Automatic Laughter Segmentation

Mary Tai Knox

May 22, 2008

Abstract

Our goal in this work was to develop an accurate method to identify laughter segments, ultimately for the purpose of speaker recognition. Our previous work used MLPs to perform frame level detection of laughter using short-term features, including MFCCs and pitch, and achieved a 7.9% EER on the ICSI Meeting Recorder vocalized test set. We improved upon our previous results by including high-level and long-term features, median filtering, and performing segmentation via a hybrid MLP/HMM system with Viterbi decoding. Upon including the long-term features and median filtering, our results improved to 5.4% EER on the vocalized test set, which was a 32% relative improvement over our short-term MLP system, and 2.7% EER on an equal-prior test set used by others, which was a 67% improvement over previous best reported results on the equal-prior test set. After attaining segmentation results by incorporating the hybrid MLP/HMM system and Viterbi decoding, we had a 78.5% precision rate and 85.3% recall rate on the vocalized test set and a 99.5% precision rate and 88.0% recall rate on the equal-prior test set. To our knowledge these are the best known laughter segmentation results on the ICSI Meeting Recorder Corpus to date.

Contents

List of Figures

List of Tables

Acknowledgements

Several people have been instrumental throughout this project and my time at Berkeley. I would first like to thank Nikki Mirghafori, who has been and continues to be an amazing mentor. Her constant support, advice, motivation, and revisions have made this work possible. I am grateful to Nelson Morgan, my official adviser, for his input and feedback throughout this project and my graduate career. I am indebted to all of the students and research scientists at ICSI, who make ICSI a fun and lively environment and whose expansive knowledge has been one of my biggest resources. In particular I would like to thank Christian Mueller, George Doddington, Lara Stoll, Howard Lei, Adam Janin, Joan Isaac Biel, Andreas Stolcke, Vijay Ullal, and Kofi Boakye for sharing their insights. I appreciate SRI's assistance with the phone and speech recognizer outputs and Khiet Truong for sharing her datasets. I am grateful to my family for always being supportive and encouraging me to pursue my interests. Finally, I thank Galen for sharing the Berkeley experience with me and making my life more well-rounded.

This work is partially supported by the NSF under grant 0329258.

Chapter 1

Introduction

Audio communication contains a wealth of information in addition to spoken words. Specifically, laughter provides cues regarding the emotional state of the speaker [1], topic changes in the conversation [2], and the speaker's identity. Therefore, automatically identifying when laughter occurs could be useful in a variety of applications. Specifically, a laughter detector incorporated with a digital camera could be used to identify an opportune time to take a picture [3]. Laughter could also be beneficial when performing a video search for humorous clips [4]. Additionally, identifying laughter could improve many aspects of speech processing. For example, identifying non-speech sounds, such as laughter, could decrease word error rate [2]. Also, in diarization, identifying overlapped segments reduces the diarization error rate [5] and for the ICSI Meeting Recorder Corpus, the corpus used in this work, 40% of laughter time is overlapped [6]. Therefore, identifying laughter may contribute to a reduction in the diarization error rate.

The motivation for this study is to enable us to use laughter for speaker recognition, as our intuition is that many individuals have distinct laughs. Currently, state-of-the-art speech recognizers include laughter as a 'word' in their vocabulary. However, since laughter recognition is not the ultimate goal of such systems, they are not optimized for laughter segmentation. For example, SRI's conversational telephone speech recognizer [1] [7] was run on the vocalized test set, which will be described in Chapter 3, and achieved a 0.1% false

[1]This recognizer was not trained on the training set used in this study.

alarm rate and 78% miss rate; in other words when it identified laughter it was usually correct, however, most of the laughter segments were not identified. Due to the high miss rate along with the fact that laughter occurs in only slightly more than 6% of the vocalized time in the 'Bmr' subset of the ICSI Meeting Recorder Corpus, the dataset used in this work, SRI's conversational telephone speech recognizer would not be useful for speaker recognition since there would be very few laughter segments recognized from which to identify speakers. Therefore, to be able to explore the utility of laughter segments for speaker recognition, it is first necessary to build a robust system to segment laughter, which is the focus of this work.

1.1 Laughter Acoustics

Previous work has studied the acoustics of laughter [8, 9, 10, 11]. The authors differed in the extent to which they characterized laughter, with some claiming laughter has very specific attributes while others emphasized that laughter is a variable signal [9]. The differences in how specific the characterization of laughter was could be due to the vastly different number of subjects and laugh bouts studied in each experiment. For example, there were 2 subjects and 15 laugh bouts analyzed in [8], 51 laugh bouts investigated in [10], and 97 subjects and 1024 laugh bouts analyzed in [9]. Not surprisingly, due to the small sample size of some of the studies, the conclusions of these works varied and sometimes contradicted one another. Many agreed that laughter has a repetitive "breathy" consonant-vowel structure (i.e. ha-ha-ha or ho-ho-ho) [8, 10, 11]. One work went further and concluded that laughter is usually a series of short syllables repeated approximately every 210 ms [10]. Yet, others found laughter to be highly variable [9, 11], particularly due to the numerous bout types (i.e. voiced song-like, unvoiced snort-like, unvoiced grunt-like, etc.) [9], and thus difficult to stereotype. These conclusions led us to believe that automatic laughter detection is not a simple task.

1.2 Related Work

Earlier work pertaining to automatic laughter detection focused on identifying whether a *predetermined* segment contained laughter using various machine learning methods including Hidden Markov Models (HMMs) [4], Gaussian Mixture Models (GMMs) [1], and Support Vector Machines (SVMs) [2]. Note that the objectives of these studies differed as described below.

