Evaluating automatic laughter segmentation in meetings using acoustic and acoustic-phonetic features

Khiet P. Truong and David A. van Leeuwen

TNO Human Factors
P.O. Box 23, 3769 ZG, Soesterberg, The Netherlands
{khiet.truong, david.vanleeuwen}@tno.nl

ABSTRACT

In this study, we investigated automatic laughter segmentation in meetings. We first performed laughter-speech discrimination experiments with traditional spectral features and subsequently used acoustic-phonetic features. In segmentation, we used Gaussian Mixture Models that were trained with spectral features. For the evaluation of the laughter segmentation we used time-weighted Detection Error Tradeoff curves. The results show that the acoustic-phonetic features perform relatively well given their sparseness. For segmentation, we believe that incorporating phonetic knowledge could lead to improvement. We will discuss possibilities for improvement of our automatic laughter detector.

Keywords: laughter detection, laughter

1. INTRODUCTION

Since laughter can be an important cue for identifying interesting discourse events or emotional user-states, laughter has gained interests from researchers from multidisciplinary research areas. Although there seems to be no *unique* relation between laughter and emotions [12, 11], we all agree that laughter is a highly communicative and social event in human-human communication that can elicit emotional reactions. Further, we have learned that it is a highly variable acoustic signal [2]. We can chuckle, giggle or make snort-like laughter sounds that may sound differently for each person. Sometimes, people can even identify someone just by hearing their laughter. Due to its highly variable acoustic properties, laughter is expected to be difficult to model and detect automatically.

In this study, we will focus on laughter recognition in speech in meetings. Previous studies [6, 13] have reported relatively high classification rates, but these were obtained with either given pre-segmented segments or with a sliding n-second window. In our study, we tried to localize spontaneous laughter in meetings more accuractely on a frame basis. We did not make distinctions between different types of laughter, but we rather tried to build a generic laughter model. Our goal is to automatically detect laughter events for the development of affective systems. Laughter event recognition implies automatically positioning the start and end time of laughter. One could use an automatic speech recognizer (ASR) to recognize laughter which segments laughter as a by-product. However, since the aim of an automatic speech recognizer is to recognize speech, it is not specifically tuned for detection of non-verbal speech elements such as laughter. Further, an ASR system employing a full-blown transcription may be a bit computationally inefficient for the detection of laughter events. Therefore, we rather built a relatively simple detector based on a small number of acoustic models. We started with laughter-speech discrimination (which was performed on pre-segmented homogeneous trials), and subsequently, performed laughter segmentation in meetings. After inspection of some errors of the laughter segmentation in meetings, we believe that incorporating phonetic knowledge could improve performance.

In the following sections we desribe the material used in this study (Section 2), our methods (Section 3) and we explain how we evaluated our results (Section 4). Subsequently, we show our results (Section 5) and discuss how we can improve laughter segmentation (Section 6).

2. DATABASE

We used spontaneous meetings from the ICSI Meeting Recorder Corpus [8] to train and test our laughter detector (Table 1). The corpus consists of 75 recorded meetings with an average of 6 participants per meeting and a total of 53 unique speakers. We used the close-talk recordings of each participant. The first 26 ICSI 'Bmr' ('Bmr' is a naming convention of the type of meeting at ICSI) meetings were used for training and the last 3 ICSI 'Bmr' meetings (10 unique speakers, 2 female and 8 male) were used for testing. Some speakers in the training set were also present in the test set. Note that the manually produced laughter annotations were not always precise, e.g., onset and offset of laughter were not always marked.

Table 1: *Amount of data used in our analyses (duration, numbers of segments in brackets).*

	Training 26 Bmr meetings	Testing 3 Bmr meetings
Speech	81 min (2422)	10 min (300)
Laughter	83 min (2680)	10 min (279)

For training and testing, we used only audible laughter events (relatively clearly as perceived by the first author). The segments consisted of solely audible laughter which means that so-called "speech-laughs" or "smiled speech" was not investigated.

