Laughter Detection in Noisy Settings

Mary Felkin, Jérémy Terrien and Kristinn R. Thórisson

Center for Analysis and Design of Intelligent Agents
Reykjavik University, Menntavegur 1, 101 Reykjavik

Abstract. Spontaneous human speech contains a lot of sounds that are not proper speech, yet carry meaning, laughter being a good example. Recognizing such sounds from speech-sounds could improve speech recognition systems as well as widen the communicative range of automatic dialogue systems. Our goal is to develop methods for automatic classification non-speech vocal sounds. As laughter varies widely between individuals and cultures it represents a nice subset for studying various detection and analysis techniques for this purpose. The approach we describe here is based on the C4.5 machine learning algorithm. We focus on finding the onset and offset of laughter using single-speaker audio recordings. Prior efforts using machine learning have not, to our knowledge, used C4.5. To the best of our knowledge, our results are the best so far detecting laughter from non-laughter sounds, using a single-speaker/single-microhpone signal with noisy background (general office environment), 89.9% at best. Here we describe our method and detail the results from two separate experiments, the first on simply detecting laughter and the second applying this method for differentiating between three different kinds of non-laughter sounds.

Keywords: Sentiment analysis, Speech processing, Spoken language processing, Machine learning, Laughter recognition

1 Introduction

The importance of laughter in human relationships can hardly be contested; its importance in communication has been pointed out by a number of authors. Like much of the prior work on laughter detection, our aim is to improve speech recognition by eliminating periods of non-speech sound. As false positives constitute a significant portion of speech recognition errors, a high-quality solution in this respect could be expected to improve speech recognition considerably. Furthermore, our ultimate goal is to apply these techniques in a robot or virtual humanoid to enable it to produce the appropriate conversational responses in realtime dialogue with people. Many prior papers on automatic laughter detection leave out details on the average duration of the laughter and only mention the length of the full recordings containing (one or more bursts of) laughter – these presumably being the recordings that got them the best results. In our corpus, laughter duration of 2.5 s produced the highest accuracy. The recordings used

here contained a fair amount of background noise, as generally found in office environments, including people talking, objects being moved around, etc. (the noise was, however, nowhere nearly as loud as the primary signal). We filtered the signal with a X-step filtering technique and applied C4.5 learning algorithm to the filtered signal. The accuracy of our laughter recognition system is comparable to results previously obtained from clean sound samples and, to the best of our knowledge, better than any previously obtained from noisy data.

The paper is organized as follows: After a review of related work we describe the signal processing algorithms employed and show how correlated their output is. Then we describe the results from training C4.5 on the corpus and present the results of applying it to new data.

2 Related Work

A number of papers have been published on the application of machine learning for detecting the difference between speech and laughter in audio recordings [12] [25] [7] [24] [9]. The work differs considerably on several dimensions including the cleanliness of data, single-person versus multiple-person soundtracks, as well as the learning methods used. Reasonable results of automatic recognition have been reported using support vector machines [25], [7], Hidden Markov Models [12] [13], artificial neural nets [9] [25] and Gaussian Mixture Models [25], [24]. Some of the studies ([20], [9], [24], [25], [7], [11], [22], [19] and [12]) relie on very expensive databases such as the ICSI meeting corpus [5] and [18]. Use of a common corpus might make one think it possible to easily compare results between studies. The studies, however, pre-process data in many ways, from manual isolation of laughter versus non-laughter segments, to completely free-form multi-party recordings. They are therefore not easily comparable. Paper [6] describes a classification experiment during which fourteen professional actors recorded themselves reading a few sentences and expressed, in each recording, an emotion chosen from {neutral, hapiness, sadness, anger}. The authors do not give their results. In paper [14] laughter versus non-laughter is detected from among 96 audio-visual sequences and visual data is used to improve accuracy. Paper [15] also describes audio-visual laughter detection, based on temporal features and using perceptual linear prediction coefficients. Among the highest reported recognition rate for audio alone was that of [9], which reported only 10% misses and 10% false positives, with a 750msecs sample length, using a neural network on clean data. We achieve comparable accuracy on noisy data. For a review of multimodal video indexing, see [21].