Cai et al. used HMMs trained with Mel Frequency Cepstral Coefficients (MFCCs) and perceptual features to model three sound effects: laughter, applause, and cheer. They used data from TV shows and sports broadcasts to classify 1 second windows overlapped by 0.5 seconds. They utilized the log-likelihoods to determine which classes the segments belonged to and achieved a 92.7% recall rate and an 87.9% precision rate for laughter [4].

Truong and van Leeuwen classified presegmented ICSI Meeting Recorder data as laughter or speech. The segments were determined prior to training and testing and had variable time durations. The average duration of laughter and speech segments were 2.21 and 2.02 seconds, respectively. They used GMMs trained with perceptual linear prediction (PLP), pitch and energy, pitch and voicing, and modulation spectrum features. They built models for each type of feature. The model trained with PLP features performed the best at 7.1% EER for an equal-prior test set, which will be described in Chapter 3 [1]. The EER in Truong and van Leeuwen's presegmented work was computed on the segment level, where each segment (which had variable duration) was weighted equally. Note that this is the only system that scored EER on the segment level. All other systems reported EER on the frame level, where each frame was weighted equally.

Kennedy and Ellis studied the detection of overlapped (multiple speaker) laughter in the ICSI Meetings domain. They split the data into non-overlapping one second segments, which were then classified based on whether or not multiple speakers laughed. They used SVMs trained on statistical features (usually mean and variance) of the following: MFCCs, delta MFCCs, modulation spectrum, and spatial cues. They achieved a true positive rate of 87% [2].

More recently, automatic laughter recognition systems improved upon the previous systems by detecting laughter with higher precision as well as identifying the start and end times of the segments. In particular, Truong and van Leeuwen utilized GMMs trained on PLP features with a Viterbi decoder to segment laughter. They achieved an 8% EER, where each frame was weighted equally, on an equal-prior test set [12].

1.3 Preliminary Work

Initially, we experimented with classifying laughter at the segment level using SVMs. We then proceeded to improve our laughter detection precision by utilizing Multi-Layer Perceptrons (MLPs) to detect laughter at the frame level, where each frame was 10 ms. These systems are described below. In Tables 1.1 and 1.2, we compare our work with the work of others for presegmented and frame level laughter detection, respectively.

Similar to Kennedy and Ellis [2], we experimented with using mean and variance MFCC features to train SVMs to detect laughter. We call this our *SVM system*. Initially, we calculated the features over a one second interval with a 0.5 second forward shift; in other words, every 0.5 seconds the features were calculated over a one second window thereby setting the precision of the classifier to be 0.5 seconds. This approach had good results (9% EER on the vocalized test set) but did not precisely detect start and end times of laughter segments since the data was rounded to the nearest half of a second. We then decreased the forward shift to 0.25 seconds. This system performed better with an EER of 8%. However, the time to compute the features and train the SVM increased significantly and the storage space needed to store the features approximately doubled. Furthermore, the resolution of laughter detection was still poor (only accurate to 0.25 seconds) [13].

The shortcomings of our SVM system (namely, the need to parse the data into segments, calculate and store to disk the statistics of the raw features, and poor resolution of start and end times) were resolved by using MLPs to detect laughter [13]. The MLPs were trained with short-term features, including MFCCs and pitch features, from a context window of input frames, thereby obviating the need to compute and store the means and standard deviations

since the raw data over multiple frames was included in the feature vector for a given frame. The MLPs were used to evaluate the data on a frame-by-frame basis, where each frame was 10 ms, thus eliminating the need to presegment the data, while at the same time achieving an 8% EER on the vocalized ICSI Meeting Recorder test set. This system was the basis of our current work and will be referred to as the *short-term MLP system*. Figure 1.1 shows an overview of the short-term MLP system.

Table 1.1: Previous work on presegmented laughter detection.

	Cai et al. [4]	Truong & van Leeuwen [1]	Kennedy & Ellis [2]	Knox & Mirghafori [13]
Machine Learning	HMM	GMM	SVM	SVM
Window Duration	1 s	∼2 s	1 s	1 s
Window Shift	0.5 s	0 s	0 s	0.25 s
Dataset	TV Programs	ICSI Meetings	ICSI Meetings	ICSI Meetings
Results	92.7% Recall 87.9% Precision	7.1% EER[2,3]	87% True Positive Rate	8% EER[4]

Table 1.2: Previous work on frame-based laughter detection.

	Truong & van Leeuwen [12]	Knox & Mirghafori [13]
Machine Learning	GMM	MLP
Frame Duration	0.016 s	0.010 s
Dataset	ICSI Meetings	ICSI Meetings
Results	8.2% EER[2]	7.9% EER[4]

[2]This EER was reported on the equal-prior test set. (See Chapter 3 for a description of the test set.)

[3]This is a segment-based EER, where each segment (which had variable duration) was equally weighted.

[4]This EER was reported on the vocalized test set. (See Chapter 3 for a description of the test set.)

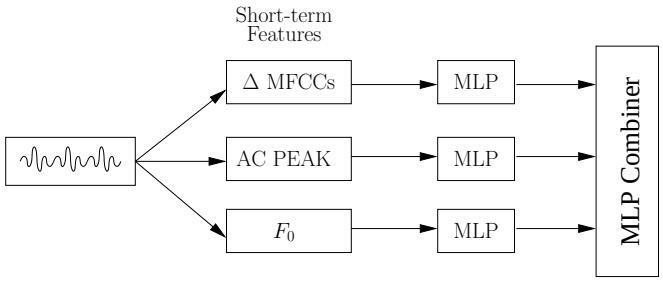

Figure 1.1: Short-term MLP system diagram.

