Multi Modal Laughter Detection in Natural Discourses

Stefan Scherer, Friedhelm Schwenker, and Nick Campbell

University of Ulm, NiST, ATR

Abstract. ABSTRACT

Key words: Echo State Networks, Gaussian Mixture Model Super Vectors, Support Vector Machines, Laughter Detection, Natural Discourse

1 Introduction

Additional to direct communication using speech paralinguistic dialog elements, such as laughter, moans and back channeling are important factors of human human interaction. They are essential to convey information such as agreement or disagreement in an efficient way. Furthermore, laughter is an indication for the positive perception of a discourse element of the laughing dialog partner, or an indication for uncertainty considering nervous or social laughters. Overall laughter is a very communicative element of discourses that is necessary for "healthy" communication and it can be used to measure engagement in discourse [1–3]. Laughter is acoustically highly variable, e.g. giggles, exhaled or inhaled laughs, or even snort like laughters exist. Therefore, it is suspected, that laughter is difficult to model [4]

However, modeling laughter and thereby detecting laughter in natural discourses has been the topic of several related work: in [3] one second large segments of speech are considered, whether somebody of the speakers laughed or not within that second 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 no accurate on- and offsets of the laughs can be detected.

Truong and Leeuwen [2, 4] 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 [4] 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. The EER on segmenting the whole meeting of 10.9%.

In future work they want to use HMMs to improve their results. This approach allows a very accurate detection of laughter on- and offsets every 16 ms. However, it does not consider the in [1] mentioned average length of a laughter segment of around 200 ms only deciding on 32 ms of speech.

In [1] the same data set as in [2, 4] 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 two different approaches are used to recognize laughter in a meeting corpus [**?**], comprising audio and video data for multi modal detection experiments. The first approach is using so called GMM super vectors [8] as input for a standard SVM classifier [5]. The super vectors are generated out of previously extracted modulation spectrum audio features. In the second approach Echo State Networks (ESN) [6], making use of the sequential characteristics of the modulation spectrum features that are scaled perceptually as MFCCs, are used. 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 [1]. In a third and final 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.

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, Sect. 3 describes the utilized features and the extraction algorithm, Sect. 4 comprises detailed descriptions of the approaches for classifying the data. Section 5 reports the obtained results of the single and multimodal approaches. Finally, Sect. 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 were 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. 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 and the sample rate was 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 [**?**]. In Fig. **??** 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 obstructions.

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 in each fold of the 10-fold cross validations. Except in the experiments in which all the dialog data is presented, comprising 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.

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.

3 Features

4 Classification Setups

In this section two relatively novel approaches towards recognizing laughter in speech are presented. The first method uses SVMs as a classifier and so called GMM super vectors. The novelty of this approach are the GMM super vectors, which are high dimensional representations generated by an universal background model (UBM) and its adaptation by an analyzed input (in this case an utterance or a frame of extracted audio features) using the maximum a posteriori (MAP) adaptation method. The approach is described in more detail in the following section.

The second approach uses low dimensional (8 dimensions) inputs to a recursive neural network (RNN) called Echo State Network (ESN). The ESN consists mainly of a large number of sparsely interconnected neurons in a dynamic reservoir and is trained efficiently using the direct pseudo inverse function to adapt the weights from the reservoir towards the output layer. The ESN is capable due to its sequential characteristics of using the past few inputs to predict upcoming events. A more detailed description of the ESN is given in Sect. 4.2.

4.1 Support Vector Machine GMM Super Vector Approach

The first step in order to obtain the super vectors for the SVM is to generate the GMM UBM. In general the UBM is a speaker, emotion, and speech independent representation of all the training data used. The UBM additionally does not contain any temporal information. It represents the general structure of the data. In order to construct the UBM the standard GMM approach is used. The estimation of the probability density function represented by a linear combination of Gaussian basis functions, a so called Mixture Model, is done using the expectation maximization approach [**?**]. After maximization the GMMs form an abstraction of the input data [8, **?**]. Once this abstraction is obtained, the GMM UBM can be used to generate the super vectors using the MAP adaptation. The super vector then is the concatenation of all the means of the GMMs in the adapted UBM. This process is represented schematically in Fig. 1. The MAP adaptation is viewed as "shifting the GMM UBM space" towards the presented utterance or feature frame, respectively. The resulting new means of the GMMs are then used to represent the utterance. The adaptation is done in three major steps [11]:

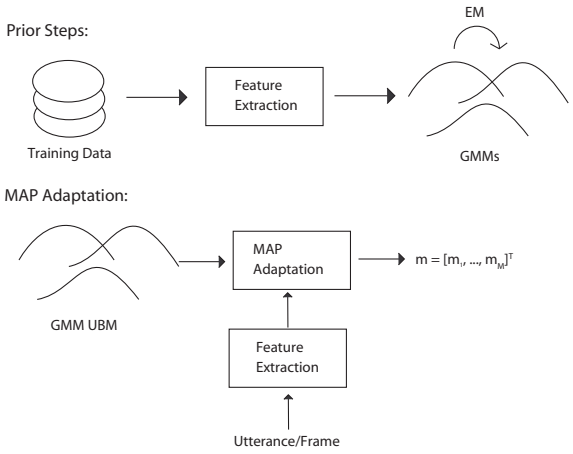

Fig. 1.

– Get probability that GMM component m is producing utterance x ($x = [x_1, ..., x_T]$): $n_m = \sum_{t=1}^{T} P(m|x_t)$

– Calculate the new mean estimate and covariance estimate for the presented input $E_m(x) + E_m(x^2)$.
– Adapt the parameters of the UBM to represent the current input using α_m as a weight to balance the adaptation according to the a posteriori probability $\alpha_m = \frac{n_m}{n_m+r}$, with r being a scalar value [12].

The super vectors of the training data are then used as input to a standard SVM. After training newly presented and accordingly adapted feature vectors are used as input for the SVM in order to be classified for testing. For a detailed description on how SVMs work, please refer to [5].

4.2 Echo State Network Approach

5 Experiments and Results

In this section the achieved results and the conducted experiments are reported. Furthermore, an analysis of the results is given. The goal of this section is to compare the two different approaches and to discuss their advantages and disadvantages, respectively. First the results of the GMM super vector approach using a standard SVM as a classifier are mentioned followed by the ESN approaches, both single- and multimodal.

5.1 GMM Super Vectors Results

In the first experiment, using the GMM super vectors as input to a standard binary classification SVM with L2-soft margin penalization [7], we formed a UBM for laughters containing no speech and randomly selected speech samples of an average length of 2 seconds in each fold of a 10-fold cross validation. Nine parts of the data were used as training and one part for testing. First, as described in Sect. 4.1, a UBM has to be generated using the training data. The training data has of corse been processed according to Sect. 3 in advance. Next the SVM is trained on the concatenated means of 20 GMMs after the UBM has been adapted using the MAP adaptation for each utterance in the training set. This results in a 160 dimensional vector since 20 GMM components are used and each of them possesses 8 dimensions according to the dimensions of the modulation spectrum features.

For testing, the test utterances are fed into the MAP adaptation algorithm in order to get concatenated mean vectors, which are then classified by the trained SVM. For each utterance only one vector is produced as mentioned in Sect. 4.1. Therefore, it is clear that segmented parts of the audio have to be categorized in advance making an online detection impossible. However, the recognition results were very accurate and resulted in as little as an average error over the 10-folds of 4%.

In a second experiment in a 10-fold cross validation 10 different UBMs were generated using only laughter segments containing no speech and randomly selected speech samples for training. For testing however, the whole conversation

containing laughters labeled as such comprising speech elements, which are either preceding, surrounding, or following a laughter in the segment, was input to the classifier to give frame-wise decisions on all the data. This experiment resulted as expected in a much higher error rate, over the 10 folds an average error of 28% was the result. The SVM clearly struggles to give correct predictions when presented with only one frame at a time and loses a lot of its accuracy in an online experiment. This could be a result of the fact that the SVM does not take any dynamics between following and preceding feature vectors into account for its decisions. In the following section a completely different approach is utilized to detect laughter in continuous speech.

5.2 Echo State Network Results

A second set of experiments was conducted using a RNN architecture, taking advantage of sequential dynamic characteristics of the features as well as the nature of conversations. The basis architecture used are the ESNs introduced in Sect. 4.2. The ESN consisted 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.

