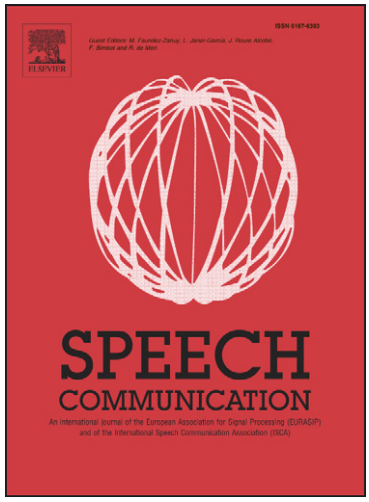
Automatic discrimination between laughter and speech

Khiet P. Truong, David A. van Leeuwen



Automatic discrimination between laughter and speech

Khiet P. Truong, David A. van Leeuwen

PII: S0167-6393(07)00002-7
DOI: 10.1016/j.specom.2007.01.001
Reference: SPECOM 1602

To appear in: *Speech Communication*

Received Date: 18 October 2005
Revised Date: 17 November 2006
Accepted Date: 4 January 2007

Please cite this article as: Truong, K.P., van Leeuwen, D.A., Automatic discrimination between laughter and speech, *Speech Communication* (2007), doi: 10.1016/j.specom.2007.01.001

Automatic Discrimination Between Laughter And Speech

Khiet P. Truong *, David A. van Leeuwen

TNO Human Factors, Department of Human Interfaces, P.O. Box 23, 3769 ZG Soesterberg, The Netherlands

Abstract

Emotions can be recognized by audible paralinguistic cues in speech. By detecting these paralinguistic cues that can consist of laughter, a trembling voice, coughs etc., information about the speaker's state and emotion can be revealed. This paper describes the development of a gender-independent laugh detector with the aim to enable automatic emotion recognition. Different types of features (spectral, prosodic) for laughter detection were investigated using different classification techniques (Gaussian Mixture Models, Support Vector Machines, Multi Layer Perceptron) often used in language and speaker recognition. Classification experiments were carried out with short pre-segmented speech and laughter segments extracted from the ICSI Meeting Recorder Corpus (with a mean duration of approximately two seconds). Equal error rates of around 3% were obtained when tested on speaker-independent speech data. We found that a fusion between classifiers based on Gaussian Mixture Models and classifiers based on Support Vector Machines increases discriminative power. We also found that a fusion between classifiers that use spectral features and classifiers that use prosodic information on pitch statistics and voicing/unvoicing patterns usually increases the performance for discrimination between laughter and speech. Our acoustic measurements showed differences between laughter and speech in mean pitch and in the ratio of the durations of unvoiced to voiced portions, which indicate that these prosodic features are indeed useful for discrimination between laughter and speech.

Key words: Automatic detection laughter, Automatic detection emotion

* Corresponding author. Tel.: +31 346 356 339; fax: +31 346 353 977.
Email address: khiet.truong@tno.nl (Khiet P. Truong).

1 Introduction

Researchers have become more and more interested in automatic recognition of human emotion. Nowadays, different types of useful applications employ emotion recognition for various purposes. For instance, knowing the speaker's emotional state can contribute to the naturalness of human-machine communication in spoken dialogue systems. It can be useful for an automated Interactive Voice Response (IVR) system to recognize impatient, angry or frustrated customers who require a more appropriate dialogue handling and to route them to human operators if necessary (see e.g., Yacoub et al., 2003). In retrieval applications, automatic detection of emotional acoustic events can be used to segment video material and to browse through video recordings; for instance, Cai et al. (2003) developed a 'hotspotter' that automatically localizes applause and cheer events to enable video summarization. Furthermore, a meeting browser that also provides information on the emotional state of the speaker was developed by Bett et al. (2000). Note that the word 'emotion' is a rather vague term susceptible to discussion. Often the term 'expressive' is also used to refer to speech that is affected by emotion. We will continue using the term 'emotion' in its broad sense.

The speaker's emotional and physical state expresses itself in speech through paralinguistic features such as pitch, speaking rate, voice quality, energy etc. In the literature, pitch is indicated as being one of the most relevant paralinguistic features for the detection of emotion, followed by energy, duration and speaking rate (see Ten Bosch, 2003). In general, speech shows an increased pitch variability or range and an increased intensity of effort when people are in a heightened aroused emotional state (Williams & Stevens, 1972; Scherer, 1982; Rothganger et al., 1998; Mowrer et al., 1987). In a paper by Nwe et al. (2003), an overview of paralinguistic characteristics of more specific emotions is given. Thus, it is generally known that paralinguistic information plays a key role in emotion recognition in speech. In this research, we concentrate on audible, identifiable paralinguistic cues in the audio signal that are characteristic for a particular emotional state or mood. For instance, a person who speaks with a trembling voice is probably nervous and a person who is laughing is most probably in a positive mood (but bear in mind that other moods are also possible). We will refer to such identifiable paralinguistic cues in speech as "paralinguistic events". Our goal is to detect these "paralinguistic events" in speech with the aim to make classification of the speaker's emotional state or mood possible.

2 Focus on automatic laughter detection

We have decided first to concentrate on laughter detection, due to the facts that laughter is one of the most frequently annotated "paralinguistic events" in recorded natural speech databases, it occurs relatively frequently in conversational, spontaneous speech and it is an emotional outburst and acoustic event that is easily identified by humans. Laughter detection can be meaningful in many ways. The main purpose of laughter detection in this research is to use laughter as an important cue to the identification of the emotional state of the speaker(s). Furthermore, detecting laughter in, e.g., meetings can provide cues to semantically meaningful events such as topic changes. The results of this research can also be used to increase the robustness of non-speech detection in automatic speech recognition. And finally, the techniques used in this study for discrimination between laughter and speech can also be used for similar discrimination tasks between other speech/non-speech sounds such as speech/music discrimination (see e.g., Carey et al., 1999).

Several studies have investigated the acoustic characteristics of laughter (e.g., Bachorowski et al., 2001; Trouvain, 2003; Bickley & Hunnicutt, 1992; Rothganger et al., 1998; Nwokah et al., 1993) and compared these characteristics to speech. Of these studies, the study by Bachorowski et al. (2001) is probably the most extensive one using 97 speakers who produce laugh sounds, while the other studies mentioned here use 2 to 40 speakers. Although studies by Bachorowski et al. (2001) and Rothganger et al. (1998) conclude that F_0 is much higher in laughter than in speech and that speech is rather monotonic, lacking a strongly varying melodic contour that is present in laughter, there are other studies that report on mean F_0 measures of laughter that are rather speech-like (Bickley & Hunnicutt, 1992). There are also mixed findings on intensity measures of laughter: while Rothganger et al. (1998) report on higher intensity values for laughter that even resemble screaming sounds, Bickley & Hunnicutt (1992) did not find large differences in amplitude between laughter and speech. Researchers did agree on the fact that the measures were strongly influenced by the gender of the speaker (Bachorowski et al., 2001; Rothganger et al., 1998) and that laughter is a highly complex vocal signal, notable for its acoustic variability (Bachorowski et al., 2001; Trouvain, 2003). Although there exists high acoustic variability in laughter, both between and within speakers, Bachorowski et al. (2001) noted that some cues of the individual identity of the laughing person are conveyed in laughter acoustics (i.e., speaker dependent cues). Furthermore, culture specific laughs may also exist: although no significant differences were found between laughter from Italian and German students (Rothganger et al., 1998), laughter transcriptions by Campbell et al. (2005) show that Japanese laughter can be somewhat different from the more typical "haha" laughs that are commonly produced in Western culture. A similarity between laughter and speech was found by Bickley & Hunnicutt (1992):

according to their research, the average number of laugh syllables per second is similar to syllable rates found for read sentences in English. However, they (Bickley & Hunnicutt, 1992) also identified an important difference between laughter and speech in the durations of the voiced portions: a typical laugh reveals an alternating voiced-unvoiced pattern in which the ratio of the durations of unvoiced to voiced portions is greater for laughter than for speech. This is one of the features that can be used for the development of a laughter detector.