Among the methods used in prior work for pre-processing are mel-frequency cepstral coefficients (MFCCs) [9] [25] (see also the seminal work introducing their use for audio processing: [23]) and Perceptual Linear Prediction features (see [24] for a an example of related use and [3] for a more general discussion).

3 Data Collection

We collected sound samples through via a simple graphical interface; subjects were a convenience sample of volunteers. Recordings were done in a relatively noisy environment (people talking and moving in the background). We used a single microphone of reasonable quality, but without noise cancellation mechanisms. The quality of the recordings was such that an average human listener could clearly distinguish between the background noise and the signal (laughter), the latter being considerably louder (as we use energy-based descriptors our method could not function if the background noise was as loud as the primary signal).

The instructions to each participant were to "Please laugh into the microphone. Every sample should last at least three seconds." For the non-laughter sounds we instructed them that these could "include anything you want. We would appreciate it if you would try to give us samples which you think may be confused with laughter by a machine but not by a human. For example, if you think the most discriminant criteria would be short and rythmic bursts of sound, you could cough. If you think phonemes are important, you could say 'ha ha ha' in a very sad tone of voice, etc.".

The volunteers were asked to produce and record 20 samples, each lasting 3 seconds:

- 5 samples of laughter
- 5 samples of spontaneous speech
- 5 samples of reading aloud
- 5 samples of other sounds (OS) of their own choice

Among the other sounds people recorded were humming, coughing, singing, animal sound imitations, etc. One volunteer thought that rythmic hand clapping and drumming could also be confused with laughter by a machine so he was allowed to produce such non-vocal sounds.

The volunteers were encouraged to record themselves speaking and reading in their native languages; a web browser was at their disposal in order to enable them to find something to read. The majority of the volunteers were Icelandic and French and we also recorded Italians, Poles, two Hungarians, one Romanian woman and one Spanish man. No native English speaker participated but some volunteers also recorded themselves speaking and reading in English for the sake of diversity. About two thirds of our volunteers were men.

4 Signal Processing Using CUMSUM

We assume that each phoneme can be defined as a stationary segment in the recorded sound samples. Several algorithms have been developed to extract the stationary segments composing a signal of interest. We chose a segmentation algorithm based on auto-regressive (AR) modeling, the cumulated sums (CUMSUM) algorithm [8] (step 1 in Fig. 2. In a *change detection* context the problem

consists of identifying the moment when the current hypothesis starts giving an inadequate interpretation of the signal, so another hypothesis (already existing or created on the fly) become the relevant one. An optimal method consists in recursive calculation, at every time step, of the logarithm of the likelihood ratio $\Lambda(x_t)$. This is done by CUMSUM algorithm [1] as follows:

H_0 and H_1 are two hypothesis
$H0 : x_t, t \in]0, k]$ where x_t follows a probability density f_0
$H1 : x_t, t \in]k, n]$ where x_t follows a probability density f_1

The likelihood ratio $\Lambda(x_t)$ is defined as the ratio of the probability densities of x under both hypothesis (equation 1).

$$\Lambda(x_t) = \frac{f_1(x_t)}{f_0(x_t)} \tag{1}$$

The instant k of change from one hypothesis to the other can then be calculated according to [1], [4] (equations 2 and 3).

$$K = inf\{n \geq 1 : max \sum_{j=1}^{t} log\Lambda(x_j) \geq \lambda_0\};\ 1 \leq t \leq n \tag{2}$$

$$K = inf\{n \geq 1 : S_n - min\ S_t \geq \lambda_0\};\ 1 \leq t \leq n \tag{3}$$

Where S_t is the cumulated sum at time t, defined according to equation 4.

$$S_t = \sum_{t=1}^{n} log\Lambda(x_t);\ S_0 = 0 \tag{4}$$

In the general case, with several hypotheses, the detection of the instant of change k is achieved through the calculation of several cumulated sums between the current hypothesis H_c and each hypothesis i of the N hypotheses already identified.

