ANALYSIS OF LAUGHTER AND SPEECH-LAUGH SIGNALS USING EXCITATION SOURCE INFORMATION

Sri Harsha Dumpala, Karthik Venkat Sridaran, Suryakanth V Gangashetty and B. Yegnanarayana

International Institute of Information Technology, Hyderabad - 500032, India

{sriharsha.dumpala, karthik.venkat}@research.iiit.ac.in, svg@iiit.ac.in, yegna@iiit.ac.in

ABSTRACT

Speech-laugh is a speech-synchronous form of laughter that often occurs in natural conversation. However, there are deviations in features of speech-laugh when compared with laughter and neutral speech individually. The objective of this study is to analyse the excitation source features to capture the deviations between laughter and speech-laughs in voiced regions. The features used in this analysis are based on instantaneous fundamental frequency and strength of excitation (β) at epochs. Modified zero frequency filtering (ZFF) method is used to extract the features. Kullback-Leibler (KL) distances obtained show that there are deviations in excitation source features which can be exploited to develop a method to discriminate speech-laughs from laughter. Experimental results show that features used are robust and speaker independent in discriminating speech-laughs from laughter. Results showing deviations of laughter and speech-laughs from neutral speech were also presented.

Index Terms— Speech-laugh, laughter, zero frequency filtering, fundamental frequency, strength of excitation.

1. INTRODUCTION

As the sophistication of automatic speech recognition (ASR) systems increases, there is more of a need to recognise speech co-occuring with different paralingusitic events. Speech-laugh is one such event where laughter co-occurs with speech.

In natural conversation, significant part of laughter co-occurs with speech which is referred to as speech-laugh. It is observed that more that 50 percent of laughs in conversation are speech-laughs [1]. They have characteristics of both laughter and speech but neither features of laughter nor speech dominates. Speech-laughs not only signifies the emotional state of a speaker but also carry the linguistic information. Inspite of the presence of laughter which is highly variable, both the lingusitic and emotional information are perceived naturally by human beings. Traditional automatic speech recognition (ASR) systems consider both laughter and speech-laugh as paralinguistic elements. This resulted in loss of information. Discriminating speech-laughs from laughter improves the accuracy of ASR systems. It also helps to know the emotion expresssed by laughter i.e., happy, sarcasm etc.

Analysis of laughter at three levels i.e., bout, call and segment levels was reported in [2]. As per further analysis based on the acoustic features, laughter sounds were differentiated into three broad categories namely, song-like, snort-like and grunt-like [3]. Many studies were carried on automatic detection of laughter [4, 5, 6]. Spectral features like MFCC's, delta MFCC's, energy of high frequency components were used to spot laughter [4]. Spotting of voiced laughter segments in continuous speech using features based on instantaneous fundamental frequency and strength of excitation at epochs is reported in [7].

In recent times, much emphasis was laid on detection of laughter but there are only few studies analysing the acoustic aspects of speech-laugh [1, 8, 9]. Interactions between mother and child were analysed and it was observed that speech-laugh is a simultaneous production of laughter and articulation. Speech-laugh is not just laughter superimposed on articulation but formed as a result of complex vocal production maintaining speech like fundamental frequency (F_0) and laugh like amplitude and rhythm [1]. Analysis of phonetic characteristics of speech-laugh showed that a reinforced expiratory activity is present in speech-laughs. This is noticed either as an increased harmonic noise during periodic portions or as stronger aspiration during unvoiced portions [8]. The position of laughter and speech-laugh in syntax and dialogue structure was analysed based on child-robot interaction [10] and observed that speech-laughs and laughter does not co-occur internally inside syntactic units.

Differences between laughter and speech-laugh were analysed using acoustic features like formant spectrum, fundamental frequency (F_0) range and voice quality [9]. It was observed that the glottal configuration of laughter is different from speech and speech-laugh due to high subglottal pressure. The objective of this study is to analyse laughter and speech-laugh in voiced regions using excitation source features. Variation in patterns of fundamental frequency (F_0) and strength of excitation (β) contours are exploited to discriminate speech-laugh from laughter.

