Detecting Laughter: Towards Building a Humorous Machine

Narsimh Kamath
National Institute of Technology Karnataka, India
narsimh@gmail.com

Abstract

Laughter is a common social gesture which often indicates the presence of humor. Detecting laughter and then understanding humor can make machines interact with us in a more natural way. This paper presents an algorithm to automatically detect laughter segments in speech. The voiced laughter of the speaker is recognized and the approximate onsets of the laughter bouts are used to annotate stored conversations. A simple algorithm based on the acoustic properties of voiced laughter is proposed and implemented for the same. The algorithm is able to detect the segments of laughter bouts in data consisting of sentences obtained from the switchboard corpus with an accuracy rate of 77.41% and a false detection rate of 12.90%.

Introduction

This paper addresses the problem of detecting spontaneous laughter in speech. Laughter is an interesting cue occurring commonly in everyday speech, and is most often an indicator of happiness. Laughter can be a stimulus to a joke being told by a friend, or to a funny incident. As such we try to detect laughter and annotate stored conversations with the intention of obtaining cues to possibly humorous sections of the conversation. Detection of laughter will assist in building a more humane, and perhaps more affable human-machine interaction system. Indeed, it will become possible to have a more natural conversation with an interactive machine that detects laughter and understands humor. Detecting laughter can also enable a machine to recognize possible humor in a conversation and thereby learn to have a sense of humor. Laughter detection might also help lower the error rates of speech to text conversion by increasing the robustness of non-speech detection.

Some previous work has attempted to study the acoustic properties of laughter (Bickley and Hunnicutt 1992) and to create techniques for automatically detecting laughter in clean speech (Carter 2000). Prof Byrant in his website outlines attempts to measure the social and aesthetic relevance of laughter.

A study of the acoustic properties of laughter (Bickley and Hunnicutt 1992) reveals that speech and laughter are often similar in the fundamental frequency ranges, formant frequencies, and the presence of voiced syllables or vowels. The authors find that the difference is an increased unvoiced region, as compared to clean speech. A study of laughter from a more sociological point of view (Provine 1996) reveals that laughter is characterized by a series of short vowel-like notes (syllables), each about 75 milliseconds long, that are repeated at regular intervals about 210 milliseconds apart, also known as a 'laughter bout'. Thus, a laughter bout of 1 second might have around 4 syllables. The author also finds that a specific vowel sound does not define laughter, but similar vowel sounds are typically used for the notes of a given laughter bout. He also reports that laughter in response to humor is most likely to be voiced.

One attempt (Carter 2000) to differentiate between laughter and non-laughter involves cross correlating the syllables in the time domain using a set of heuristics. The results mentioned appear encouraging, although it is pointed out that it does not differentiate between laughter, and any other periodic sound such as a series of dog barks, or loud footsteps. In one case (Kennedy and Ellis 2004) a laughter detection system for meetings, using a Support Vector Machine classifier with MFCC feature vectors is described. Some other works also attempt to detect a very generic set of features in speech, including laughter, with conventional Hidden Markov Models using MFCCs (Zhang and Perez 2005). However, very little work has been done on a specialized laughter detection scheme.

In this paper, we use the fact that voiced laughter is harmonic, due to the presence of vowels, and conduct experiments to find that although the pitch, or fundamental frequency ranges are similar for both speech and laughter, the variation of pitch in laughter is more than that in speech and that this variation occurs in a much shorter duration of time. We first detect the onsets of all vowels in speech. Then we find out the variance of pitch estimates within each detected vowel. Finally we use a high pass filter to extract rapid changes in the variance of pitch values between adjacent vowels to detect the onsets of laughter bouts.

Experiments

In this section we present the experiments conducted to identify effective features for detecting voiced laughter.

For the initial experiments, sentences from two male and two female speakers were recorded at a sampling rate of 10000 Hz. Each sentence was then appended by a voiced laughter segment of the speaker, which was also recorded at a sampling rate of 10000 Hz. The four sentences were of approximate duration 5 to 6 seconds, out of which voiced laughter constituted nearly 3 seconds. The sentences were then analyzed with respect to harmonicity, and pitch.

