Synthesis of laughter by modifying excitation characteristics

Sathya Adithya Thati,[a] Sudheer Kumar K, and B. Yegnanarayana
International Institute of Information Technology, Hyderabad 500032, India

(Received 8 June 2012; revised 1 February 2013; accepted 8 March 2013)

In this paper, a method to synthesize laughter by modifying the excitation source information is presented. The excitation source information is derived by extracting epoch locations and instantaneous fundamental frequency using zero frequency filtering approach. The zero frequency filtering approach is modified to capture the rapidly varying instantaneous fundamental frequency in natural laugh signals. The nature of variation of excitation features in natural laughter is examined to determine the features to be incorporated in the synthesis of a laugh signal. Features such as pitch period and strength of excitation are modified in the utterance of vowel /a/ or /i/ to generate the laughter signal. Frication is also incorporated wherever appropriate. Laugh signal is generated by varying parameters at both call level and bout level. Experiments are conducted to determine the significance of different features in the perception of laughter. Subjective evaluation is performed to determine the level of acceptance and quality of synthesis of the synthesized laughter signal for different choices of parameter values and for different input types.
 [http://dx.doi.org/10.1121/1.4798664]

PACS number(s): 43.72.Ja [MAH] Pages: 3072–3082

I. INTRODUCTION

Nonverbal vocalization plays a key role in expressing emotions in natural human communication. Laughter is one such vocalization that is mostly used to express joyous mood. It induces a positive emotive state on listeners. To a lesser extent, laughter is also used in other emotional contexts such as sarcasm, humiliation, etc., making it an important indicator of emotion/mood.

Recent trends in speech synthesis seem to focus on improving the expressive quality of speech and also on synthesizing non-normal (emotion) speech (Cadic and Segalen, 2008). The extent of emotion in speech is manipulated by including nonverbal cues (Robson and MackenzieBeck, 1999). Also, nonverbal vocalization, such as laughter, adds naturalness to the synthesized speech by bringing the synthesized speech closer to natural human conversation.

Laughter is categorized into three basic types: voiced "song-like" laughter, "snort-like" laughter with perceptually salient nasal turbulence, and "grunt-like" laughter with laryngeal and oral-cavity frication (Bachorowski *et al.*, 2001). Although only about 30% of the analyzed laughs are predominantly voiced, they induce significantly more positive emotional responses in listeners than unvoiced laughs (Bachorowski and Owren, 1995). Trouvain (2003) has segmented laughter at different levels (phrasal, syllabic, segmental, phonation, and respiration) for understanding the structure of a typical laugh. An instance of laughter is referred to as an *episode*. The segment of the laughter episode produced between two inhalation gaps is known as *laughter bout* or *bout*. An entire laugh can have several bouts separated by inhalation (Trouvain, 2003). *Calls* are the discrete acoustic events that together constitute a bout (Bachorowski *et al.*, 2001). Each call of a voiced laughter consists of voiced part followed by an unvoiced/silence part (*intercall* interval). Provine concluded that laughter is usually a series of short syllables repeated approximately every 210 ms (Provine, 1996). Different acoustic descriptions of laughter have been used in the literature for different studies (Bachorowski *et al.*, 2001; Trouvain, 2003; Bickley and Hunnicut, 1992).

The main sound feature of laughter is aspiration /h/. Laughter sounds like a sequence of syllables, which are consonants followed by vowels (open mouthed laughter), or vocalic nasals (closed mouthed laughter; Edmonson, 1987). They are typically perceived as *ha-ha-ha* or *hi-haha* sequence in the case of open-mouthed laughter. The open-mouthed laughter causes lowering of the jaw, resulting in an /a/-colored sound for all vowel categories (Caroline and Igarashi, 2006). It sounds like a sequence of breathy consonant-vowel syllables (/hV/) as in *ha-ha-ha* or *heh-heh* (Bickley and Hunnicut, 1992). Bachorowski *et al.* (2001) found that the vowel-like laughs generally contained central sounds. Ruch and Ekman (2001) also mentioned that the laughter "vowel" is the central vowel schwa or /e/.

Speech production is a controlled process that is guided by a set of rules. The movement of articulators is dictated by the sequence of subword units to be uttered. But unlike speech, there are no rules guiding the process of production of laughter. Laughter is typically produced by a series of sudden bursts of air, released by the lungs, keeping the vocal tract almost steady. Lungs and vocal folds (source of excitation) play a major role in laughter production. Due to high air pressure built up in the lungs, there is larger than normal air flow per unit time through the vocal tract. This results in rapid vibration of the vocal folds. Since vocal folds cannot maintain/sustain that unusual high pitch frequency, their vibration tends to decrease to reach the normal pitch frequency. There is also turbulence generated at the vocal folds which results in the signal being breathy (noisy) when compared to normal speech (Caroline and Igarashi, 2006). All

[a]Author to whom correspondence should be addressed. Electronic mail: sathya.adithya@research.iiit.ac.in

this is in the production of one *call*. The process of call production repeats itself with certain intercall variations to produce a bout.

Synthesis of laughter appears to be more involved as compared to synthesis of speech. This is because of the difficulty in modeling the highly complex mechanism of production of laughter. The variability in laughter production also makes analysis and synthesis of laughter a challenging task. Waveforms of units of speech are available for concatenation for speech synthesis after the text is analyzed to determine the units (unit-selection synthesis; Hunt and Black, 1996). There are no such units in laughter that can be concatenated. Laughter cannot be broken down into distinct units like in speech. Hence, methods like unit-selection cannot be applied to synthesis of laughter. Laughter needs to be synthesized at the signal level, as natural laughter samples are difficult to collect, and organize them into units.

Analysis of signals of natural laughter is needed to understand the characteristics of laughter at both call level and bout level. This will help to bring the synthesized laugh closer to a natural laugh, both at segmental and suprasegmental levels. Laughter has been analyzed based on both source and system characteristics of production. Since laughter is produced by the human speech production mechanism, the laugh signal is also analyzed like a speech signal in terms of the acoustic features of speech production. Typically, the acoustic analysis of laughter is carried out using duration, fundamental frequency of voiced excitations (F_0), and spectral features (Bachorowski *et al.*, 2001; Provine, 1996; Vettin and Todt, 2004). Conventional methods of analysis were used to derive the features of the glottal vibration by Bachorowski *et al.* (2001) and Bickley and Hunnicut (1992). Spectrum-based features like harmonics, spectral tilt, and formants were used to analyze laughter (Bickley and Hunnicut, 1992; Caroline and Igarashi, 2006; Szameitat *et al.*, 2011). The acoustic structure in human laughter was discussed by Kipper and Todt (2007) and Vettin and Todt (2004). Observations were also made on the number of calls per bout and number of bouts in a laughter episode. The problem of extracting rapidly varying instantaneous fundamental frequency (F_0) was addressed by Sudheer *et al.* (2009). Sudheer *et al.* (2009) measured the following features: (i) rapid changes in the instantaneous fundamental frequency (F_0) within a call, (ii) strength of excitation (SoE) within each glottal cycle and its relation to F_0, and (iii) temporal variability of F_0 and SoE across calls within a bout. These features were used for spotting laughter in continuous speech.