Automatically separating laughter from speech is not as straightforward as one may think since both sounds are created by the vocal tract and therefore share characteristics. For example, laughter usually consists of vowel-like laugh syllables that can be easily mistaken for speech syllables by an automatic speech recognizer. Additionally, there are different vocal-production modes that produce different types of laughter (e.g., voiced, unvoiced) which causes laughter to be a very variable and complex signal. Furthermore, laughter events are typically short acoustic events of approximately 2 seconds (according to our database). Several researchers have already focused on automatic laughter detection; usually these studies employed spectral/cepstral features to train their models. Cai et al. (2003) tried to locate laughter events in entertainment and sports videos: they modeled laughter with Hidden Markov Models (HMM) using Mel-Frequency Cepstral Coefficients (MFCCs) and perceptual features such as short-time energy and zero crossing rate. They achieved average recall and precision percentages of 92.95% and 86.88% respectively. In the LAFCam project (Lockerd & Mueller, 2002), a system was developed for recording and editing home videos. The system included laughter detection using Hidden Markov Models trained with spectral coefficients. They classified presegmented laughter and speech segments correctly in 88% of the test segments. For automatic segmentation and classification of laughter, the system identified segments as laughter correctly 65% of the time. Kennedy & Ellis (2004) developed their laugh detector by training a Support Vector Machine (SVM) with Mel-Frequency Cepstral Coefficients, their deltas, spatial cues or modulation spectra coefficients. Their ROC (Receiver Operating Characteristic) curve showed a Correct Accept rate of approximately 80% at a 10% False Alarm rate. However, when the laughter detector was applied to data that was recorded on a different location, the performance decreased substantially. Recently, Campbell et al. (2005) used Hidden Markov Models to distinguish between four types of laughter and achieved a identification rate of greater than 75%.

In a previous study (see Truong & Van Leeuwen, 2005), we have also investigated the detection of laughter, making use of Gaussian Mixture Models (GMMs) and different sets of spectral and prosodic features. In the current laughter detection study, we extend the use of classification techniques (e.g., Support Vector Machines, Multi Layer Perceptron) and try to fuse different

classifiers (trained with different types of features). We aim at detection of individual laughter (as opposed to simultaneous laughter, i.e., where multiple speakers are laughing at the same time) in the first place. In second place, we will also explore far-field recordings, where the microphones are placed on the table, to detect laughter events in which more than one person is laughing (i.e., simultaneous laughter). Furthermore, we investigate promising features for laughter detection, as in contrast to the more conventional spectral/cepstral features used in speech/speaker recognition and in previous laughter detection studies, and employ these in different classification techniques.

In this paper, we describe how we developed, tested and compared a number of different laughter detectors developed in the current study and we report on the results achieved with these laughter detectors. Firstly, we define the laughter detection problem addressed in this study and describe the task of the detector in Section 3. Section 4 deals with the material used to train and test the classifiers. In Section 5, the different sets of features and the different methods are described. Subsequently, in Section 6 we test the laugh detectors and show the results. Finally, we conclude with a summary of the results, a discussion and some recommendations for future research in Section 7.

3 Defining the laughter detection problem addressed in this study

In this study, we develop an automatic laughter detector whose task is to discriminate between laughter and speech, i.e., to classify a given acoustic signal as either laughter or speech. We decided to keep the discrimination problem between laughter and speech clear and simple. Firstly, we use *presegmented* laughter and speech segments whose segment boundaries are determined by human transcribers. Providing an automatic time-alignment of laughter, which is a somewhat different problem that can be tackled with other techniques such as Hidden Markov Modeling and Viterbi decoding, is thus not part of the task of the laughter detector. Therefore, we can use a detection framework, which is often used in speaker and language recognition. Secondly, we only use (homogeneous) signals containing solely audible laughter or solely speech; signals in which laughter co-occurs with speech are not used. Consequently, "smiling speech" is not investigated in this study. And thirdly, we use close-talk recordings from head-mounted microphones rather than far-field recordings from desktop microphones. With close-talk recordings, we can analyze a clearer signal uttered by one single speaker and thereby aiming at detection of individual laughter uttered by one single person.

Previous laughter detection studies (Cai et al., 2003; Lockerd & Mueller, 2002; Kennedy & Ellis, 2004; Campbell et al., 2005; Truong & Van Leeuwen, 2005) usually investigated one classification technique using spectral features for

laughter detection. In the current study, we will investigate at least two different classification methods and four different feature sets (e.g., spectral and prosodic) for laughter detection and compare these to each other. Classification experiments will be carried out on speaker-dependent and speaker-independent material, and on material from an independent database with a different language background. Equal Error Rate (where the False Alarm rate is equal to the Miss rate) is used as a single-valued evaluation measure. A Detection Cost Function (DCF) will be used to evaluate the actual decision performance of the laughter detector. Summarizing, we investigate features and methods in order to automatically discriminate presegmented laughter segments from presegmented speech segments, uttered by individual speakers with the goal to enable emotion classification.

4 Material

In order to obtain realistic results, we decided to look for a speech database that contains natural emotional speech that is not acted. Furthermore, for practical reasons, the database should also include some paralinguistic or emotional annotation. Therefore, for training and testing, we decided to use the ICSI Meeting Recorder Corpus (Janin et al., 2004) since it meets 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. We included material from the Spoken Dutch Corpus (Corpus Gesproken Nederlands, CGN, Oostdijk, 2000) as an independent test set. The two databases and the material used to train and test our classifiers will be described below in Section 4.1 and Section 4.2.

4.1 ICSI Meeting Recording Corpus

The ICSI Meeting Recording Corpus consists of 75 recorded meetings with an average of six participants per meeting and a total of 53 unique speakers. Among the participants are also non-native speakers of English. There are simultaneous recordings available of up to ten close-talking microphones of varying types and four high quality desktop microphones. Using far-field desktop recordings brings along many additional problems such as background noise and the variation of the talker position with respect to the microphone. Therefore, we performed classification experiments with both types of recordings, but we focused on the close-talking recordings and used these in our main classification experiments. In a subsequent experiment, tests were carried out with models trained with far-field recordings to detect simultaneous laughter

(see Section 6.4). The speech data was divided in training and test sets: the first 26 ICSI 'Bmr' ('Bmr' is a naming convention which stands for the type of meeting, in this case Berkeley's Meeting Recorder weekly meeting) subset recordings were used for training and the last three ICSI 'Bmr' subset recordings were used as testing (these are the same training and test sets used as in Kennedy & Ellis, 2004). The 'Bmr' training and test sets contain speech from sixteen (fourteen male and two female) and ten (eight male and two female) speakers respectively. Because the three ICSI 'Bmr' test sets contained speech from speakers who were also present in the 26 ICSI 'Bmr' training sets, thus another test set was investigated as well to avoid biased results caused by overlap between speaker identities in the training and test material. Four ICSI 'Bed' (Berkeley's Even Deeper Understanding weekly meeting) sets with eight (six male and two female) unique speakers that were not present in the 'Bmr' training material were selected to serve as a speaker-independent test set.

All laughter and speech segments selected were *presegmented* (determination of onset and offset was not part of the task of the classifier) that were cut from the speech signal. Laughter segments were in the first place determined from laughter annotations in the human-made transcriptions of the ICSI corpus. The laughter annotations were not carried out in detail; labelers labeled whole vocal sounds as laughter which is comparable to word-level annotation. 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. Therefore, we decided to listen to all of the annotated laughter segments and made a quick and rough selection of laughter segments that do not contain speech or inaudible laughter. Furthermore, although we know that there are different types of laughter, e.g., voiced, unvoiced, 'snort-like' (Bachorowski et al., 2001; Trouvain, 2003), we decided not to make distinctions between these types of laughter because our aim was to develop a generic laughter model. Speech segments were also determined from the transcriptions: segments that only contain lexical vocalized sounds were labeled as speech.

In total, we used 3264 speech segments with a total duration of 110 minutes (with mean duration $\mu = 2.02$ seconds and standard deviation $\sigma = 1.87$ seconds) and 3574 laughter segments with a total duration of 108 minutes (with mean duration $\mu = 1.80$ seconds and standard deviation $\sigma = 1.25$ seconds, for more details, see Table 1).

Table 1
Amount (duration in *min*, number of segments in N) of laughter and speech data used in this research

	Training	**Test**		
	26 ICSI 'Bmr' meetings	**3 ICSI 'Bmr' meetings**	**4 ICSI 'Bed' meetings**	**14 CGN conversations**
	dur/N	*dur*/N	*dur*/N	*dur*/N
Speech segments	81 min/2422	10 min/300	15 min/378	4 min/164
Selected laughter segments	83 min/2680	10 min/279	11 min/444	4 min/171

Table 2
Similarities between the 26 'Bmr' training sets and the 3 'Bmr', 4 'Bed' and 14 CGN test sets

	Test material		
Compared to Bmr trainingset	**3 ICSI 'Bmr'**	**4 ICSI 'Bed'**	**14 CGN**
Same speaker identities?	Yes	No	No
Same acoustic conditions?	Yes	Yes	No
Same language?	Yes	Yes	No

4.2 Spoken Dutch Corpus (Corpus Gesproken Nederlands, CGN)

In addition to the ICSI Meeting Recorder Corpus, the Spoken Dutch Corpus was used as an independent test set. The Spoken Dutch Corpus contains speech recorded in the Netherlands and Flanders (a total of approximately nine million words) and comprises a variety of speech types such as spontaneous conversations, interviews, broadcast recordings, lectures and read speech. We used speech data from the spontaneous conversations ('face-to-face') recordings and selected laughter segments by listening to the annotated non-speech sounds. After listening to the data, the CGN recordings (table-top microphones) were perceived as somewhat clearer and less noisy than the ICSI far-field recordings. Testing on this independent CGN test set would be a challenging task for the classifiers since there are notable differences (see Table 2) between training (26 ICSI 'Bmr' recordings) and test set (CGN): the location, acoustic and recording conditions of the recordings are different and even the language is different (some studies report on the existence of culture and/or language specific paralinguistic patterns in vocal emotion expression).

