Automatic Laughter Detection

Mary Knox
18037207
knoxm@eecs.berkeley.edu

December 15, 2006

Abstract

We built a system to automatically detect laughter from acoustic features of audio. To implement the system, we trained a neural network on four classes of features: mel frequency cepstral coefficients (MFCCs), fundamental frequency, rms frequency, and ac peak. We used the ICSI\textsuperscript{1} Meeting Recorder corpus to train and test our system. The combined system of MFCCs and ac peak features performed best, with an equal error rate (EER)\textsuperscript{2} of 8.15%.

1 Introduction

There are many facets to spoken communication other than the words that are used. One of the many examples of this is laughter. Laughter is a powerful cue in human interaction because it provides additional information to the listener. Laughter communicates the emotional state of the speaker. This information can be beneficial in human-machine interaction. For example, it could be used to automatically detect an opportune time for a digital camera to take a picture (Carter, 2000). Laughter can also be utilized in speech processing by identifying jokes or topic changes in meetings and improve speech-to-text accuracy by recognizing non-speech sounds (Kennedy, L. and Ellis, D., 2004). Furthermore, many people have unique laughs so hearing a person’s laugh helps in the identification of others. The goal of this project was to build an automatic laughter detector in order to examine the many uses of laughter in the future.

Previous work has been done in both studying the characteristics of laughter (Bickley and Hunnicutt, 1992) (Bachorowski, J., Smoski, M., and Owren, M., 2001) (Provine, 1996) and building automatic laughter detectors (Carter, 2000) (Kennedy, L. and Ellis, D., 2004) (Truong, K.P. and Van Leeuwen, D.A., 2005) (Cai, R., Lu, L., Zhange, H., and Cai,L., 2003). Many researchers have found a wide range of results with respect to characterizing laughter. Many agree that laughter has a “breathy” consonant-vowel structure (Bickley and Hunnicutt, 1992) (Trouvain, 2003). Some researchers go on to make generalizations about laughter. For instance, Provine concluded that laughter is usually a series of short syllables repeated approximately every 210 ms (Provine, 1996). However, many have found that

\textsuperscript{1}International Computer Science Institute; Berkeley, CA
\textsuperscript{2}EER is the rate at which the percentage of misses is equal to the percentage of false alarms
laughter is highly variable (Trouvain, 2003) and thus difficult to stereotype (Bachorowski, J., Smoski, M., and Owren, M., 2001).

Kennedy and Ellis (Kennedy, L. and Ellis, D., 2004) ran experiments to detect when multiple people laughed using a support vector machine (SVM) classifier trained on four types of features: mel frequency cepstral coefficients (MFCCs), delta MFCCs, modulation spectrum, and spatial cues. The data was split into one second windows, which were classified as multiple speaker laughter or non-laughter events. They achieved a true positive rate (percentage of laughter that was correctly classified) of 87%.

Truong and van Leeuwen (Truong, K.P. and Van Leeuwen, D.A., 2005) used gaussian mixture models (GMM) trained with perceptual linear prediction (PLP) features, pitch and energy, pitch and voicing, and modulation spectrum. The experiments were run on presegmented laughter and speech segments. Determining the start and end time of laughter was not part of the experiment. They built models for each of the four features. The model trained with PLP features performed the best, at 13.4% equal error rate (EER) for a data set similar to the one used in our experiments.

The goal of this experiment is to automatically detect the onset and offset of laughter. In order to do so, a neural network with one hidden layer was trained with MFCC and pitch features, which will be described in Section 3.2. These features were chosen based on the results of previous experiments. Although Kennedy and Ellis and Truong and van Leeuwen used modulation spectrum, in both cases this feature did not perform well in comparison to the other features. This task is slightly different from both Kennedy and Ellis, who used 1 second windows, and Truong and van Leeuwen, who tested on presegmented data.

In Section 2, we discuss the data that was used in this experiment. Section 3 describes our neural network. Results from the experiment are given in Section 4 and in Section 5 we conclude and discuss our results.

2 Data

We used the ICSI Meeting Recorder Corpus (Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., Wooters, C, 2003) to train and test the detector. It is a hand transcribed corpus of multi-party meeting recordings, in which each of the speakers wore close-talking microphones. Distant microphones were also recorded; however, they were not used in this experiment. The full text was transcribed as well as non-lexical events, including coughs, laughs, lip smacks, etc. There were a total of 75 meetings in this corpus. In order to compare our results to the work done by Kennedy and Ellis (Kennedy, L. and Ellis, D., 2004) and Truong and van Leeuwen (Truong, K.P. and Van Leeuwen, D.A., 2005), we used the same training and testing sets, which were from the ‘Bmr’ subset of the corpus. This subset contains 29 meetings. The first 26 were used for training and the last 3 were used to test the detector.