In a first experiment we conducted a 10-fold cross validation on the speech data and laughter comprising no speech. As mentioned before one tenth was kept for testing and 9 were used for training. The ESN, in contrast to the SVM in the first experiment, 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 misclassification rate of around 13% was achieved by the ESN.

In a second series of a 10-fold cross validation again 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 including speech. Therefore, a more subjective view on the results is necessary. In Fig. **??**, 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 Sect. 4.2 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 line XXXX at around ZZZZ of Fig. **??**. In this particular case the laughter in the dialog 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. For a comparison between the output before and after the post-processing step please

refer to Fig. **??**. Here the red line is the binary target speech vs. laughter, the blue line corresponds to the ESN output after smoothing using a median filter and the green crosses correspond to the binary decisions received by taking the sign of the output into account.

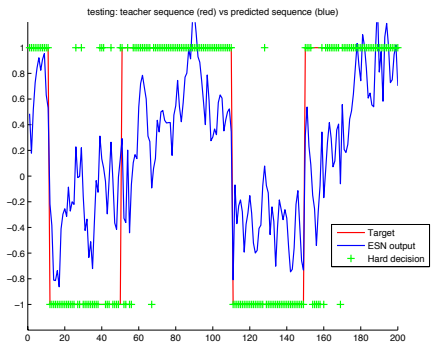

Fig. 2.

In a third and 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. Since, we wanted to keep the system as universal as possible we omitted the x and y pixel shifts, since they are sensitive to changes like the position of the camera etc. We only 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 can be seen in Fig. **??**. Where two separate ESNs are trained in the two modalities. After training in the test phase the outputs are added using different weighting for each ESN. Thorough testing 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 was not significantly better since the improvement only resulted in an error of 9%. In Fig. **??** 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.

Overall the SVM with the GMM super vectors outperforms the ESN in the closed training and test experiment. However, it is inapt for an online classification since it only makes use of one frame at a time, but a target that is that variable as laughter may not be possible to detect correctly using this approach. The ESN on the other side is a RNN and memorizes previous states and is able to model dynamics. Therefore, it is capable of online detection and in bootstrapping tasks to label unlabeled or partially labeled conversations.

6 Summary and Discussion

References

1. Knox, M., Mirghafori, N.: Automatic laughter detection using neural networks. Proc. of Interspeech. pp. 2973-2976. ISCA, (2007)
2. Truong, K. P., Van Leeuwen, D. A.: Automatic detection of laughter. Proc. of Interspeech. pp. 485-488. (2005)
3. Kennedy, L., Ellis, D.: Laughter detection in meetings, Proc. of NIST ICASSP, Meeting Recognition Workshop. (2004)
4. Truong, K. P., Van Leeuwen, D. A.: Evaluating laughter segmentation in meetings with acoustic and acoustic-phonetic features. Workshop on the Phonetics of Laughter. Saarbrücken, (2007)
5. Bishop, C. M.: Pattern Recognition and Machine Learning. Springer, (2006)
6. Jaeger, H., Haas, H.: Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. Sience 304. pp, 78-80. (2004)
7. Rossi, F., Villa, N.: Support vector machine for functional data classification. Neurocomputing, 69, issues 7-9. pp. 730-742. Elsevier, (2006)
8. Zhao, J., et al.: Advances in SVM-based system using GMM super vectors for text-indipendent speaker verification. Tsinghua Science and Technology, 13, issue 4. pp. 522-527. (2008)
9. Mao, X., Zhang, B., Luo, Y.: Speech emotion recognition based on a hybrid of HMM/ANN. Proc. of AIC. pp. 367-370. WSEAS, (2007)
10. Bilmes, J. A.: A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. Technical report International Computer Science Institute and Computer Science Division, Department of Electrical Engineering and Computer Science, U.C. Berkeley. (1998)
11. Bimbot, F., et al.: A tutorial on text-independent speaker verification. EURASIP Journal on Applied Signal Processing. pp. 430-451. (2004)
12. Kandovan, R. S., Lashkari, M. R. K., Etemad, Y.: Optimization of speaker verification using adapted Gaussian mixture models for high quality databases. Proc. of SPPR. pp. 264-268. ACTA Press, (2007)