The paper is organised as follows: Section 2 explains the data used for analysis. Section 3 explains the modified ZFF method and the features used. In Section 4, deviations between laughter and speech-laughs are analysed using the features extracted. In Section 5, experimental results are discussed. Finally, Section 6 provides summary along with conclusions.

.

2. DATABASE

The main challenge in analysis of laughter speech is data collection. Eliciting spontaneous laughter, especially speech-laugh is difficult. To collect data in natural scenario, data was recorded from a group of 13 male and 7 female speakers. Two speakers who knew each other from the group were asked to discuss on a funny topic which helped in eliciting both laughter and speech-laugh. After the conversation, each speaker was asked to repeat the speech-laugh utterances in his neutral speech. Data was collected in Telugu (Indian) language, at a sampling frequency of 48 KHz in a recording room (clean environment) using high quality zoom recorder. For analysis purpose, overlapped speech segments were discarded. The data was perceptually evaluated by 10 subjects. The subjects were asked to rate the speech-laugh utterances based on their perception between 1 and 5, where 5 refers to best and 1 refers to worst. The analysis

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

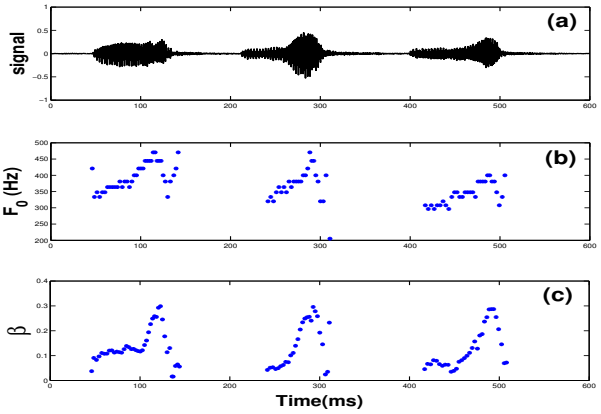

Fig. 1. *(a) A segment of laughter signal, (b) F_0 contour, (c) strength of excitation (β) contour.*

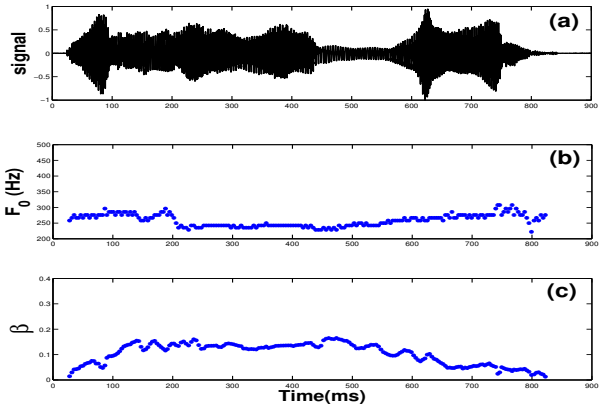

Fig. 2. *(a) A segment of speech-laugh signal, (b) F_0 contour, (c) strength of excitation (β) contour.*

was done using 90 speech-laugh utterances which were rated above 4. The analysis data also consists of 120 laughter segments.

3. METHOD USED FOR FEATURE EXTRACTION

The Subsegmental features related to excitation source information used are (a) Fundamental frequency (F_0) and (b) Strength of excitation (β). These features are extracted using modified ZFF method [7]. In ZFF method [11, 12], the critical factor is choice of window for trend removal. If the window size is too small compared to the average pitch period, then too many zero crossings occur in filtered signal. If the window is too large, then the short pitch periods corresponding to high F_0 may be missed. So, to capture the rapid variations of F_0 in laughter and speech-laugh, modified ZFF method is used. The signal is passed through zero frequency resonator with a window length of 3 msec for trend removal. Positive zero crossings of filtered signal gives epoch locations and slope of the signal calculated at each epoch location is β. Based on mean of β over a window of length 10 msec, the segments are divided into voiced and unvoiced regions. Each voiced segment is separately passed through a zero frequency resonator. The window length for trend removal in this case is equal to average pitch period of that voiced segments. The resultant signal is called ZFF signal. The negative to positive zero crossing instants in ZFF signal are called epochs. The distance between two successive epochs is called the pitch period (T_0). The reciprocal of interval between two successive epochs gives the fundamental frequency (F_0) at the epoch. The slope of ZFF signal at epochs correspond to strength of impulse like excitation (β) around epochs.