In order to “clean” the data, we only trained and tested on data that was transcribed as pure laughter or pure non-laughter. Cases in which the hand transcribed documentation had both speech and laughter listed under a single start and stop time were disregarded. Furthermore, if a speaker was silent over a period of time then their channel at that time was not used. This reduced cross-talk (other speakers appearing on the designated speaker’s channel) and allowed us to train and test on channels only when they were in use. Table
1 has the statistics of the “clean” data. The average laugh duration was 1.615 seconds with a standard deviation of 1.241 seconds. Figure 1 shows the histogram of the laughter durations.

<table>
<thead>
<tr>
<th>Bnr statistics</th>
<th>Training Data</th>
<th>Testing Data</th>
<th>All Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pure Laughter (seconds):</td>
<td>5865.069</td>
<td>739.707</td>
<td>6604.776</td>
</tr>
<tr>
<td>Pure Non-Laughter (seconds):</td>
<td>90954.894</td>
<td>7766.517</td>
<td>98721.411</td>
</tr>
<tr>
<td>Percentage Pure Laughter (%):</td>
<td>6.058</td>
<td>8.696</td>
<td>6.271</td>
</tr>
</tbody>
</table>

![Figure 1: Histogram of Laugh Duration](image)

### Method

#### 3.1 Neural Network

A neural network is a statistical learning method which consists of an input layer, hidden layer(s), and output layer. The intuition behind it is that the output is a nonlinear function of a weighted sum of the inputs. A neural network with one hidden layer was used to classify feature vectors as either laughter or non-laughter. A schematic of a neural network is shown in Figure 2. In this case the input units are the features. The input units are linked to each of the hidden units. The hidden units then take a weighted sum of the input units to get $b_i = \sum_j W_{i,j} F_j$, where $W_{i,j}$ is the weight associated with feature $F_j$ and hidden unit $A_i$. These weights are determined via training data. An activation function, $g$, is applied to the sum, $b_i$, to determine the value of $A_i$. Similarly, to compute the output of the neural net a weighted sum is taken of the hidden units and a softmax activation function is applied to

---

For more information on neural networks, see (Russell, S. and Norvig, P., 2003)
the sum in order to determine the output values, the posterior probability of laughter and non-laughter. Using the softmax activation function ensures that the two outputs sum to one.

To prevent over-fitting (or tailoring the system to the training data such that it performs poorly on testing data), the data used to train the neural network was split into two groups: training (the first 21 Bmr meetings) and cross validation (the rest of the original training set). The weights were adjusted based on the first 21 Bmr meetings via the back-propagation algorithm, which modifies the weight based on the partial derivative of the error with respect to each weight. After each epoch (pass through the training data) the cross validation frame accuracy (CVFA) is evaluated. The CVFA is the ratio of true negatives and true positives to all cross validation data. Using the default values used for training neural networks at ICSI (Ellis, 2000), the learning rate was initially set to 0.008. Once the CVFA increases by less than 0.5% from the previous epoch, the learning rate is halved at the beginning of subsequent epochs. The next time the CVFA increases by less than 0.5%, training is stopped.

Figure 2: A neural network with n input units, 200, hidden units, and 2 output units

3.2 Features

Feature selection is an essential part of machine learning. In this case, we used MFCC and pitch features to distinguish laughter from non-laugher.

3.2.1 Mel Frequency Cepstral Coefficients

Mel frequency cepstral coefficients (MFCCs) are coefficients obtained by taking the fourier transform of a signal, converting it to the mel scale, and finally taking the discrete cosine transform of the mel scaled fourier transform (wik, 2006). The mel scale is a perceptual scale used to more accurately portray what humans hear. In this experiment, MFCCs were used to capture the spectral features of laughter and non-laughter. The first order regression coefficients of the MFCCs (delta MFCCs) and the second order regression coefficients (delta-delta MFCCs) were also computed and used as features for the neural network. We used
the first 12 MFCCs (including the energy component), which were computed over a 25 ms window with a 10 ms forward shift, as features for the neural network. MFCC features were extracted using the Hidden Markov Model Toolkit (HTK)(htk, 2006).

3.2.2 Pitch

From laughter characterization papers, it was concluded that pitch features were different for laughter than for speech. Previous studies showed that the fundamental frequency ($F_0$) pattern of laughter did not decline as it typically does in speech (Bickley and Hunnicutt, 1992) and that the $F_0$ has a large range during laughter(Bachorowski, J., Smoski, M., and Owren, M., 2001).

Using the ESPS pitch tracker get_f0 (Entropic Research Laboratory, 1993), we extracted the $F_0$, rms value, and ac peak (the highest normalized cross correlation value found to determine $F_0$) for each frame. The delta and delta-delta coefficients were computed for each of these features as well.

4 Experiments and Results

Before running our system, we needed to tweak a couple of parameters: the input window size and the number of hidden units. Since the frames were so short in duration and in this data set people generally laughed for a consecutive 1.615 seconds at a time, we realized we should feed multiple frames into the neural net to determine whether or not a person is laughing. After trying 0.25, 0.50, 0.75, and 1 second windows (which would be equivalent to 25, 50, 75, and 100 frames respectively) we found that a window of 0.75 seconds worked best, based on the CVFA. We wanted the classification of laughter to be based on the middle frame so we set the offset to be 37 frames.

