LAUGHTER DETECTION IN MEETINGS

Lyndon S. Kennedy [1] *and Daniel P.W. Ellis* [1,2]

[1] LabROSA, Dept. of Electrical Engineering, Columbia University, NY NY
[2]International Computer Science Institute, Berkeley CA
{lyndon,dpwe}@ee.columbia.edu

ABSTRACT

We build a system to automatically detect laughter events in meetings, where laughter events are defined as points in the meeting where a number of the participants (more than just one) are laughing simultaneously. We implement our system using a support vector machine classifier trained on mel-frequency cepstral coefficients (MFCCs), delta MFCCs, modulation spectrum, and spatial cues from the time delay between two desktop microphones. We run our experiments on the 'Bmr' subset of the ICSI Meeting Recorder corpus using just two table-top microphones and obtain detection results with a correct accept rate of 87% and a false alarm rate of 13%.

1. INTRODUCTION

Meeting recordings are rich in information that stretches beyond just the words that are being spoken. Much of this information comes from cues that are elicited and expressed naturally as part of the structure and style of naturally occurring conversational speech. One interesting such cue is laughter, which can provide cues to semantically meaningful occurrences in meetings, such as jokes or topic changes. The desire to automatically uncover these high-level structures and occurrences motivates the need for the ability to automatically detect laughter in meetings, which may also help lower error rates in speech-to-text transcription by increasing robustness of non-speech detection.

Some previous work has attempted to uncover the characteristics of laughter [2, 4] and create techniques for automatically distinguishing laughter from other sounds in clean recordings [3, 6, 5]. Bickley and Hunnicutt [2] study the acoustic properties of laughter and compare and contrast with the acoustic properties of general speech using a few examples of male and female laughter. They find that in many ways speech and laughter are similar: in the fundamental frequency range, the formant frequencies, and the breathy quality of the onset and offset of voiced portions of the laugh syllable. They find that the dissimilarities between speech and laughter are that the ratio of unvoiced to voiced segments is greater in laughter than in speech and that in laughter there tends to be noise present in the region of the third formant and an enhanced amplitude of the first harmonic.

Carter [3] conducts a study to attempt to distinguish between laughter and non-laughter noises (including speech and other non-speech sounds) using clean sound examples. Carter attempts to capture the repetitive vowel sounds in laughter by segmenting each sound clip into syllables and then cross-correlate the syllables in the time domain using a set of heuristics.

Several other previous works make attempts to detect laughter using hidden Markov models with MFCCs and other features [6, 5].

In this work, we attempt to detect laughter events in naturally occurring conversational speech (meetings, more specifically). We define our ground truth laughter events as one-second windows which are labeled as being either a laughter event or a non-laughter event where a laughter event is defined as a window in which more than a certain percentage of the meeting participants are laughing. We then train a support vector machine classifier on the data using some features to capture the perceptual spectral qualities of laughter (MFCCs, Delta MFCCs), the spatial cues available when multiple participants are laughing simultaneously, and the temporal repetition of syllables in laughter (modulation spectrum).

In Section 2 we discuss our experimental data. In Section 3 we discuss the details of the features that we use to detect laughter. And in Sections 4 and 5 we discuss the experiments that we have conducted and draw conclusions from the results.

2. EXPERIMENTAL DATA

We conduct our experiments on the Bmr subset of the ICSI Meeting Recorder Corpus [1]. This set of data contains 29 meetings (about 25 hours) with different subsets of 8 participants who met regularly to discuss the Meeting Recorder project itself. Each of the participants is fitted with a high-quality, close-talking microphone. Additionally, there are 4 high-quality tabletop microphones and 2 lower-quality table-

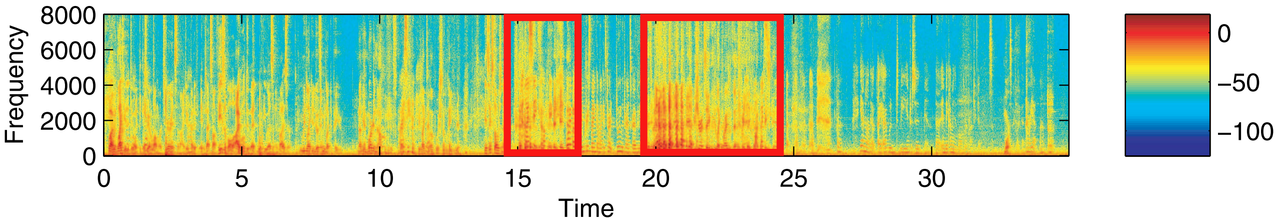

Fig. 1. Spectrogram of sample audio clip with non-laughter and laughter events (highlighted in red).

