INTERSPEECH 2009 BRIGHTON

Analysis of Laugh Signals for Detecting in Continuous Speech

Sudheer Kumar K.[1], *Sri Harish Reddy M.*[1] , *K. Sri Rama Murty*[2] *and B. Yegnanarayana* [1]

[1]International Institute of Information Technology, Hyderabad, India
[2]Indian Institute of Technology Madras, Chennai, India
{sudheerk,sriharsham}@research.iiit.ac.in, ksrmurty@gmail.com, yegna@iiit.ac.in

Abstract

Laughter is a nonverbal vocalization that occurs often in speech communication. Since laughter is produced by the speech production mechanism, spectral analysis methods are used mostly for the study of laughter acoustics. In this paper the significance of excitation features for discriminating laughter and speech is discussed. New features describing the excitation characteristics are used to analyze the laugh signals. The features are based on instantaneous pitch and strength of excitation at epochs. An algorithm is developed based on these features to detect laughter regions in continuous speech. The results are illustrated by detecting laughter regions in a TV broadcast program.
Index Terms: Laughter detection, epoch, strength of excitation

1. Introduction

The phenomenon of laughter is common in human communication as a way of expressing the emotion of happiness. It is produced by the speech production mechanism using a highly variable physiological process. The vocalized expression of laughter varies across gender, individuals and context. Despite its variability, laughter is perceived naturally by human listeners. In recent years much of the research done in the area of speech recognition has been mainly concentrated on natural data. This requires data collected in natural environment which contain many non-speech elements like laughter and other non-linguistic sounds. Automatic detection of such elements helps in increasing the accuracy of recognition. It also helps us to know the emotional state of the speaker which makes us easy to converse with them.

Laughs were analyzed at three levels: bout, call and segment levels [1]. The entire laugh is referred to as an episode which consists of laughter bouts that are produced during one exhalation. Calls are the discrete acoustic events that constitute a bout, and each call of a voiced laughter consists of a voiced part followed by an unvoiced/silence part. Segments are the audibly reflected changes in the production within a call. It is assumed that each laughter bout contains several calls, so that isolated calls are not considered as laughter.

Since laughter is produced by the human speech production mechanism, the laughter signal is also analyzed like a speech signal in terms of the acoustic features of the speech production. Analysis of laughter could be done for synthesis, where perceptually important characteristics need to be preserved, or for studying the acoustic features during its production. Based on analysis of large database of laughter sounds, Bachorowski and colleagues have differentiated three broad categories, namely, song-like (consisting primarily of voiced sounds), snort-like (consisting largely unvoiced calls with perceptually salient nasal-cavity turbulence) and grunt-like (with turbulence from laryngeal or oral cavities) [2]. Typically, the acoustic analysis of laughter is carried out using duration (between onset and offset of acoustic events), F_0 (fundamental frequency of voiced excitation) and spectral features. All of these are used to describe the temporal variability, source variability and variability in production modes [2]. Formants, pitch and voice quality analysis are used to discriminate speech, speech-laughs and laugh [3].

The variability in laughter production is complex in the sense that it is not guided by the production rules of speech. Hence it is difficult to describe the phenomenon of laughter precisely, although it can be perceived by the listeners. The analysis and description is also limited by the available tools for analysis of laughter signals. The objective of this study is to show that some important features of laughter acoustics can be highlighted using some new tools for analysis proposed in this paper. It is likely that these new features may help to spot the laughter regions in continuous speech communication. Some of these features are: (a) Rapid changes in the instantaneous fundamental frequency (F_0) within calls of a laughter bout; (b) Strength of excitation within each glottal cycle and its relation to F_0; (c) Loudness of speech derived from the excitation information; (d) Temporal variability of F_0, strength and formants across calls within a bout. Some of these features were studied using conventional methods of analysis for F_0 and voiced quality, but using mostly spectrum-based features, like harmonics, spectral tilt and formants [3]. The difficulty in deriving the features of excitation source using short-time spectral analysis limits the analysis significantly, especially due to the choice of the size, shape and position of the segment in relation to the acoustic events in speech production. The main problem is to extract the rapidly varying instantaneous F_0. Moreover, traditional short-time spectrum analysis masks several important subsegmental (less than pitch period) features of the glottal source that are unique in the production of laughter. There are also many studies done on automatic spotting of laughter [4] [5] [6]. Kennedy and Ellis [4] tried to spot laughter by using some spectral features like MFCC's, delta MFCC's, energy of the high frequency components etc. Mary Knox [6] used similar features and also some more like phone information (using phone recognizer), shimmer, jitter etc. The analysis in this paper is focused mainly on using source and excitation information of laughter.