5 Method

5.1 Features

We make a distinction between frame-level and utterance-level features that can be used in different modeling techniques to develop laughter classifiers. Frame-level features refer to features extracted each 16 ms of the utterance, so the length of the resulting feature vector is variable and depends on the length of the utterance. These features were normalized by applying a z-normalization where mean and standard deviation are calculated over the utterance: $\hat{x}_{frame} = (x_{frame} - \mu_{utt})/\sigma_{utt}$. Utterance-level features refer to features extracted per whole utterance, so the resulting feature vector has a fixed length which is independent of the length of the utterance (in this paper, the term 'utterance' is also used to refer to a 'segment'). Utterance-level features were normalized by applying a z-normalization where mean and standard deviation are calculated over the whole training data set: $\hat{x}_{utt} = (x_{utt} - \mu_{train})/\sigma_{train}$. In addition to the more conventional spectral features used in speech/speaker recognition, we also investigated three other sets of features. All features were used in different classification techniques described below (summarized in Table 3).

5.1.1 Frame-level features

Spectral features (PLP) Spectral or cepstral features, such as Mel-Frequency Cepstrum Coefficients (MFCCs) and Perceptual Linear Prediction Coding features (PLP, (Hermansky, 1990)), are often successfully used in speech and speaker recognition to represent the speech signal. We chose PLP features (for practical reasons, but MFCCs would also have been good candidates) to model the spectral properties of laughter and speech. PLP features use an auditorily-inspired signal representation including Linear-Predictive smoothing on this psychophysically-based short-time spectrum. Each 16 ms, twelve PLP coefficients and one energy feature were computed for a frame of 32 ms. In addition, delta features were computed by calculating the deltas of the PLP coefficients (by linear regression over five consecutive frames) and z-normalization was applied, which resulted in a total of 26 features.

Pitch & Energy features (P&E) Several studies (e.g. Williams & Stevens, 1972) have shown that with a heightening of arousal of emotion, for example laughter, speech shows an increased F_0 variability or range, with more source energy and friction accompanying increased intensity of effort. Furthermore, Bachorowski et al. (2001) found that the mean pitch in both male

and female laughter was considerably higher than in modal speech. Therefore, pitch and energy features were employed as well: each 10 ms, pitch and Root-Mean-Square (RMS) energy were measured over a window of 40 ms using the computer program Praat (Boersma & Weenink, 2005). In Praat, we set the pitch floor and ceiling in the pitch algorithm at 75 Hz and 2000 Hz respectively. Note that we changed the default value of the pitch ceiling of 600 Hz, which is appropriate for speech analysis, to 2000 Hz since studies have reported pitch measurements of over 1000 Hz in laughter. If Praat could not measure pitch for a particular frame (for example if the frame is unvoiced), we set the pitch value at zero to ensure parallel pitch feature streams and energy feature streams. The deltas of pitch and energy were calculated and a z-normalization was applied as well which resulted in a total of four features.

5.1.2 Utterance-level features (Fixed-length feature vectors)

Pitch & Voicing features (P&V) In addition to pitch measurements per frame, we also measured some more global, higher-level pitch features to capture better the fluctuations and variability of pitch in the course of time: we employed the mean and standard deviation of pitch, pitch excursion (maximum pitch − minimum pitch) and the mean absolute slope of pitch (the averaged local variability in pitch) since they all carry (implicit) information on the behaviour of pitch over a period of time. Furthermore, Bickley & Hunnicutt (1992) found that the ratio of unvoiced to voiced frames is greater in laughter than in speech and suggest this as a method to separate laughter from speech: "*. . . A possible method for separating laughter from speech, a laugh detector, could be a scan for the ratio of unvoiced to voiced durations . . .*". Therefore, we also used two relevant statistics as calculated by Praat: the fraction of locally unvoiced frames (number of unvoiced frames divided by the number of total frames) and the degree of voice breaks (the total duration of the breaks between the voiced parts of the signal divided by the total duration of the analyzed part of the signal). A total of six global z-normalized pitch & voicing features per utterance were calculated using Praat (Boersma & Weenink, 2005).

Modulation spectrum features (ModSpec) We tried to capture the rhythm and the repetitive syllable sounds of laughter, which may differ from speech: Bickley & Hunnicutt (1992) and Bachorowski et al. (2001) report syllable rates of 4.7 syllables/s and 4.37 syllables/s respectively while in normal speech, the modulation spectrum exhibits a peak at around 3−4 Hz, reflecting the average syllable rate in speech (Drullman et al., 1994). Thus, according to their studies, it appears that the rate of syllable production is higher in laughter than in conversational speech. Modulation spectrum features for laughter detection were also previously investigated by Kennedy & Ellis (2004) who

Table 3
Features used in this study, their abbreviations and the number of features extracted per utterance

Features			
Frame-level		**Utterance-level**	
Perceptual Linear Prediction	**Pitch & Energy**	**Pitch & Voicing**	**Modulation Spectrum**
PLP	**P&E**	**P&V**	**ModSpec**
26 per 16 ms	4 per 10 ms	6 (per utterance)	16 (per utterance)

found that the modulation spectrum features they used did not provide much discriminative power. 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 was then calculated and the first 16 spectral coefficients (modulation spectrum range up to 25.6 Hz) were normalized (z-normalization) and used as input features.

5.2 Modeling techniques

In this subsection, we describe the different techniques used to model laughter and speech employing the features as described above.

5.2.1 Gaussian Mixture Modeling

Gaussian Mixture Modeling concerns modeling a statistical distribution of Gaussian Probability Density Functions (PDFs): a Gaussian Mixture Model (GMM) is a weighted average of several Gaussian PDFs. We trained 'laughter' GMMs and 'speech' GMMs with different sets of features (frame-level and utterance-level). The GMMs were trained using five iterations of the Expectation Maximization (EM) algorithm and with varying numbers of Gaussian mixtures (varying from 2 to 1024 Gaussian mixtures for different feature sets) depending on the number of extracted features. In testing, a maximum likelihood criterion was used. A 'soft detector' score is obtained by determining the log-likelihood ratio of the data given the 'laughter' and 'speech' GMMs respectively.

5.2.2 Support Vector Machines

Support Vector Machines (SVMs, Vapnik, 1995, 1998) have become popular among many different types of classification problems, for instance face identification, bioinformatics and speaker recognition. The basic principle of this discriminative method is to find the best separating hyperplane between groups of datapoints that maximizes the margins. We used SVMTorch II, developed by the IDIAP Research Institute (Collobert & Bengio, 2001) to model the SVMs using different sets of features, and tried several kernels (linear, Gaussian, polynomial and sigmoidal) that were available in this toolkit.

SVMs typically expect fixed-length feature vectors as input which in our case means that the frame-level features (PLP and P&E) have to be transformed to a fixed-length vector while the utterance-level features (P&V and ModSpec) do not require this transformation since these feature vectors already have a fixed length. This transformation was carried out using a Generalized Linear Discriminant Sequence (GLDS) kernel (Campbell, 2002) which resulted in high-dimensional expanded vectors with fixed lengths for PLP and P&E features (GLDS kernel performs an expansion into a feature space explicitly). Subsequently, a linear kernel (a Gaussian kernel was also tested) was used in SVMTorch II to train the SVM GLDS.

5.2.3 Multi Layer Perceptron

For fusion of our classifiers, a Multi Layer Perceptron (MLP) was used (which is often used for fusion of classifiers, e.g. El Hannani & Petrovska-Delacretaz, 2005; Adami & Hermansky, 2003; Campbell et al., 2004). This popular type of feedforward neural network consists of an input layer (the input features), possibly several hidden layers of neurons and an output layer. The neurons calculate the weighted sum of their input and compare it to a threshold to decide if they should "fire". We used the LNKnet Pattern Classification software package, developed at MIT Lincoln Laboratory (Lippmann et al., 1993), to train and test our MLP classifiers. Z-normalization was applied to obtain mean and standard deviation values of zero and one respectively in all feature dimensions.