top microphones. The meetings are hand-transcribed and include additional markings for microphone noise and human produced non-speech sounds (laughter, heavy breathing, etc.). In our experiments we use only 2 of the high-quality tabletop microphones and disregard the other available channels.

We further test our system using the development data for the NIST Rich Transcription 2004 Spring Meeting Recognition Evaluation which consists of eight 10-minute excerpts from meetings from several different sites (CMU, ICSI, LDC, and NIST). Each site has a close-talking and distant microphone set-up comparable to the ICSI Meeting Recorder corpus, described above. There are also hand-transcriptions available for the meetings which also have additional markings for microphone noise and human produced non-speech sounds (including laughter).

Ground truth laughter events are determined from laughter annotations in the human-generated meeting transcripts. For each speaker turn containing a laughter annotation, the active speaker is labeled as laughing. Using one-second time frames, we calculate the level of laughter in the meeting as a percentage of the total participants who are laughing during the given frame. We apply a threshold to the percentage of laughing participants to determine which frames are part of a laughter event and which are not.

3. FEATURES

3.1. Cepstral Features

We calculate the cepstral features by first calculating the mel-frequency cepstral coefficients for each 25ms window with a 10ms forward shift. We then take the mean and variance of each coefficient over the 100 sets of MFCCs that are calculated for each one-second window.

3.2. Delta Cepstral Features

The delta cepstral features are determined by calculating the deltas of the MFCCs with a 25ms window and 10ms forward shift. We then take the standard deviation of each coefficient over the 100 sets of Delta MFCCs that are calculated for each one-second window.

3.3. Spatial Cues

We calculate the spatial features by cross-correlating the signals from two tabletop microphones over 100ms frames with a 50ms forward shift. For each cross-correlation, we find the normalized maximum of the cross-correlation (where a maximum of 1 is a perfect fit) and the time delay where the maximum occurs. We then take the mean and variance of the normalized maximum and the variance of the delay of the maximum over the 40 sets of cross-correlations that are calculated for each one-second window.

The intuition behind this feature is that in segments where several participants are laughing, the direction from which the signal comes will not be consistent. We expect that in multi-participant laughter events, the quality and delay of the best-fit cross-correlation between two tabletop microphones will be highly varied, whereas they will be mostly steady during single-speaker, non-laughter segments.

3.4. Modulation Spectrum

We calculate the modulation spectrum features by taking a one-second waveform, summing the energy in the 1000-4000Hz range in 20 ms windows, applying a Hanning window, and then taking the DFT. We use the first 20 coefficients of the DFT as our modulation spectrum features.

With this feature we are trying to catch the coarse repetition of vowel sounds, which is characteristic of most laughter. We expect to be able to capture the repeated high-energy pulses which occur roughly every 200-250ms in laughter [2].

4. EXPERIMENTS

We evaluate the performance of each of our feature sets by training a support vector machine classifier for each feature set on 26 of the available meetings and testing on the remaining 3 meetings. By rotating the test meetings through all the available meetings, each of the 1926 ground truth laughter events are included in the test results.

We evaluate each of the feature sets separately and summarize the performance results in Figure 2. We observe that

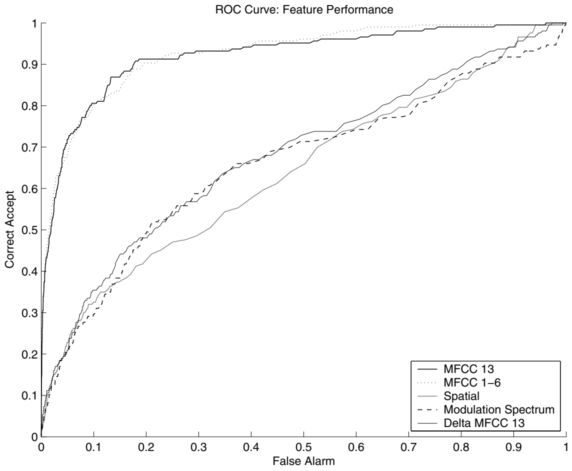

Fig. 2. Laughter detection performance of different features on the ICSI set.