We also had to determine the number of hidden units. Since MFCCs were the most valuable features for Kennedy and Ellis (Kennedy, L. and Ellis, D., 2004), we used MFCCs as the input units and modified the number of hidden units to be 50, 100, 200, and 300 while keeping all other parameters the same. Based on the accuracy on the cross validation set, we saw that 200 hidden units performed best.

The neural networks were first trained on the four classes of features: MFCCs, $F_0$, rms frequency, and ac peak. The detection error trade-off (DET) curves for the each of the classes is shown in Figures 3, 4, 5, and 6. Each plot shows the DET curve for a neural network trained on the feature itself (blue line), the deltas (green line), the delta-deltas (cyan line), and the combination of the feature, delta, and delta-delta (red line). The EERs for each system are shown in Table 2.

Since the “All” systems, which combined the feature, delta, and delta-delta, usually performed better than the individual systems we combined them to improve our results. Since the system using MFCC features had the lowest EER, we combined MFCCs with each of the other classes of features. Table 3 shows that after combining, the MFCC+AC Peak system performed the best. We then combined MFCC, ac peak, and rms frequency features. Finally, we combined all of the systems and computed the EER.
Figure 3: DET curve for MFCC features

Figure 4: DET curve for ac peak features
Figure 5: DET curve for F0 features

Figure 6: DET curve for rms features
Figure 7: DET curve for combined systems

Table 2: Equal Error Rates (%)

<table>
<thead>
<tr>
<th>Feature</th>
<th>MFCCs</th>
<th>$F_0$</th>
<th>RMS</th>
<th>AC Peak</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feature</td>
<td>11.35</td>
<td>23.26</td>
<td>32.22</td>
<td>16.75</td>
</tr>
<tr>
<td>Delta</td>
<td>9.62</td>
<td>24.42</td>
<td>26.52</td>
<td>22.37</td>
</tr>
<tr>
<td>Delta-Delta</td>
<td>11.23</td>
<td>27.83</td>
<td>26.62</td>
<td>27.61</td>
</tr>
<tr>
<td>All</td>
<td>10.66</td>
<td>22.80</td>
<td>26.01</td>
<td>16.72</td>
</tr>
</tbody>
</table>

Table 3: Equal Error Rates for Combined Systems (%)

<table>
<thead>
<tr>
<th>Feature</th>
<th>EER</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC+$F_0$</td>
<td>9.38</td>
</tr>
<tr>
<td>MFCC+RMS</td>
<td>8.59</td>
</tr>
<tr>
<td>MFCC+AC Peak</td>
<td>8.15</td>
</tr>
<tr>
<td>MFCC+AC Peak+RMS</td>
<td>8.19</td>
</tr>
<tr>
<td>MFCC+AC Peak+RMS+$F_0$</td>
<td>8.80</td>
</tr>
</tbody>
</table>

5 Discussion and Conclusions

From Table 2, it is clear that MFCC features outperformed pitch related features. This is consistent with Kennedy and Ellis’ (Kennedy, L. and Ellis, D., 2004) results. For Truong and van Leeuwen (Truong, K.P. and Van Leeuwen, D.A., 2005), PLP features outperformed the other features. PLP, like MFCCs, are perceptually scaled spectrums so it is not surprising that they performed well for the task of laughter detection too.

AC peak values also performed well, which suggests that the cross correlation of an audio signal helps in detecting laughter. This seems reasonable since laughter is repetitive (Provine, 1996). However Provine found that the repetitions were 210 ms apart, which
exceeds the time used to compute the cross correlation.

Comparing the combined systems, we see that the combination of MFCC and ac peak features performed the best. This shows that by combining different systems we can gain more information about the data and achieve better EERs. However, when we added rms frequency and fundamental frequency features, the EER increased. A possible reason for this increase is that since the ac peak, rms frequency, and fundamental frequency are so related to one another the addition of the features did not contribute new information about the data and instead only cluttered the system with more inputs.

In the future we plan to use additional features to detect laughter. Trouvain noted the repetition of a consonant-vowel syllable structure (Trouvain, 2003). Based on this model, we could run a phoneme recognizer on the audio and detect when a phoneme is repeated. Another approach would be to compute the modulation spectrum. This should also portray the repetitive nature of laughter. Although previous experiments have shown it to perform worse than other features such as MFCCs and pitch, it may improve the EER if combined with other features.

In conclusion, we have shown that neural networks can be used to automatically detect the onset and offset of laughter with an EER of 8.15%. These results are slightly better than previous experiments which classified presegmented data. By adding more features, we hope to further improve these results.

6 Acknowledgements

We would like to thank Nikki, Chuck, Howard, Lara, Khiet, Mathew, and Galen for their help with this experiment.
References


Trouvain, J. (2003). Segmenting phonetic units in laughter. in Proc ICPHS.