3.1. Fundamental Frequency (F_0)

It was observed that the fundamental frequency (F_0) values are higher for laughter compared to speech-laugh [9]. In case of laughter, there is more airflow through the vocal tract which results in faster vibration of vocal folds. Due to this effect, there is an increase in the F_0 values of laughter. In speech-laughs, the articulatory configurations for speaking are continuously maintained [8] and this voicing cuts the air flow. This results in decrease of F_0 compared to laughter. It is observed that within a call of laughter there is a rapid rise and fall in F_0 contour [7]. This pattern is not observed in speech-laugh. The F_0 contour for speech-laugh represents the pattern similar to neutral speech. It is also observed that F_0 values of speech laugh are higher than neutral speech. The difference in F_0 contour for laughter and speech-laugh are shown in Fig. 1(b) and Fig. 2(b) respectively.

3.2. Strength of Excitation (β)

The strength of excitation (β) contour follows a similar trend observed in F_0 contour as shown in Fig. 1(c) and Fig. 2(c) for laughter and speech-laugh respectively. The large amount of air pressure which is built during laughter is reduced because of the articulatory configuration present in the speech-laugh. It is observed that, due to increase in the closing phase of the vocal folds there is a decrease of β in speech-laugh compared to laughter as shown in Table 1.

Table 1. *Average values of F_0 and β for 'neutral', 'laughter' and 'speech-laugh' utterances for 13 male and 7 female speakers.*

	Laughter		Speech-laugh		Neutral	
	F_0	β	F_0	β	F_0	β
Male	334	85.6	215	61.8	156	53.69
Female	405	131.36	298	92.6	237	85.45

As mentioned in previous sections, there is a difference in patterns of F_0 and β for laughter and speech-laugh. The following features are derived from F_0 and β to capture these patterns.

3.3. Slope of F_0 Contour (α)

The slope of F_0 is used to capture the pattern of F_0 contour. α values are obtained at each epoch location by calculating the difference between maximum and minimum F_0 values in a window of 5 consecutive epochs. It is then divided by the duration of the window. As there is a sudden variation in F_0 contour of laughter, the α values are higher for laughter compared to speech-laugh as shown in the Fig. 3(b) and Fig. 4(b).

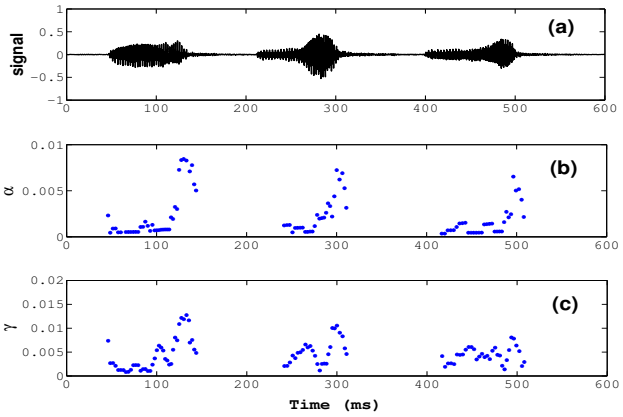

Fig. 3. *(a) A segment of laughter signal, (b) slope of F_0 (α) contour, (c) slope of strength of excitation (γ) contour.*

