Automatic Detection of Laughter

Khiet P. Truong and David A. van Leeuwen

Department of Human Interfaces
TNO Human Factors, Soesterberg, The Netherlands
{khiet.truong, david.vanleeuwen}@tno.nl

Abstract

In the context of detecting 'paralinguistic events' with the aim to make classification of the speaker's emotional state possible, a detector was developed for one of the most obvious 'paralinguistic events', namely laughter. Gaussian Mixture Models were trained with Perceptual Linear Prediction features, pitch&energy, pitch&voicing and modulation spectrum features to model laughter and speech. Data from the ICSI Meeting Corpus and the Dutch CGN corpus were used for our classification experiments. The results showed that Gaussian Mixture Models trained with Perceptual Linear Prediction features performed best with Equal Error Rates ranging from 7.1%-20.0%.

1. Introduction

Traditional speech technologies have always concentrated on extracting linguistic content from the speech signal (*what* is said) without focusing on the emotional content (*how* it is said). However, in human-machine interaction it becomes more and more important and useful to identify the speaker's emotional state. Knowing the speaker's emotional state contributes to the naturalness of human-machine communication processes. For instance, emotion recognition can be important for Interactive Voice Response Systems (IVR) with specific applications to call centers [1]: impatient or frustrated customers require a more appropriate dialogue handling and angry customers should be automatically routed to human operators. Emotion recognition is also useful in the field of multimedia retrieval or video summarization [2] and in automatic meeting transcriptions [3].

The speaker's emotional state expresses itself in speech through paralinguistic features such as pitch, speaking rate, voice quality etc. For example, Nwe et al. [4] report on studies that have shown that speaking rate is higher for the state *anger* than for *sadness*. In our research, we concentrate on audible, identifiable cues in the audio signal that are characteristic for a particular emotional state or mood. Some examples of these cues are laughter which is characteristic for joy or a humorous state, raised voice which is characteristic for anger, and trembling voice which is characteristic for nervousness. We will refer to such cues as 'paralinguistic events'. Our goal is to automatically detect these 'paralinguistic events' with the aim to make classification of the speaker's emotional state or mood possible.

In search of a suitable emotional speech database with paralinguistic annotations we found that laughter was one of the most often annotated paralinguistic events which occurred relatively frequently in recorded natural speech. On the basis of these observations we decided to focus on the automatic detection of laughter. Several studies have focused on the characteristics of laughter [5, 6, 7] and on automatic detection of laughter [2, 3]. Bachorowski et al. [5] and Trouvain [6] both found that laughter is a highly variable and complex signal whose characteristics are not yet unveiled. They found that there are many different types of laughter: voiced, unvoiced, song-like, grunt-like etc., and that although some aspects of laughter resemble speech, there are some notable differences between the two signals. For instance, compared to speech, laughs have longer unvoiced portions than voiced portions [7].

Cai et al. [2] have attempted to locate laughter events in entertainment and sports videos. They modeled laughter with Hidden Markov Models (HMM) in combination with Mel-Frequency Cepstral Coefficients (MFCCs) and perceptual features such as short-time energy and zero crossing rate. Another attempt to automatic laughter detection was made by Kennedy & Ellis [3]. They used a Support Vector Machine classifier trained with MFCCs and their deltas, spatial cues and modulation spectra to detect laughter events in meetings.

In this study we examine the use of Perceptual Linear Prediction (PLP) features, pitch&energy features, global pitch&voicing features and modulation spectrum features in Gaussian Mixture Models (GMMs) with the goal to automatically discriminate laughter from speech, thus to automatically detect laughter. PLP features were used to capture the perceptual and spectral characteristics of laughter. Pitch and energy are popular features in emotion recognition and since Bachorowski et al. [5] report on higher pitch levels for laughter than for speech, we also employed pitch&energy features in our study. Statistics of pitch and voicing features were used to capture global pitch information and information on the degree of voicing. Finally, we tried to model the repetitive syllable sounds in laughter by exploring the use of the modulation spectrum.

We will describe in Section 2 the speech data that we used in our classification experiments. The modelling technique and the features used to develop the classifier are discussed in section 3. In section 4 we describe the classification experiments and their results. Finally, in section 5 we discuss the results and draw conclusions.