1.4 Overview of the Current Work

In this work, we extend upon the short-term MLP system [13] in two ways: including additional features which capture the longer duration characteristics of laughter and using the output of the MLP (the posterior probabilities) to calculate the emission probabilities of the HMM. The reasons for pursuing these approaches are:

- Laughter has temporal qualities different from speech, namely a repetitive disposition [8, 10, 11]. By including long-term features we expect to improve upon the accuracy attained by the short-term MLP system.

- The short-term MLP system scored well, as mentioned in Section 1.3. Yet, its downfall was that since it classified laughter at the frame level, even small differences between the posteriors (the MLP output) of sequential frames could result in the abrupt end or start of a segment. By incorporating an HMM with Viterbi decoding, the transition probabilities can be adjusted to reflect distinct transitions from laughter to non-laughter and vice versa and the output of our system would be segments of (non-)laughter instead of frame based scores.

- An HMM alone typically assumes conditional independence between sequential acoustic frames, which may not be a good assumption for laughter (or speech). However, our MLP is set up to estimate the posterior conditioned on the features from a context window of successive frames. By including the MLP outputs in the HMM, we introduced additional temporal information without complicating the computation of the

HMM.

In summary, both short-term and long-term features were extracted from the audio. We trained MLPs, which used the softmax activation function in the output layer to compute the posterior probabilities of laughter and non-laughter, on each class of features and then performed a posterior level combination. The output of the posterior level combination was used to calculate the emission probabilities in the HMM. Finally, Viterbi decoding produced parsed laughter and non-laughter segments, which were the desired results of the processing.

1.5 Outline of Chapters

The outline for this report is as follows: in Chapter 2 we describe our hybrid MLP/HMM system set up, in Chapter 3 we describe the data used in this study, in Chapters 4 and 5 we provide and discuss our results, and in Chapter 6 we provide our conclusions and ideas for future work.

Chapter 2

Method

We extracted short-term and long-term features from our data. Similar to the short-term MLP system, we trained an MLP on each feature class to output the posterior probabilities of laughter and non-laughter. We then used an MLP combiner, with a softmax activation function to perform a posterior level combination. The softmax activation function guaranteed that the sum of the two MLP outputs (the probabilities that the frame was (non-)laughter given the acoustic features) was equal to one. The output of the posterior level combiner was then median filtered to smooth the probability of laughter for sequential frames. The median filtered posterior level combination will be referred to here as the *short- and long-term median filtered MLP system* or the *S+L-term MF MLP system*, which is shown in Figure 2.1. The outputs of the S+L-term MF MLP system (the 'smoothed' posterior probabilities of (non-)laughter) were then used in the hybrid MLP/HMM system [14] to calculate the emission probabilities for the HMM. A trigram language model was included in the HMM. Finally, the output of the hybrid MLP/HMM system was laughter segmentation. An overview of the hybrid MLP/HMM system is shown in Figure 2.2.

2.1 Features

We will now describe the short-term and long-term features used to train the MLPs. Note that not all of the extracted features were used in the final system.

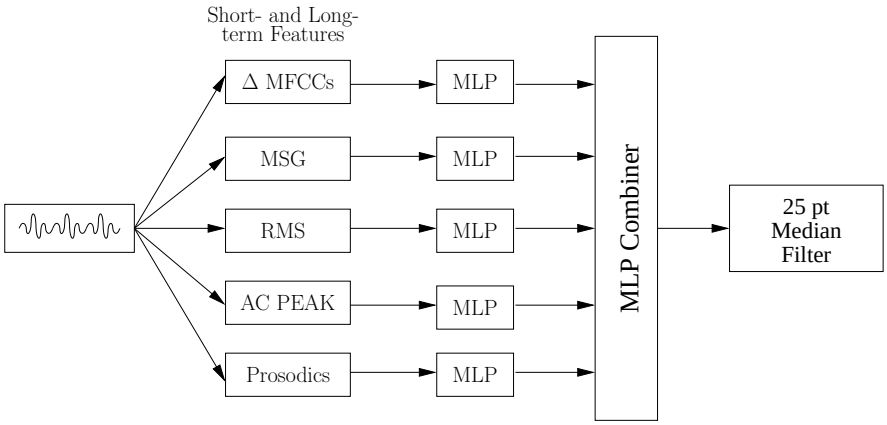

Figure 2.1: Short- and long-term median filtered MLP system diagram.

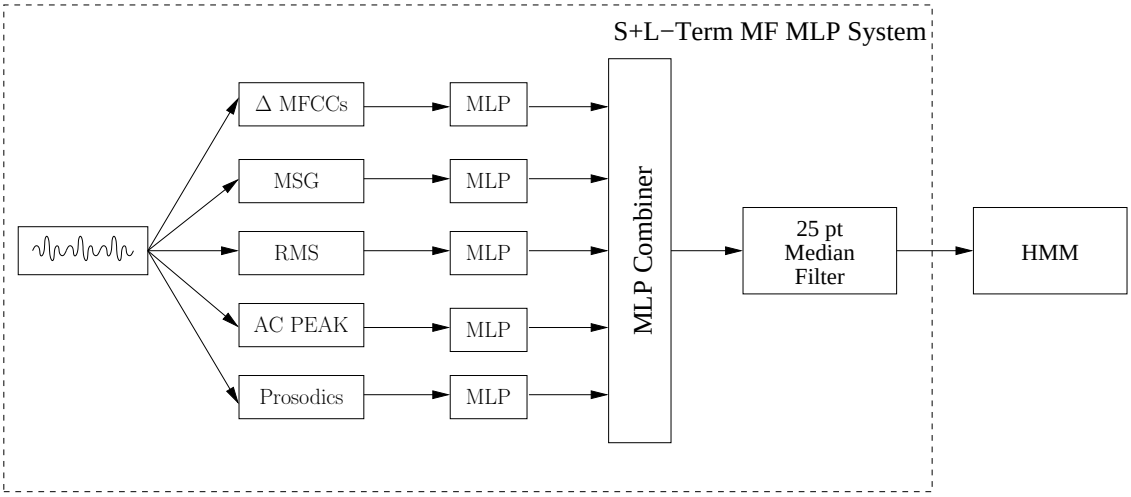

Figure 2.2: Hybrid MLP/HMM system diagram.