3. METHOD

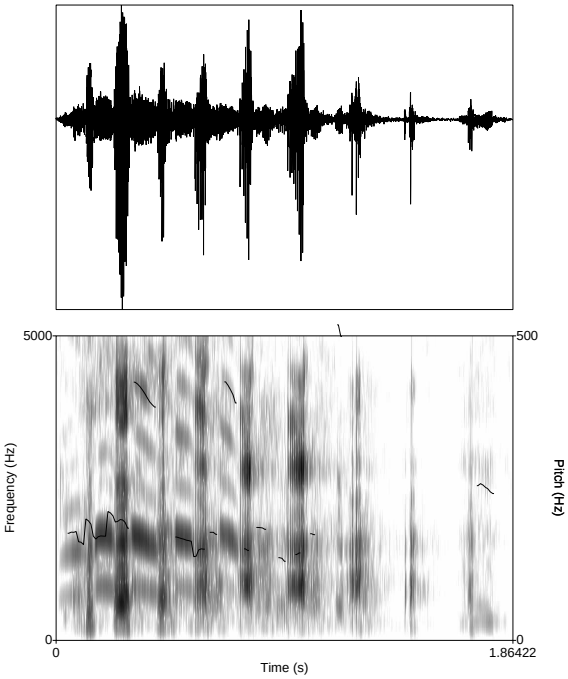
3.1. Acoustic modeling

3.1.1. Laughter-speech discrimination

For laughter-speech discrimination, we used cepstral and acoustic-phonetic features. Firstly, Gaussian Mixture Models (GMMs) were trained with Perceptual Linear Prediction Coding (PLP) features [5]. Twelve PLP coefficients and one log energy component, and their 13 first order derivatives (measured over five consecutive frames) were extracted each 16 ms over a window with a length of 32 ms. A 'soft detector' score is obtained by determining the log likelihood ratio of the data given the laughter and speech GMMs respectively.

Secondly, we used utterance-based acoustic-phonetic features that were measured over the whole utterance, such as mean log F_0, standard deviation of log F_0, range of log F_0, the mean slope of F_0, the slope of the Long-Term Average Spectrum (LTAS) and the fraction of unvoiced frames (some of these features have proven to be discriminative [13]). These features were all extracted with PRAAT [4]. Linear Discriminant Analysis (LDA) was used as a discrimination method which has as advantage that we can obtain information about the contribution of each feature to the discriminative power by examining the standardized discriminant coefficients which can be interpreted as feature weights. The posterior probabilities of the LDA classification were used as 'soft detector' scores. Statistics of F_0 were chosen because some studies have reported significant F_0 differences between laughter and speech [2] (although contradictory results have been reported [3]). A level of 'effort' can be measured by the slope of the LTAS: the less negative the slope is, the more vocal effort is expected [9]. And the fraction of unvoiced frames was chosen since due to the characteristic alternating voicing/unvoicing pattern which is often present in laughter, it is expected that the percentage of unvoiced frames is larger in laughter than in speech (which was suggested by [3]), see Fig. 1. Note that measures of F_0 can only be measured in the vocalized parts of laughter. A disadvantage of such features is that they cannot easily be used for a segmentation problem because these features describe relatively slow-varying patterns in speech that require a larger time-scale for feature extraction (e.g., an utterance). In segmentation, a higher resolution of extracted features (e.g., frame-based) is needed because accurate localization of boundaries of events is important.

Figure 1: *Example of laughter with typical voiced/unvoiced alternating pattern, showing a waveform (top) and a spectrogram (bottom).*



3.1.2. Laughter segmentation

For laughter segmentation, i.e., localizing laughter in meetings, we used PLP features and trained three GMMs: laughter, speech and silence. Silence was added because we encountered much silence in the meetings, and we needed a way to deal with it. In order to determine the segmentation of the acoustic signal into segments representing the N defined classes (in our case $N = 3$) we used a very simple Viterbi decoder [10]. In an N-state parallel topology the decoder finds the maximum likelihood state sequence. We used the state sequence as the segmentation result. We controlled the number of state transitions, or the segment boundaries, by using a small state transition probability. The state transi-

tion probability a_{ij} from state i to state $j \neq i$ were estimated on the basis of the average duration of the segments i and the number of segments j following i in the training data. The self probabilities a_{ii} were chosen so that $\sum_j a_{ij} = 1$. After the segmentation into segments $\{s_i\}$, $i = 1, \ldots N_s$, we calculated the average log-likelihoods L_{im} over each segment i for each of the models m. We defined a log-likelihood-ratio as $L_{laugh} - \max(L_{speech}, L_{silence})$. These log-likelihood-ratios determine final class-membership.

4. EVALUATION METRIC

For laughter-speech discrimination, we used the Equal Error Rate (EER) as a single performance measure, adopted from the detection framework. In laughter-speech discrimination, we can identify two types of errors: a *false alarm*, i.e., a speech segment is falsely detected as laughter, and a *miss*, i.e., a laughter segment is incorrectly detected as speech. The EER is defined as the error rate where the false alarm rate is equal to the miss rate.