5.2.4 Fusion techniques

With the aim to achieve better performance, we tried to combine some of the best separate classifiers with each other. The idea behind this is that classifiers developed with different algorithms or features may be able to complement each other. The fusions applied in this study were all carried out on score-level. Fusion 'on score-level' means that we use the output of a classifier which can be considered 'scores' (e.g. log likelihood ratios, posterior probabilities)

given for test segments and combine these (for example by summation) with scores from other classifiers. We will refer to scores that are obtained when tested on laughter segments as *target scores* and scores that are obtained when tested on speech segments as *non-target scores*. There are several ways to fuse classifiers; the simplest one is by summing the scores using a linear combination, i.e. adding up the scores obtained from one classifier with the scores obtained from the other classifier (see Fusion A1 and B1 in Table 6), which is a natural way of fusion:

$$S_f = \beta S_A + (1 - \beta) S_B \tag{1}$$

where β is an optional weight that can be determined in the training phase. We used $\beta = 0.5$ so that the classifiers A and B are deemed equally important. For this sort of linear fusion to be meaningful, the scores must have the same range. If the scores do not have the same range, which can be the case when scores obtained with different classifiers are fused with each other (e.g., fusing GMM and SVM scores with each other), then normalization of the scores is required before they can be added up. We applied an adjusted form of T(est)-normalization (see Auckenthaler et al., 2000; Campbell et al., 2004) before summing GMM and SVM scores. This was done by using a fixed set of non-target scores as a basis (we decided to use the non-target scores of the 'Bmr' test set as a basis) from which μ and σ were calculated; these were used to normalize the target and non-target scores of the other two test sets ('Bed' and CGN) by subtracting μ from the score and subsequently dividing by σ: $\hat{S} = (S - \mu)/\sigma$.

Another way to combine classifiers is to apply a second-level classifier to the scores. This second-level classifier must also be trained on a fixed set of scores (again we used the scores obtained with the 'Bmr' test set as a training set) which serve as feature input to the second-level classifier. Fig. 3 gives an overview of the fusions of classifiers that we have performed.

6 Classification experiments and results

The performances of the GMM, SVM, and fused classifiers, each trained with different feature sets (PLP, Pitch&Energy, Pitch&Voicing and Modulation Spectrum features) were evaluated by testing them on three ICSI 'Bmr', four ICSI 'Bed' subsets and fourteen CGN conversations. We use the Equal Error Rate (EER) as a single-valued measure to evaluate and to compare the performances of the different classifiers.

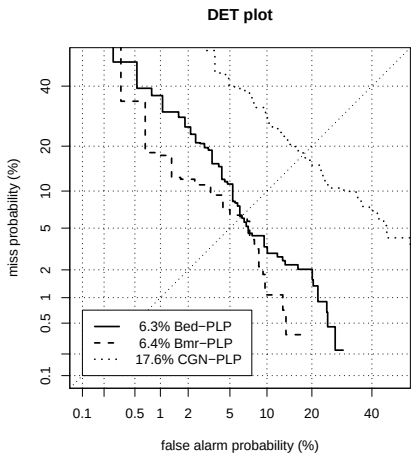


Fig. 1. DET plot of best-performing single GMM classifier, trained with PLP features and 1024 Gaussian mixtures

6.1 Results of separate classifiers

We started off training and testing GMM classifiers. Each GMM classifier was trained with different numbers of Gaussian mixtures since the optimal number of Gaussian mixtures depends on the amount of extracted datapoints of each utterance. So for each set of features, GMMs were trained with varying numbers of Gaussian mixtures (we decided to set a maximum of 1024) to find a number of mixtures that produced the lowest EERs. This procedure was repeated for the other three feature sets. The results displayed in Table 4 are obtained with GMMs trained with the number of mixtures that produced the lowest EERs for that particular feature set.

Table 4 shows that a GMM classifier trained with spectral PLP features outperforms the other GMM classifiers trained with P&E, P&V or ModSpec features. Also note that the ModSpec features produce the highest EERs. A Detection Error Tradeoff (DET) plot (Martin et al., 1997) of the best performing GMM classifier is shown in Figure 1. Note that, as expected, the EERs increase as the dissimilarities (see Table 2) between training material and test material increase (see Table 4). We also tried to extend the use of GMMs by training a Universal Background Model (UBM) which is often done in speaker recognition (e.g. Reynolds et al., 2000). The performance did not improve which was probably due to the small number of non-target classes:

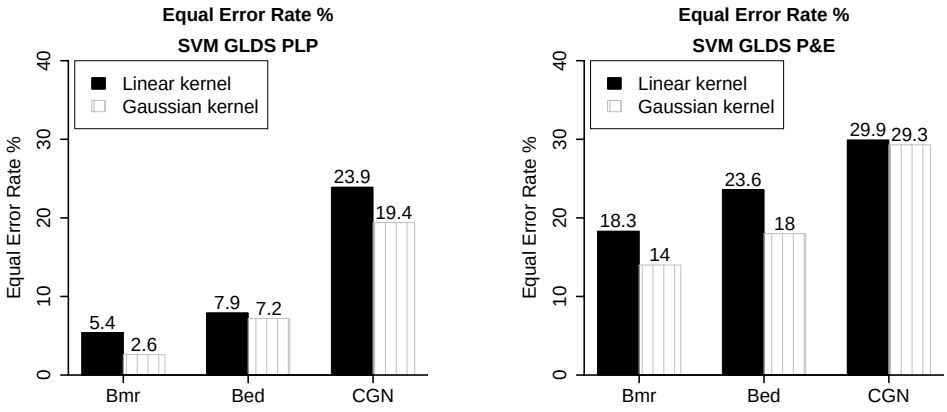
Table 4
Equal Error Rates (in %) of GMM classifiers trained with frame-level or utterance-level features and with different numbers of Gaussians

	Frame-level features		**Utterance-level features**	
	GMM PLP	**GMM P&E**	**GMM P&V**	**GMM ModSpec**
	1024 Gauss.	64 Gauss.	4 Gauss.	2 Gauss.
Bmr	6.4	14.3	20.0	37.7
Bed	6.3	20.4	20.9	38.7
CGN	17.6	32.2	28.1	44.5

the UBM in our case is trained with only twice as much data compared to the class-specific GMMs.

An SVM classifier typically expects fixed-length feature vectors as input. For the frame-level features PLP and P&E, we used a GLDS kernel (Campbell, 2002) to obtain fixed-length feature vectors. Subsequently, the SVMs were further trained with the expanded features using a linear kernel, which is usually done in e.g., speaker recognition. Since preliminary classification experiments showed good results with a Gaussian kernel, we also trained the expanded features using a Gaussian kernel: this improved the EERs considerably for the frame-level PLP and P&E features as can be seen in Figure 2. The results of both frame-level and utterance-level features used in SVMs are shown in Table 5, where we can observe that SVM GLDS using spectral PLP features outperforms the other SVMs. The second-best performing feature set for SVM is the utterance-level P&V feature set. Taking into consideration the number of features, 26 PLP features per frame per utterance as opposed to 6 P&V features per utterance, and the fact that we obtain relatively low EERs with P&V features, we may infer that P&V features are relatively powerful discriminative features for laughter detection. Further, our utterance-level features perform considerably better with SVMs than with GMMs (compare Table 4 to Table 5).

To summarize, comparing the results of the classifiers and taking into account the different feature sets, we can observe that SVM in most of the cases, performs better than GMM. The best performing feature set for laughter detection appears to be frame-level spectral PLP. Concentrating only on the PLP-based features, we can observe that GMMs generalize better over different test cases than SVMs do. Utterance-level prosodic P&V features are promising features since the number of features is relatively small and they produce relatively low EERs. Further, we have seen that utterance-level features, such as P&V and ModSpec, perform better with a discriminative classifier SVM than with GMM. The next step is to fuse some of these classifiers to investigate whether performance can be improved by combining different



(a) SVM GLDS trained with expanded PLP features

(b) SVM GLDS trained with expanded P&E features

Fig. 2. Results of SVM GLDS trained with (a) PLP or (b) P&E features and a linear or Gaussian kernel

Table 5
Equal Error Rates (in %) of SVM classifiers trained with frame-level or utterance-level features and with a Gaussian kernel

	Frame-level features		**Utterance-level features**	
	SVM GLDS PLP	**SVM GLDS P&E**	**SVM P&V**	**SVM ModSpec**
	Gaussian kernel			
Bmr	2.6	14.0	9.0	28.7
Bed	7.2	18.0	11.4	32.9
CGN	19.4	29.3	23.3	29.3

classifiers and different features.