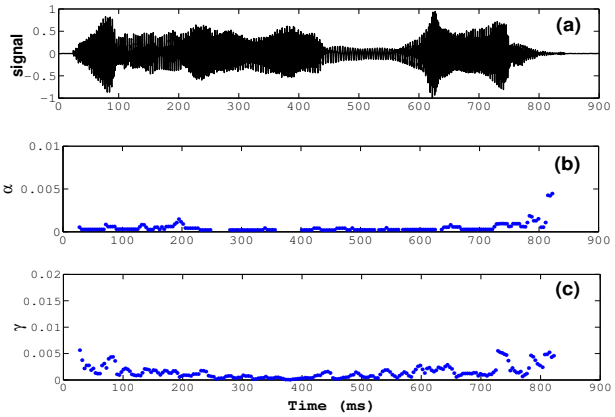

Fig. 4. *(a) A segment of speech-laugh signal, (b) slope of F_0 (α) contour, (c) slope of strength of excitation (γ) contour.*

Table 2. *Average values of α, γ, η for 'laughter' and 'speech-laugh' utterances of 13 male and 7 female speakers.*

	Laughter			Speech-laugh		
	α	γ	η	α	γ	η
Male	190	62	36	102	34	14
Female	270	74	58	145	37	33

3.4. Slope of Strength of Excitation (γ)

Similar to F_0 Contour, the β values also vary rapidly at epochs for laughter compared to speech-laugh. Hence γ values are higher for laughter compared to speech-laugh as shown in the Fig. 3(c) and Fig. 4(c). These γ values are obtained at each epoch in a similar way as calculated for slope of F_0 contour (α)

3.5. Ratio of Strength of Excitation and Pitch Period (η)

Ratio of strength of excitation (β) and pitch period (T_0) is used as an approximate measure of the opening phase of the vocal folds [7]. Higher values of η for laughter specifies that the opening phase of vocal folds is more for laughter compared to speech-laugh. The values of η for laughter and speech-laugh are shown in Table 2. η is calculated at every epoch location using eq.(1).

$$\eta = \frac{\beta}{T_0} = \beta F_0 \tag{1}$$

where β is strength of excitation, T_0 is pitch period and F_0 is fundamental frequency, at that epoch location.

4. ANALYSIS OF FEATURES

Distributions of the features F_0, β, α, γ and η for laughter and speech-laugh for 13 male and 7 female speakers are shown in Fig. 5. It can be observed that, for all the features used, distributions of laughter are concentrated more at one region and that of speech-laughs are concentrated more at another region. This difference shows that the features are reliable to discriminate speech-laugh from laughter. Apart from distribution of features, there are also differences in mean and variance of the features used. In order to capture these variations, the Kullback-Leibler (KL) distance is calculated using eq. (2). Distributions of F_0, β, α, γ and η are used in pairs (10 2-D distributions) to compute KL distance measure between laughter and speech-laugh.

$$D = \frac{1}{2}(tr(\Sigma_1^{-1}\Sigma_0) + (\mu_1 - \mu_0)^T \Sigma_1^{-1}(\mu_1 - \mu_0) - k - ln(\frac{det\Sigma_0}{det\Sigma_1})) \tag{2}$$

where D is the KL distance, k is the dimension of the distribution, Σ_0, Σ_1 are covariance matrices and μ_0, μ_1 are corresponding means of the 2-D distributions of reference and test segments respectively.

Table 3. *Average KL distance of the 10 distributions between laughter and speech-laugh for 4 male and 3 female speakers.*

	Male		Female	
	$l_1\ V_s\ l_2$	$l\ V_s\ sl$	$l_1\ V_s\ l_2$	$l\ V_s\ sl$
F_0 and β	44.35	168.47	50.39	144.31
F_0 and α	41.26	170.62	52.48	233.90
F_0 and γ	53.45	183.44	59.35	164.81
F_0 and η	49.38	193.41	55.43	194.36
β and α	20.87	141.13	22.07	134.13
β and γ	22.45	154.13	23.20	120.53
β and η	79.12	181.55	91.18	134.58
α and γ	31.87	154.26	16.35	81.93
α and η	32.55	168.59	18.23	116.55
γ and η	40.11	161.25	21.39	120.95

Here, l_1 and l_2 are the laughter segments of the same speaker in different utterances. The KL distance measure was computed between $l_1\ V_s\ l_2$ considering features in pair as shown in Table 3. The same process is repeated for $l\ V_s\ sl$, where l refers to laughter and sl refers to speech-laugh. While comparing $l\ V_s\ sl$, the average of $l_1\ V_s\ sl$ and $l_2\ V_s\ sl$ is computed and represented as $l\ V_s\ sl$. Based on the KL distance measures obtained for 4 male and 3 female speakers as shown in Table 3, a threshold is placed for each pair of features separately.