2. Data

Emotional speech databases that contain realistic natural data are sparse. Most studies use speech databases that contain elicited emotional speech from hired actors. For realistic results, we would like to use a spontaneous speech database that has some paralinguistic or emotional tagging included. For training and testing our classifiers, we decided to use the ICSI Meeting Recorder Corpus [8] since it met our requirements: the corpus contains text-independent, speaker-independent realistic, natural speech data and it contains human-made annotations of non-lexical vocalized sounds including laughter, heavy breath sounds, coughs etc. The corpus consists of 75 recorded meetings with an average of 6 participants per meeting and a total

of 53 unique speakers. Each partipant wore a head-mounted microphone and additionally, six tabletop microphones simultaneously recorded the audio. For our experiments, we used the audio recorded with the head-mounted microphones. The data was divided in training and test sets: the first 26 ICSI 'Bmr' subset recordings were used for training and the last 3 ICSI 'Bmr' recordings were used for testing (these are the same data sets as used in Kennedy & Ellis [3]). The training and test sets contained speech from 16 and 11 speakers respectively. Additionally, 4 'Bed' subset recordings were used as test set to avoid biased results caused by overlap between speaker identities in the training and test material. Furthermore, as an independent test set we used spontaneous laughter and speech data from the Dutch CGN corpus [9] which is recorded on a different location and under different acoustic conditions than the ICSI corpus (note that the speech is from a different language and that some studies report on the existence of culture- and/or language specific paralinguistic patterns in vocal emotion expression). Testing on this independent data set would yield the most realistic results. Laughter segments from this corpus were selected by listening to a set of annotated non-speech sounds.

The experiments were conducted on *presegmented* laughter and speech segments (determination of onset and offset was not part of the task of the classifier) that were extracted from the speech signal. Laughter segments were in the first place determined from laughter annotations in the human-made transcriptions. After closer examination of some of these annotated laughter segments in the ICSI corpus, it appeared that not all of them were suitable for our classification experiments: for example, some of the annotated laughs co-occurred with speech and sometimes the laughter was not even audible. These non-suitable 'bad' laughter segments were later discarded (by 1 person who listened to the laughter segments) from the training and test data and new models were trained and tested on this selected data. Furthermore, no distinctions were made between different types of laughter, e.g. voiced, unvoiced, 'snort-like' laughter [5, 6]. Speech segments were also determined from the orthographic transcriptions: segments that did not contain any non-lexical vocalized sounds were labeled as speech.

In total we used 3264 speech segments with a total duration of 110 minutes (mean = 2.02s, sd = 1.87s) and 5917 laughter segments with a total duration of 218 minutes (mean = 2.21s, sd = 1.79s) (for more details, see Table 1).

	Training	**Test**		
	Bmr	**Bmr**	**Bed**	**CGN**
	min/N	**min/**N	**min/**N	**min/**N
Speech	81/2422	10/300	15/378	4/164
Laughter	177/4655	18/467	19/614	-/-
Selected Laughter	83/2680	10/279	11/444	4/171

Table 1: *Amount of laughter and speech data used in experiments, min=minutes, N=number of segments, where 'selected' means: removal of 'bad' laughter segments.*

3. Method

3.1. Modelling technique

Gaussian Mixture Models were used to train a laughter and a speech model. The models are trained using varying numbers of Gaussian components (varying from 2 - 256) depending on the available extracted features and 5 iterations of the Expectation Maximization (EM) algorithm. In testing, a maximum likelihood criterion was used. A 'soft detector' score is obtained by determining the likelihood ratio of the speech data given the 'laughter' and 'speech' GMMs respectively.

3.2. Features

3.2.1. Perceptual Linear Prediction features

Perceptual Linear Prediction features were used to train the GMMs. PLP coding is similar to Linear Predictive Coding (LPC) analysis based on the short-term spectrum of speech with the advantage that PLP is more consistent with human hearing [10]. PLP modifies the short-term spectrum of the speech by several psychophysically based transformations. For each frame, 13 PLP coefficients were computed with a forwardshift of 0.016s. Additionally, delta features were determined by calculating the deltas of the PLP coefficients (by linear repression over 5 consecutive frames) which resulted in a total of 26 features.