6.2 Results of fused classifiers

Since each of our separate classifiers were trained with a different set of features, it would be interesting to investigate whether using a combination of these classifiers would improve performance. We will focus on the fusions between the classifiers based on spectral PLP features and the classifiers based on prosodic P&V features (the two best performing classifiers so far). Fusions were carried out on score-level using fusion techniques described in Section 5.2.4 (see Fig. 3 for a schematic overiew of all fusions applied). In Table 6, we indicate whether the performances of the classifiers fused with PLP and P&V are significantly better than the performance of the single classifier trained

with only spectral features (PLP). The significance of differences between EERs was calculated by carrying out a McNemar test with a significance level of 0.05 (Gillick & Cox, 1989). Table 6 shows that in many cases, the addition of the P&V-based classifier to the PLP-based classifier decreases EERs significantly, especially in the case of the SVM classifiers (B1, B2, B3) and in the CGN test set. For SVMs, the method of fusion does not appear to influence the EERs significantly differently. However, for GMMs, the linear fusion method performs significantly worse than the other two fusion methods.

We also combined GMM and SVM classifiers since the different way of modeling that GMM (generative) and SVM (discriminative) employ may complement each other; GMM models data generatively and SVM models data discriminatively. We first tried to combine these scores linearly: GMM and SVM scores from a spectral PLP classifier were first normalized using T-normalization and then summed (see Section 5.2.4). This resulted in relatively low EERs for Bed: 3.4% and CGN: 12.8%. Other normalization techniques for the scores could be used but this was not further investigated in the current study. We continued fusion with the use of a 2nd-level classifier that functions as a sort of 'merge/fuse-classifier'. As we can observe in Table 7, a fused GMM and SVM classifier (C1, C2) performs indeed significantly better than a single GMM or SVM classifier. When P&V is added to the fused GMM-SVM classifier, performances are only significantly better in the case of the CGN test set (see Table 7: compare D1 and D2 to C1 and C2). According to the classification experiments carried out in this study, the fused classifiers both D1 *and* D2 (fused with GMM *and* SVM scores) perform the best with the lowest EERs: D1 performs significantly better than B2 (without GMM scores), D2 performs significantly better than B3 (without GMM scores) but there is no significant difference between D1 and D2. Note that the reason for the missing results of the 'Bmr' test set in Table 6 and 7 is that the scores of this set were used as a training set (to train the 2ndSVM or MLP fuse-classifier).

Instead of using a 2nd-level classifier to fuse the output of classifiers, we have also tried to fuse classifiers directly on feature-level, i.e. feeding PLP and P&V features all together in a single classifier, in our case SVM. We could only perform this for SVM since the GLDS kernel expanded the frame-level PLP features to a fixed-length feature vector that was fusible with the utterance-level P&V features. We compared the obtained EERs (Bmr: 1.7%, Bed: 6.9%, CGN: 18.8%) with the EERs of the single SVM, trained with only PLP features (Table 6, B0) and found that the differences between the EERs were not significant, meaning that the addition of P&V features to PLP features on feature-level, in these cases, did not improve performance. This could be explained by the fact that the PLP feature vector for SVM already has 3653 dimensions (expanded by GLDS kernel); one can imagine that the effect of adding six extra dimensions (P&V features) to a vector that already consists of 3653 dimensions can be small.

Table 6
EERs of fused classifiers of the *same* type on decision level (∗ indicates whether the difference in performance is significant with respect to the single classifier, A0 or B0 displayed in the last column, e.g., A1 is a fusion between 2 first-level classifiers).

Label	**Classifiers**	**Features**	**Fusion method**	**EERs (%)**			**Compare to**
				Bmr	**Bed**	**CGN**	
A0	**GMM**	**PLP**	**none**	**6.4**	**6.3**	**17.6**	-
A1	GMM	PLP, P&V	linear	8.6	11.7*	22.7	A0
A2	GMM	PLP, P&V	2ndSVM	-	5.8	13.4*	A0
A3	GMM	PLP, P&V	MLP	-	6.1	12.8*	A0
B0	**SVM**	**PLP**	**none**	**2.6**	**7.2**	**19.4**	-
B1	SVM	PLP, P&V	linear	2.6	5.6	12.2*	B0
B2	SVM	PLP, P&V	2ndSVM	-	5.2*	12.2*	B0
B3	SVM	PLP, P&V	MLP	-	4.7*	11.6*	B0

Table 7
EERs of fused classifiers of *different* types (∗ indicates whether the difference in performance is significant with respect to another classifier displayed in the last column, e.g. D1 is a fusion between 4 first-level classifiers)

Label	**Classifiers**	**Features**	**Fusion method**	**EERs (%)**			**Compare to**
				Bmr	**Bed**	**CGN**	
C1	GMM, SVM	PLP	2ndSVM	-	3.2*	11.6*	A0, B0
C2	GMM, SVM	PLP	MLP	-	3.7*	11.0*	A0, B0
D1	GMM, SVM	PLP, P&V	2ndSVM	-	3.2	8.7*	C1
D2	GMM, SVM	PLP, P&V	MLP	-	2.9	7.5*	C2

To summarize, using a combination of the output of classifiers based on spectral and prosodic features rather than using a single classifier based on spectral features solely, improves performance significantly in many cases and increases robustness. The lowest EERs were obtained by fusing different types of classifiers, namely GMM and SVM classifiers, which performed significantly better than classifiers that do not use scores from another type of classifier. Finally, both SVM and MLP can be used as a fusion method; no significant differences in performance of the two fusion methods were found.

As an example of how such a classifier could work in practice, we have divided an 8-seconds long sentence in 0.5-seconds segments and classified each segment as either speech or laughter. We can see in Fig. 4 that the classifier is able to identify laughter in this short utterance, although it is done in a rather cumbersome way. HMM techniques and Viterbi decoding techniques

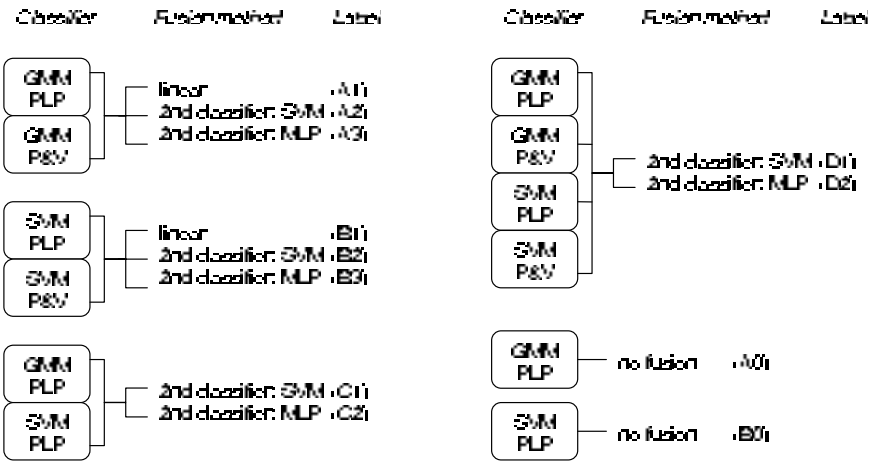


Fig. 3. Fusion scheme of classifiers

are probably more suitable to tackle this segmentation problem which can be investigated in the future.

6.3 Actual decision performance of classifier

We have used Equal Error Rate (EER) as a single-valued measure to evaluate and compare the performances of the classifiers. However, the EER is a point on the DET curve that can only be determined after all samples have been classified and evaluated. The EER can only be found *a posteriori* while in real life applications, the decision threshold is set *a priori*. As such, EER is not suitable for evaluating the *actual* decision performance. An a priori threshold can be drawn by evaluating the detection cost function (DCF, Doddington et al., 2000) which is defined as a weighted sum of the Miss and False Alarm probabilities:

$$\begin{aligned} C_{det} = & C_{Miss} \times P(Miss|Target) \times P(Target) \\ & + C_{FA} \times P(FA|NonTarget) \times P(NonTarget) \end{aligned} \quad (2)$$

where C_{Miss} is the cost of a Miss and C_{FA} is the cost of a False Alarm, $P(Miss|Target)$ is the Miss rate, $P(FA|NonTarget)$ is the False Alarm rate and $P(Target)$, $P(NonTarget)$ are the a priori probabilities for a target and non-target respectively ($P(NonTarget) = 1 - P(Target)$). We chose $C_{Miss} =$