2.1.1 Mel Frequency Cepstral Coefficients (MFCCs)

In this study, first order regression coefficients of the MFCCs (delta MFCCs) were used to capture the short-term spectral features of (non-)laughter. The delta features were calculated for the first 12 MFCCs as well as the log energy, which were computed over a 25 ms window with a 10 ms forward shift using the Hidden Markov Model Toolkit [15]. From our short-term MLP system results [13], we found that delta MFCCs performed better than both MFCCs and delta-delta MFCCs. Moreover, results degraded when using delta MFCCs in combination with one or both of the aforementioned features. Thus, we only used delta MFCCs in this work.

2.1.2 Pitch and Energy

Studies in the acoustics of laughter [8, 9] and in automatic laughter detection [1] investigated the pitch and energy of laughter as potentially important features for distinguishing laughter from speech. Thus, we used the ESPS pitch tracker `get_f0` [16] to extract the fundamental frequency (F_0), local root mean squared energy (RMS), and the highest normalized cross correlation value found to determine F_0 (AC PEAK) for each frame (10 ms). The delta coefficients were computed for each of these features as well.

2.1.3 Phones

Laughter has a repeated consonant-vowel structure [8, 10, 11]. We hoped to exploit this attribute of laughter by extracting phone sequences. We used SRI's unconstrained phone recognizer to extract the phones. However, the phone recognizer annotates nonstandard phones including a variety of filled in pauses and laughter. Although this was not the original information we intended to extract, it seemed plausible for the 'phone' recognition to improve our previous results. Each frame produced a binary feature vector of length 46 (the number of possible 'phones'), where the only non-zero value was the 'phone' associated with the frame.

2.1.4 Prosodics

Our previous system, the short-term MLP system, included only short-term features. However, laughter has a distinct repetitive quality [8, 10, 11]. Since prosodic features are extracted over a longer interval of time, they likely would help differentiate laughter from non-laughter. We used 18 prosodic features, which were standard measurements and statistics of jitter, shimmer, and long-term average spectrum. We included 5 features of jitter (local, local absolute, relative average perturbation (RAP), 5-point period perturbation quotient, and a function of the RAP), which measures the duration differences of sequential periods. The local, local in dB, 3-, 5-, and 11-point amplitude perturbation quotients (APQ), and a function of the 3-point APQ of shimmer, which calculates the differences in amplitudes of consecutive periods, were also included as features. Moreover, statistics of the long-term average spectrum (mean, min, max, range, slope, standard deviation, and local peak height) were included. Many of these features included temporal information about the signal, which could be beneficial in identifying laughter. These features were extracted over a moving window of 0.5 seconds and a forward shift of 0.01 seconds using PRAAT [17].

2.1.5 Modulation-Filtered Spectrogram (MSG)

Modulation-filtered spectrogram (MSG) features were calculated using `msgcalc` [18]. The MSG features compute the amplitude modulations at rates of 0-16 Hz. Similar to Kennedy and Ellis [2], we used modulation spectrogram features, which capture both temporal and spectral information, to characterize the repetitiveness of laughter. Furthermore, MSG features have been shown to perform well in adverse acoustic settings [18] which could improve the robustness of our system.

2.2 MLP

A multi-layer perceptron (MLP) with one hidden layer was trained using Quicknet [19] for each of the 7 feature classes (delta MFCCs, RMS, AC PEAK, F_0, phones, prosodics, and

MSG), resulting in a total of 7 MLPs. Similar to the short-term MLP system, the input to the MLP was a context window of feature frames where the center frame was the target frame as shown in Figure 2.3 [13]. Since features from neighboring frames were included in the feature vector for a given frame, we calculated features for the entire meeting, even during times in which the speaker was silent. However, the MLP was only trained and tested on target frames that were vocalized, since only vocalized audio was included in our dataset which will be described in Chapter 3. We used the softmax activation function at the output layer to compute the probability that the target frame was laughter.

The development set was used to prevent over-fitting the MLP parameters. Specifically, the MLP weights were updated based on the training set via the back-propagation algorithm and then the development set was scored after every training epoch resulting in the cross validation frame accuracy (CVFA). The learning rate, as well as deciding when to conclude training, was determined by the CVFA improvement between epochs.

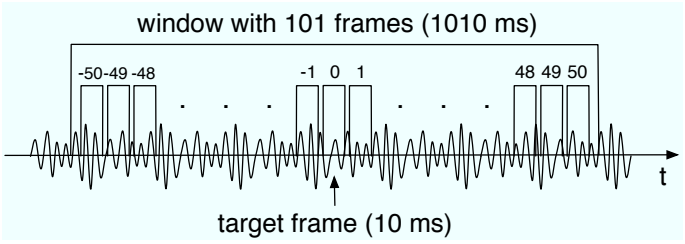

Figure 2.3: For each frame evaluated, the inputs to the MLP were features from a context window of 101 frames.

2.3 Posterior Level Combination

We performed a posterior level combination of the 7 scores, the posterior probabilities of laughter, attained from the MLPs for each feature class using an additional MLP with the softmax activation function. As in [13], because the input to the combiner was computed over a large context window (101 frames), we reduced the context window input to the MLP combiner to 9 frames. We also reduced the number of hidden units to 1 in order to keep the

complexity of the combination MLP small.