3.2.2. Pitch and energy

Other features that are often used in emotion recognition research are prosodic features, such as pitch. Studies also mention energy among the most useful features for emotion recognition. Both of these features were examined as well. For each frame with a shift of 0.01s, pitch and RMS energy were measured using Praat [11]. The deltas of pitch and RMS energy were calculated as well which resulted in a total of 4 features.

3.2.3. Global pitch and voicing-related features

In addition to pitch measurements per frame, we examined the use of more global pitch features such as mean and standard deviation of pitch, pitch excursion (maximum pitch−minumum pitch) and the averaged local variability in pitch (mean absolute slope of pitch). Bickley & Hunnicutt [7] found that the ratio of unvoiced to voiced frames is greater in laughter than in speech. Therefore, we also calculated the fraction of locally unvoiced frames and the degree of voice breaks, which is the total duration of the breaks between the voiced parts of the signal divided by the total duration of the analysed part of the signal. A total of 6 global features per segment were calculated using Praat [11].

3.2.4. Modulation spectrum

We tried to capture the rhythm and the repetitive syllable sounds of laughter, which may differ from speech, with the help of the modulation spectrum. The modulation spectra of speech and laughter were calculated by first obtaining the amplitude envelope via a Hilbert transformation. The envelope was further low-pass filtered and downsampled. The power spectrum of the envelope is then calculated and the first 16 spectral coefficients (modulation spectrum range up to 25.6 Hz) are used as features.

4. Experiments and results

The performances of the classifiers, each trained with different feature sets (PLP, pitch&energy, pitch&voicing and modulation spectrum features) were evaluated by testing them on the 'Bmr029', 'Bmr030' and 'Bm031' subsets of the ICSI Corpus. We used the same data sets as in Kennedy & Ellis [3] to make a better comparison between the results possible. In addition to the 3 'Bmr' subsets, we applied the laughter detection model to

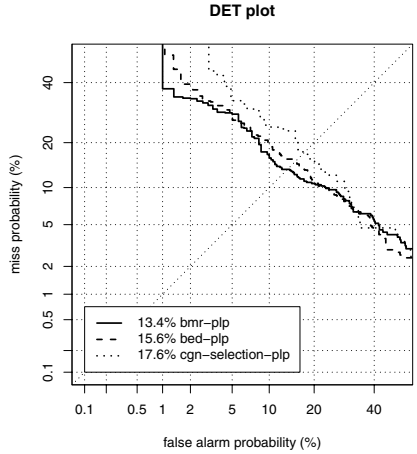

Figure 1: *Results of GMM laughter detection model trained with PLP features (256 Gaussians), trained with Bmr data and tested on Bmr (bmr-plp), Bed (bed-plp) and selected CGN (cgn-selection-plp) data, where 'selection' means: removal of 'bad' laughter segments.*

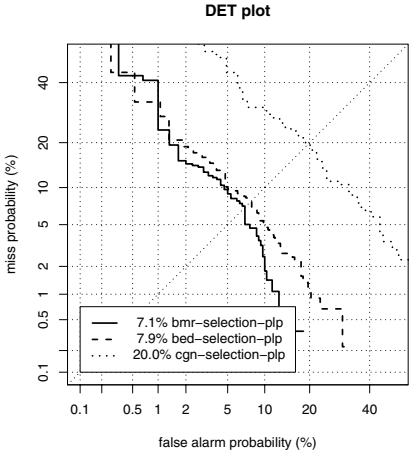

Figure 2: *Results of GMM classifier trained with PLP features (256 Gaussians), trained with selection of Bmr and tested on selection of Bmr (bmr-selection-plp), Bed (bed-selection-plp) and CGN (cgn-selection-plp) data.*

4 'Bed' subsets of the ICSI corpus, which contained speakers that were not present in the training data. And for even more realistic results, we applied the model to data from the Dutch CGN corpus. Fig. 1 shows performances of the laughter detection model trained with PLP features, applied to 3 ICSI 'Bmr' sets (*bmr-plp*), 4 ICSI 'Bed' sets (*bed-plp*) and a small selection of CGN data (*cgn-selection-plp*).

