs introduced in Section 4. The ESN consist of a dynamic reservoir with 1500 neurons that are sparsely interconnected with each other. The probability for a connection between neuron x and y is 2%. Recursive connections are also set at the same probability. The last parameter of the ESN that has to be set is the spectral width influencing the dynamics of the reservoir. The spectral width α was set to 0.15.

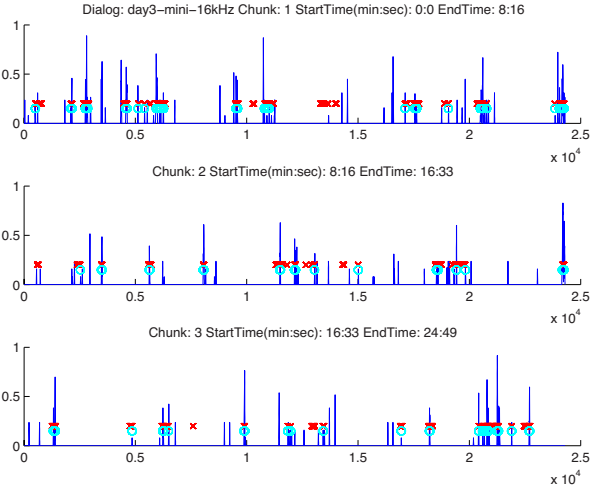

Fig. 3 Fusion Echo state network output after post processing. The blue line corresponds to the output, blue circles depict hits and red crosses correspond to human labels.

In a first experiment a 10-fold cross validation was conducted on the speech data and laughter comprising no speech. The ESN recognizes laughter on each frame provided by the feature extraction. The knowledge of on- and offsets of utterances or laughter provided by the labels is not utilized for the classification. An average error rate of around 13% was achieved by the ESN.

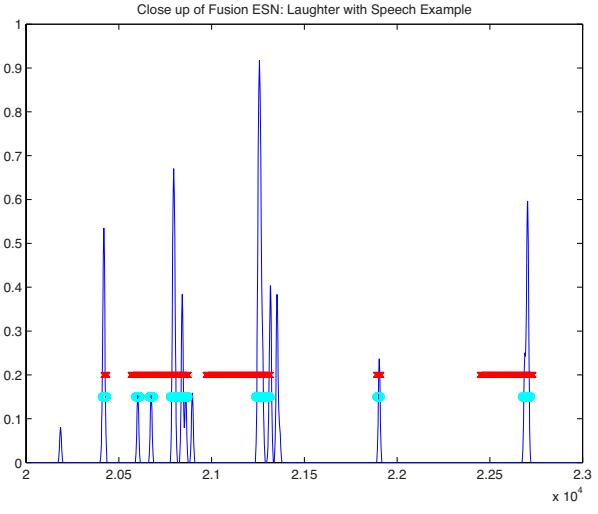

Fig. 4 Close up look to the Echo state network output after post processing. The blue line corresponds to the output, blue circles depict hits and red crosses correspond to human labels.

In a second series of a 10-fold cross validation the whole dialog is given as input to the ESN. The classification accuracy was again around 90%. An average misclassification rate of 10.5% was achieved over the 10 folds. The increase in accuracy can be explained as a result of overlapping training and test data. However, these percentages are biased as already mentioned before, since the ESN was only trained on laughter containing no speech and the whole dialog contains laughters that are labeled together with speech in some cases. Therefore, a more subjective view on the results is necessary. In Figure 3, the first three parts out of ten of the conversation are seen. Laughter labels are indicated by red crosses and the output of the ESN after post-processing as described in Section 4 is displayed in blue. The ESN clearly peaks most of the time at the labeled laughters. Only a few laughters are omitted and some are inserted. Furthermore, it is interesting that some labels are quite long, but only at the end or beginning the ESN peaks as in Figure 4 at around 22000 on the x axis. In this particular case the laughter in the conversation appears at the end of the utterance labeled by the human labelers. Therefore, the system could be used for post-processing of the manual labels and refine them.

