# Getting the Last Laugh: Automatic Laughter Segmentation in Meetings

*Mary Tai Knox*[1,2], *Nelson Morgan*[1,2], *Nikki Mirghafori*[1]

[1]International Computer Science Institute, Berkeley, California, USA
[2]Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, USA
{knoxm, morgan, nikki}@icsi.berkeley.edu

## 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 our 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 our test set and 2.7% EER on an equal-prior test set used by others. 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 our test set. To our knowledge these are the best known laughter detection results on the ICSI Meeting Recorder Corpus to date.

**Index Terms**: laughter segmentation, emotion recognition, hybrid MLP/HMM

## 1. Introduction

As the sophistication of speech systems increases, there is more of a need to recognize features other than words. One area of research that could improve speech and speaker recognition is the study of nonverbal sounds like laughter.

Laughter recognition could be useful in many aspects of speech processing. For example, identifying laughter could decrease the word error rate by identifying nonspeech sounds [10]. Also, in diarization, identifying overlapped speech reduced the diarization error rate [3] and for the ICSI meeting recorder corpus 40% of laughter time was overlapped [13]. Therefore, identifying laughter may help reduce 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 [16] was run on the same test set used in this study and achieved a 0.1% false 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 occurred in only slightly more than 6% of the evaluated time in this dataset, 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 paper. Note that SRI’s conversational telephone speech recognizer was not trained on the training set used in this study.

Earlier work pertaining to automatic laughter detection focused on identifying whether a *predetermined* segment (usually 1 second or longer) contained laughter using various machine learning methods including Hidden Markov Models (HMMs) [6], Gaussian Mixture Models (GMMs) [19], and Support Vector Machines (SVMs) [10]. 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, we previously used Multi-Layer Perceptrons (MLPs) trained on short-term features to classify each frame (10 ms) as laughter or non-laughter and achieved an 8% Equal Error Rate (EER) on our test set [12]. This system was the basis of our current work and will be referred to as the *short-term MLP system*. Also, Truong and van Leeuwen utilized GMMs with a Viterbi decoder to segment laughter and achieved an 8% EER on an equal-prior test set [20], which will be described in Section 3.

In this work, we extend upon the short-term MLP system [12] 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 than speech, namely a repetitive disposition [2, 14, 18]. 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. Yet, its downfall was that since it classified laughter at the frame level, even small differences between the posteriors (MLP outputs) of sequential frames could result in the abrupt start or end 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.

This paper is outlined as follows: in Section 2 we explain our laughter segmentation system, in Section 3 we describe the data used in this work, in Section 4 we provide the results of our systems, in Section 5 we discuss our results, and in Section 6 we give our conclusions as well as areas of future work.

## 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 (non-)laughter. We then used an MLP combiner, with a softmax activation function, to perform a posterior level combination. The softmax activation function guarantees that the sum of the two MLP outputs (the probabilities that the frame was (non-)laughter given the acoustic features) is 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 *MF MLP system*. The outputs of the MF MLP system (the 'smoothed' posterior probabilities of (non-)laughter) were then used in the hybrid MLP/HMM system [5] 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.

### 2.1. Features

We will 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.

#### *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 [7]. From our short-term MLP system results [12], we found that delta MFCCs performed better than both MFCCs and delta-delta MFCCs. Moreover, the 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 [1, 2] and in automatic laughter detection [19] 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` [17] 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 [2, 14, 18]. 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 annotated 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' label 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 [2, 14, 18]. 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. These features were extracted over a moving window of 0.5 seconds and a forward shift of 0.01 seconds using PRAAT [4].

#### *2.1.5. Modulation-Filtered Spectrogram (MSG)*

Modulation-filtered spectrogram (MSG) features were calculated using `msgcalc` [11]. The MSG features compute the amplitude modulations at rates of 0-16 Hz. Similar to Kennedy and Ellis [10], we used modulation spectrogram features to characterize the repetitiveness of laughter. Furthermore, MSG features have been shown to perform well in adverse acoustic settings [11] which could improve the robustness of our system.

### 2.2. MLP

A multi-layer perceptron (MLP) with one hidden layer was trained using Quicknet [9] 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 [12], the input to the MLP was a context window of feature frames where the center frame was the target frame as shown in Figure 1. 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.

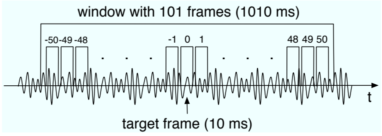

Figure 1: For each frame evaluated, the inputs to the MLP were features from a context window of 101 frames.