There have been attempts to insert natural laugh samples in speech for simulating natural conversation (Campbell, 2006). To insert laughter into conversational speech, laugh samples from a corpus were selected and incorporated in concatenative synthesis (Campbell, 2006). There have also been attempts to model laughter (Trouvain and Schroeder, 2004; Sundaram and Narayanan, 2007; Lasarcyk and Trouvain, 2007). Trouvain and Schroeder (2004) superimposed the duration and F_0 of natural laugh samples onto recordings of diphones ("*hehe*") to generate laughter. Results showed that careful control of the laugh intensity is required for better perception. An attempt to synthesize laughter has been made by Sundaram and Narayanan (2007) using the principle of a damped simple harmonic motion of a mass-spring model to capture the overall temporal behavior of laughter at episode level. The voicing pattern of laughter was seen as an oscillatory behavior, and was observed in most laughter bouts. The behavior of alternate voiced and unvoiced segments was modeled with equations that described the simple harmonic motion of a mass attached to the end of a spring. By varying a set of parameters of the mass-spring system, the authors were able to synthesize the temporal behavior of variety of laughter. At the call level, the signal was synthesized using the standard linear prediction based analysis-synthesis model. The pitch variations were described in terms of minimum, mean, and maximum values within each call. The amplitude and duration information within each call was also specified by the user. Lasarcyk and Trouvain (2007) used an articulatory speech synthesizer to model laughter. A real laugh signal was taken from a spontaneous speech database, and synthetic versions of it were created. Features like breathing noise were also approximated, as they do not normally occur in speech. It was reported that synthesis of laughter, taking into account the variations in durational patterns, intensity, and F_0 contours, improves the perception of naturalness (Lasarcyk and Trouvain, 2007).

Many laughs consist of a mix of voiced and unvoiced types (Bachorowski and Owren, 2004). Voiced laughs are the versions that are commonly thought of as typical laughter, and can have a song-like quality if F_0 happens to fluctuate in a melodic way over the course of several bursts. Unvoiced laughs (snort-like and grunt-like) can be similar to voiced versions, but they lack regular vocal fold vibration, and hence are noisy and atonal in comparison with voiced laughs. As mentioned earlier, voiced laughs are significantly more likely to elicit positive responses from listeners than unvoiced laughs. Moreover, voiced laughter can be studied systematically due to the contribution of the vocal fold vibration as the main source of excitation of the vocal tract system. Hence, in this paper, the focus is on the synthesis of voiced laughter.

In the synthesis procedure given in Sundaram and Narayanan (2007), the overall temporal variations are conditioned by the control parameters of the model. Likewise, the variations of the pitch and amplitude parameters within a call are specified in the form of range of these parameters. In this paper, the patterns of variations of pitch and SoE that occur in natural laughter are captured in the form of a model at both bout level and call level by analyzing several samples of laugh signals from several speakers. In particular, the rapid variations of the instantaneous fundamental frequency and the associated variation in the SoE are modeled and incorporated in this study. Samples of voiced laughter are analyzed to derive the parameters to control the synthesizer at the call level and bout level. The synthesis can be performed using either a natural steady vowel segment from a speaker or a synthetic vowel segment.

The paper is organized as follows: In Sec. II, the basic analysis tool to extract the information of the excitation component of the signal is described. In Sec. III, analysis of

laughter is discussed to derive the characteristics of natural laugh signals. Section IV describes how the features of a speech signal can be modified to incorporate the desired features of laughter, followed by the description of the synthesis procedure. Experiments to determine the perceptual significance of features are discussed in Sec. V. Subjective evaluation of the results of laughter synthesis is presented in Sec. VI. Section VII gives a summary of this paper.

II. METHOD TO EXTRACT INSTANTANEOUS FUNDAMENTAL FREQUENCY AND SoE AT EPOCHS

A method was proposed (Murty and Yegnanarayana, 2008; Murty *et al.*, 2009) for extraction of the instantaneous F_0, epochs, and strength of impulselike excitation at epochs. 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.

A zero-frequency filtered (ZFF) signal is derived as follows:

(a) The speech signal $s[n]$ is differenced to remove unwanted very low frequency components

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

(b) The differenced speech signal is passed through a cascade of zero-frequency resonators (digital resonators having poles at zero frequency) given by the following equation:

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

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

(c) The trend in $y_0[n]$ is removed by subtracting the mean computed over a window at each sample. The resulting signal $y[n]$ is the ZFF signal, given by

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

where $(2N+1)$ is the size of the window, which is in the range of 1 to 1.5 times the average pitch period in samples.

This method does not capture the rapid variations of F_0 that appear in the calls of a laughter episode. To capture the rapid variations of a laugh signal, the method was modified using the following steps to derive the epochs and their strengths from the ZFF signals (Sudheer *et al.*, 2009):

(1) Pass the signal through the zero-frequency resonator with a window length of 3 ms for trend removal. The ZFF signal has high energy in the regions of voiced speech and laughter, and low energy in the nonvoiced regions, which includes unvoiced and silence regions.
(2) Voiced and nonvoiced segments of the signal are determined using the ZFF signal. Samples of normalized ZFF signal are squared and their running mean over a window of 10 ms is calculated to estimate the envelope of the signal. It is then normalized by using the following equation:

$$s_2 = 1 - e^{-10*s_1}, \tag{4}$$

where s_1 is the estimated envelope and s_2 is the normalized envelope. The set of samples in s_2 having a value above the threshold of 0.3 is marked as voiced regions in the signal. The value 10 in Eq. (4) and the threshold 0.3 are determined based on study on large amount of speech data.
(3) After finding the voiced segments, the signal in each voiced region is passed separately through a zero-frequency resonator with window length for trend removal derived from that segment. The location of the maximum peak in the autocorrelation function of the segment is used to determine the window length for trend removal in that region. Due to rapid changes in the pitch period values, the window size for trend removal is chosen adaptively for each segment.
(4) 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 (slope) gives the SoE.