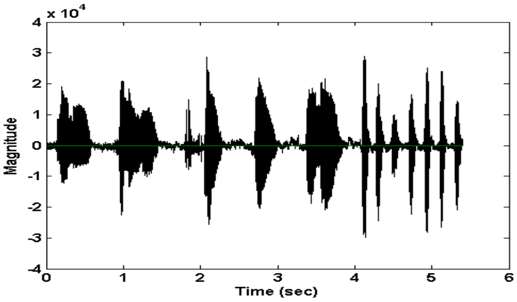

Figure 1: Waveform of a spoken sentence with a laughter bout appended at the end of the sentence (after approximate time 4 seconds).

It was observed that voiced laughter was harmonic with the syllables in each laughter bout showing a formant structure similar to that in speech.

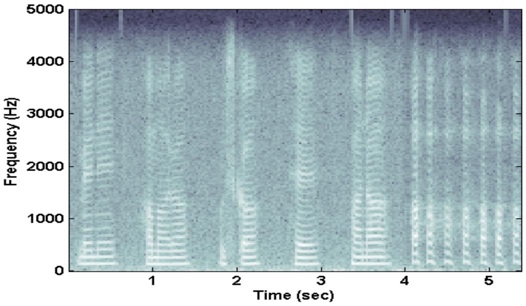

Figure 2: Spectrogram of the spoken sentence shown in fig. 1.The harmonic nature of laughter, within each small voiced region of laughter can be observed.

The time domain waveform showed that each laughter bout consists of several small duration syllables. An important observation was made in the pitch contour of the sentence. The pitch contour remained fairly stable within each vowel in the speech region. However, there was observed a lot of variation in the pitch contour even within an individual vowel in the laughter region. Moreover this variation occurred within a very short interval of time. This revealed a distinguishing factor between speech and voiced laughter.

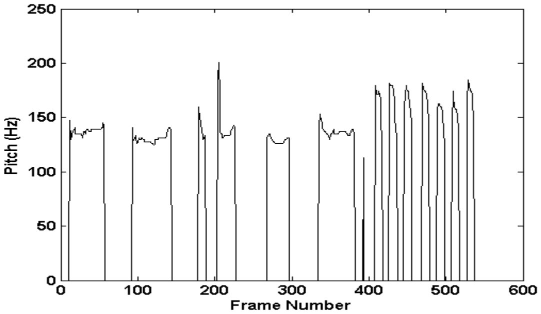

Figure 3: The pitch estimates for the voiced regions of the spoken sentence shown in fig 1.The fast variations of pitch in the laughter region can be observed. There is a more gradual variation in most parts of the pitch contour in the speech region. The frame length used for the pitch detection was 20ms and the overlap fraction between two frames was 0.5.

While the speaker was observed to be maintaining a more or less steady pitch during each vowel in voiced speech, the pitch in the voiced laughter segments was observed to be varying drastically, or at least the variation was found to be more than that in the speech region.

Algorithm

In this section we describe the algorithm used for the detection of voiced laughter in speech. First we briefly describe the features used.

Pitch

Pitch can be broadly defined to be the fundamental frequency of speech waveform. We estimate pitch values at every 10 ms. This temporal resolution ensures the estimation of even rapidly varying fundamental frequencies. For the actual task of pitch estimation, we use the subharmonic summation (SHS) method (Hermes 1988) which was deemed to be suitable as it estimates pitch values based on the harmonicity of speech, and we wanted pitch estimates predominantly in the voiced harmonic regions. The values set for some of the parameters are as follows: Number of points per octave: 48.Number of harmonics: 10.Compression strength 0.85.

Harmonicity

Voiced laughter is harmonic in nature. As such we attempt to identify such harmonic natures by estimating vowel onsets using a vowel onset detection method (Hermes 1990). The intuition behind detecting vowel onsets is that a voiced laughter bout, due to its harmonic structure will also be detected, along with all other vowels in speech. We can then use the pitch variation parameter to detect voiced laughter.