### 2.3. Posterior Level Combination/Median Filter

We performed a posterior level combination of the 7 scores attained from the MLPs for each feature class using an additional MLP with the softmax activation function. As in [12], because the input to the combiner was computed over a large context window (101 frames), we reduced the context window of the combiner to 9 frames. We also reduced the number of hidden units to 1 in order to keep the complexity of the MLP small.

We found that although from one frame to the next the MLP inputs minimally changed, the outputs of the posterior level combination varied more than expected. To discourage erroneously small (non-)laughter segments, we used a median filter to smooth the posterior level combination.

### 2.4. Hybrid MLP/HMM

The MF MLP system, described above, computed the probability that each frame was laughter given the acoustic features over a context window. While the MF MLP system performed well, it was not addressing the goal of this paper, which is segmenting laughter. In order to segment laughter, we implemented the hybrid MLP/HMM system [5], where the posteriors from the MF MLP system 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 [15]. 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.

## 3. Data

We trained and tested the segmenter on the ICSI Meeting Recorder Corpus [8], 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 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 speaker's identity. 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 [10, 12, 19, 20], 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 detector.

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 were disregarded since we could not be sure which exact time interval(s) contained laughter. Also, unannotated time was excluded. This exclusion 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. This data is consistent with the results shown in [13], which found that over all 75 meetings in the ICSI Meeting Recorder Corpus 9% of vocalized time was spent laughing.

We also reported results on an *equal-prior test set* in order to compare to the work of others. Similar to our test set, the equal-prior test set used in [20] contained data from the last 3 meetings of the 'Bmr' subset. However, for the equal-prior test set, the number of non-laughter segments was reduced to be roughly equivalent to the number of laughter segments. Since the data was roughly equalized between laughter and non-laughter, this test set is referred to as the equal-prior test set. A summary of the datasets is shown in Table 1.

Table 1: *'Bmr' dataset statistics.*

| | Train | Develop-<br>ment | Test | Eq-Prior<br>Test |
|---|---|---|---|---|
| Laughter (s) | 4479 | 1418 | 744 | 596 |
| Non-Laughter (s) | 75470 | 15582 | 7796 | 593 |
| % Laughter | 5.6% | 8.3% | 8.7% | 50.2% |

## 4. Experiments and Results

### 4.1. Development Set Results

Delta MFCC features performed best in our short-term MLP system [12]. Therefore, we experimented using 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. 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 2 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.

Table 2: *Feature class results on development set.*

| Feature (#) | Hidden Units | Normalized | EER (%) |
|---|---|---|---|
| $\Delta$MFCCs (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 |

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 3, 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 3: *Posterior level combination results on development set.*

| System | EER (%) |
|---|---|
| $\Delta$MFCCs + MSG | 7.2 |
| $\Delta$MFCCs + MSG + RMS | 7.0 |
| $\Delta$MFCCs + MSG + RMS + AC | 7.0 |
| **$\Delta$MFCCs + MSG + RMS + AC + PROS** | **6.5** |
| $\Delta$MFCCs + MSG + RMS + AC + PROS + $F_0$ | 7.0 |
| $\Delta$MFCCs + MSG + RMS + AC + PROS + $F_0$ + Phones | 7.8 |

After examining the output of the posterior level combination, we discovered that for sequential frames the 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 MF MLP system.

The segmentation results were scored 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 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.

### 4.2. Test Set Results

After tuning on the development set, we evaluated our systems on our withheld test set. The EER was calculated for the MF MLP system. Its output was the probability that a frame was laughter given the features and demonstrated the advantages of the MF MLP system over the short-term MLP system, which were adding the long-term features and smoothing the output via median filtering. Our EER reduced from 7.9% for the short-term MLP system to 5.4% for the MF MLP system, which was a 32% relative improvement. Moreover, we wanted to compare our MF MLP system with the work of others studying laughter recognition, namely [20]. 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 [20].

We then ran our 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 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.

## 5. Discussion

The inclusion of long-term and temporal features significantly improved our results on our test set (from 7.9% reported in [12] to 5.4% EER for the MF MLP system). We believe these features exploited the repetitive consonant-vowel structure of laughter to distinguish laughter from non-laughter.

Furthermore, we found that our results dramatically improved when we used the MF MLP system on the equal-prior test set previously used in [20]. Specifically, the 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. Note that although we evaluated this system on the equal-prior test set, we never modified the priors of our training data which is summarized in Table 1. Our hypothesis for the better EER for the equal-prior test set compared to our test set is that the equal-prior test set focused on discriminating laughter from speech whereas our 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 MF MLP system; however, we could compare the segmentation operating point with the results of the MF MLP system. The segmentation had a 14.7% miss rate and a 2.2% false alarm rate for our test set. When the 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 similar to the MF MLP system for the more difficult task of marking start and stop times of laughter. We feel that laughter segmentation and diarization have similar structures. Thus, similar to diarization, we report the precision and recall rates on our test set to be 78.5% and 85.3%, respectively.