2.4 Median Filter

We found that although from one frame to the next the MLP inputs minimally changed — only one frame out of 101 frames and one frame out of 9 frames were different in the context windows for the MLPs trained on each feature class and the combination MLP, respectively — the outputs of the posterior level combination varied more than expected. Since we wanted to discourage erroneously small (non-)laughter segments, we used a median filter to smooth the posterior level combination. We experimented to empirically determine an appropriate length median filter, which will be described in Chapter 4.

2.5 Hybrid MLP/HMM

The short- and long-term median filtered MLP system, described above, computed the probability that each frame was (non-)laughter given the acoustic features over a context window. While the S+L-term MF MLP system performed well, it was not addressing the goal of this work, which is segmenting laughter. In order to segment laughter, we implemented the hybrid MLP/HMM system (see Figure 2.2), where the posteriors from the MLP combiner were used to determine the emission probabilities of the HMM using Bayes' rule and the training data was used to build a trigram language model. Viterbi decoding was performed to label the data as laughter and non-laughter segments using Noway [20]. In order to speed up Noway runtime we concatenated the vocalized data, the data evaluated in this work, leaving out audio that contained crosstalk and silence.

Chapter 3

Data

We trained and tested the segmenter on the ICSI Meeting Recorder Corpus [21], a hand transcribed corpus of multi-party meeting recordings, in which the participants were recorded individually on close-talking microphones and together on distant microphones. Since our main motivation for this work was to investigate the discriminative power of laughter for speaker recognition, we only used the close-talking microphone recordings. By doing so, we could be more sure of the identity of the speaker. The full text was transcribed in addition to non-lexical events (including coughs, lip smacks, mic noise, and most importantly, laughter). There were a total of 75 meetings in this corpus. Similar to previous work [1, 2, 12, 13], we trained and tested on the 'Bmr' subset of the corpus, which included 29 meetings. The first 21 were used in training, the next 5 were used to tune the parameters (development), and the last 3 were used to test the system.

We trained and tested only on data which was hand transcribed as vocalized. Cases in which the hand transcribed documentation had both speech and laughter listed under a single start and stop time, or laughter-colored speech, were disregarded since we could not be sure which exact time interval(s) contained laughter. Also, unannotated time was excluded. These exclusions reduced training and testing on crosstalk and allowed us to train and test on channels only when they were in use. Ideally, a silence model would be trained in this step instead of relying on the transcripts; however, due to the abundance of crosstalk in this dataset, the training of a silence model becomes more difficult. This dataset was

consistent with the results shown in [6], which found that in all of the 75 meetings in the ICSI Meeting Recorder Corpus 9% of vocalized time was spent laughing. Figures 3.1, 3.2, and 3.3 show normalized histograms of laughter, non-laughter, and both laughter and non-laughter segment durations, respectively. The segment start and end times were marked in the transcriptions. As visualized in the histograms, the variance of the segment duration was lower for laughter (1.6) than non-laughter (12.4). Furthermore, the median laughter segment duration, 1.24 s, was less than the median non-laughter segment duration, 1.51 s.

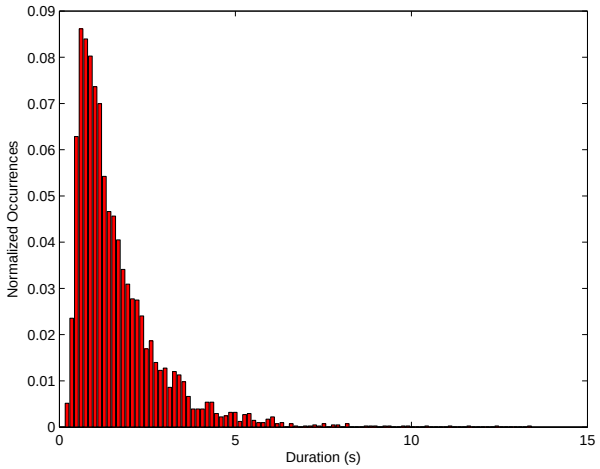

Figure 3.1: Normalized histogram of laughter segment duration in seconds.

We reported results on two test sets: the one described in the previous two paragraphs, which contained the hand transcribed vocalized data and hence is referred to as the *vocalized test set*, and an *equal-prior test set*. The vocalized and equal-prior test sets both contained data from the last 3 meetings of the 'Bmr' subset. However, for the equal-prior test set, the number of non-laughter segments used was reduced to be roughly equivalent to the number of laughter segments. Since the data was roughly equalized between laughter and non-laughter, this is referred to as the equal-prior test set. A summary of the datasets is shown in Table 3.1.

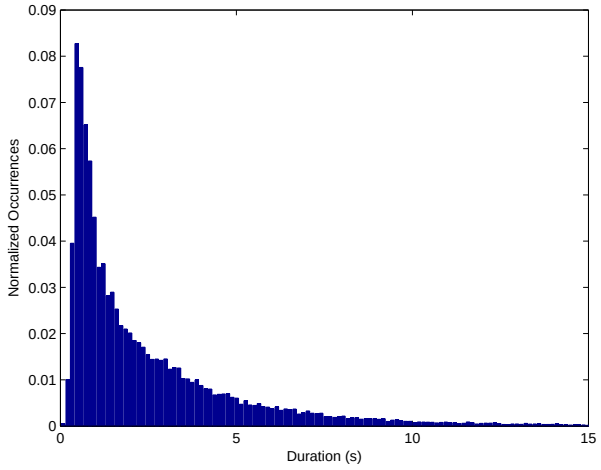

Figure 3.2: Normalized histogram of non-laughter segment duration in seconds.