Steps Involved in the algorithm

Step1: Divide the speech waveform into overlapping frames, each of duration 20 ms, with an overlapping fraction of 0.5

Step2: Estimate pitch values for each of the frames obtained in step 1, using the SHS method.

Step3: Detect the time instants when a single vowel starts, and ends.

Step4: Find the mean ($\overline{p}$) of the estimated pitch values over all the frames contained within each vowel.

Step 5: Find the variance of the pitch estimates within each detected vowel. We call this quantity the pitch variance.

$$\text{PitchVariance}_i = \frac{\sum_j (p_{i,j} - \overline{p_i})^\wedge 2}{n_i} \qquad (1)$$

Where i=vowel number; j=frame number and n_i is the length in number of frames belonging to the i_{th} vowel.

Table 1: Average values of pitch and pitch variance (1) for the four sentences (two male and two female) used in the experiments, each consisting of a spoken sentence with a voiced laughter bout appended at the end of the sentence.

Speech/ Laughter	Average Pitch (Hz)	Average Pitch Variance (Hz^2)
Voiced Speech	203.41	44.12
Voiced Laughter	221.58	183.62

Step 6: Apply a high pass filter to the signal consisting of the pitch variance values of the vowels. This is done to emphasize the rapid variations in the pitch variance from one vowel to another.

$$H_{hp}(z) = \frac{1}{(1 + 0.9z^{-1})} \qquad (2)$$

This high pass filter was selected for the ease of implementation, and due to the fact that we merely wanted to identify the approximate instants of rapid variation in the pitch variance parameter.

Step 7: Identify local maxima in the filtered signal obtained in Step 6, and store the corresponding frame numbers. Find the minimum value amongst all the local maxima.

Step 8: Identify the local maxima having value at least h times the minimum value. Here h is a threshold value and we set it at 12 based on best performance. The frames at which these local maxima occur correspond to frames of voiced laughter. If two such frames occur within an interval of 1 second then keep only the frame with the higher value. Finally annotate the stored conversation at the time instants so identified.

The results of some of the steps of the algorithm for the spoken sentence shown in fig. 1 are shown in fig. 4 and fig. 5. This sentence, spoken by a male speaker has lower than average pitch variance values in both voiced speech and voiced laughter regions. The vowel onset detection algorithm detected 15 vowels, and the last 7 of these correspond to the vowels present in the voiced laughter region.

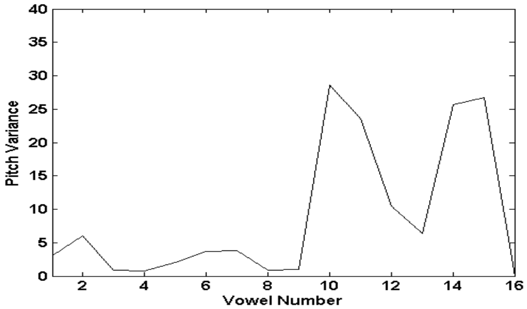

Figure 4: The pitch variance values (1) for each of the 15 vowels detected in the spoken sentence shown in fig. 1.The sudden jump in values for some of the last 7 vowels, corresponding to the laughter region can be observed.

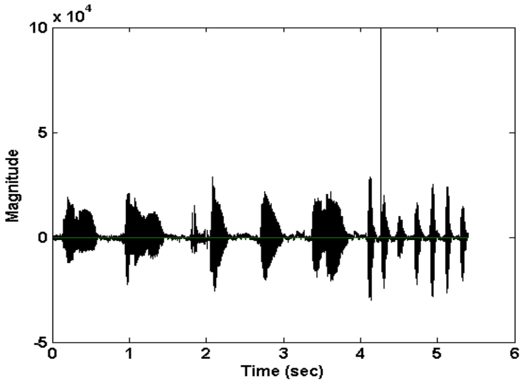

Figure 5: The annotated waveform corresponding to detected voiced laughter segment for the spoken sentence shown in fig. 1.The thin vertical line indicating the voiced laughter region is not at the onset of the laughter bout but is a sufficient indicator of the presence of a voiced laughter region.