We define a detection function $D(t, i) = max\ S(n, i) - S(t, i)\ for\ i \in \{1, ..., N\}$. This function is then compared to a threshold λ in order to determine the instant of change between both hypotheses.

In several instances the distribution parameters of random variable x, under the different hypothesis, are unknown. As a workaround, the likelihood ratios used by CUMSUM are set according to either signal parameters obtained from AR modeling or the decomposition of the signal by wavelet transform [8]. In this paper we used the AR modeling approach.

When the different samples xi of a signal are correlated, these samples can be expressed by an AR model (equation 5).

$$x_i + \sum_{k=1}^{q} a_k x_{i-k} = \epsilon_i;\ \epsilon_i \in N(0, \sigma) \tag{5}$$

Where :

ϵ_i is the prediction error

$a_1, ..., a_k$ are the parameters of the AR model
q is the order of the model

If x follows a Gaussian distribution the prediction errors ϵ_i also follow a Gaussian distribution and are not correlated. In this case the logarithm of the likelihood ratio of the prediction errors $\Lambda(\epsilon_i)$ can be expressed under H_0 and H_1 hypothesis as in [8] (equation 6).

$$log(\Lambda(\epsilon_i)) = \frac{1}{2} log \frac{\sigma_0^2}{\sigma_1^2} + \frac{1}{2}(\frac{(\epsilon_{i,0})^2}{\sigma_0^2} - \frac{(\epsilon_{i,1})^2}{\sigma_1^2}) \tag{6}$$

Where :

σ_j^2 is the variance of the prediction error under the j^{th} hypothesis
$\epsilon_{i,j}$ is the prediction error under the j^{th} hypothesis

When several hypotheses exist, the likelihood ratio between the current hypothesis H_c and every already identified hypothesis is calculated. The cumulated sum $S(n, i)$ at time n between the current hypothesis and the i^{th} hypothesis is calculated according to equation 7.

$$S(n, i) = S(n-1, i) + \frac{1}{2} log \frac{\sigma_c^2}{\sigma_i^2} + \frac{1}{2}(\frac{(\epsilon_{t,c})^2}{\sigma_c^2} - \frac{(\epsilon_{t,i})^2}{\sigma_i^2}) \tag{7}$$

The detection function D(t, i) is defined:
$D(t, i) = max\ S(t, i)$ - $S(n, i) for\ 1 \leq t \leq n$
The instant of change is detected whenever one of the M detection functions reaches a λ_0 threshold.

As a final step we detected silence (only background noise is heard) by energy-thresholding all other hypothesis.

5 Attribute Construction for Chunks

To separate audio segments from silence segments (step 2 in Fig. 2) we applied an energy threshold on each detected stationary segment. We chose to keep all segments that represent 80% of the energy of the original signal. All non-selected segments where considered silence and discarded from further analysis. All contiguous phonemes where then mixed to form a *burst* (step 3 in Fig. 2).

For each burst W_i we first computed their fundamental frequency, defined as the frequency of maximal energy in the burst's Fourier power spectrum. The power spectrum of the burst i ($Pxx_i(f)$) was estimated by averaged modified periodogram. We used a Hanning window of one second duration with an overlap of 75%. The fundamental frequency F_i and the associated relative energy $Erel_i$ are then obtained according to equations 8 and 9.

$$F_i = argmax_f\ Pxx_i\ (f) \tag{8}$$

$$Erel_i = \frac{max\ (Pxx_i(f))}{\sum_{f=0}^{\frac{F_s}{2}}\ Pxx_i\ (f)} \tag{9}$$

where F_s is the sampling frequency.

We also considered the absolute energy E_i, the length L_i and the time instant T_i of each burst.