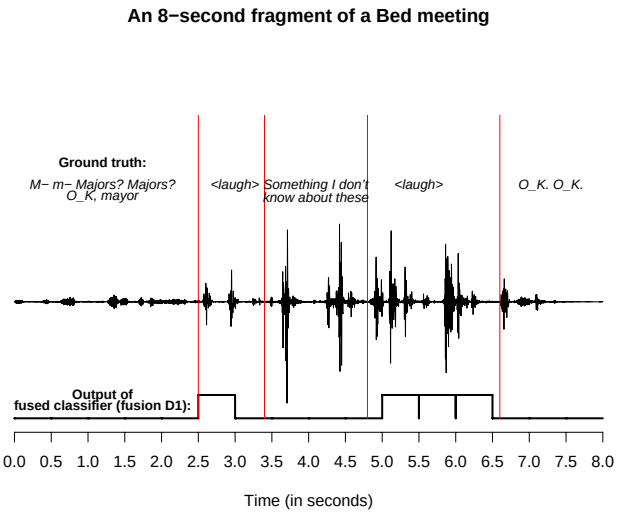


Fig. 4. Fused classifier applied to a fragment of a Bed meeting

$C_{FA} = 1$ and $P(Target) = P(NonTarget) = 0.5$; this particular case of DCF is also known as the half total error rate (HTER) which is in fact the mean of the Miss rate and False Alarm rate. A threshold can be determined by choosing the score threshold where the probabilities of error are equal as this this should lead to minimum costs (under the assumption of a unit-slope DET curve); this threshold is then used to classify new samples resulting in an evaluation of the actual performance of the system. We used the scores obtained with the Bmr test set to determine (calibrate) thresholds for the single GMM classifier trained with PLP features (see A0 in Table 6) and the fused classifier D2 (see Table 7), which can be considered one of the parameters in these cases of fusions. The actual decision performances obtained with thresholds at EER are shown in Table 8 where we can see that the HTERs are usually higher than the EERs; this shows the difficulty of determining a threshold based on one data set and subsequently applying this threshold to another data set. The difference between EER and HTER is larger for the CGN test set than for the Bed test set (see Table 8), illustrating that the Bed test set is more similar to the Bmr set on which we calibrated the thresholds, then the CGN test set is. Further, the unequal error rates, especially in the CGN case, are an indication of mistuned thresholds.

Table 8
Actual decision performances of GMM classifier, trained with PLP features (1024 Gaussians) and of fused classifier obtained by Fusion D2, see Table 7

Classifier	**Test set**	**EER**	**HTER**	**Actual Miss rate**	**Actual False Alarm rate**
GMM-PLP (A0, Table 6)	Bed	6.3%	6.2%	6.1%	6.3%
	CGN	17.6%	36.0%	70.2%	1.8%
Fusion D2 (Table 7)	Bed	2.9%	3.0%	2.5%	3.4%
	CGN	7.5%	25.8%	50.3%	1.2%

6.4 Results of far-field recordings

So far, we have only used close-talk microphone recordings to train our GMM and SVM classifiers. This went relatively well, especially for the ICSI 'Bmr' and 'Bed' meetings, but there was always a performance gap between the results of these two meetings and the 14 CGN conversations caused by dissimilarities between the two databases (see Table 2). Although the quality of the table-top recordings in these 14 CGN conversations was close to the quality of the close-talk recordings in the ICSI corpus, the differences in acoustics of close-talk and distant recordings is most probably one of the factors that caused this performance gap. To train new GMM models based on table-top recordings, adjusted definitions for laughter and speech events were applied because in the case of table-top microphones, it is possible that more than one person is laughing or speaking at the same time. A laughter event was defined as an event where more than one person is laughing aloud. Laughter events where one person is laughing aloud were usually hardly audible in the far-field recordings; therefore we only concentrated on audible, simultaneous laughter from multiple persons. It appeared that "speaking at the same time" did not occur as often as "laughing at the same time" did (speaking simultaneously can be perceived by people as rude while the opposite holds for laughing), so a speech event was defined as an event where at least one person is speaking. So, the task of the classifier is slightly changed from detecting individual human laughter to simultaneous human laughter. We used one of the four available high-quality desktop microphone recordings. The signal was divided into 1 second frames and for each 1 second frame we determined automatically from the transcriptions whether there was more than one person laughing or not. New segments were only extracted for the ICSI material since in the case of CGN material we were already using table-top recordings. With these segments we trained new 'laughter' and 'speech' GMM models with PLP features. Fig. 5 shows DET curves of this classification experiment.

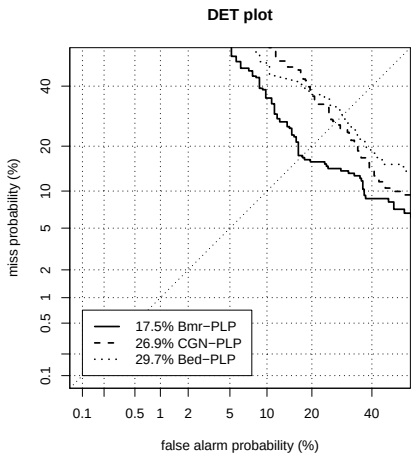


Fig. 5. DET plot of GMM classifier (1024 Gaussians) trained with PLP features applied to 'Bmr', 'Bed' and CGN, trained and tested on far-field recordings

6.5 A closer look on the prosodic Pitch&Voicing features

The features extracted from the signals reveal some differences between laughter and speech which were also reported in previous studies on laughter. Table 9 shows mean F_0 measurements from previous studies on laughter, while Table 10 displays mean values of several features, measured in the current study (for F_0 we report values in Hertz for comparison with previous studies, but we also report log F_0 which are more Gaussian-like). We can observe some differences between laughter and speech in Table 10 and Fig. 6, for instance, mean F_0 is higher in laughter than in speech (which was also found in Bachorowski et al., 2001; Rothganger et al., 1998, see Table 9) but there is still some overlap. Furthermore, the measurements also indicate that laughter contains relatively more unvoiced portions than speech which is in agreement with what was found by Bickley & Hunnicutt (1992).

It may be imaginable that not all of the P&V features presented here are equally important. We carried out a feature selection procedure to determine which individual features are relatively important. This was done by using the Correlation based Feature Selection procedure in the classification toolkit WEKA (Witten & Frank, 2005). According to this selection procedure, 'mean pitch' and 'fraction unvoiced frames' are the most important features (relative to the six P&V features) that contribute to the discriminative power of the

Table 9
Mean F_0 measurements in laughter from previous studies, standard deviations in parentheses, table adopted from Bachorowski et al. (2001)

	Mean F_0 (Hz)	
Study	Male	Female
Bachorowski et al. (2001)	284 (155)	421 (208)
Bickley & Hunnicutt (1992)	138	266
Rothganger et al. (1998)	424	472

Table 10
Mean measurements in laughter and speech from current study, no distinction between male and female, with standard deviations in parentheses

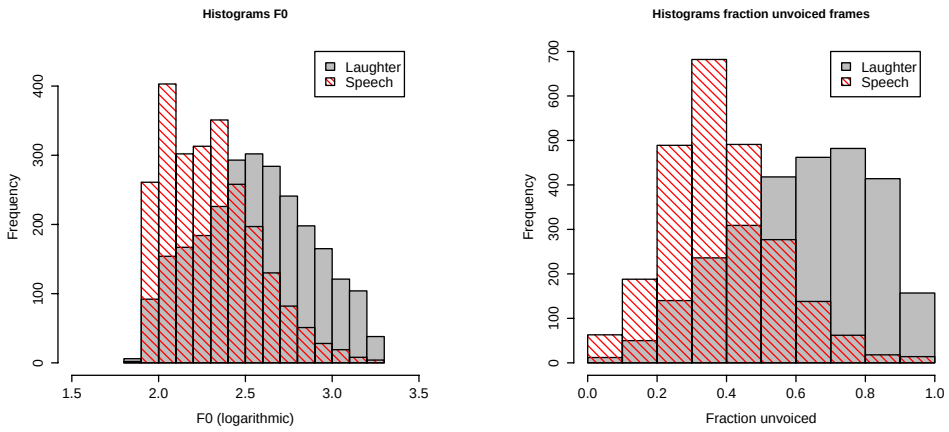
	Laughter	Speech
Mean F_0 (Hz)	475 (367)	245 (194)
Mean F_0 (log)	2.56 (0.32)	2.30 (0.26)
Mean fraction unvoiced frames (%, number of unvoiced frames divided by the number of total frames)	62 (20)	38 (16)
Mean degree of voice breaks (%, total duration of the breaks between the voiced parts of the signal divided by the total duration of the analysed part of the signal)	34 (22)	25 (17)

model. With these two features trained in an SVM, we achieve relatively low EERs (EERs Bmr: 11.4% Bed: 12.9% CGN: 29.3%). Although these EERs are significantly higher than those of the SVM trained with all six P&V features (see Table 5), the results of the SVM trained with only two features can be considered good taking into account the small number of features used to train this SVM.