The results at various stages to obtain the pitch period contour and SoE from a segment of speech signal are shown graphically in Fig. 1. Speech signal and the ZFF signal obtained in the first step are plotted in Figs. 1(a) and 1(b), respectively. Figure 1(c) illustrates the voicing and nonvoicing decision on the speech signal using the ZFF signal. Figure 1(d) shows the filtered signal obtained after passing the voiced segments through a zero-frequency resonator with an adaptive window length. Figures 1(e) and 1(f) show the contours of the SoE and pitch period obtained for the segment of speech signal.

III. ANALYSIS OF LAUGH SIGNALS FOR SYNTHESIS

In this section, natural laugh signals from several speakers (over five male and five female) are analyzed to derive the characteristics of laughter. Laughter samples for analysis were taken from "The AVLaughterCycle Database" (Urbain *et al.*, 2010), "The AMI Meeting Corpus" (Bachorowski and Owren, 2004), and online repository of Bachorowski *et al.* (2001). Analysis of laugh signals is done in terms of the excitation characteristics of the production mechanism to determine the features needed to synthesize laughter (Sudheer *et al.*, 2009). Features that are taken into consideration are: (i) rapid changes in F_0 within calls of a laughter bout, (ii) SoE at each epoch, (iii) durations of different calls in a bout, and (iv) breathiness/fricative segments in the laugh signal. Following are the main features that are observed in the laugh signals.

A. Pitch period

Fundamental frequency of laughter is observed to be significantly higher than that for normal speech. As described

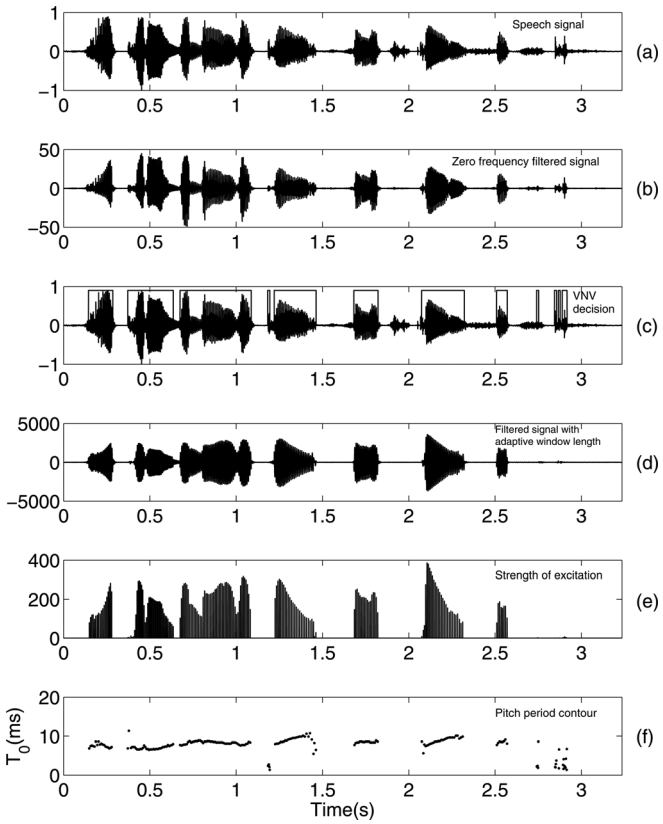

FIG. 1. (a) A segment of speech signal, (b) ZFF signal using a window length of 3 ms for trend removal, (c) voiced/nonvoiced decision based on ZFF, (d) filtered signal obtained with adaptive window length for trend removal, (e) SoE, and (f) pitch period (T_0) obtained from epoch locations.

earlier, during laughter production, there will be more airflow due to high sub-glottal pressure through the vocal tract. This results in faster vibration of the vocal folds, and hence reduction in the pitch period. There is a raising pattern observed in the pitch period contour of a call. The general pattern that is observed in the pitch period contour within a call is that it starts with some low value, decreases slightly, and then increases nonlinearly to a high value, with vocal folds tending to reach the normal pitch frequency. This is because it is not normal for the vocal folds to maintain the initial high fundamental frequency (F_0). The higher this slope, the more intense is the laughter. With the progress of the calls, the slope of the pitch period (T_0) contour also tends to fall. The rate of fall across calls is assumed to be linear. Figure 2(c) shows the pattern of pitch period contour for a segment of a typical laugh signal. We can observe from the figure that the pitch period values change nonlinearly. A quadratic approximation seems to fit the pitch period contour well for a majority of the laugh signals based on observation of characteristics of laugh signals for over ten different speakers in various contexts. The following quadratic polynomial is used to approximate the shape of the pitch period (T_0) contour within the voiced region of a call:

$$y[n] = T_{\min} + \frac{\left(n - \frac{L}{3}\right)^2}{\left(\frac{2L}{3}\right)^2} * (T_{\max} - T_{\min}), \tag{5}$$

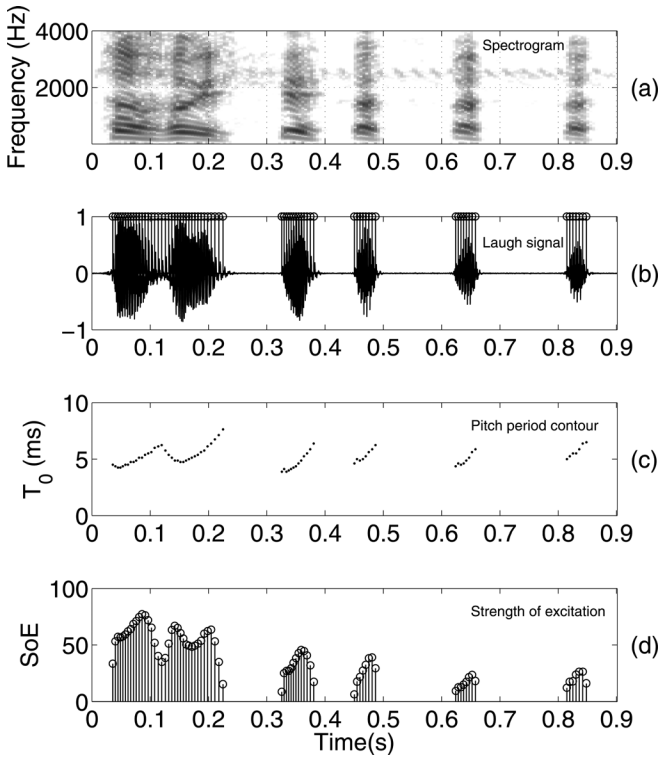

FIG. 2. (a) Spectrogram of laugh signal, (b) a segment of laugh signal, (c) pitch period derived from the epoch locations, and (d) SoE at the epochs.

where L is the length of the voiced region, and $T_{\min}$ and $T_{\max}$ are the specified minimum and maximum values of the pitch period within the call region, respectively. This relation has been obtained after examining several T_0 contours in natural laugh signals of several speakers. While this is only an approximation, it reflects the changes in the pitch period contour of natural laugh signals. Figure 3 shows the actual pitch period contour and the modeled (using quadratic approximation) pitch period contour.

