Laughter Detection in Noisy Settings

Mary Felkin, Jérémy Terrien and Kristinn R. Thórisson

CADIA

Abstract. 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 some of those working on analysis of audio data before us, the goal of our work is to be able to classify many types of non-speech vocal sounds. The approach we use in this work relies upon machine learning techiques, as it could take centuries to hand-code algorithms for detecting laughter and other sounds, which have high variability both between cultures and between individuals. Here we describe our application of C4.5 to find the onset and offset of laughter using single-speaker audio recordings. Prior efforts using machine learning have not, to our knowledge, used C4.5. We got the best results so far on noisy[1] data as compared to the literature.

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

1 Introduction

Unlike much of the prior work on laughter detection our ultimate aim is not simply the detection of laughter but the use of this information – by a robot or virtual humanoid – to produce the appropriate conversational responses in real-time dialogue with people. Such a system could also be used 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. 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 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.

[1] By "noisy" we mean the sound tracks used were raw recordings, most of which included background noises such as people talking further away from the microphone or objects being moved. This noise was however never as loud as the primary (intended) recording.

2 Related Work

A number of papers have been published on the application of learning for detecting the difference between speech and laughter in audio recordings [9] [15] [5] [14] [7]. 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 [15], [5], Hidden Markov Models [9] [10], artificial neural nets [7] [15] and Gaussian Mixture Models [15], [14]. 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. Some use clean data while others use real-life recordings (as we did). The results are therefore not easily comparable. Among the methods used for pre-processing are mel-frequency cepstral coefficients [13] and Perceptual Linear Prediction features [3]. The wide spectrum of laugh-related studies, of which these are but a small sample, encompass, without being restricted to, pychology, cognitive science and philosophy as well as acoustics, giving to our topic an important place in any field related to intelligence and to communication.

3 Data Collection

Sound samples were collected through a user-friendly interface; subjects were volunteers from Reykjavik University's staff and student pool. Recordings were done in a relatively noisy environment (people talking and moving in the background, and often people hanging around while the recording was achieved). We used no noise cancellation mechanisms. A human listener could however clearly distinguish between the background noise and the primary recording, the latter being louder. We use energy-based descriptors so our method could not function if the background noise was as loud as the primary recording.

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

- 5 samples of him/herself laughing
- 5 samples of him/herself speaking spontaneously
- 5 samples of him/herself reading aloud
- 5 samples of him/herself making other sounds (OS)

The other noises recorded included humming, coughing, singing, animal sound imitations, etc. One volunteer thought that rythmic hand clapping and drumming could also be confused with laughter so he was allowed to produce such non-vocal sounds.

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 University cosmopolitan environment allowed us to record speech and reading in several different languages, the volunteers were encouraged to record themselves speaking and reading in their native languages.

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. In a first approach, we chose a segmentation algorithm based on auto-regressive (AR) modeling, the CUMSUM (CUMulated SUMs) algorithm [6]. The pupose is classification according to the genre of the movie (science fiction, western, drama, etc.).

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 the CUMSUM algorithm [1]:

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 [6]. 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 [6] (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 pre-processing step, we detected hypothesis corresponding to silence (only background noise is heard) by energy thresholding of all hypothesis.

5 Attribute Construction for Chunks

To separate audio segments from silence segments 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*.

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. Their use can be seen in the decision tree.

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:

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 2. but not normalised according to the total energy of the burst.

To differentiate records corresponding to a laugh or a non-laugh utterance, we characterised each burst series by the regularity of each parameter. This approach allowed 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 evaluated the median of the absolute instantaneous difference of the parameters. For the parameter T_i, we evaluated the standard deviation of the instantaneous emission period, i.e. $T_{i+1} - T_i$.

5.2 Machine Learning Tools

No single descriptor on it’s own is sufficient to differentiate laughter vs. non laughter samples. This indicates that there is no trivial method to differenciate laughter from non-laughter samples, using our descriptors, and supervised classification techniques are required. We solved this problem with the decision tree inducer C4.5 [11] [12].