7 Discussion and Conclusions

7.1 Discussion

During the analyses, some aspects were encountered that may require some attention or more investigation. For example, as can be seen in Table 9, there are somewhat large differences between pitch measurements of laughter in different studies, including the current study. As Bachorowski et al. (2001) already noted in their paper, automated pitch algorithms do not always work



(a) mean pitch (b) fraction unvoiced frames

Fig. 6. Histograms of (a) mean pitch and (b) fraction of unvoiced frames of laughter and speech

well in the analysis of laughter since laughter can exhibit F_0 values that are considerably higher than in speech; F_0 values of over 1000 Hz have been reported for laughter. Due to this observation, we have set our pitch algorithm in Praat to analyze F_0 within the range of 75 Hz−2000 Hz. Since most pitch algorithms aim at analyzing human speech with a usual pitch range of 80−200 Hz for male and 150−350 Hz for female speakers, it is not unlikely that pitch errors have occurred that may have influenced our results (see also large F_0 span). However, we do think that F_0 can be an important cue for laughter detection.

Also note that we selected a homogeneous set of laughter and speech segments. The task of the detector was to discriminate between isolated laughter and speech segments. Consequently, phenomena like "smiled speech" were not investigated in this study. In "smiled speech", speech is uttered with spread lips which has acoustic properties that probably are more speech-like than laughter-like. Therefore, modeling "smiled speech" will probably require a modified approach and is a seperate detection problem that can be investigated in future research.

Furthermore, note that our laughter and speech segments under analysis are relatively short with an average duration of 1.80 seconds ($\sigma = 1.25$ seconds) and 2.02 seconds ($\sigma = 1.87$ seconds) respectively. To capture the syllable repetitions and rhythm structure that are characteristic for laughter, we computed modulation spectra of these short segments. However, it is possible that the segments were too short to compute reliable modulation spectra, and as a possible consequence, we obtained relatively high EERs for this feature set.

Finally, although previous research report on significant differences between male and female laughter (Bachorowski et al., 2001; Rothganger et al., 1998),

we developed gender-independent classifiers with which we achieved relatively low EERs. The question is whether it is necessary to train gender-dependent classifiers, and how much improvement can be achieved with gender-dependent models since for gender-dependent models, one needs a gender classifier as well.

7.2 Conclusions and recommendations for future research

7.2.1 Conclusions

Our goal was to develop a laugh detector by investigating features and methods in order to automatically discriminate presegmented laughter segments from presegmented speech segments (close-talk recordings) to enable emotion classification. We concentrated on a set of segments containing either laughter solely or speech solely. By using more conventional features often used in speech/speaker recognition (spectral PLP features) and other features (P&E, P&V, ModSpec) in classification techniques, we were able to automatically discriminate laughter segments from speech segments with relatively low error rates for short utterances. Using classification techniques that are also often used in speech and speaker recognition (GMM, SVM and MLP) we have developed different gender-independent classifiers and we have tested these on three different test sets: speaker-dependent ICSI 'Bmr', speaker-independent ICSI 'Bed' and a separate independent test set taken from the CGN corpus. Fusion (on score-level) of classifiers trained with different techniques and different feature sets were also investigated. We obtained results from which we can draw interesting conclusions.

Firstly, our results show that spectral features alone (in our case PLP features) contain much useful information for discrimination between laughter and speech since they outperform all other features investigated in this study (compare Table 4 and 5). Thus, we can conclude that spectral features alone can be used to discriminate between laughter and speech.

However, according to our detection results, prosodic P&V features are also very promising features since with a smaller amount of feature data (six P&V features per segment, as in contrast to 26 PLP features per frame) we obtain second-best EERs. Moreover, our acoustic measurements indeed show that there are measurable differences between laughter and speech in mean pitch and the pattern of voiced/unvoiced portions (see Table 10 and Fig. 6). Thus, the differences in the voicing/unvoicing pattern (as previously suggested by Bickley & Hunnicutt, 1992) and the differences in pitch between laughter and speech are promising cues for discrimination between these two sounds.

Based on the previous two conclusions, we combined classifiers that use spectral features together with classifiers that use P&V features and found that

this combination improves the performance of the laugh detector considerably. Using a 2nd-order SVM or MLP classifier, we fused scores obtained with frame-level spectral features together with scores obtained with utterance-level prosodic P&V features which usually resulted in significantly lower EERs than when only scores from frame-level spectral features are used (see fusions A2, A3, B2 and B3 in Table 6). Thus, we can conclude indeed that a classifier based on spectral features alone can be used to discriminate between laughter and speech, but a significant improvement can be achieved in most cases when it is fused with a classifier that is based on prosodic features. Furthermore, fusing scores from SVM and GMM is also very fruitful; our classifiers that did use both SVM and GMM scores (see fusions D1 and D2 in Table 7) performed significantly better than classifiers that only used SVM scores (see fusions B2 and B3 in Table 6).

Further, the other two feature sets investigated in this study, P&E and ModSpec features, did not provide as much discriminative power as PLP and P&V features did. We also experienced that discriminative classification techniques such as SVM can model better utterance-level features than GMMs do (compare Table 5 to Table 4). And finally, our laughter and speech models can cope relatively well with laughter and speech segments from CGN which is a completely different database in a different language (Dutch) and which is recorded under different acoustic conditions than the ICSI corpus that was used for training. It appears that our models trained with close-talk recordings (thus laughter from one speaker) in English can cope with CGN's desktop recordings in Dutch in which simultaneous laughter is also present (after fusion, we obtain EERs $< 10\%$).

7.2.2 Future research

For our laughter classification experiments, we used only segments that were already segmented (based on a human transcription) and segments that contained either laughter solely or speech solely (thus a homogeneous set of segments); segments in which speech co-occurred with laughter were discarded. In other words, detection of onset and offset of laughter was not investigated in this study but can be addressed in a follow-up study. Detection of onset and offset of laughter (laughter segmentation) can be seen as a separate problem, resembling a speech recognition problem, that gives rise to other interesting questions such as how to define the beginning and end of laughter, and what kind of evaluation measures to use: these are problems that can be typically addressed within a Hidden Markov Model framework. Note that individual laughter is often modeled in speech recognition systems. There, the objective is to properly recognize words and specific 'laughter' phones or words have been introduced to deal with laughter as an interruption of the speech flow. These systems, we may assume, are not tuned to detect the laughter events

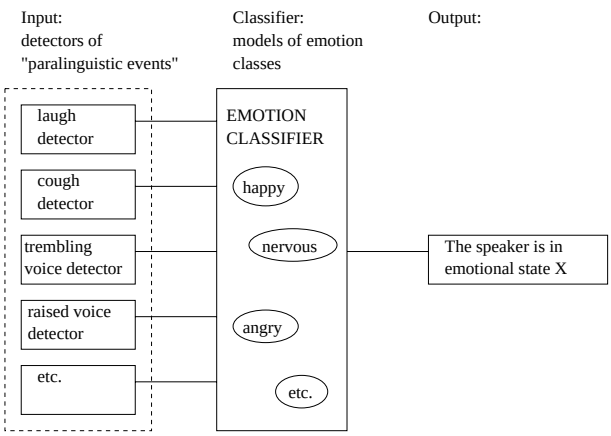


Fig. 7. Plan for emotion detector

correctly, but rather, to minimize word error rate. But they do in fact solve the laughter segmentation problem. It would be interesting to evaluate a state-of-the-art LVCSR (Large Vocabulary Continuous Speech Recognition) system on its 'off-the-shelf' laughter detection capabilities on the similar data.

As Bachorowski et al. (2001) noted, although laughter is a highly variable and complex signal, it would be interesting to investigate as to what extent laughter is dependent on speaker, culture and to investigate the different types of laughter that can be used in different contexts. The observation that the difference between the results of the speaker-dependent 'Bmr' test set and the speaker-independent 'Bed' test set are relatively small (especially when using PLP or P&V features) could indicate that the models were able to cover some speaker-independent properties of laughter. Furthermore, different types of laughter can reveal different states or emotions of the speaker: Ohara (2004) distinguished a hearty laugh, an amused laugh, a satirical laugh, and a social laugh, so laughter is not always uttered by a speaker in a joyful or humorous state. Campbell et al. (2005) developed a detector that can automatically recognize and distinguish between these four types of laughter in Japanese (identification rate $> 75\%$). Although we have also tested with laughter segments taken from a Dutch corpus (as opposed to an English corpus), it is difficult to draw conclusions about culture-dependency from these results since the acoustic and recording conditions differ.