In Section 2 we present new method of analysis of excitation source features, especially the instantaneous F_0 and the strength of excitation at the epochs. In Section 3 laugh signals are analyzed to extract the source features to describe the laughter acoustics. Spotting these unique regions in continuous speech is proposed in Section 4 and its performance is examined for some TV broadcast data in Section 5. Finally Section 6 summarizes the results presented in this paper, and discusses issues to be explored further in the study of laughter acoustics.

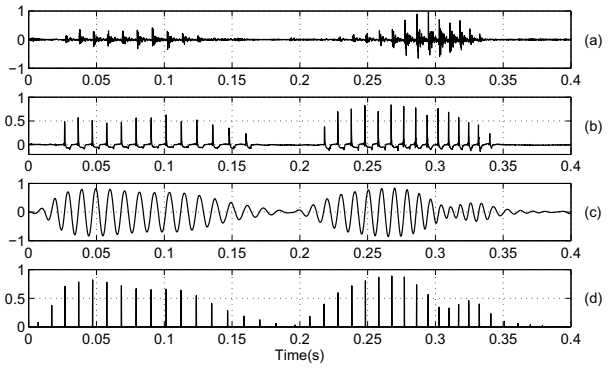

Figure 1: Illustration of epoch extraction and their strengths from the zero-frequency filtered signal. (a) A segment of speech signal. (b) DEGG signal. (c) Filtered signal. (d) Strength of excitation (α).

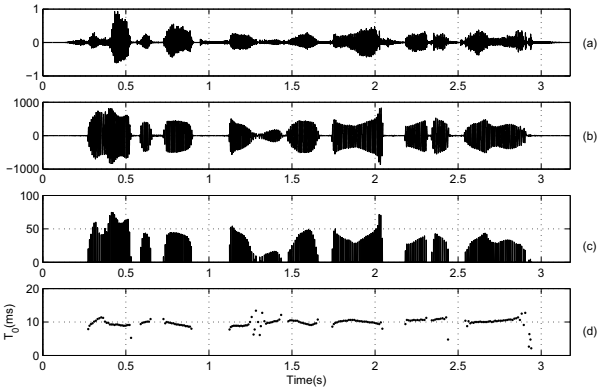

Figure 2: (a) A segment of speech signal. (b) Filtered signal obtained with adaptive window length for trend removal. (c) Strength of excitation (α). (d) Pitch period.

2. Method to extract instantaneous frequency and epoch strength

Recently a new method is proposed for extraction of the instantaneous F_0 [7], for epoch extraction [7] and for strength of excitation at epochs [8]. The method uses the zero-frequency filtered signal derived from speech to obtain the epochs (instants of significant excitation of the vocal tract system) and the strength at the epochs. The following steps are involved in processing the speech signal to derive the epochs and their strengths from the filtered signals [8]. (a) Difference the speech signal $s[n]$ to remove any very low frequency component introduced by the recording device.

$$x[n] = s[n] - s[n-1]. \tag{1}$$

(b) Pass the differenced speech signal $x[n]$ through a cascade of two ideal zero-frequency resonators. That is

$$y_0[n] = -\sum_{k=1}^{4} a_k y_0[n-k] + x[n], \tag{2}$$

where $a_1 = -4, a_2 = 6, a_3 = -4$ and $a_4 = 1$.

(c) Compute the average pitch period using the autocorrelation function for every 30 ms speech segments.

(d) Remove the trend in $y_o[n]$ by subtracting the local mean computed over a window obtained from (c) at each sample. The resulting signal $y[n]$ is the zero-frequency filtered signal, given by

$$y[n] = y_0[n] - \frac{1}{2N+1} \sum_{m=-N}^{N} y_0[n+m]. \tag{3}$$

Here $2N + 1$ corresponds to the number of samples in the window used for mean subtraction. The choice of the window size is not critical as long as it is in the range of one to two pitch periods.

(e) The instants of positive zero crossings of the filtered signal give the locations of the epochs.

(f) The strength of the epoch (denoted as α) is obtained by taking the slope of the filtered signal around the epoch. The slope is measured by taking the difference between the positive and negative sample values around each epoch. Fig. 1 illustrates extraction of epochs and their strengths from the filtered signal derived from the speech signal. The strength values are compared with the amplitudes of the peaks around the epochs in the differenced electro GlottoGraph (DEGG). While this method works well for the variations of F_0 in normal speech signals, it cannot capture the rapid changes of F_0 that occur in the calls of a laughter episode or cycle.