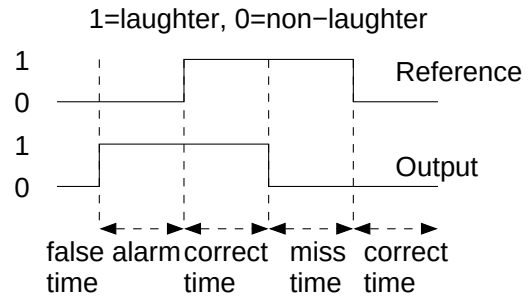
The evaluation of the automatic laughter *segmentation* was not so straightforward. One of the reasons to define log-likelihood ratios for the segments found by the detector, is to be able to compare the current results based on segmentation to other results that were obtained with given pre-segmented segments and that were evaluated with a trial-based DET analysis (Detection Error Tradeoff [7]). In this analysis we could analyze a detector in terms of DET plots and post-evaluation measures such as Equal Error Rate and minimum decision costs. In order to make comparison possible we extended the concept of the trial-based DET analysis to a time-weighted DET analysis for two-class decoding [14]. The basic idea is (see Fig. 2) that each segment in the hypothesis segmentation may have sub-segments that are either

- correctly classified (hits and correct rejects)
- missed, i.e., classified as speech (or other), while the reference says laughter
- false alarm, i.e., classified as laughter, while the reference says speech (or other)

We can now form tuples (λ_i, T_i^e) where T_i^e is the duration of the sub-segment of segment i and e is the evaluation over that sub-segment, either 'correct', 'missed' or 'false alarm'. These tuples can now be used in an analysis very similar to the DET analysis. Define θ as the treshold determining the operating point in the DET plot. Then the false alarm probability is estimated from the set $\mathcal{T}_\theta$ of all tuples for which $\lambda_i > \theta$

$$(1) \quad p_{\text{FA}} = \frac{1}{T_{\text{non}}} \sum_{i \in \mathcal{T}_\theta} T_i^{\text{FA}}$$

Figure 2: *Definitions of correct classifications and erroneous classifications in time.*



and similarly the miss probability can be estimated as

$$(2) \quad p_{\text{miss}} = \frac{1}{T_{\text{tar}}} \sum_{i \notin \mathcal{T}_\theta} T_i^{\text{miss}}$$

Here T_{tar} and T_{non} indicate the total time of target class (laughter) and non-target class (e.g., speech) in the reference segmentation.

5. RESULTS

We tested laughter-speech discrimination and laughter segmentation on a total of 27 individual channels of the close-talk recordings taken from three ICSI 'Bmr' meetings. For laughter-speech discrimination we tested with pre-segmented laughter and speech segments, while for laughter segmentation, full-length channels of whole meetings were applied. The scores (log-likelihood ratios or posterior probabilities) obtained in these audio channels were pooled together to obtain EERs, Table 2. In order to enable better comparison between the laughter-speech discrimination and the laughter segmentation results, we have also performed a segmentation experiment in which we concatenated the laughter and speech segments (used in the discrimination task) randomly to each other and subsequently performed laughter segmentation on this chain of laughter-speech segments. Thus the difference in performance in Fig. 3 is mainly caused by the presence of other sounds, such as silence, in meetings. A disadvantage of the time-weighted DET curve (used for laughter-segmentation) is that it does not take into account the absolute number of times there was an error.

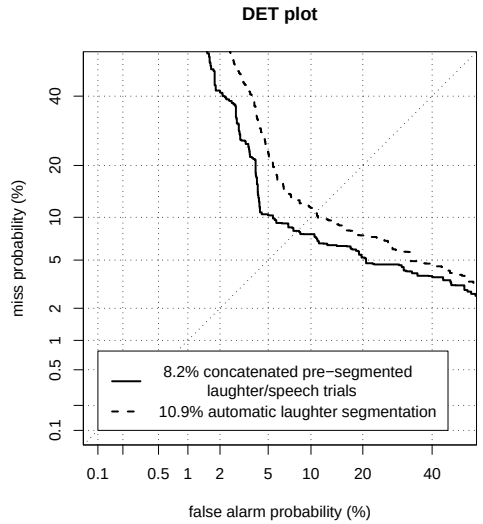
Many of the errors in laughter segmentation were introduced by sounds like, e.g., breaths, coughs, background noises or crosstalk (softer speech from other participants). It seems that, especially, unvoiced units in laughter can be confused with these type of sounds (and vice versa).