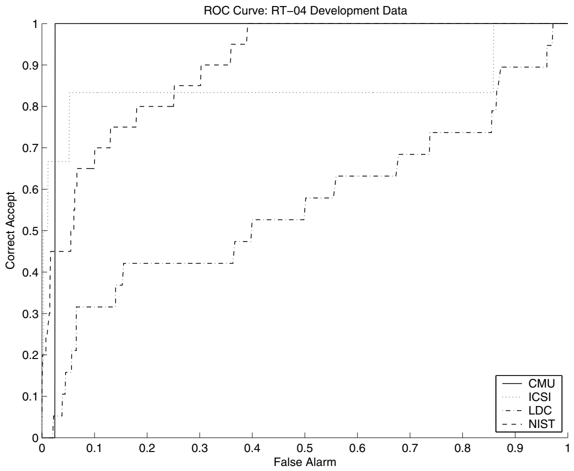

Fig. 3. Laugther detection performance trained on the ICSI set and tested on the RT-04 evaluation set.

the MFCC features considerably outperform the other features and investigate the contributions of each of the cepstral coefficients when each is used independently to train an SVM. We observe that the first cepstral coefficient gives the largest performance contribution by far and that the sixth coefficient is the next-best performing coefficient. The second, third, fourth and fifth coefficients all provide some non-trivial contribution, while coefficients beyond the sixth coefficient tend to perform no better than random guessing. With these observations taken into account, we evaluate the performance of an SVM trained on only the first six MFCCs and find that the performance is on par with an SVM trained on all 13 coefficients. The performance of this feature set is also shown in Figure 2.

We also evaluate the performance gains received by training with different combinations of features and find that no feature combination significantly outperforms MFCCs alone.

We test the generality of our approach and features by taking our model trained on the 26-meeting ICSI training set, using MFCCs and spatial features, and testing it on the RT-04 development data. The results for this evaluation are summarized in Figure 3. The CMU data contains only one positive laughter events. The ICSI, LDC, and NIST sets contain 6, 18 and 19 positive laughter events, respectively. It should be noted that the data in the ICSI set of the RT-04 development data is also in our training set and is not meaningful. It is included only for the sake of completeness.

5. DISCUSSION AND CONCLUSIONS

Our results show that of the features tried, MFCCs are by far the best-performing feature for laughter event detection. Using only the first six MFCCs, instead of all 13 coefficients, gives performance on par with all 13 coefficients. Other feature sets, including Delta MFCCs, Modulation Spectrum, and Spatial Cues provide non-trivial detection performance on their own, but do not provide much complementary discriminative power when used in conjunction with the MFCC coefficients.

We observe that training a laughter detector on data from meetings recorded at one location and applying the laughter detection model to meetings recorded at a different location with different participants yields mixed results. We see in Figure 3 that a model trained on data from the ICSI site achieves reasonable detection results on data from the CMU and NIST recordings. The detection results on the LDC recordings, however, does not appear to be significantly better than random guessing.

An interesting result that we observe is that our model is able to match the intensity of a given laughter event to a certain extent. Figure 4 demonstrates this ability, using a meeting in the ICSI test set as an example. In the top figure, we see the ground truth percentage of laughing participants in the meeting versus time. in the bottom figure, we see the certainty of our detection results as a distance from the SVM decision boundary versus time. Comparing the percentage of laughing participants to the distance from SVM decision boundary, we can see that these two values are highly correlated. In our detection, then, it becomes apparent that when we make errors, it may be due largely to thresholds which are not optimally set. We may detect a laughter event which is counted as a False Alarm, but examining the ground truth would show that there was, indeed, some laughter happening in the data and the detection error is not as bad as it may appear. Figure 5 shows a plot of detection results against the ground truth percentages and provides further evidence that the two are correlated.