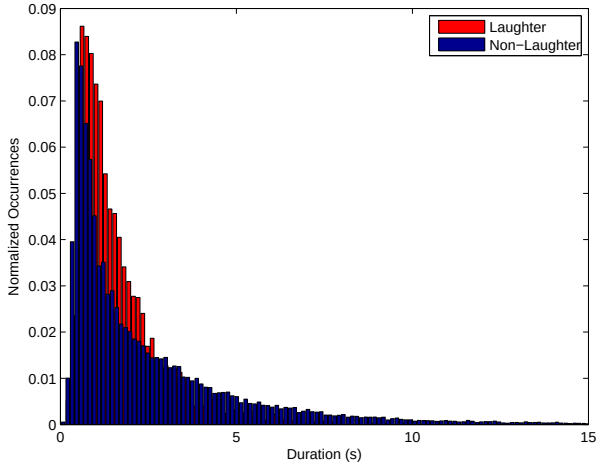

Figure 3.3: Normalized histogram of (non-)laughter segment duration in seconds.

Table 3.1: 'Bmr' dataset statistics.

	Train	Development	Vocalized Test	Eq-Prior Test
Laughter (s)	4479	1418	744	596
Non-Laughter (s)	75470	15582	7796	593
% Laughter	5.6%	8.3%	8.7%	50.2%

Chapter 4

Results

4.1 Development Set Results

Delta MFCC features performed best in our short-term MLP system [13]. Therefore, we experimented with these features to determine an appropriate context window size. We trained many MLPs varying the context window size as well as the number of hidden units. By doing so, we were able to compare systems with similar training examples to parameters ratios. The results are shown in Table 4.1 and in Table 4.2 we show the respective training examples to parameters ratios. We found that on our development set, a window size of 101 frames and 200 hidden units performed best. We then continued to use a context window of 101 frames (1.01 seconds) for each of our other features and varied the number of hidden units to see what performed best. We also experimented with mean-and-variance normalization for each of the features over the close-talking microphone channels. In Table 4.3, we show the parameters for our best systems for each feature class along with the lengths of the feature vectors, the number of hidden units, whether or not it was mean-and-variance normalized, and the achieved EER.

The MLP described in Section 2.3 was used to combine the posterior probabilities from each feature class using forward selection. As shown in Table 4.4, delta MFCCs, MSG, RMS, AC PEAK, and prosodic features combined to achieve a 6.5% EER on the development set, which was the best posterior level combination.

Table 4.1: ΔMFCC EERs (%) for various window sizes and hidden units.

Hidden Units	Window Size					
	51	75	101	125	151	175
100	11.4	9.7	9.9	9.6	10.7	11.1
200	10.9	9.8	**9.3**	9.5	10.2	10.8
300	10.7	9.86	9.6	9.7	10.2	11.0
400	10.8	10.0	9.6	9.9	10.3	11.2
500	10.9	10.0	9.7	9.8	10.3	10.8

Table 4.2: Training examples to parameters ratios for ΔMFCCs.

Hidden Units	Window Size					
	51	75	101	125	151	175
100	120	82	61	49	41	35
200	60	41	30	25	20	18
300	40	27	20	16	14	12
400	30	20	15	12	10	9
500	24	16	12	10	8	7

After examining the output of the posterior level combination, we discovered that for sequential frames the combination output posteriors still sometimes varied. In order to smooth the output and subsequently attain more segment-like results, we median filtered the best posterior level combination output. Empirically, we found that a median filter of 25 frames worked well. After applying the median filter, our EER reduced to 6.1% for the S+L-term MF MLP system. The segmentation results were evaluated in a similar manner to the MLP results in that we did frame-by-frame scoring. We calculated the false alarm and miss rates for the Viterbi decoder output, which was the output of the hybrid MLP/HMM system, and found them to be 1.8% and 20.8%, respectively. Despite the high miss rate, the hybrid MLP/HMM system was incorrect only 3.4% of the time due to the large number of non-laughter examples in the dataset.

Table 4.3: Feature class results on development set.

Feature (#)	Hidden Units	Normalized	EER (%)
ΔMFCCs (13)	200	No	9.3
MSG (36)	200	No	10.5
Prosodic (18)	50	No	13.9
AC PEAK (2)	1000	No	14.4
Phones (46)	50	No	17.3
RMS (2)	1000	Yes	20.1
F_0 (2)	1000	Yes	22.5

Table 4.4: Posterior level combination results on development set.

System	EER (%)
ΔMFCCs + MSG	7.2
ΔMFCCs + MSG + RMS	7.0
ΔMFCCs + MSG + RMS + AC	7.0
ΔMFCCs + MSG + RMS + AC + PROS	**6.5**
ΔMFCCs + MSG + RMS + AC + PROS + F_0	7.0
ΔMFCCs + MSG + RMS + AC + PROS + F_0 + Phones	7.8

4.2 Test Set Results

After tuning on the development set, we evaluated our systems on our withheld test sets. The EER was calculated for the S+L-term MF MLP system. Its output was the probability that a frame was laughter given the features and demonstrated the advantages of the S+L-term MF MLP system, which were adding the long-term features and smoothing the output via median filtering, over the short-term MLP system. Our EER reduced from 7.9% for the short-term MLP system [13] to 5.4% for the S+L-term MF MLP system on the vocalized test set, which was a 32% relative improvement. Moreover, we wanted to compare our S+L-term MF MLP system with the work of others studying laughter recognition, namely [12]. When

we evaluated our system on the equal-prior test set, we found that the EER reduced to 2.7%, which was a 67% relative improvement from the 8.2% EER reported in [12].