The critical factor in the above method is the choice of the window for trend removal from the output of the zero-frequency resonator. If the window size is too small compared to the average pitch period, then too many zero crossings occur in the filtered signal. If it is too large, then the short pitch periods corresponding to high F_0 may be missed. In order to capture the rapid variations in F_0 between speech and laughter the following procedure is adopted:

(a) Pass the signal through the zero-frequency resonator with window length of 3 ms for trend removal. This window length has been chosen in such a way that it gives high energy in the filtered signal in case of speech and laughter and low energy in the nonvoiced and non-speech regions.

(b) Positive zero crossings of the filtered signal gives the epoch locations, and the slope calculated as the difference of values of the samples after and before the epochs gives the strength of excitation. Mean of the strength of excitation over a window of 10 ms is calculated, and if this value is more that 30 percent of the maximum strength value of the complete signal then that segment is considered as a voiced segment, otherwise it is a nonvoiced segment.

(c) After finding the voiced segments, each voiced region is separately passed through a zero-frequency resonator with window length for trend removal derived from that segment. This is done by first computing the autocorrelation of the segment with a frame size of 20 ms and a frame shift of 10 ms. Then the maximum occurring peak in the autocorrelated signals is chosen as the window length for that region.

(d) The positive zero crossings of the final filtered signal give the epoch locations, and the difference in the values of the samples after and before each epoch gives the strength of excitation.

Fig. 2 illustrates the epochs and strength of excitation for a segment of speech signal using the modified epoch extraction method.

3. Analysis of laughter signals

The source characteristics of laughter signals are analyzed using features like pitch period (T_0), strength of excitation (α), and

some parameters derived from them which are explained below in detail.

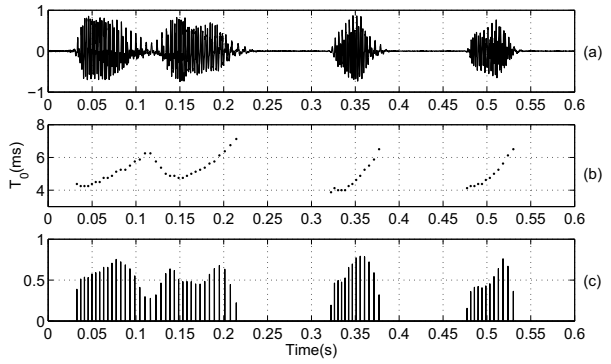

Figure 3: (a) A segment of laughter signal. (b) Pitch period derived from the epoch locations. (c) Strength of excitation (α) at the epochs.

Pitch period and its variation: It is observed that pitch frequency for laughter is more than that for normal speech [2]. For normal speech the pitch frequency typically ranges between 80 Hz and 200 Hz for male speakers and 200 Hz to 400 Hz for female speakers, whereas for laughter the mean pitch frequency for males is above 250Hz, and for females it is above 400 Hz [2]. As mentioned earlier, there will be more air flow through the vocal tract in the case of laughter. This will result in faster vibration of the vocal folds, and hence reduction in the pitch period. Apart from lower pitch period, there is also a raising pattern in the pitch period contour within a call. In some cases the pitch period may even start with some large value, decreases to some minimum and then increases again. This may be because this high pitch frequency (F_0) is not normal for the vibration of the vocal folds to maintain that frequency, and hence it tends to decrease. Fig. 3(b) shows this general trend of T_0 in the calls within a bout. The main issue here is extracting the pitch period accurately. It is not easy as in case of normal speech, since the pitch variation is large in laughter. The pitch period is extracted as explained in Section 2.

Strength of excitation: Since there is large amount of air pressure build up in the case of laughter, (as large amounts of air is exhaled), the closing phase of the vocal folds is very fast. This will result in an increase in the strength of excitation. Strength of excitation (α) at every epoch is computed as the difference between two successive samples of the filtered signal in the vicinity of the epoch. Fig. 3(c) shows this general trend of α in the calls within a bout.