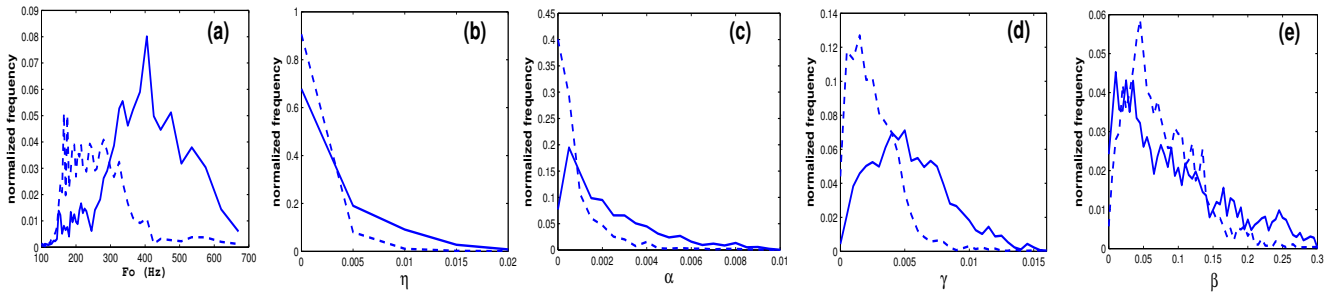

Fig. 5. *The normalized distributions for 13 male and 7 female speakers using (a) F_0, (b) η, (c) α, (d) γ and (e) β. In the illustrations dotted lines represents speech-laugh and solid line represents laughter.*

5. EXPERIMENTS AND RESULTS

As observed from Table 3, the lower values of KL distance measures for laughter1 and laughter2 (l_1 V_s l_2) compared to laughter and speech-laughs (l V_s sl) show that, the variations across laughter are less for the features used compared to variations between laughter and speech-laughs. Based on this observation experiments are performed to discriminate laughter and speech-laugh in a speaker dependent manner.

5.1. Experiment 1

1. A sample laughter segment is taken as reference. A test segment (either laughter or speech-laugh) of the same speaker is taken for comparison with the reference
2. All the 5 features mentioned are extracted for both reference and test segment.
3. The KL distance for 10 2-D distributions between reference and test segments are calculated i.e., 10 KL distance measures are obtained
4. The KL distance obtained for each distribution is compared with the pre-defined thresholds
5. If five or more than five KL distance measures exceed the pre-defined threshold then the test segment is considered as speech-laugh otherwise it is a laughter segment.

5.2. Experiment 2

It is evident from Table 1 that feature space of neutral speech is closer to speech-laugh compared to laughter i.e., the KL distance measures computed between neutral speech and laughter is more compared to distance obtained between neutral speech and speech-laughs. Based on this observation another analysis was made using neutral speech as reference.

1. Instead of collecting a sample laughter segment, neutral speech of a speaker is taken as reference. A test segment (either laughter or speech-laugh) of the same speaker is taken for comparison with the reference
2. All the 5 features are extracted and KL distance measures are computed for 10 2-D distributions between test and reference segments
3. The KL distance obtained for each distribution is compared with the pre-defined thresholds. Here thresholds are obtained by comparing the KL distances obtained between neutral speech and laughter and between neutral speech and speech-laughs, for 4 male and 3 female speakers.
4. If five or more than five KL distance measures exceed the pre-defined threshold, the test segment is considered as laughter otherwise it is a speech-laugh.