5.1 Burst Series Parametrisation

A burst series is defined as a succession of n sound burst bursts. The number of bursts is not constant from one series to another. Our approach to pre-processing for audio stream segmentation was based on the following hypotheses:

Fig. 1. The values of the attributes with respect to the binary class "laughter" vs. "not laughter"

1. **F.** Maximum energy frequency: The fundamental frequency of each audio burst is constant or slowly varying. No supposition has been made concerning the value of this parameter since it could vary according to the gender of the speaker (we performed no normalisation to remove these gender-related differences in vocal tract length). It could also vary according to the particular phoneme pronounced during the laugh, i.e."hi hi hi" or "ho ho ho", or, as some native Greenanders' laugh, "t t t".
2. **Erel.** Relative energy of the maximum: The relative energy of the fundamental frequency of each burst is constant or slowly varying. This parameter should be high due to the low complexity of the phoneme.
3. **E.** Total energy of the burst: The energy of each burst is slowly decreasing. The laugh is supposed to be involuntary and thus no control of the respiration to maintain the voice level appears. This is, as we will see, a useful criterion because when a human speaks a sentence, he or she is supposed to control the volume of each burst in order to maintain good intelligibility and this control for the most part only breaks down when expressing strong emotions.
4. **L.** Instant of the middle of the burst: The length of each burst is low and constant due to the repetition of the same simple phoneme or group of such.
5. **T.** Length of the burst: The difference between consecutive burst occurence instants is constant or slowly varying. A laugh is considered as an emission of simple phonemes at a given frequency. No supposition concerning the frequency was done since it could vary strongly from one speaker to the other. At the opposite, a non laughing utterance is considered as a "random" phoneme emission.
6. **Te.** Total energy of the spectre's summit: Same as **Erel** but not normalised according to the total energy of the burst.

To differentiate laughter from a non-laughter we characterise each burst series by the regularity of each parameter. This approach allows us to be independent of the number of bursts in the recorded burst series. For the parameters F_i, $Erel_i$, E_i and L_i, we evaluate the median of the absolute instantaneous difference of the parameters. For the parameter T_i, we evaluate the standard deviation of the instantaneous emission period, i.e. $T_{i+1} - T_i$.

As can be seen in Fig. 1, the distribution of the attribute values with respect to the class is very similar for both laughter and non-laughter. This means that none of these attributes, taken on its own, would be a very good indicator of the class (though low values of T and high values of Te do somewhat discriminate between laughter and non-laughter). It is only through their induced combination that success is achieved.

Fig. 2. The complete algorithm

5.2 Machine Learning Tools

We found that no single descriptor on its own is sufficient to differentiate laughter vs. non-laughter samples, meaning that no trivial method exists to differentiate them using our descriptors. Supervised classification techniques are therefore required. We solved this problem with the decision-tree inducer *C4.5* [16] [17].

Fig. 3. Comparison: top is speech, bottom is laughter.

Fig.3 illustrates the difference between laughter and speech. These curves were produced by Audacity[1]. The tall and rather regular bursts in the lower curve are characteristic of laughter (the smaller bumps are loud breathing sounds). Speech, illustrated above, has more variety. On top of the 6 attributes described above, we tried to find other descriptors illustrating the regularity of laughter, for example setting a very high threshold to cover breathing sounds and other noises and then taking the standard deviation of the duration of the top part of the bursts and of the distance between them. In fig.3 the parts which were taken in consideration are the parts sticking out of the central area and drawn against a white background. These intuitive descriptors, and some others, detracted from the efficiency of our classification with C4.5, so they we removed them

[1] Audacity is free, open source software for recording and editing sounds: http://audacity.sourceforge.net/

from the database. They did not, however, detract from the efficiency of other classification algorithms such as Naive Bayes (the version accepting numerical attributes)[2]. In the following section we only present the results obtained with C4.5. We do not present results of other classification algorithms. We cannot. Naive Bayes results, the second bests, were only a few points below C4.5 results but were obtained from a different database, a database with the same examples but more descriptors. So they are not strictly comparable. The six descriptors we present here are the subset of the set of all implemented descriptors which is best suited for C4.5 in this experiment.