Ratio of strength of excitation and pitch period: Since the closing phase of the vocal folds is fast for laughter, the corresponding opening phase will be larger in duration. So we have used the ratio (η) of the strength to excitation (α) at the epoch location and the pitch period (T_0) as an approximate measure of the opening phase.

$$\eta = \alpha/T_0 \tag{4}$$

Slope of pitch period contour: The pitch period contour of laughter has a unique pattern of rising rapidly at the end of a call. So, we use the slope of the pitch period contour to capture this pattern. First the pitch period contour is normalized between 0 and 1. At every epoch location the slope of the pitch period contour is obtained using a window width of 5 successive epochs. The slope is calculated by dividing the difference between the maximum and minimum of the 5 pitch period values within each window by the duration of the window. We denote this slope by β.

Slope of strength of epochs: As in the case of the pitch period, the strength of excitation at epochs also changes rapidly. Hence the slope of the normalized strengths is calculated by dividing the difference between maximum and minimum of the normalized strength values within 5 epochs window by the duration of the window. We denote this slope by γ.

4. Proposed method for voiced laughter identification

As mentioned earlier, the production of laughter and speech are different in many ways. As a result, source features like pitch period (T_0), strength of excitation (α) differ. Distributions of the features T_0, α, η, β, γ for laughter and speech samples of 5 male and 5 female speakers are shown in the Fig. 4. We can see from the distributions that there are certain regions where the laughter feature values are more concentrated, and there are regions where the speech feature values are more concentrated. This difference in distribution of features show that they could be used to discriminate between speech and laughter. A 'value threshold' is placed for each of the features separately. Since it is not possible to put a single threshold on all the epochs, a 'fraction threshold' is used to determine the percent of epochs that exceed the 'value threshold' for the segment to be laughter. This is done by observing these features for several laughter and speech samples. Table 1 gives the value and fraction thresholds for each of the features. The proposed method for laughter spotting consists of the following steps:

(a) The signal is first segmented into voiced and nonvoiced regions by passing the signal through the zero-frequency resonator using a window length of 3 ms for trend removal.

(b) For every voiced region, epochs are extracted using the zero-frequency filtering method with a window size for trend removal derived adaptively from the signal. (Explained in detail in Section 2.)

(c) The five features described in Section 3 are extracted for every epoch in the voiced region.

(d) If a voiced segment has more epochs than determined by the 'fraction threshold' for all the features, then that segment is considered as a laughter segment.

Table 1: Value and fraction thresholds for each feature.

	T_0	α	η	β	γ
Value threshold	5	0.1	0.002	0.003	0.005
Fraction threshold (%)	40	20	20	20	20

5. Results of laughter detection in continuous speech

The proposed algorithm for laughter detection was tested on a data collected from a TV program. Each episode is typically about 30 minutes of informal interview with a celebrity. It contains laughter in between naturally occurring speech. The data is not clean, as it is mixed with a low amplitude background music and noise.

The laughter segments were manually labeled with start time and end time by listening to the show. The manually labeled laughter segments has the time stamps of the complete

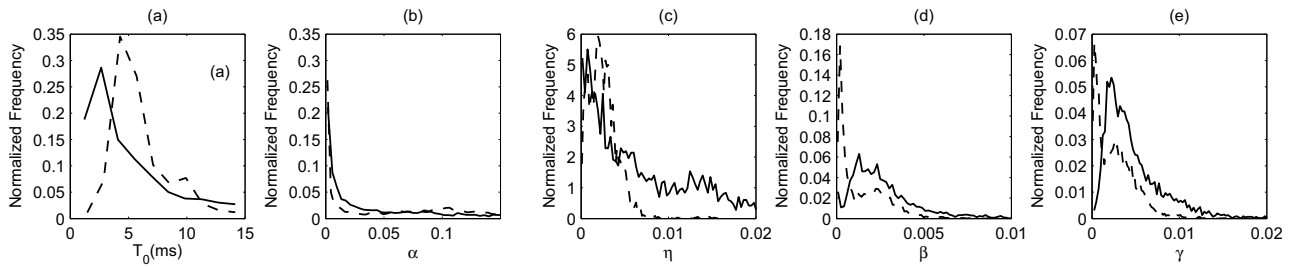

Figure 4: Illustration of differences in the excitation source features of Laughter and Normal speech using distributions of (a) T_0, (b) α, (c) η, (d) β and (e) γ . In each plot the distributions of corresponding features of Laughter and Normal speech were represented with solid and dashed lines respectively.

Table 2: Results of laughter spotting from 30 minutes of TV show.