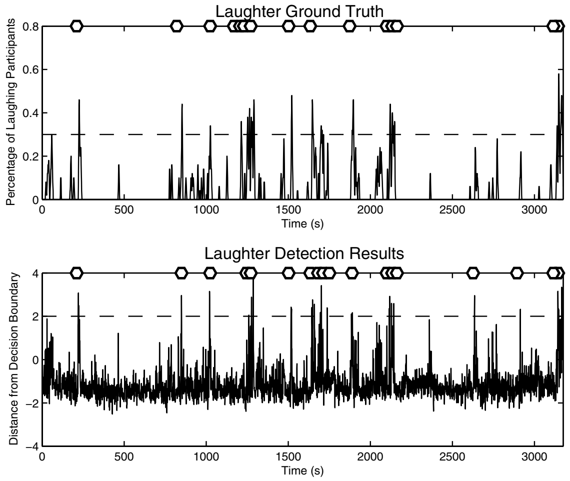

Fig. 4. Laughter ground truth and detection results for a sample meeting in the test set. Top: laughter ground truth expressed as a percentage of laughing participants; Hexagons: laughter ground truth events found by thresholding the laughing percentage at .3. Bottom: laughter detection results expressed as the distance from the SVM decision boundary; Hexagons: laughter detection results found by thresholding the distance from the SVM decision boundary at 2.

We have shown that MFCCs and SVMs provide a discriminative system for detecting laughter events in meetings. The laughter detection model derived from one set of recordings can be ported to another set of recordings with varied success. And finally, we observe that ‘certainty’ of the classification results may be used as a predictor of the intensity (percentage of laughing participants) in a given laughter event.

6. ACKNOWLEDGMENT

This work was supported in part by the NSF under grant IIS-0212396 “Mapping Meetings”, and by the European Union under the EU IST project “M4: multi-modal meeting manager” (IST-2001-34485).

7. REFERENCES

[1] N. Morgan, D. Baron, J. Edwards, D. Ellis, D. Gelbart, A. Janin, T. Pfau, E. Shriberg, , and A. Stolcke, “The meeting project at ICSI,” in *Proc. HLT*, 2001, pp. 246–252.

[2] C. Bickley and S. Hunnicutt, “Acoustic analysis of laughter.” in *Proc. Intern. Confer. on Spoken Language Processing*, Banff, 1992, pp. 927–930.

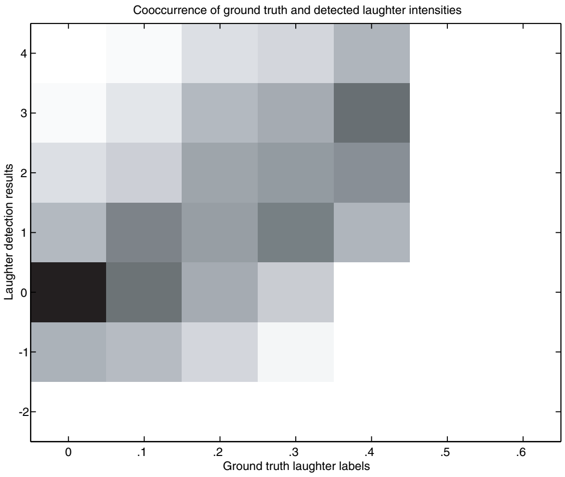

Fig. 5. Laughter detection results as a distance from the SVM decision boundary versus laughter ground truth labels. Each axis is a continuous value quantized into 7 bins. Each cell shows the probability of a particular output bin given the appearance of a particular ground truth bin. Darker cells have higher probabilites.

[3] A. Carter, “Automatic acoustic laughter detection.” Masters Thesis, Keele University, 2000.

[4] J. Trouvain, “Segmenting phonetic units in laughter” in *Proc. Intern. Confer. on the Phonetic Sciences*, Barcelona, 2003 pp. 2793-2796.

[5] R. Cai, L. Lu, H.-J. Zhang, and L.-H. Cai, “Highlight sound effects detection in audio stream.” in *Proc. Intern. Confer. on Multimedia and Expo*, Baltimore, 2003.

[6] W. Burleson, “Humor modeling in the interface position statement.” in *Proc. Conference on Humor Modeling in the Interface*, Ft. Lauderdale, 2003.