B. SoE

Similar to pitch period, the SoE at epochs also changes rapidly. It increases nonlinearly, and then decreases almost in a similar fashion. The slope of the SoE contour typically falls with the progress of the calls. Figures 2(c) and 2(d) illustrate the general trend of the contours of the pitch period and SoE for a segment of a typical laugh signal. The pattern of the nonlinear increase and decrease in the strengths can be observed in Fig. 2(d). Note also the somewhat inverse relation in the variation of T_0 and SoE contours. The approximate inverse quadratic relation of the SoE contour is given by

$$y[n] = 1 - \frac{\left(n - \frac{4L}{7}\right)^2}{\left(\frac{4L}{7}\right)^2}, \tag{6}$$

where L is the length of the voiced region of the call. Note that this approximation has been derived at after examining several natural SoE contours from several speakers. Note

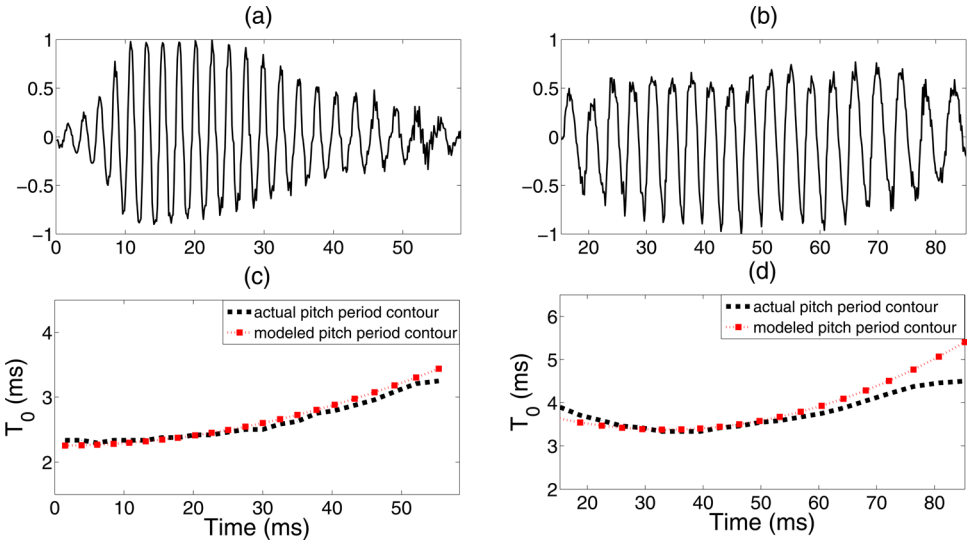

FIG. 3. (Color online) Illustration of original and modeled pitch period contours. Two laugh calls are shown in (a) and (b) with their corresponding pitch period contours in (c) and (d), respectively. In (c) and (d), the actual pitch period contour is shown using a dashed line and the modeled one is shown using a dotted line.

also that the relation in Eq. (6) is not an exact inverse of Eq. (5).

C. Duration

The gap between two calls of a laughter is referred to as the intercall gap. The duration of the intercall gap is called intercall duration (ICD), and the duration of the call is called call duration (CD). CDs are typically observed to be in the range of 0.08 to 0.20 s. For synthesis, any value in that range could be used for the CD. ICDs are generally in the range of 0.5 to 1.5 times the CD. The ratio of the duration of unvoiced to voiced segments in a laugh signal was reported to be >1 by Bickley and Hunnicut (1992). ICD in a laughter bout was observed to increase with the progress of the calls. This was also confirmed by Kipper and Todt (2007), where the duration of call was reported to decrease and the duration of the interval increases within a bout. In general, no pattern was observed for CDs. The CDs vary depending on the speaker and the kind of laughter.

D. Frication

The airflow during the open phase of the glottis is very high. This results in a strong turbulent noise source at the glottis (Caroline and Igarashi, 2006). Within a call, the volume velocity of air typically reduces from left to right. As a result the amount of breathiness also falls within a call. As the calls progress, the amount of breathiness decreases in successive calls. Also, because of high amount of air flow, glottal fricative /h/ (aspiration) is produced in the intercall intervals.

IV. SYNTHESIS OF LAUGHTER

In this work, laughter is synthesized by modifying the features mentioned in Sec. III for a natural vowel uttered by a speaker or for a synthetic vowel. The synthetic vowel is generated by exciting the all-pole model corresponding to the vowel with a sequence of glottal pulses at a specified pitch frequency. Each glottal pulse is approximated by the Liljencrants-Fant model (Fant *et al.*, 1985; Fant, 1995). The synthesis process involves modifying the characteristics of the source without changing the characteristics of the system. The following are the main stages involved in generating a laugh signal.

A. Incorporation of feature variations

1. Pitch period modification

Pitch period of the input natural/synthetic vowel signal is modified using the method discussed in Rao and Yegnanarayana (2006). The input signal of the vowel is passed through the zero-frequency resonator for deriving the epoch locations as described in Sec. II. The interval between the epoch locations gives the pitch period. Note that throughout this study speech samples at 8 kHz sampling rate are considered. Hence, a 10th order pitch synchronous linear prediction analysis is used to separate approximately the source (LP residual) and system (LP coefficients) components. The LP residual and LP coefficients are associated with every epoch location. The desired pitch period contour for laughter is generated from the specification of the pitch period modification. The pitch period contour within each call follows a quadratic polynomial [see Eq. (5)]. New epoch locations are derived from the modified pitch period contour. The LP residual and LP coefficients are copied for each epoch in this new epoch sequence from the corresponding nearest epochs of the input signal. The residual in each epoch interval of the new epoch sequence is resampled according to the pitch modification factor at that epoch. The procedure for generating the new residual signal was described in Rao and Yegnanarayana (2006).