We can observe in Fig. 1 that the performance of the classifier decreases as the dissimilarity between training and test data increases. However, from Fig. 1 it appears that our loss of performance caused by dissimilarities between training and test data is not as large as was reported in a previous study on laughter detection [3]: we obtain Equal Error Rates (EERs) of 13.4%, 15.5% and 17.6% on 'Bmr', 'Bed', and CGN data respectively.

The GMM laughter and speech models were also trained and tested on a selection of the data where 'bad' laughter segments (that contained speech or inaudible laughter) were discarded. Expectedly, the performance of the classifier increased (compare Fig. 1 to Fig. 2) because the data contained less 'noise' after selection. However, in this case, the dissimilarities between training ('Bmr') and test set ('Bed', CGN) did lead to a considerable loss in performance for the CGN test set but not for the 'Bed' test set (see Fig. 2). There are pros and cons to such a clearer laughter detection model; the choice for it depends on the type of task or application that the laughter detection model will be used for.

In the subsequent classification experiments we decided to use the selected laughter and speech material for training and testing. Fig. 3 and Fig. 4 show results that were achieved with a laughter and speech model trained with pitch&energy and pitch&voicing features respectively. Although there are less data points available for training (the pitch&voicing features are extracted per segment and not per frame), the models trained with global pitch&voicing features (EERs 19.0% - 37.3%) perform better than the models trained with pitch&energy features (EERs 23.2% - 59.7%). Fig. 5 shows that the modulation spectrum features do not perform well (EERs 37.7% - 44.5%) in discriminating laughter from speech: the use of these features requires further investigation.

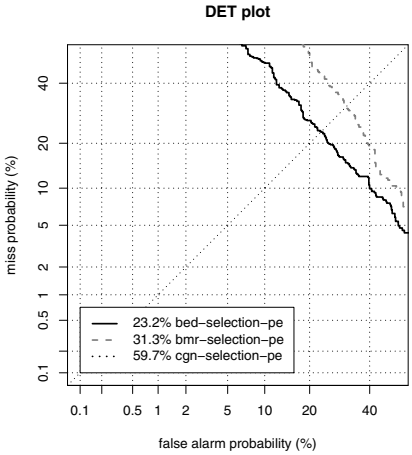

Figure 3: *Results of GMM classifier trained with pitch&energy features (2 Gaussians), trained with selection of Bmr and tested on selection of Bmr (bmr-selection-pe), Bed (bed-selection-pe) and CGN (cgn-selection-pe) data.*

5. Discussion and conclusions

We can conclude that GMMs trained with spectral PLP features outperform other GMMs trained with pitch&energy, pitch&voicing and modulation spectrum features in automatic detection of laughter. Moreover, compared to the other features that we have tried, PLP features are relatively robust when the models are applied to data that was recorded on a different location (CGN data). The models trained with pitch&voicing features measured globally per segment perform better than pitch&energy features measured per frame;

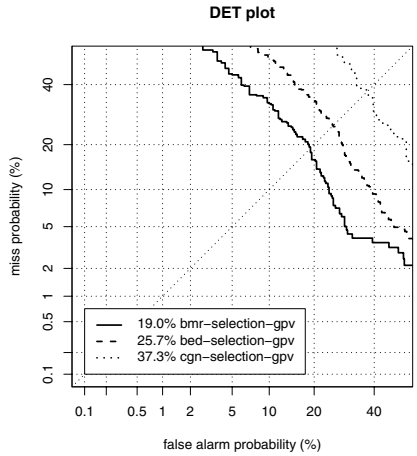

Figure 4: *Results of GMM classifier trained with global pitch and voicing-related features (4 Gaussians), trained with selection of Bmr and tested on selection of Bmr (bmr-selection-gpv), Bed (bed-selection-gpv) and CGN (cgn-selection-gpv) data.*