In a the final experiment we made use of the available video data in order to test the networks performance to detect laughter in a multimodal approach. The available video data only comprises basic features such as head and body movement for each speaker and x and y coordinate changes of the head frame by frame. As input we made use of the body and head movement and normalized them for each speaker towards an average of 0 and variance 1. Using this approach we receive 8 dimensional features at a sampling rate of 10 Hz due to the output of the Viola Jones face recognition algorithm. In order to be able to use the same ESN architecture for the movement data we had to adapt the sampling rate of the data. This is done by simply memorizing the movement ESN output for 5 frames instead of only one. The final architecture incorporates the output of two separate ESNs in the two modalities in a weighted sum before post processing steps are taken. After training in the test phase the outputs are added using different weighting for each ESN. Thorough testing and considering the coarseness of the video data and the related bias, resulted in a weighting of 0.7 for the audio related ESN and a weight of 0.3 for the movement ESN. This fusion is then post-processed as the audio ESN output in the first two experiments. Using this fusion we obtain less false alarms and less misses. Therefore, the overall performance got better. However, the result might not be significantly better since the improvement only resulted in an error of 9%. Further, numerical studies need to be carried out to test statistical significance. In Figure 5 a comparison of the three ESN modalities speech, movement, and fusion is given. It is seen that the fusion output is less unstable and calmer in comparison to the other two. The output for the movement data is the most unstable output and resulted in around 18% error.

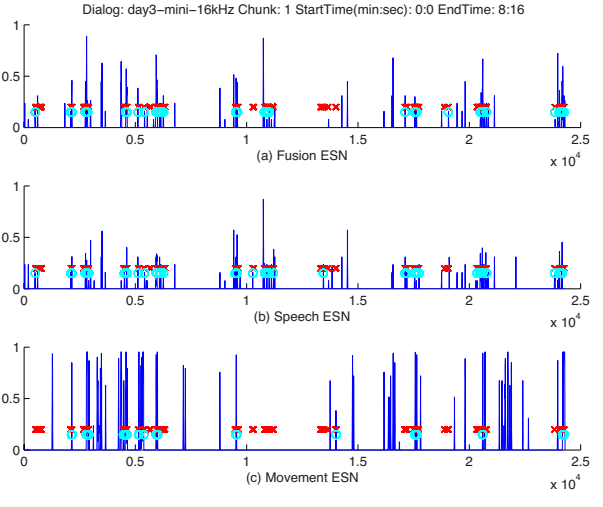

Fig. 5 Comparison between the three possible modalities: (a) fusion approach, (b) speech only and (c) movement only

6 Summary and Discussion

The work presented in this paper aims towards the online detection of laughter in speech. Multi-party natural discourse data was used as data, consisting of three videos of around 90 minutes in length. The conversations were not restricted by any constraints and were therefore as natural as possible, comprising all sorts of gestures, laughters, and noise sources such as drinking a cup of coffee or people entering the room. This naturalness of the conversation was supported further by unobtrusive recording devices placed centrally on the table. The participants themselves were not wired with any close up microphone nor intimidated by any camera facing them.

This multimodal data was then used in several recognition experiments. Results comparable to related work were achieved using multimodal data and ESNs. On the other hand the ESN takes the extracted features covering 200 ms of audio corresponding to the average length of a laughter syllable [9] directly as input. In a series of experiments in which the task was to detect laughter in a continuous stream of input the ESN compensates the lack of information by memorizing previous decisions and features and utilizing this information in order to predict upcoming events. The ESN approach single- and multimodal can be used as a tool for the online recognition of laughter. The detected laughter may then be used to directly control robots in human robot interaction scenarios. Since, a robot that may react appropriately in a timely fashion to laughs or smiles seems more natural than a robot that is incapable of reacting to such events in real time [12].

For future work we aim towards a method to measure "engagement" in discourse, which can be applied to measure the quality of interaction between humans as well as between a human and an artificial agent. Previously recorded data, such as in [15], will be labeled accordingly and an automatic recognition approach using multimodal data as input to ESN ensembles will be constructed. Since laughter is an indicator for the positive perception of a discourse element, we consider the detection of laughter an essential part aiming for the goal of measuring the quality of interaction, and the degree of interaction, as laughter is one of the main features to detect participant involvement [10]. An agent that keeps you engaged in a conversation will be perceived more positive than one that bores you.