2. Modifying the residual by the SoE

SoE is an estimate of the strength of the impulse-like excitation at the epoch. In order to establish the relation between the SoE and the amplitude of the peaks of the Hilbert envelope of the residual signal at each instant of glottal closure (i.e., epoch), the following experiment was conducted (Murty *et al.*, 2009). A sequence of impulses with varying durations between consecutive impulses and with

different amplitudes is generated. The sequence is passed through an all-pole filter with LP coefficients corresponding to different vowels. The output signals are passed through the zero-frequency resonator, and the values of the SoE are obtained. The resulting values of the SoE are compared with the amplitudes of the impulses. There is an approximate linear relation observed between them. Note that the energy of the LP residual corresponds to the gain term and hence the strength of impulse in the all-pole model approximation in linear prediction. In order to maintain the gain of the residual signal to the same level of the SoE, the amplitudes of the residual samples in an epoch interval are modified by multiplying them with a scaling factor according to the desired SoE contour. The inverse of the quadratic approximation as given in Eq. (6) is used for the desired SoE contour. The general trend of the SoE contour in each call is to first increase, and then decrease nonlinearly [see Fig. 2(d)].

3. Incorporation of frication

Frication or breathiness is incorporated at two levels: within the voiced region of each call and within the intercall intervals. To generate frication, white Gaussian noise equal to the length of the voiced region of a call is generated. The noise samples are scaled to obtain an energy equal to about 80% to 120% of the energy of the residual signal within a call. The desired amount of noise depends on the call number in the bout, as the amount of noise energy is reduced with the progress of the calls. The frequency response of the noise samples is modified to generate a band limited Gaussian noise. Hence the noise samples are passed through a bandpass filter with a resonance frequency of 2500 Hz and bandwidth of 500 Hz. The choice of the resonance frequency and its bandwidth is not critical, except that the noise is somewhat band limited in the higher frequency range. The sequence is then multiplied with a weighing function $w[n] = 1 - n/L$, where n is the sample number and L is the total number of samples in the voiced region of a call, so as to obtain a linearly decreasing effect of frication. The resulting noise samples are added to the residual samples to generate modified residual, which in turn is used to excite the corresponding all-pole filter to synthesize the call region of laugh signal. Separately, band limited random noise with low amplitude (i.e., about 0.1% of the energy of the call) is used to fill the intercall region.

B. Steps in the synthesis of laughter

The block diagram in Fig. 4 shows the steps involved in the synthesis of a voiced region of a call in the laugh signal.

(1) A segment of the signal (segment of natural or synthetic vowel) corresponding to the length of a CD is chosen.
(2) The input signal is passed through a zero-frequency resonator for deriving the epoch locations. Pitch period is obtained by computing the interval between successive epoch locations.
(3) A 10th order linear prediction analysis is performed on the signal (sampled at 8 kHz) to derive the source (LP residual) and system (LP coefficients) components. The LP residual between epochs and the LP coefficients are associated for each epoch.
(4) The pitch period contour and SoE contour are determined as described in Sec. III, according to the desired prosody modification as described in Sec. IV A.
(5) New LP residual is generated according to the pitch period contour as explained in Sec. IV A 1.
(6) The residual signal is scaled according to the SoE contour as described in Sec. IV A 2.
(7) Frication is then incorporated in the residual of the voiced region of a call as explained in Sec. IV A 3.
(8) The modified residual signal is then used to excite the all-pole filter of the vowel to synthesize the voiced region of the call.
(9) Band limited random noise with very low amplitude (about 0.1% of the energy of the call) is generated and is used to fill the intercall interval.
(10) The above steps are repeated for synthesizing different calls in the laughter, and to finally obtain a laughter bout.

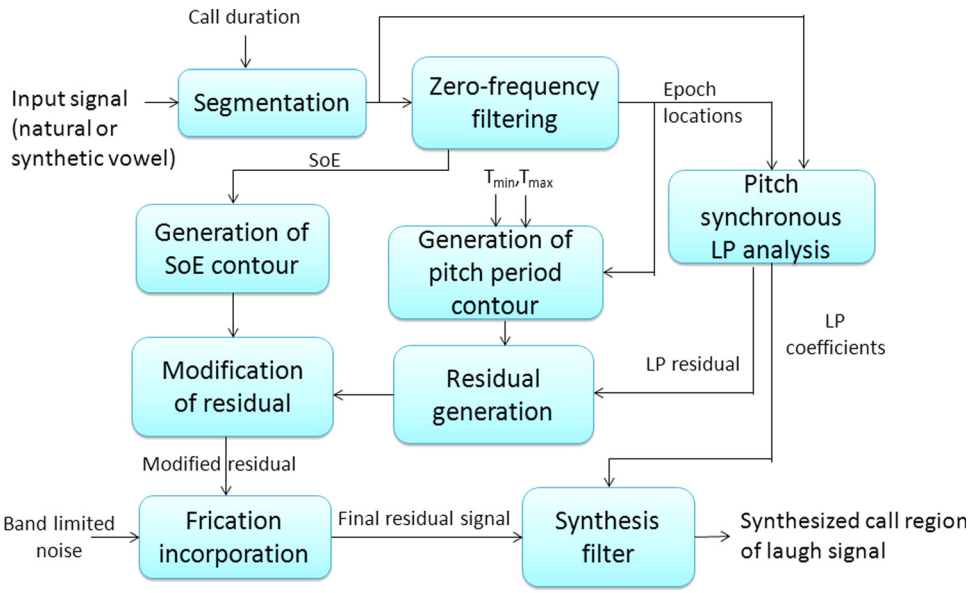

FIG. 4. (Color online) Block diagram for generation of one call region (voiced) of laugh signal.

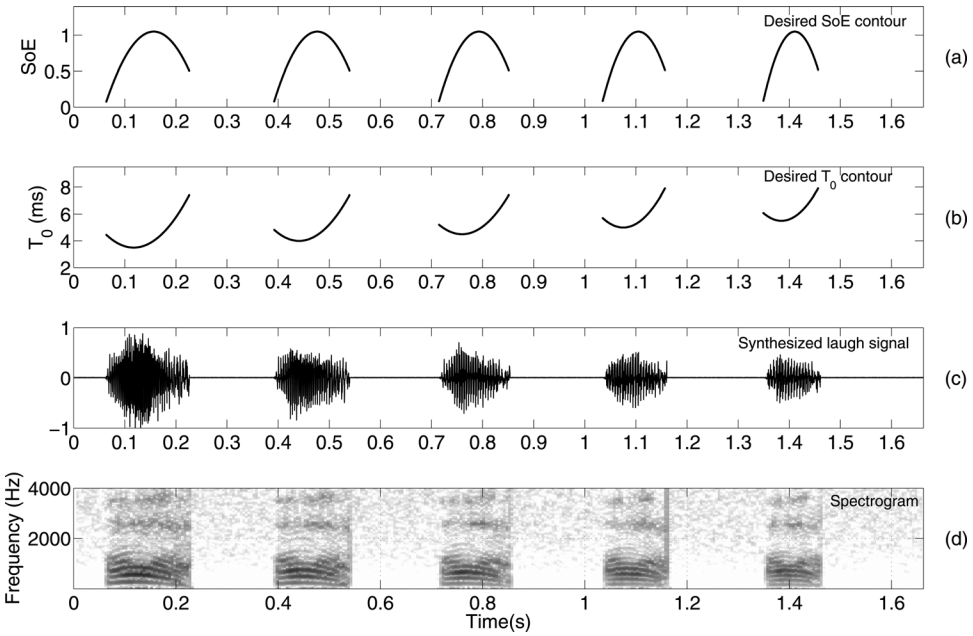

FIG. 5. Illustration of synthesized laugh signal. (a) Desired SoE contour, (b) desired pitch period (T_0) contour, (c) synthesized laugh signal, and (d) spectrogram of the synthesized laugh signal.