In order to find the weaknesses of our segmentation system, we listened to the errors for our test set. Similar to [20], many of the errors occurred due to breathing sounds.

## 6. Conclusions and Future Work

We have significantly improved results in laughter segmentation by including high-level and long-term features. We achieved a 5.4% EER for our test set using a median filtered posterior level combination of short and long-term features (the MF MLP system). After performing Viterbi, we segmented laughter as opposed to making a frame level decision. Our hybrid system had a 78.5% precision rate and 85.3% recall rate. To our knowledge, these are the best results reported on the ICSI Meeting Recorder Corpus.

In the future, we intend to include silence in our detection system in order to process all of the data instead of only vocalized segments. We also plan on investigating the gains of using laughter features in speaker recognition.

## 7. Acknowledgments

The authors would like to thank Christian Mueller and Andreas Stolcke for their help with the prosodic and phonetic features, respectively, SRI for the phone and speech recognizer outputs, Kofi Boakye for assisting in the implementation of Noway, and Khiet Truong for sharing her datasets. This work is partially supported by the NSF under grant 0329258 and the Swiss National Science Foundation through the research network IM2.

## 8. References

[1] J. Bachorowski, M. Smoski, M. Owren, “The Acoustic Features of Human Laughter”, Acoustical Society of America, pp. 1581–1597, 2001.

[2] C. Bickley, S. Hannicutt, “Acoustic Analysis of Laughter”, Proc. ICSLP, Banff, Canada, 1992.

[3] K. Boakye, B. Trueba-Hornero, O. Vinyals, G. Friedland, “Overlapped Speech Detection For Improved Diarization in Multiparty Meetings”, Proc. ICASSP, Las Vegas, Nevada, 2008.

[4] P. Boersma, D. Weenink, Praat: Doing Phonetics By Computer, http://www.praat.org/.

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

[6] R.Cai, L.Lu, H. Zhang, L. Cai, “Highlight Sound Effects Detection in Audio Stream”, Proc.IEEE ICME, Baltimore, Maryland, USA, 2003.

[7] Hidden Markov Toolkit (HTK), Cambridge University Engineering Department, http://htk.eng.cam.ac.uk/.

[8] A. Janin et al., “The ICSI Meeting Corpus”, Proc. ICASSP, Hong Kong, 2003.

[9] D. Johnson, “QuickNet3”, http://www.icsi.berkeley.edu/Speech/qn.html.

[10] L. Kennedy, D. Ellis, “Laughter Detection in Meetings”, Proc. ICASSP Meeting Recognition Workshop, Montreal, Canada, 2004.

[11] B. Kingsbury, N. Morgan, S. Greenberg, “Robust Speech Recognition Using the Modulation Spectrogram”, Speech Communication, 25, pp. 117–132, 1998.

[12] M. Knox, N. Mirghafori, “Automatic Laughter Detection Using Neural Networks”, Proc. INTERSPEECH, Antwerpen, Belgium, 2007.

[13] K. Laskowski, S. Burger, “Analysis of the Occurrence of Laughter in Meetings”, Proc. INTERSPEECH, Antwerpen, Belgium, 2007.

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

[15] S. Renals, “Noway”, http://www.icsi.berkeley.edu/∼dpwe/projects/sprach/sprachcore.html.

[16] A. Stolcke et al., “Recent Innovations in Speech-to-Text Transcription at SRI-ICSI-UW”, IEEE TASLP, Volume 14, Issue 5, pp. 1729–1744, 2006.

[17] D. Talkin, A Robust Algorithm for Pitch Tracking (RAPT), In W.B. Kleijn and K.K. Paliwal, Speech Coding and Synthesis, New York, 1995.

[18] J. Trouvain, “Segmenting Phonetic Units in Laughter”, Proc. ICPhS, Barcelona, Spain, 2003.

[19] K. Truong, D. VanLeeuwen, “Automatic Detection of Laughter”, Proc. INTERSPEECH, Lisbon, Portugal, 2005.

[20] K. Truong, D. VanLeeuwen, “Evaluating Laughter Segmentation in Meetings with Acoustic and Acoustic-Phonetic Features”, in Workshop on the Phonetics of Laughter, Saarbrucken, Germany, 2007.