The LDA analysis with the PRAAT- features in the laughter-speech discrimination indicated that

Table 2: *EERs of laughter-speech discrimination and laughter segmentation (tested on 3 ICSI Bmr meetings). The lower the EERs, the better the performance.*

Discrimination Pre-segmented		Segmentation Concatenated laughter/speech	Whole meetings
GMM PLP	LDA PRAAT	GMM PLP	GMM PLP
0.060	0.118	0.082	0.109

Figure 3: *Time-weighted DET curves of laughter segmentation, tested on 3 ICSI Bmr meetings.*



mean log F_0 and the fraction of unvoiced frames had the highest weights, which means that these two features contributed the most discriminative power to the model. The LDA model in combination with these features seem to perform relatively well, given the small number of features used.

6. DISCUSSION AND CONCLUSIONS

We believe that the performance of the laughter segmenter can be improved by incorporating phonetic knowledge into the models. In a previous study [13], a fusion between spectral and acoustic-phonetic features showed significant improvement in laughter-speech discrimination. However, acoustic-phonetic features are usually measured over a longer timescale which makes it difficult to use these for segmentation. Currently, we are modeling laughter as a whole with GMMs that are basically one-state Hidden Markov Models (HMMs). The results of the LDA analysis indicate that we could employ phonetic information about the voiced (where we can measure F_0) and unvoiced parts of laughter (the fraction of unvoiced frames appeared to be discriminative). We could use HMMs to model subcomponents of laughter which are based on phonetic units, e.g., a VU (voiced-unvoiced) syllable could be such a phonetic unit. With HMMs, we can then better model the time-varying patterns of laughter, such as the characteristic repeating /haha/ pattern by the HMM state topology and state transition probabilities. However, for this purpose, a large database containing different laughter sounds which are annotated on different phonetic levels is needed. In addition, our laughter segmentation model may be too generic. We could build more specific laughter models for, e.g., voiced laughter, which appears to be perceived as 'more positive' by listeners [1]. Further, we have used a time-weighted DET analysis which has as an important advantage that it has a DET-like behavior so that comparisons between other studies that use DET analyses are easier to make. Disadvangtages are that it does not take into account the number of times that a detector has made an error, and our time-weighted evaluation could have been too strict (it is not clear what exactly defines the beginning and end of laughter).

We are currently implementing an online laughter detector which will be used in an interactive affective application. Additional challenges arose during the development of our online laughter detector, such as how to perform online normalization. In the future, we intend to improve our laughter detector by employing more phonetic properties of laughter.

ACKNOWLEDGEMENTS

This work was supported by a Dutch Bsik project: MultimediaN.

7. REFERENCES

[1] Bachorowski, J.-A., Owren, M. 2001. Not all laughs are alike: Voiced but not unvoiced laughter readily elicits positive affect. *Psychological Science* 12, 252–257.

[2] Bachorowski, J.-A., Smoski, M., Owren, M. 2001. The acoustic features of human laughter. *J.Acoust.Soc.Am.* 110, 1581–1597.

[3] Bickley, C., Hunnicutt, S. 1992. Acoustic analysis of laughter. *Proc. ICSLP* 927–930.

[4] Boersma, P. 2001. Praat: system for doing phonetics by computer. *Glot International*.

[5] Hermansky, H. 1990. Perceptual linear predictive (PLP) analysis of speech. *J.Acoust.Soc.Amer.* 87, 1738–1752.

[6] Kennedy, L., Ellis, D. 2004. Laughter detection in meetings. *NIST ICASSP 2004 Meeting Recognition Workshop* 118–121.

[7] Martin, A., Doddington, G., Kamm, T., Ordowski, M., Przybocki, M. 1997. The DET curve in as-

sessment of detection task performance. *Proc. Eurospeech* 1895–1898.
[8] Morgan, N., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Janin, A., Pfau, T., Shriberg, E., Stolcke, A. 2001. The meeting project at ICSI. *Proc. Human Language Technologies Conference* 1–7.
[9] Pittam, J., Gallois, C., Callan, V. 1990. The long-term spectrum and perceived emotion. *Speech Communication* 9, 177–187.
[10] Rabiner, L., Juang, B. 1986. An introduction to Hidden Markov Models. *IEEE ASSP Magazine* 3, 4–16.
[11] Russell, J., Bachorowski, J., Fernandez-Dols, J. 2003. Facial and vocal expressions of emotion. *Annu.Rev.Psychology* 54, 329–349.
[12] Schröder, M. 2003. Experimental study of affect bursts. *Speech Communication* 40, 99–116.
[13] Truong, K., Van Leeuwen, D. 2005. Automatic detection of laughter. *Proc. Interspeech* 485–488.
[14] Van Leeuwen, D., Huijbregts, M. 2006. The AMI speaker diarization system for NIST RT06s meeting data. *Proc. MLMI* 371–384.