(11) Multiple bouts are synthesized, each with a different number of calls and with different values for control parameters.

Figure 5 shows a synthesized laugh signal along with the desired SoE and T_0 contours that are used to generate it. Figure 5(a) shows the desired SoE contour, which follows the inverse of a quadratic polynomial given in Eq. (6).

Figure 5(b) shows the desired pitch period contour. The LP residual is modified to incorporate the desired pitch period (T_0) contour. The contour is generated by using the Eq. (5) for given values of $T_{\min}$ and $T_{\max}$.

Figure 5(c) shows the synthesized laugh signal, and Fig. 5(d) shows its spectrogram. The T_0 of the first call ranges from 3.5 ms to 7.5 ms. The T_0 of the last call ranges from 5.5 ms to 8 ms. The minimum T_0 is increased as the calls progress. Also, the CD decreases with calls. The CD for the first call is chosen as 0.165 s. Durations of the remaining calls are decreased gradually. The first ICD is chosen to be the same as the duration of the first call, and the ICD is increased progressively. After the calls are generated, intensity of the calls is decreased as desired.

The proposed laughter synthesis system is a flexible system, where the parameters to generate laughter can be controlled by the user. The parameters that can be manually set by the user along with their preferred range are given in Table I. The range values of the parameters are obtained after examining several examples of natural laugh signals from several speakers and in several contexts. Although any value chosen in the mentioned range will work, a proper combination of the values produces good quality of the synthesized laughter. Following are a few examples of the many subtle and important interdependencies among the parameters that need to be taken into account to avoid generating poor quality laugh signal.

(i) Long bouts are associated with higher values of mean F_0 for calls.
(ii) Calls are longer in duration when they are less in number.
(iii) ICD depends on the call number.

There are several such interdependencies that need to be taken into account to produce natural sounding synthetic laughter.

TABLE I. Parameters and preferred range of values for laughter synthesis.

Parameter	Preferred range of values
Number of bouts	1–3
Number of calls in each bout	4–7 (depends on bout number)
Duration of each call	50–250 ms (depends on call number)
Duration of each intercall	50–250 ms (0.5 to 1.5 times of CD)
Maximum T_0 of each call	5–8 ms (for male)
	4–6 ms (for female)
Minimum T_0 of each call	2–4 ms (for male)
	1–2 ms (for female)
Amount of frication in each call (in terms of residual energy)	80% to 120%
Intensity ratio of first call to last call	1 to 100 ($<1 \Rightarrow$ increasing intensity)

V. EXPERIMENTS

A. Perceptual significance of features

An experiment based on analysis-by-synthesis approach was conducted to determine the perceptual significance of

TABLE II. Perceptual evaluation scores obtained for the modified versions of an original laugh signal.

					Acceptability	
Sample	T_0	SoE	Breathiness	CDs and ICDs	Mean	Standard deviation
1	0	0	0	0	4.35	0.47
2	0	0	0	1	4.05	0.46
3	0	0	1	0	3.7	0.47
4	0	1	0	0	4.1	0.57
5	1	0	0	0	3.0	0.48

TABLE III. Performance of laughter synthesis system in terms of MOS.

Combination	First to last call intensity ratio	ICD to CD ratio	Breathiness	Quality of synthesis		Acceptability	
				mean	Standard deviation	mean	Standard deviation
1	2	0.5	0	2.92	0.84	2.83	0.89
2	2	0.5	1	3	0.82	2.79	0.84
3	2	1	0	3.06	0.83	2.88	0.77
4	2	1	1	2.93	0.72	2.93	0.85
5	10	0.5	0	3.27	0.86	3.31	0.94
6	10	0.5	1	3.07	0.87	3	1.12
7	10	1	0	3.23	0.73	3	0.89
8	10	1	1	3.12	0.85	2.79	0.97
9	10	1.5	0	3.21	0.8	2.93	0.91
10	10	1.5	1	2.93	0.8	2.63	0.98

the features described in Sec. III. For this experiment, original laugh signals are taken, and the following features are modified: T_0 ($=1/F_0$), SoE, amount of breathiness, and CDs and ICDs. For each laugh signal, one of the four features is modified to suppress the effect of that feature.

The modification of the original laugh signal is performed as follows: The laugh signal is segmented using the voiced/nonvoiced decision to extract the calls and intercalls. For every laugh syllable (call plus intercall), the following changes are made:

(1) T_0: Rising pitch period contour is replaced with a constant average pitch period.
(2) SoE: Strength of excitation contour is replaced with a constant average value.
(3) Breathiness: Breathiness is reduced by a factor of 10, by decreasing the relative amplitudes of the samples which are more than 1 ms away from the epoch within each pitch period and in the intercall intervals.
(4) Duration: CDs and ICDs are changed randomly using different proportions of the voiced and unvoiced regions.

The modified signals were played randomly to 20 subjects who were asked to score the samples for acceptability according to their preference on a scale of 1 to 5 where 1 is very poor, 2 is poor, 3 is average, 4 is good, and 5 is excellent. Table II gives the results of the experiment. The table gives mean opinion scores (MOSs) of acceptability for modifications of different features and for the original signal. In Table II, a "0" in the feature column indicates that the feature is not modified, and a "1" indicates that it is modified in the given sample. We can see from the table that the first sample, which is the original version (all 0s), has high score. The case (sample 5), where the pitch changes are suppressed, gives the least score, indicating the significance of pitch contour over other features.