6 Results

In the following we use $10-folds$ cross validation; cross-validation is the practice of partitioning a set of data into subsets to perform the analysis on a single subset while the others are used for training the classification algorithm. This operation is repeated as many times as there are partitions, which means we train on 90 of the samples and test on the remaining 10. We do this 10 times and average the results. In this way, our accuracy is a good (if slightly pessimistic [10]) estimator of what our accuracy would be upon unknown examples.

6.1 Laugh detection

The first column below indicates the length of the samples used in the corresponding experiment as a percentage of the 3 seconds total length. It also happens with spontaneous speech and other noises. As can be seen below the presence or absence of these other noises (OS means "Other Sounds") does not have a great impact upon the accuracy (Acc).

Length	Acc. with OS	Acc. without OS
75%	88.6%	86.4%
80%	88.1%	88.8%
85%	89.5%	89.6%
90%	86.1%	85.2%
95%	84.4%	87.6%
100%	86.4%	85.2%

Table 1: Results according to relative sample length

We found that pragmatically, 3 seconds is a bit too long: In many samples people had not been able to sustain laughter for a full 3 seconds, giving a tail of the sound file as noise. This may explain why our best results were at 2.5 seconds.

[2] The classification algorithms we experimented with are these of the Weka platform: www.cs.waikato.ac.nz/ml/weka/

6.2 Multi-class values experiments

In the first experiment we were trying to distinguish between laughter and non-laughter, but were not interested in the difference between the different kinds of non-laughter sounds. In two new experiments, we tested the ability of our system to differentiate between the three non-laughter types. In the first experiment we ran our classifier on a database where the samples were labeled according to 3 possible values: Laughter, Reading and Speech. We call this the ternary experiment. The "Other Sounds" samples were excluded. In the second experiment all samples were included and so the class had four possible values, laughter, Reading, Speech and Other. We call this the quaternary experiment. For comparison purposes, all multi-class-valued results were transformed into their binary equivalent according to equation 10 [2] where N is the number of possible class values, acc_N the accuracy obtained on the N class values problem and acc_2 the equivalent binary accuracy.

$$acc_2 = acc_N^{\frac{log(2)}{log(N)}} \tag{10}$$

.

In fig.4, the X axis is the lengh of the samples (as a percentage of the full 3 seconds length) and the Y axis is the classification accuracy, using 10-folds cross-validation on the given dataset. The lines are colour-coded. Dark blue is the first (binary) experiment. Light blue is the ternary experiment. Dotted green is the quaternary experiment. The X axis is labeled in percentage of total file length. When we speak about for example 80% of the file length we always mean the first 80%.

The results show that our system, which was designed specifically for laughter detection, performs poorly on these other tasks.

Fig. 4. Comparisons: dark blue is binary, light blue is ternary and dotted green is quaternary experiment

7 Conclusions

Among all possible non-verbal sounds, laughing and crying are those which carry the strongest emotional-state related information. Their utterance predates language skills acquisition in newborn babies. In the framework of inter-adult communication, laughter could be the non-verbal sound which is the most meaningfull while still being relatively common. In the machine learing community the C4.5 algorithm is well known as being a robust and multi-purpose. Using

this algorithm we have designed a system specifically for the purpose of recognising laughter; our preprocessing is appropriate for laughter detection, and our results for single-speaker / single-microphone with background noise is better than reported state of the art. However, we have also found that it is not very good for other discrimination, an example being detecting the difference between reading aloud and spontaneous speech. However, as laughter is important we are currently working on optimising our algorithm for real-time uses in speech recognition systems, in order to improve its quality in automatic dialogue systems.

In this study, noisy settings means real uncontrolled ambiant noise. In order to prove the robustness of the the proposed parameters and the associated decision making tool (C4.5) an evalution of the proposed processing chain in controlled noisy situations has to be performed. Moreover, we hypothesized a constant noisy environment (stationnary noise) justifying a simple energy threshold method for differenciating silent from non silent segments. We chose to differenciate silent to non silent segments according to a certain percentage of the original signal energy, here 80%. This choice was arbitrary. This should be validated in a further study and compared to more sophisticated selection algorithms. In real situations, we could acknowledge that this hypothesis might be not valid and more complicated selection methodologies have to be evaluated. This will be done in further work.