We then ran the vocalized test set through the hybrid MLP/HMM system and the output segmentation had a 2.2% false alarm rate and 14.7% miss rate (or incorrect 3.3% of the evaluated time). The precision and recall rates were 78.5% and 85.3%, respectively. For the equal-prior test set, we had a 0.4% false alarm rate and 12.0% miss rate, resulting in being incorrect 6.2% of the time. We calculated the precision to be 99.5% and the recall to be 88.0% on the equal-prior test set.

Chapter 5

Discussion

The inclusion of long-term and temporal features significantly improved our results on our vocalized test set (from 7.9% reported in [13] to 5.4% EER for the S+L-term MF MLP system). We believe these features exploited the repetitive consonant-vowel structure of laughter to distinguish non-laughter from laughter.

Furthermore, we found that our results dramatically improved when we used the S+L-term MF MLP system on the equal-prior test set previously used in [12]. Specifically, the S+L-term MF MLP system had a 2.7% EER on the equal-prior test set, which was a 67% improvement over the previous best reported results on the equal-prior test set. The S+L-term MF MLP system incorporated both short-term and long-term features over a context window of frames whereas the previous best reported segmentation results on the equal-prior test set included only short-term spectral features, namely PLPs [12]. Note that although we evaluated this system on the equal-prior test set, we never modified the priors of our training data which contained laughter only 5.6% of the time, as shown in Table 3.1. Our hypothesis for the better EER for the equal-prior test set compared to the vocalized test set is that the equal-prior dataset focused on discriminating laughter from speech whereas the vocalized test set was discriminating between laughter and all other vocalized sounds. The frequency of misclassification for laughter and vocalized sounds other than speech appears to be higher, particularly for annotated heavy breathing.

Our results after segmentation were also promising. We were not operating near the EER

so we could not compare the EER of the hybrid MLP/HMM system to that of the S+L-term MF MLP system; however, we could compare the segmentation operating point with the results from the S+L-term MF MLP system. The segmentation had a 14.7% miss rate and a 2.2% false alarm rate for the vocalized test set. When the S+L-term MF MLP system had a 14.7% miss rate, the false alarm rate was 2.3%. Thus, at a 14.7% miss rate, the hybrid MLP/HMM system performed similarly for the more difficult task of marking start and stop times of laughter. We feel that laughter segmentation and diarization (which segments which speaker is speaking when) have similar structures. Thus, similar to diarization error reporting, we report the precision and recall rates to be 78.5% and 85.3%, respectively.

In order to find the weaknesses of our segmentation system, we listened to the miss and false alarm errors for the vocalized test set. Similar to [12], many of the errors occurred due to breathing sounds. A breakdown of the errors and their durations are shown in Tables 5.1 and 5.2. In Figures 5.1 and 5.2, we show the percentage that each error type contributed to the false negative and false positive rates, respectively. As shown in Figure 5.1, more than half of the false negative errors were in fact not laughter at all. A large portion of this discrepancy arose due to annotated breath-laughs, which often times was simply a single breath. Thus, in actuality the false negative rate is lower than previously reported. From Figure 5.2, it is clear that breathing is often mistaken for laughter. This could be the case for a couple of reasons. First, portions of laughter do often sound like breathing, particularly when the microphone is located close to the mouth. Second, the annotated breath-laughs mentioned earlier, which are more similar to breathing than laughter, were used to train the laughter detector; therefore, the laughter training was contaminated with examples which were not laughter.

In order to see how training on the breath-laughs affected the error rate, we trained our S+L-term MF MLP system after removing all annotated breath-laughs and scored the output using the equal-prior test set, which did not include the annotated breath-laughs. Surprisingly, the EER increased from 2.7% to 3.1%. This increase in EER leads us to believe more in the validity of our first hypothesis, that laughter often sounds like breathing in this dataset especially due to the close-talking microphones.

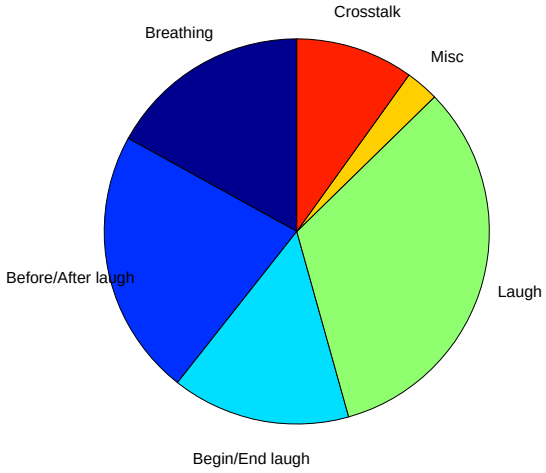

Figure 5.1: Breakdown of false negative errors.

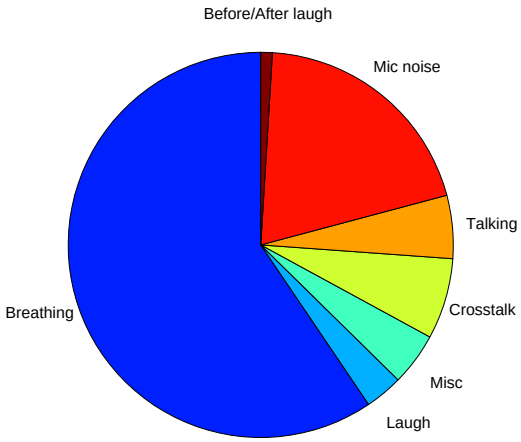

Figure 5.2: Breakdown of false positive errors.

Table 5.1: False negative error types and durations.

Error Description	Duration (s)
Laugh	35.13
Begin/End laugh	15.97
Before/After laugh	23.86
Breathing	18.08
Crosstalk	10.55
Misc	2.96
Total	106.55

Table 5.2: False positive error types and durations.