B. Speaker identification from laughter

An experiment was conducted to explore the possibility of identifying a person from laughter. As part of the experiment, subjective tests were conducted. Subjects were presented with ten isolated laughter samples comprised of both synthesized and natural laughter. Synthesized laughs belonged to four different speakers. Natural laughs included laughs of three celebrities and four others, whom the listeners are familiar with. Identity of the speakers was hidden from the subjects, and they were asked to identify the speakers from the laughter samples. Speech produced by the same speakers was also presented to the subjects. None of the subjects were able to identify any of the speakers from their isolated laughter samples, though they were able to identify all the speakers from their respective speech samples. This implies that laughter does not carry significant speaker-specific information.

The results are not really surprising. We can identify a person from his/her laughter only if the laughter contains any special/unique attention-gathering characteristics, or if we have had enough exposure to the person's laughter in isolation. It is even more difficult to identify a person from synthesized laughter, probably because a person's natural way/style of laughter might not match with the prosodic features of the synthesized laughter.

VI. RESULTS OF SYNTHESIS

Subjective tests were performed to evaluate the quality and acceptability of the synthesized laughter. The laughter samples were synthesized for six different types of inputs and also for ten different combinations of the parameter values given in Table III. The different combinations of parameter values were chosen to demonstrate the flexibility in synthesizing laughter. Note that these combinations of parameters are only illustrative, but not exhaustive. The following are the six different types of inputs considered for this study:

TABLE IV. Perceptual evaluation scores (MOS) obtained for different input types.

Input type	Quality of synthesis	Acceptability
Natural female /i/	3.16	2.94
Natural male /i/	3.57	3.22
Natural female /a/	3.22	3.1
Natural male /a/	3.42	3.38
Synthetic male /i/	2.36	2.17
Synthetic male /a/	2.73	2.68

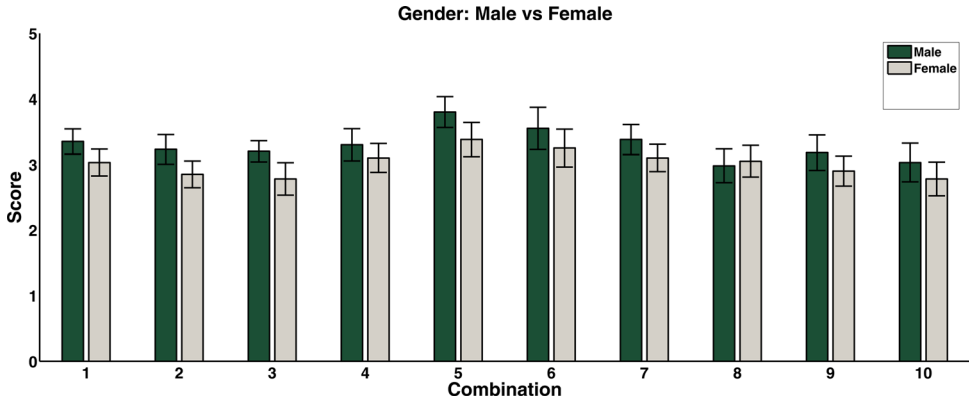

FIG. 6. (Color online) Bar graph illustrating the mean scores of acceptability (with 95% confidence interval) for synthesis of male and female laughter.

(i) 1 female speaker and natural vowel /i/,
(ii) 1 male speaker and natural vowel /i/,
(iii) 1 female speaker and natural vowel /a/,
(iv) 1 male speaker and natural vowel /a/,
(v) 1 male voice pitch and synthetic vowel /i/,
(vi) 1 male voice pitch and synthetic vowel /a/.

With these 6 different types of inputs and 10 different combinations, 60 different laughter samples are synthesized. These samples were played in random order to 20 subjects to elicit their subjective impression on the quality of synthesis and acceptability. They were asked to give their opinion on a scale of 1 to 5 where 1 is very poor, 2 is poor, 3 is average, 4 is good, and 5 is excellent. While the quality and acceptability are somewhat related, in general, it appears that the average score for acceptability may be lower than the average score for quality, although there may be a few exceptions.

Table III gives the MOS for quality and acceptability over all the six cases of input for the ten different combinations of the parameter values. Note that for call intensity ratios (from the first to the last call), one small (i.e., 2) value and one large (i.e., 10) value are chosen to study its effect on perception. It appears that this parameter does not have a significant effect on the MOS. Similarly, the ICD to CD ratios (0.5, 1.0, 1.5), and the presence (i.e., "1") and absence (i.e., "0") of breathiness also do not seem to significantly affect the quality and acceptability scores. The results in Table III show that any reasonable choice of combination of parameter values seems to produce laughter samples of reasonable quality and acceptability, i.e., with an average score over 3 in most cases.

The MOS for different types of input voices, averaged over all the ten combinations of parameter values, are given in Table IV. It is interesting to note that, in general, the scores are lower for female voices than for male voices, and the scores are higher for natural vowels than for synthetic vowels. Also, it appears that the scores for the vowel /a/ seem to be higher than for the vowel /i/.

For the three cases (gender, vowel, and natural/synthetic) in Table IV, the bar graphs for each combination of the parameter values are shown in Figs. 6–8. The graphs show the mean values of the opinion scores of acceptability from the 20 subjects. The graphs also show the 95% confidence interval for each combination of parameter values. It can be clearly seen that the MOS are similar across all the ten combinations considered here. The differences in MOS between natural and synthetic vowels is high compared to the difference between the two vowels (/a/ and /i/), and also between male and female speakers.

VII. SUMMARY

In this paper, features of source of excitation of a vowel utterance (natural or synthetic) have been modified to synthesize laughter. Analysis was done to study the excitation characteristics in laugh signals. Excitation source characteristics such as epoch locations, instantaneous F_0, and SoE features were derived using the zero-frequency resonator output of the speech signal. These features were manipulated to synthesize a laugh signal.

Some advantages of the proposed method are that users have control over the parameters, and that the method can be

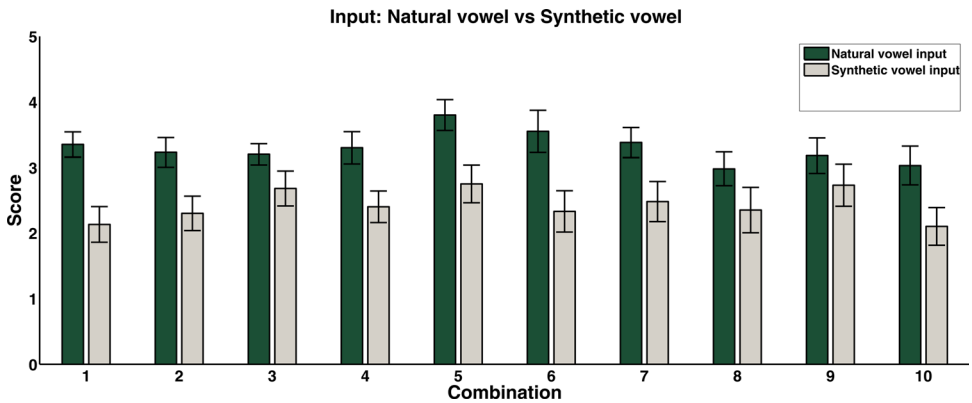

FIG. 7. (Color online) Bar graph illustrating the mean scores of acceptability (with 95% confidence interval) for synthesis using natural and synthetic input vowels.