	MDR (%)	FAR (%)
Segment level	10.2	29.1
Segment level after Post-processing	11.1	24.1
Bout level	4.1	32.1
Bout level after Post-processing	4.1	27.4

laughter bout, but not the individual calls. For obtaining the MDR (Missed Detection Rate) and FAR (False Alarm Rate) on voiced segments, these laughter regions are segmented into voiced and nonvoiced regions. This gives the start and end times of the voiced regions (calls) in a laughter. We assume that the laughter calls cannot occur in isolation, and hence the detected laughter segments that occurred in isolation are removed. Laughter segments of duration less than 0.5 sec and having atleast 3 seconds of non-laughter segments on either side are treated as isolated segments. This assumption reduced the FAR as can be seen from Table 2. Table 2 shows the MDR and FAR at segment level and at bout level detection of laughter before and after post-processing. Fig.5 shows the features and their corresponding decisions on a sample of the data.

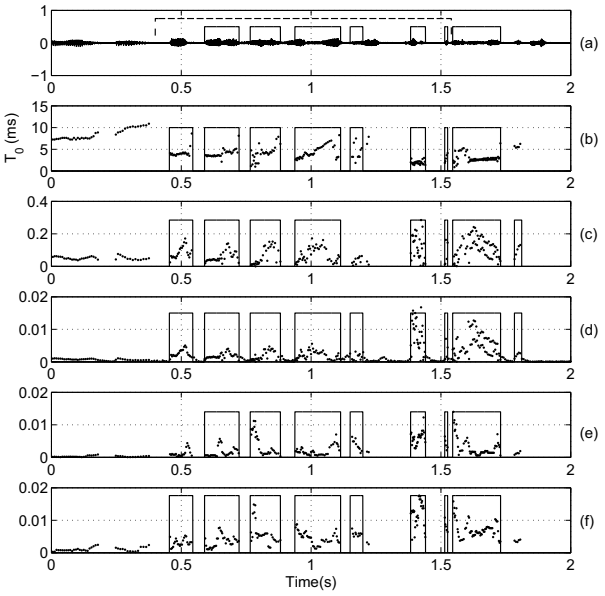

Figure 5: (a) Signal containing both laughter and speech, with solid lines representing the automatic detected laughter and dashed line representing the manually labeled laughter (b) Pitch period (T_0), (c) Strength of Excitation (α), (d) η, (e) β and (f) γ

6. Summary and conclusions

In this paper we have presented features based on excitation source for analyzing and characterizing the laugh signals. The excitation features are derived from the zero-frequency resonator output of speech signals in the form of instantaneous pitch and the strength of excitation at epochs. These features were used to develop an algorithm to discriminate laughter and speech regions in continuous speech. The performance of the algorithm was studied on a TV broadcast data.

The main signal processing issue is in analyzing laugh signals to capture the rapidly changing glottal activity, especially the rapidly changing pitch and strength of excitation. Since laughter in practice occurs along with other acoustic disturbances, extracting the rapidly changing glottal activity of laughter in practical environments is a challenging task. Currently we are exploring other source-related features, together with traditional spectral features to improve the performance of laughter detection in continuous speech.

7. References

[1] Trouvain, J, "Segmenting phonetic units in laughter," in *Proc. 15th ICPhS*, Barcelona, 2003, pp. 2793–2796.

[2] Bachorowski J., Smoski M and Owren M., "The acoustic features of human laughter," *Journal of the Acoustical Society of America*, vol. 111, pp. 1582–1597, 2001.

[3] Menezes Caroline and Yosuke Igarashi, "The speech laugh spectrum," in *Proceedings of the 6th International Seminar on Speech Production (ISSP)*, Dec. 13-15 2006, pp. 157–524.

[4] L. Kennedy and D. Ellis, "Laughter Detection in Meetings," in *Proc. ICASSP Meeting Recognition Workshop*, Montreal, Canada, 2004.

[5] Truong K.P. and Van Leeuwen D.A., "Automatic detection of laughter," in *Proc. of the INTERSPEECH 2005*, Lisbon, Portugal, 2005, pp. 485–488.

[6] M. Knox and N. Morgan and N. Mirghafori, "Getting the Last Laugh: Automatic Laughter Segmentation in Meetings," Brisbane, Australia, 2008, pp. 797–800.

[7] K. Sri Rama Murthy and B. Yegnanarayana., "Epoch Extraction from Speech Signals," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 16, no. 8, pp. 1602–1613, Nov. 2008.

[8] K. Sri Rama Murty and B. Yegnanarayana and Anand Joseph M., "Characterization of Glottal Activity from Speech Signals," *IEEE signal processing letters*, vol. 16, no. 6, June 2009.