Error Description	Duration (s)
Laugh	5.41
Before/After laugh	1.65
Breathing	101.18
Crosstalk	11.54
Talking	9.01
Mic noise	33.83
Misc	7.54
Total	170.16

Chapter 6

Conclusions and Future Work

Automatic laughter detection has the potential to influence computer-human interaction, largely due to the additional emotional knowledge the computer gains. Based on this information, computer responses can be adapted appropriately.

In our preliminary study, we used SVMs trained on statistics of MFCCs over a one second window to classify a 0.25 second segment as containing laughter or not. Due to the time and storage space required to compute and store such features and the low precision, we determined that using an MLP (which can use features from a context window of frames to score a single frame) was a more valuable modeling tool. Our short-term MLP system had a 7.9% EER on the vocalized test set and was trained on MFCCs and pitch features.

Although the EER of the short-term MLP system was relatively low, as reported in this work, we have since significantly improved results in the area by including high-level and long-term features, which capture more of the temporal features of laughter, as well as incorporating an HMM, which factors in state transitions which are beneficial to segmentation. We achieved a 5.4% EER on the vocalized test set and a 2.7% EER on the equal-prior test set using the short- and long-term median filtered MLP system. After incorporating an HMM and performing Viterbi, we segmented laughter as opposed to making a frame level decision. The hybrid MLP/HMM system had a 78.5% precision rate and 85.3% recall rate on the vocalized test set and a 99.5% precision rate and 88.0% recall rate on the equal-prior test set. To our knowledge, these are the best results reported on the ICSI Meeting Recorder

Corpus to date.

In the future, the results of this work could be used in speaker recognition and emotion recognition. As mentioned previously, the motivation for this work was to investigate the discriminative power of laughter for speaker recognition. Using the hybrid MLP/HMM system, features could be extracted over the identified laughter segments and used in speaker recognition. Also, silence, in addition to laughter and other vocalized sounds, could be included in the hybrid MLP/HMM detection system in order to process all of the data instead of only vocalized segments. To evaluate the benefits of using laughter features in speaker recognition, the NIST Speaker Recognition Evaluation (SRE) datasets would be a good resource. In addition to having numerous speakers, most of the SRE data was recorded on phones which have limited crosstalk and tend to have less audible breathing, which should make laughter detection easier. Another related area of interest is to identify types of laughter. By doing so, one could get a more detailed perspective of the interactions that are occurring. This could also be used to improve laughter detection by pooling data across the different laughter types.

Bibliography

[1] K. Truong and D. van Leeuwen, "Automatic detection of laughter," in *INTERSPEECH*, 2005.

[2] L. Kennedy and D. Ellis, "Laughter detection in meetings," in *ICASSP Meeting Recognition Workshop*, 2004.

[3] A. Carter, "Automatic acoustic laughter detection," Master's thesis, Keele University, 2000.

[4] R. Cai, L. Lu, H. Zhang, and L. Cai, "Highlight sound effects detection in audio stream," in *IEEE ICME*, 2003.

[5] K. Boakye, B. Trueba-Hornero, O. Vinyals, and G. Friedland, "Overlapped speech detection for improved diarization in multiparty meetings," in *ICASSP*, 2008.

[6] K. Laskowski and S. Burger, "Analysis of the occurrence of laughter in meetings," in *INTERSPEECH*, 2007.

[7] A. Stolcke, B. Chen, H. Franco, V. Ramana Rao Gadde, M. Graciarena, M.-Y. Hwang, K. Kirchhoff, A. Mandal, N. Morgan, X. Lei, T. Ng, M. Ostendorf, K. Sonmez, A. Venkataraman, D. Vergyri, W. Wang, J. Zheng, and Q. Zhu, "Recent innovations in speech-to-text transcription at SRI-ICSI-UW," *IEEE TASLP*, vol. 14, pp. 1729–1744, September 2006.

[8] C. Bickley and S. Hannicutt, "Acoustic analysis of laughter," in *ICSLP*, 1992.

[9] J. Bachorowski, M. Smoski, and M. Owren, "The acoustic features of human laughter," *Acoustical Society of America*, pp. 1581–1597, 2001.

[10] R. Provine, *Laughter: A Scientific Investigation*. New York: Viking Penguin, 2000.

[11] J. Trouvain, "Segmenting phonetic units in laughter," in *ICPhS*, 2003.

[12] K. Truong and D. van Leeuwen, "Evaluating laughter segmentation in meetings with acoustic and acoustic-phonetic features," in *Workshop on the Phonetics of Laughter*, 2007.

[13] M. Knox and N. Mirghafori, "Automatic laughter detection using neural networks," in *INTERSPEECH*, 2007.

[14] H. Bourlard and N. Morgan, *Connectionist Speech Recognition: A Hybrid Approach*. Boston: Kluwer Academic Publishers, 1994.

[15] "Hidden markov model toolkit (HTK)," http://htk.eng.cam.ac.uk/.

[16] D. Talkin, "A robost algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis (W.B. Kleijn and K.K. Paliwal)*. New York: Elsevier, 1995, pp. 495–518.

[17] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer," http://www.praat.org/.

[18] B. Kingsbury, N. Morgan, and S. Greenberg, "Robust speech recognition using the modulation spectrogram," *Speech Communication*, vol. 25, pp. 117–132, August 1998.

[19] D. Johnson, "Quicknet3," http://www.icsi.berkeley.edu/Speech/qn.html.

[20] S. Renals, "Noway," http://www.icsi.berkeley.edu/~dpwe/projects/sprach/sprachcore.html.

[21] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "The ICSI meeting corpus," in *ICASSP*, 2003.