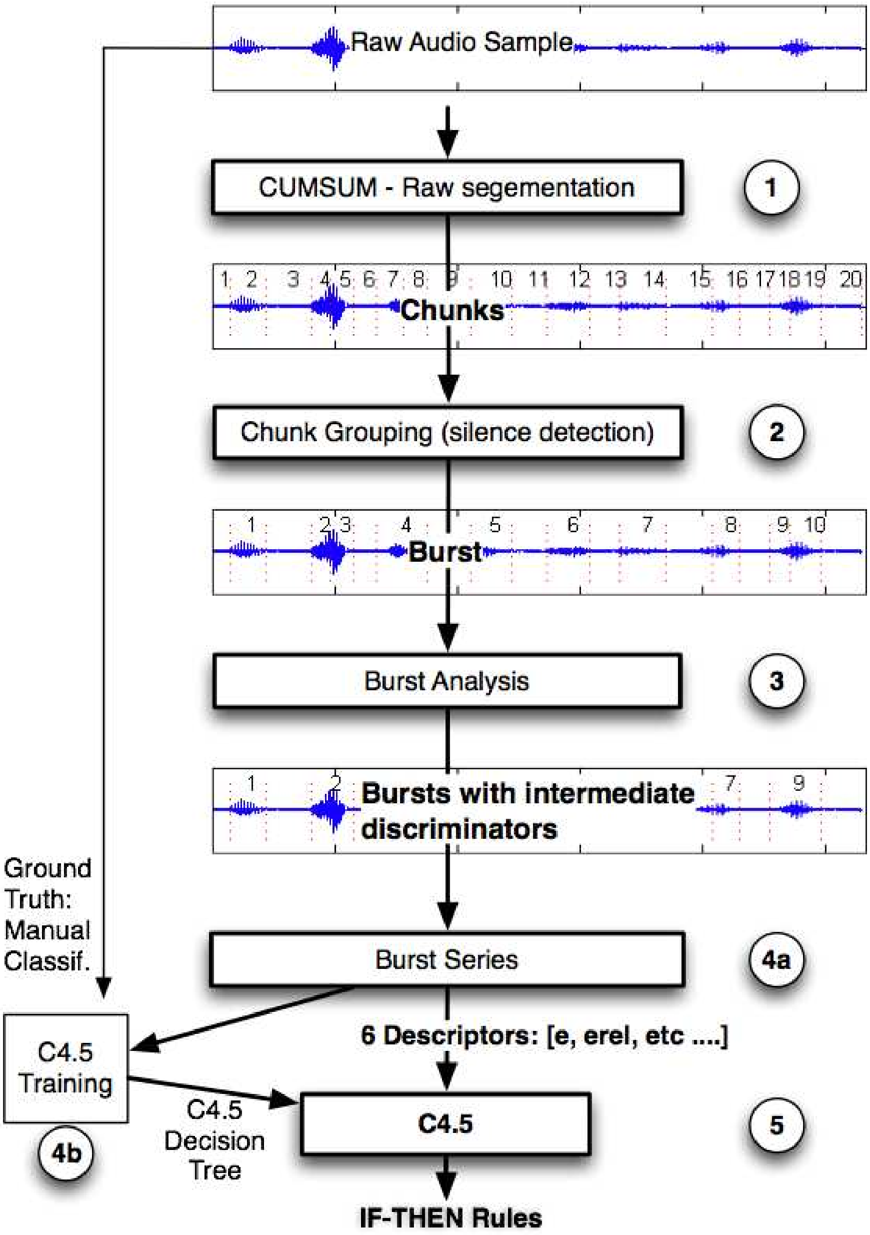

Fig.1: The complete algorithm

6 Results

In the following we use $10 - folds$ cross validation[2]. 3 seconds is too long: In many samples, people had not been able to laugh during 3 seconds, so the tail of the sound file is noise.

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

6.2 Multi-class values experiments

In two further 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 one all samples were included and so the class had four possible values, laughter, Reading, Speech and Others. We call this the quaternary experiment. It should be noted that during the first experiment we were only trying to distinguish between laughter and non-laughter and not between the different kinds of non-laughter. 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.

[2] 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 [8]) estimator of what our accuracy would be upon unknown examples.

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

.

In fig.2, 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 thus:

Dark blue is the first (binary) experiment
Light blue is the ternary experiment
Dotted green is the quaternary experiment

It shows our system, designed specifically for laughter detection, performs poorly on other tasks. In particular, our system is not meant to differentiate between Reading and Speech.

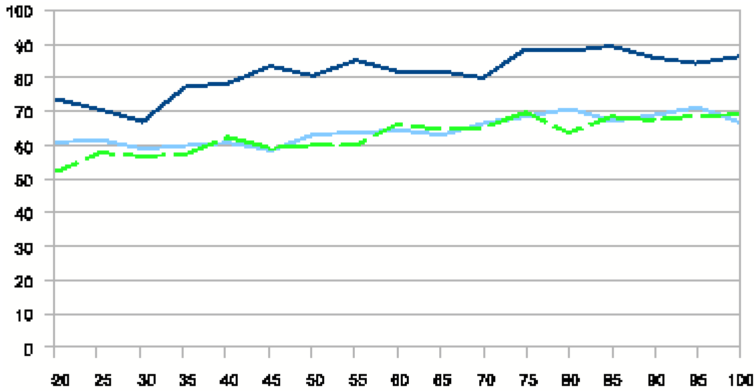

Fig. 2. Comparisons: dark blue is binary, light blue is ternary and dotted green is quaternary experiment

7 Conclusions

Laughter is important. Among all possible non-verbal sounds, laughing and crying are these which carry the strongest emotional-state related information. Their utterance predates language skills acquisition by newborn babies. Laughter is typically human, with the possible inclusion of some other primates. In the framework of inter-adult communication, laughter could be the non-verbal sound which is the most meaningfull while still being relatively common. C4.5 is well known as being a robust multi-purpose algorithm. What has been designed specifically for the purpose of recognising laughter are our preprocessing formulas and we have shown that our preprocessing is appropriate for laughter detection, but useless for other tasks such as distinguishing between reading aloud and spontaneous speech. We have shown that we do better than the state of the art on audio data, and we are now working on optimising our algorithm for real-time uses. One approach is to experiment with fewer descriptors to reduce computation cost.

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. Lyndon S. Kennedy and Daniel P.W. Ellis. Laughter detection in meetings. *Proc. NIST Meeting Recognition Workshop*, 2004.
6. 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.
7. Mary Knox. Automatic laughter detection. *Final Project (EECS 294)*, 2006.
8. 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.
9. 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.
10. Hideki Kashioka Nick Campbell and Ryo Ohara. No laughing matter. *Interspeech'2005 - Eurospeech*, 2005.
11. J. R. Quinlan. *C4.5: Programs for Machine Learning.* Morgan Kaufmann Publishers, 1993.
12. J. R. Quinlan. Improved use of continuous attributes in c4.5. *Journal of Artificial Intelligence Research*, 4:77–90, 1996.
13. Jonathan T.Foote. Content-based retrieval of music and audio. *Multimedia Storage and Archiving Systems II, Proc. of SPIE.*, 1997.
14. Khiet P. Truong and David A. van Leeuwen. Automatic detection of laughter. *Interspeech'2005 - Eurospeech*, 2005.
15. Khiet P. Truong and David A. van Leeuwen. Automatic discrimination between laughter and speech. *Speech Communication*, 49:144–158, 2007.