Results

In this section we describe the data used for evaluating this algorithm, and state the results. Data in the form of sentences from the switchboard-1, Release-2 corpus was used for the purpose of evaluation. The corpus contains 2430 conversations averaging 6 minutes in length. We use a subset of the corpus consisting of 28 sentences having a total of 31 laughter bouts from conversations having laughter bouts of duration more than 1 second. The parts of a conversation transcribed laughter and having duration less than 1 second are often unvoiced laughter segments and hence we ignore them for the moment as humor is mostly accompanied by voiced laughter. For purposes of evaluation we consider an estimate of laughter detection by the algorithm to be correct if the estimated time instant of the laughter bout falls within the time interval of the concerned segment marked laughter in the transcriptions. We consider an estimate false otherwise. Some of the parameters chosen for evaluation were: frame length 20ms, overlapping fraction between frames 0.5, and threshold value for the pitch variance parameter 12.

Table 2: Summary of results for data obtained from the switchboard corpus.

Number of voiced laughter events	Accuracy Rate	False detection Rate
31	77.90 %	12.90 %

Discussion

Our results show that using a combination of harmonicity and pitch variation, it is possible to achieve voiced laughter detection. The algorithm we have proposed has a limitation that when parts of a sentence might be spoken with greater than usual emphasis, the pitch contours in the concerned utterances show greater variation, and this might lead to a false detection of voiced laughter. A post processing algorithm for this might be developed where the occasional discrepancies in the pitch contour within the vowels in the speech region, are taken care of. Also a more elaborate evaluation can be taken up on a database prepared with detection of laughter in mind.

It is interesting that although voiced laughter is similar to speech in that it too consists of vowels having identical formant structures, the fundamental frequency shows distinctive variation in very short periods of time. This feature of voiced laughter has been used to identify voiced laughter regions. The pitch variance parameter has been found to be an effective measure of the amount of pitch variation for detecting voiced laughter. Phonetically, laughter has been found to be of several types (Trouvain, 2003) and it might be possible to detect and identify each type of laughter.

Future work can involve a post processing algorithm for removing the false laughter detections, as well as use in human machine interaction systems which might be able to understand and respond to humor. For this it is also necessary to understand how we comprehend humor. It is important that we build machines that have the ability to understand human emotions, and the work we have done is an attempt in this direction.

Acknowledgements

I am grateful to the Digital Audio Processing Lab, IIT Bombay headed by Dr. Preeti Rao, for having introduced me to the exciting field of speech processing through my summer internship. The internship gave me the opportunity of working on pitch detection algorithms and vowel onset detection. I thank the reviewers for their suggestions and advise.

References

Bickley, C. and Hunnicutt, S. 1992. Acoustic analysis of laughter. *In Proceedings of the International Conference On Spoken Language Processing*, Banff, pp. 927–930.

Carter, A. 2000. Automatic acoustic laughter detection. Masters Thesis, Keele University.

Website: http://gabryant.bol.ucla.edu/research.html.

Provine, R.1996.Laughter. *American Scientist.* 84. 1 38-47.

Kennedy, L.S. and Ellis, D.P.W. March 2004. Laughter Detection In Meetings. *In Proceedings of the NIST. Meeting Recognition Workshop*, Montreal.

Zhang, D. and Perez, G. *2005.* Semi-supervised Adapted HMMs for Unusual Event Detection. *In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition.*

Hermes, D. 1988. Measurement of pitch by subharmonic Summation. *Journal of the Acoustical Society of America*. 83(1).

Hermes, D. 1990. Vowel-onset detection. *Journal of the Acoustical Society of America.* 87 (2).

Trouvain, J. (2003). Segmenting phonetic units in laughter. *In Proceedings of the 15th International Conference of the Phonetic Sciences* (ICPhS).Barcelona, (Spain), pp. 2793-2796.