Table 4. *Confusion matrix using laughter as reference.*

	Laugh	Speech-laugh
Laugh	86.67%	13.33%
Speech-laugh	22.67%	77.33%

Table 5. *Confusion matrix using neutral speech as reference.*

	Laugh	Speech-laugh
Laugh	88.89%	11.11%
Speech-laugh	25.33%	74.67%

Test data was collected from 9 male speakers and 4 female speakers (Speakers are different from those used in calculating the thresholds). Data used for testing consists of 90 laughter segments and 75 speech-laugh segments. In Table 4 and Table 5, the confusion matrix obtained for experiment 1 and experiment 2 are presented respectively. Even though the inter-speaker and intra-speaker variations are high for laughter, results show that the features used in this study are robust and speaker independent. The lower accuracy in discriminating speech-laughs is due to the transition regions after fricatives and stops, exhibiting features similar to laughter.

6. SUMMARY AND CONCLUSIONS

In this paper, excitation source features are analysed to discriminate laughter and speech-laugh in voiced regions. Distribution functions and KL distance measures show that the features are robust to intra-speaker and inter-speaker variations present in laughter and speech-laugh segments. Based on the analysis, experiments were performed which do not require any training to discriminate laughter and speech-laughs in a speaker dependent manner. Results show that laughter segments are well discriminated. The transition regions after fricatives and stops in speech-laughs exhibit features similar to laughter, reducing accuracy of discriminating speech-laughs. Segregation of these regions from laughter is still a challenge and will be analysed in the future work.

7. REFERENCES

[1] Nwokah, E.E., Hsu, H-C., Davies, P. and Fogel, A, "The integration of laughter and speech in vocal communication: a dynamic system perspective," *Journal of Speech, Language and Hearing Research*, vol. 42, pp. 880 - 894, 1999.

[2] Trouvain. J, "Segmenting phonetics units in laughter," in *Proc. 15th ICPhS*, Barcelona, pp. 2793 - 2796, 2003.

[3] 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.

[4] Kennedy, L. S. and Ellis, D.P.W., "Laughter detection in meetings," in *Proc. NIST ICASSP 2004 Meeting Recognition Workshop*, Montreal, Canada, 2004.

[5] K. Troung and D. VanLeeuwen, "Automatic Detection of Laughter," in *Proc. INTERSPEECH*, pp. 485 - 488, Lisbon, Portugal, 2005.

[6] Knox, Mary Tai and Morgan, Nelson and Mirghafori, Nikki, "Getting the last laugh: automatic laughter segmentation in meetings," in *Proc. INTERSPEECH*, Brisbane, Australia, pp. 797 - 800, Sept. 2008.

[7] Sudheer K., Sri Harish Reddy M., Sri Rama Murty K. and B. Yegnanarayana, "Analysis of laugh signals for detecting in continuous speech," in *Proc. INTERSPEECH*, Brighton, UK, pp. 1591 - 1594, Sept. 2009.

[8] Trouvain, Jürgen, "Phonetic aspects of speech laughs.," *Oralité et Gestualité: Actes du colloque ORAGE, Aix-en-Provence. Paris: LHarmattan*, pp. 634 - 639, 2001.

[9] Menzezes Caroline and Yosuke Igarashi "The speech laugh spectrum," in *Proc. of the 6th International Seminar on Speech Production (ISSP),*, pp. 157 - 524, Dec. 2006.

[10] Batliner, Anton and Steidl, Stefan and Eyben, Florian and Schuller, Björn, "On laughter and speech laugh, based on observations of child-robot interaction," *The phonetics of laughing, trends in linguistics. de Gruyter, Berlin, to appear*, 2010.

[11] Sri Rama Murty K. and B. Yegnanarayana, "Epoch Extraction From Speech Signals," *IEEE Trans. Audio, Speech and Language Processing*, vol. 16, no. 8, pp. 1602 - 1613, Nov. 2008.

[12] Sri Rama Murty K., B. Yegnanarayana and Anand Joseph Xavier M, "Characterization of Glottal Activity From Speech Signals," *IEEE Signal Processing Letters*, vol. 16, no. 6, pp. 469 - 472, June 2009.