Finally, the aim of this study was to develop a laugh detector to enable emotion classification. Our plan for emotion recognition is based on the fact that emotion is expressed in speech by audible paralinguistic features or events,

which can be seen as the 'building blocks' or 'emotion features' of a particular emotion. By developing separate detectors for these paralinguistic events, e.g. laughter, crying, trembling voice, raised voice etc., and by using the output of each detector, we can train certain 'emotion profiles' (see Figure 7). Adding visual information could further help to improve emotion detection. Our plan is thus to perform emotion classification via the detection of audible paralinguistic events. In this study, we have developed a laugh detector to provide a first step in this plan towards a future emotion classifier.

Acknowledgements

We would like to thank the two reviewers for their useful comments on this work. This research was supported by the BSIK-project MultimediaN (Multimedia Netherlands).

References

Adami, A.G., Hermansky, H., 2003. Segmentation of Speech for Speaker and Language Recognition. In: Proc. of Eurospeech 2003, Geneva, Switzerland, 841–844.

Auckenthaler, R., Carey, M., Lloyd-Thomas, H., 2000. Score normalization for text-independent speaker verification systems. Digital Signal Processing 10, 42–54.

Bachorowski, J.-A., Smoski, M.J., Owren, M.J., 2001. The acoustic features of human laughter. J. Acoust. Soc. Amer. 110 (3), 1581–1597.

Bett, M., Gross, R., Yu, H., Zhu, X., Pan, Y., Yang, J., Waibel, A., 2000. Multimodal Meeting Tracker. In: Proc. of RIAO 2000, Paris, France.

Bosch ten, L., 2003. Emotions, speech and the ASR framework. Speech Communication 40, 213–225.

Bickley, C., Hunnicutt, S., 1992. Acoustic analysis of laughter. In: Proc. of ICSLP 1992, Banff, Canada, 927–930.

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

Cai, R., Lie L., Zhang, H-J., Cai, L-H., 2003. Highlight sound effects detection in audio stream. In: Proc. of the IEEE International Conference on Multimedia and Expo 2003, Baltimore, USA, 37–40.

Campbell, N., Kashioka, H., Ohara, R., 2005. No Laughing Matter. In: Proc. of Interspeech 2005, Lisbon, Portugal, 465–468.

Campbell, W. M., Reynolds, D. A., Campbell, J. P., 2004. Fusing Discriminative and Generative Methods for Speaker Recognition: Experiments on

Switchboard and NFI/TNO Field Data. In: Proc. Odyssey: The Speaker and Language Recognition Workshop 2004, Toledo, Spain, 41–44.

Campbell, W.M., 2002. Generalized linear discriminant sequence kernels for speaker recognition. In: Proc. of the IEEE International Conference on Acoustics Speech and Signal Processing 2002, Orlando, USA, 161-164.

Carey, M.J., Parris, E.S., Lloyd-Thomas, H., 1999. A comparison of features for speech, music discrimination. In: Proc. of ICASSP 1999, Phoenix, USA, 1432–1435.

Collobert, R., Bengio, S., 2001. SVMTorch: Support Vector Machines for Large-Scale Regression Problems. Journal of Machine Learning Research 1, 143–160.

Doddington, G., Przybocki, M., Martin, A., Reynolds, D., 2000. The NIST speaker recognition evaluation - overview, methodology, systems, results, perspective. Speech Communication 31, 225-254.

Drullman, R., Festen, J.M., Plomp, R., 1994. Effect of temporal envelope smearing on speech reception. J. Acoust. Soc. Amer. 95 (2), 1053–1064.

Gillick, L., Cox, S., 1989. Some Statistical Issues In The Comparison Of Speech Recognition Algorithms. In: ICASSP 1989, Glasgow, Scotland.

Goldbeck, T., Tolkmitt, F., Scherer, K.R., 1988. Experimental studies on vocal affect communication. In: Scherer, K.R. (Ed.) Facets of Emotion:Recent Research. Erlbaum, Hillsdale, 119–253.

El Hannani, A., Petrovska-Delacretaz, D., 2005. Exploiting High-Level Information Provided by ALISP in Speaker Recognition. In: Proceedings of Non Linear Speech Processing Workshop (NOLISP05), Barcelona, Spain, 19–24.

Hermansky, H., 1990. Perceptual Linear Predictive (PLP) analysis of speech. J. Acoust. Soc. Amer. 87 (4), 1738–1752.

Houtgast, T., Steeneken, H.J.M., 1985. A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. J. Acoust. Soc. Amer. 77 (3), 1069–1077.

Janin, A., Ang, J., Bhagat, S., Dhillon, R., Edwards, J., Macias-Guarasa, J., Morgan, N., Peskin, B., Shriberg, E., Stolcke, A., Wooters, C., Wrede, B., 2004. The ICSI Meeting Project: Resources and Research. In: NIST ICASSP 2004 Meeting Recognition Workshop, Montreal, Canada.

Kennedy, L.S., Ellis, D.P.W., 2004. Laughter detection in meetings. In: NIST ICASSP 2004 Meeting Recognition Workshop, Montreal, Canada.

Kiefte, M., 1999. Discriminant Analysis Toolbox. [Computer program]. Retrieved from http://www.mathworks.com/matlabcentral/fileexchange.

Lippmann, R.P., Kukolich, L., Singer, E., 1993. LNKnet: Neural Network, Machine Learning, and Statistical Software for Pattern Classification. Lincoln Laboratory Journal 6 (2), 249–268.

Lockerd, A., Mueller, F., 2002. LAFCam - Leveraging Affective Feedback Camcorder. In: Proc. of the CHI 2002 Conference on Human Factors in Computing Systems, Minneapolis, USA, 574–575.

Martin, A., Doddington, G., Kamm, T., Ordowski, M., Przybocki, M., 1997. The DET curve in assessment of detection task performance. In: Proc. of

Eurospeech 1997, Rhodes, Greece, 1895–1898.
McGilloway, S., Cowie, R., Douglas-Cowie, E., Gielen, S., Westerdijk, M., Stroeve, S., 2000. Approaching automatic recognition of emotion from voice: a rough benchmark. In: ISCA Workshop on Speech and Emotion, Belfast, Ireland.
Mowrer, D.E., LaPointe, L.L., Case, J., 1987. Analysis of five acoustic correlates of laughter. Journal of Nonverbal Behaviour 11, 191–199.
Nwe, T.L., Foo, S.W., De Silva, L.C., 2003. Speech emotion recognition using hidden Markov models. Speech Communication 41, 603–623.
Nwokah, E.E., Davies, P., Islam, A., Hsu, H-C, Fogel, A., 1993. Vocal affect in three-year-olds: A quantitative acoustic analysis of child laughter. J. Acoust. Soc. Amer. 94 (6), 3076–3090.
Ohara, R., 2004. Analysis of a laughing voice and the method of laughter in dialogue speech. Unpublished Masters Thesis, Nara Institute of Science & Technology.
Oostdijk, N., 2000. The Spoken Dutch Corpus: Overview and first evaluation. In: Proc. of LREC 2000, Athens, Greece, 887–894.
Reynolds, D.A., Quatieri, T.F., Dunn, R., 2000. Speaker verification using adapted Gaussian mixture models. Digital Signal Processing 10, 19–41.
Rothganger, H., Hauser, G., Cappellini, A.C., Guidotti, A., 1998. Analysis of laughter and speech sounds in Italian and German students. Naturwissenschaften 85 (8), 394–402.
Scherer, K.R., 1982. Methods of research on vocal communication: paradigms and parameters. In: Scherer, K.R., Ekman, P. (Eds Handbook of methods in nonverbal behavior research. Cambridge U.P., NewYork, 136–198.
Trouvain, J., 2003. Segmenting phonetic units in laughter. In: Proc. of ICPhS, Barcelona, Spain, 2793–2796.
Truong, K.P., Van Leeuwen, D.A., 2005. Automatic detection of laughter. In: Proc. of Interspeech 2005, Lisbon, Portugal, 485–488.
Vapnik, V.N., 1995. The Nature of Statistical Learning Theory. Springer, New York.
Vapnik, V.N., 1998. Statistical Learning Theory. Wiley.
Williams, C.E., Stevens, K.N., 1972. Emotions and speech: Some acoustical correlates. J. Acoust. Soc. Amer. 52, 1238–1250.
Witten, I.H., Frank, E., 2005. Data mining: Practical machine learning tools and techniques. 2nd Edition, Morgan Kaufmann, San Francisco.
Yacoub, S., Simske, S., Lin, X., Burns, J., 2003. Recognition of Emotions in Interactive Voice Response Systems. In: Proc. of Eurospeech 2003, Geneva, Switzerland, 729–732.