References

1. Nikiforov I Basseville M. *Detection of Abrupt Changes, Theory and Application*. Prentice-Hall, Englewood Cliffs, NJ, 1993.
2. Mary Felkin. Comparing classification results between n-ary and binary problems. *Quality Measures in Data Mining, book edited by F. Guillet and H. J. Hamilton*, 2007.
3. H Hermansky. Perceptual linear predictive (plp) analysis of speech. *J. Acoust. Soc. Am.*, 87:1738–1752, 1990.
4. Nikiforov IV. A generalized change detection problem. *IEEE Trans. Inform. Theory*, 41:171–171, 1995.
5. Adam Janin and al. The icsi meeting corpus. *Acoustics, Speech, and Signal Processing*, 1:364– I–367, 2003.
6. Eero Väyrynen Juhani Toivanen and Tapio Seppánen. Automatic discrimination of emotion in spoken finnish: Research utilizing the media team speech corpus. *Language and Speech*, 47:383–412, 2004.
7. Lyndon S. Kennedy and Daniel P.W. Ellis. Laughter detection in meetings. *Proc. NIST Meeting Recognition Workshop*, 2004.
8. Duchene J Khalil M. Detection and classification of multiple events in piecewise stationary signals: Comparison between autoregressive and multiscale approaches. *Signal Processing*, 75:239–251, 1999.
9. Mary Knox. Automatic laughter detection. *Final Project (EECS 294)*, 2006.
10. Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI)*, 2:11371143, 1995.
11. Kornel Laskowski and Susanne Burger. Analysis of the occurrence of laughter in meetings. *Proc. INTERSPEECH*, 2007.

12. Kornel Laskowski and Tanja Schultz. Detection of laughter in interaction in multichannel close talk microphone recordings of meetings. *Lecture Notes in Computer Science*, 5237, 2008.
13. Hideki Kashioka Nick Campbell and Ryo Ohara. No laughing matter. *Interspeech'2005 - Eurospeech*, 2005.
14. Stavros Petridis and Maja Pantic. Audiovisual discrimination between laughter and speech. *Acoustics, Speech and Signal Processing (ICASSP 2008)*, pages 5117–5120, 2008.
15. Stavros Petridis and Maja Pantic. Audiovisual laughter detection based on temporal features. *International Conference on Multimodal Interfaces*, pages 37–44, 2008.
16. J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, 1993.
17. J. R. Quinlan. Improved use of continuous attributes in c4.5. *Journal of Artificial Intelligence Research*, 4:77–90, 1996.
18. Chad Kuyper ad Patrick Menning Rebecca Bates, Elisabeth Willingham. *Mapping Meetings Project: Group Interaction Labeling Guide*. Minnesota State University, 2005.
19. Boris Reuderink. Fusion for audio-visual laughter detection. Technical Report TR-CTIT-07-84, Centre for Telematics and Information Technology, University of Twente, Enschede, 2007.
20. Boris Reuderink and al. Decision-level fusion for audio-visual laughter detection. *Lecture Notes in Computer Science*, 5237, 2008.
21. Cees G.M. Snoek and Marcel Worring. Multimodal video indexing: A review of the state-of-the-art. *Multimedia Tools and Applications*, 2005.
22. Andrey Temko and Climent Nadeu. Classification of meeting-room accoustic events with support vector machines and variable feature set clustering. *Acoustics, Speech, and Signal Processing*, 5:505–508, 2005.
23. Jonathan T.Foote. Content-based retrieval of music and audio. *Multimedia Storage and Archiving Systems II, Proc. of SPIE.*, 1997.
24. Khiet P. Truong and David A. van Leeuwen. Automatic detection of laughter. *Interspeech'2005 - Eurospeech*, 2005.
25. Khiet P. Truong and David A. van Leeuwen. Automatic discrimination between laughter and speech. *Speech Communication*, 49:144–158, 2007.