Acknowledgements. This research would not have been possible without the generous funding and the kind support of the German Academic Exchange Agency (DAAD). Furthermore, the presented work was developed within the Transregional Collaborative Research Centre SFB/TRR 62 "Companion-Technology for Cognitive Technical Systems" funded by the German Research Foundation (DFG).

References

1. Campbell, N., Kashioka, H., Ohara, R.: No laughing matter. In: Proceedings of Interspeech, ISCA, pp. 465–468 (2005)
2. Campbell, W.N.: Tools and resources for visualising conversational-speech interaction. In: Proceedings of the Sixth International Language Resources and Evaluation (LREC 2008), ELRA, Marrakech, Morocco (2008)

3. Drullman, R., Festen, J., Plomp, R.: Effect of reducing slow temporal modulations on speech reception. Journal of the Acousic Society 95, 2670–2680 (1994)
4. Hermansky, H.: Auditory modeling in automatic recognition of speech. In: Proceedings of Keele Workshop (1996)
5. Hermansky, H.: The modulation spectrum in automatic recognition of speech. In: Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 140–147. IEEE, Los Alamitos (1997)
6. Jaeger, H.: Tutorial on training recurrent neural networks, covering bppt, rtrl, ekf and the echo state network approach. Tech. Rep. 159, Fraunhofer-Gesellschaft, St. Augustin Germany (2002)
7. Jaeger, H., Haas, H.: Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. Science 304, 78–80 (2004)
8. Kennedy, L., Ellis, D.: Laughter detection in meetings. In: Proceedings of NIST ICASSP, Meeting Recognition Workshop (2004)
9. Knox, M., Mirghafori, N.: Automatic laughter detection using neural networks. In: Proceedings of Interspeech 2007, ISCA, pp. 2973–2976 (2007)
10. Laskowski, K.: Modeling vocal interaction for text-independent detection of involvement hotspots in multi-party meetings. In: Proceedings of the 2nd IEEE/ISCA/ACL Workshop on Spoken Language Technology (SLT 2008), pp. 81–84 (2008)
11. Maganti, H.K., Scherer, S., Palm, G.: A novel feature for emotion recognition in voice based applications. In: Paiva, A.C.R., Prada, R., Picard, R.W. (eds.) ACII 2007. LNCS, vol. 4738, pp. 710–711. Springer, Heidelberg (2007)
12. Pugh, S.D.: Service with a smile: Emotional contagion in the service encounter. Academy of Management Journal 44, 1018–1027 (2001)
13. Scherer, S., Hofmann, H., Lampmann, M., Pfeil, M., Rhinow, S., Schwenker, F., Palm, G.: Emotion recognition from speech: Stress experiment. In: Proceedings of the Sixth International Language Resources and Evaluation (LREC 2008). European Language Resources Association (ELRA), Marrakech, Morocco (2008)
14. Scherer, S., Oubbati, M., Schwenker, F., Palm, G.: Real-time emotion recognition from speech using echo state networks. In: Prevost, L., Marinai, S., Schwenker, F. (eds.) ANNPR 2008. LNCS (LNAI), vol. 5064, pp. 205–216. Springer, Heidelberg (2008)
15. Strauss, P.M., Hoffmann, H., Scherer, S.: Evaluation and user acceptance of a dialogue system using wizard-of-oz recordings. In: 3rd IET International Conference on Intelligent Environments, IET, pp. 521–524 (2007)
16. Truong, K.P., Van Leeuwen, D.A.: Automatic detection of laughter. In: Proceedings of Interspeech, ISCA, pp. 485–488 (2005)
17. Truong, K.P., Van Leeuwen, D.A.: Evaluating laughter segmentation in meetings with acoustic and acoustic-phonetic features. In: Workshop on the Phonetics of Laughter, Saarbrücken, pp. 49–53 (2007)