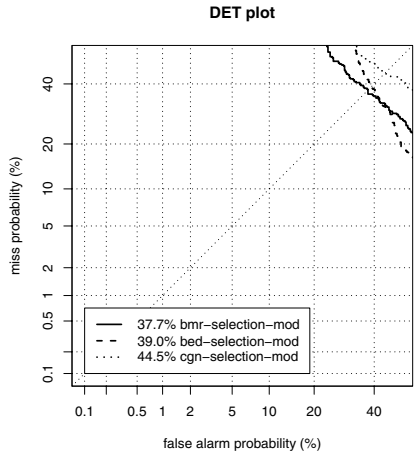

Figure 5: *Results of GMM classifier trained with modulation spectrum features (4 Gaussians), trained with selection of Bmr and tested on selection of Bmr (bmr-selection-mod), Bed (bed-selection-mod) and CGN (cgn-selection-mod) data.*

with only 6 global measurements per segment (as opposed to 4 pitch&energy measurements per frame per segment and 13+deltas PLP features per frame per segment), we may infer that pitch&voicing features are relatively strong features for discrimination between laughter and speech. For instance, one of the extracted pitch&voicing features is the fraction of unvoiced frames which appears to be larger for laughter (mean = 0.62, sd = 0.20) than for speech (mean = 0.38, sd = 0.16); this was also concluded in Bickley & Hunnicutt [7]. Further research is needed to optimize the use of these promising pitch&voicing features and to examine whether for example, a combination of PLP features and global pitch&voicing features would improve detection accuracy. Optimization is also needed for the Gaussian Mixture Models trained with modulation spectrum features (e.g. other choice of modulation spectrum features).

Another suggestion for future research is to develop a laughter detection model that is also able to determine the beginning and end of laughter. So far, we have used presegmented data to detect laughter. A laughter detection model that also provides an automatic time alignment of laughter is a more complex task that gives rise to additional problems such as: how do we decide when a laughter starts or ends and how do we evaluate the performance of such a detection model? These are typical problems that can be addressed within an HMM framework.

We have shown that it is possible to automatically distinguish human laughter from speech. Laughter is only one example of paralinguistic information that can be extracted from the speech signal. In the future, we hope to use similar methods as described in this paper for automatic detection of other paralinguistic events to make classification of emotion in speech possible.

6. Acknowledgements

We would like to thank Rob Drullman for his help on calculating the modulation spectrum and the other colleagues at TNO Human Factors for their useful comments on this paper. This research was supported by MultimediaN (Multimedia Netherlands), a BSIK-project that involves the knowledge creation and transfer on handling of video, pictures, audio, and language in information communication technologies (http://www.multimedian.nl).

7. References

[1] Yacoub, S., Simske, S., Lin, X. and Burns, J., "Recognition of Emotions in Interactive Voice Response Systems", in Proc. Eurospeech, Geneva, Switzerland, 2003.

[2] Cai, R., Lu, L., Zhang, H.J. and Cai, L.H., "Highlight sound effects detection in audio stream", in Proc. Intern. Confer. on Multimedia and Expo, Baltimore, 2003.

[3] Kennedy, L.S. and Ellis, D.P.W., "Laughter detection in meetings", NIST ICASSP 2004 Meeting Recognition Workshop, Montreal, 2004.

[4] Nwe, T.L., Foo, S.W. and De Silva, L.C., "Speech emotion recognition using hidden Markov models", Speech Communication 41, 603-623, 2003.

[5] Bachorowski, J.-A., Smoski, M.J., and Owren, M.J., "The acoustic features of human laughter", J. Acoust. Soc. Amer., Vol. 110 (3), 1581-1597, 2001.

[6] Trouvain, J., "Segmenting phonetic units in laughter", in Proc. ICPhS, Barcelona, 2003.

[7] Bickley, C. and Hunnicutt, S., "Acoustic analysis of laughter", in Proc. ICSLP, Banff, 927-930, 1992.

[8] Morgan, N., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Janin, A., Pfau, T., Shriberg, E. and Stolcke, A., "The Meeting Project at ICSI", in Proc. Human Language Technologies Conference 2001, San Diego, 2001.

[9] Oostdijk, N., "The Spoken Dutch Corpus. Overview and first evaluation", in Proc. LREC, 887-894, 2000.

[10] Hermansky, H., "Perceptual Linear Predictive (PLP) analysis of speech", J. Acoust. Soc. Amer., Vol. 87 (4), 1738-1752, 1990.

[11] Boersma, P. and Weenink, D., Praat: doing phonetics by computer (Version 4.3.01) [Computer program]. Retrieved from http://www.praat.org/, 2005.