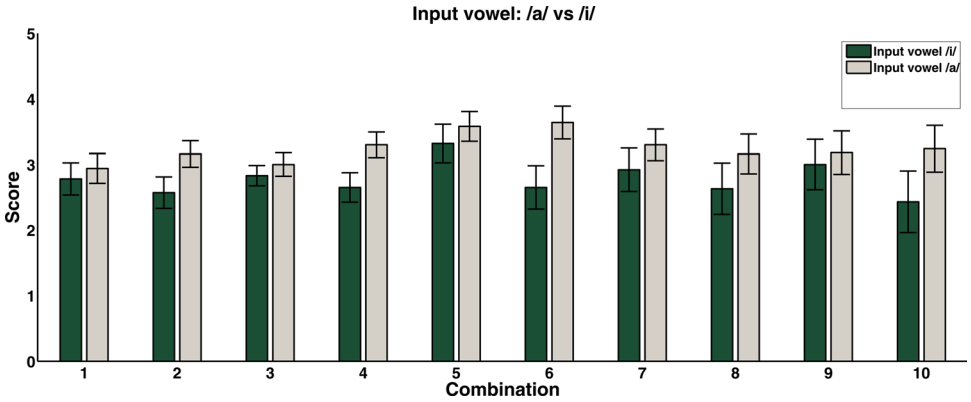

FIG. 8. (Color online) Bar graph illustrating the mean scores of acceptability (with 95% confidence interval) for synthesis using vowels /i/ and /a/ as inputs.

used to generate a wide range of variability that may be present in the physiological process of laughter production. The features and parameters that can be controlled are: duration of calls, duration of intercalls, pitch contour, range of pitch period (maximum and minimum T_0) for each call, SoE contour, amount of frication, number of calls in a bout, number of bouts, and call intensity ratios (envelope of the intensity of laughter calls at bout level). The method can produce different types of laughter, with bouts of different lengths, where each bout may contain a different number of calls. The method is also flexible enough to produce laughter with different vowels. Note that in this study only the differences in the source of excitation were taken into account. No modification was done to the system parameters. Also, the method accounts predominantly for synthesizing voiced laughter.

Experiments were conducted for studying the perceptual significance of different features of laughter. The experiments indicate that pitch contour is the most significant factor in contributing to naturalness of laughter. It was observed that it is not normally possible to identify a person from laughter. It is difficult to determine suitable combination of parameter values to generate a perfectly acceptable laughter. Hence, a range is suggested for values of the parameters that can be controlled by the user. Better results may be obtained by suitably modifying the system features also along with the source parameters. Samples of the output of the laughter synthesis system are available at http://speech.iiit.ac.in/svldemos/laughtersynthesis/index.html (last viewed March 28, 2013).

In this work, knowledge of speech production characteristics have been used for synthesizing laugh signals. The proposed signal processing methods could be used to synthesize other nonverbal cues as well, provided sufficient analysis is performed to determine the parameters to control. The major challenges in synthesizing a laugh signal is to incorporate many rapidly varying features, while maintaining the quality of the laughter.

Bachorowski, J. A., and Owren, M. J. (**1995**). "Not all laughs are alike: Voiced but not unvoiced laughter elicits positive affect in listeners," Speech Transm. Lab. Q. Prog. Status Rep. **36**(2–3), 119–156.

Bachorowski, J. A., and Owren, M. J. (**2004**). "Laughing matters," Psychol. Sci. Agenda Am. Psychol. Assoc. Online **18**(9), 1–5.

Bachorowski, J. A., Smoski, M. J., and Owren, M. J. (**2001**). "The acoustic features of human laughter," J. Acoust. Soc. Am. **110**(3), 1581–1597.

Bickley, C., and Hunnicut, S. (**1992**). "Acoustic analysis of laughter," in *Proceedings of International Conference on Spoken Language Processing*, Banff, Alberta, Canada, pp. 927–930.

Cadic, D., and Segalen, L. (**2008**). "Paralinguistic elements in speech synthesis," in *Proceedings of Interspeech*, Brisbane, Australia, pp. 1861–1864.

Campbell, N. (**2006**). "Conversational speech synthesis and the need for some laughter," IEEE Trans. Audio, Speech, Lang. Process. **14**, 1171–1178.

Caroline, M., and Igarashi, Y. (**2006**). "The speech laugh spectrum," in *Proceedings of the 7th International Seminar on Speech Production (ISSP)*, Ubatuba-SP, Brazil, pp. 517–524.

Edmonson, M. S. (**1987**). "Notes on laughter," Anthropol. Linguist. **29**, 23–33.

Fant, G. (**1995**). "The LF model revisited. Transformations and frequency domain analysis," Speech Transm. Lab. Q. Prog. Status Rep. **36**(2-3), 119–156.

Fant, G., Johan, L., and Qi-guang, L. (**1985**). "A four-parameter model of glottal flow," Speech Transm. Lab. Q. Prog. Status Rep. **26**(4), 1–13.

Hunt, A., and Black, A. (**1996**). "Unit selection in a concatenative speech synthesis system using large speech database," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, Atlanta, Georgia, pp. 373–376.

Kipper, S., and Todt, D. (**2007**). "Series of similar vocal elements as a crucial acoustic structure in human laughter," in *Proceedings of the Interdisciplinary Workshop on The Phonetics of Laughter*, Saarbrucken, Germany, pp. 3–7.

Lasarcyk, E., and Trouvain, J. (**2007**). "Imitating conversational laughter with an articulatory speech synthesizer," in *Proceedings of the Interdisciplinary Workshop on The Phonetics of Laughter*, Saarbrucken, Germany, pp. 43–48.

Murty, K. S. R., and Yegnanarayana, B. (**2008**). "Epoch extraction from speech signals," IEEE Trans. Audio, Speech, Lang. Process. **16**(8), 1602–1613.

Murty, K. S. R., Yegnanarayana, B., and Joseph, M. A. (**2009**). "Characterization of glottal activity from speech signals," IEEE Signal Process. Lett. **16**(6), pp. 469–472.

Provine, R. (**1996**). "Laughter," Am. Sci. **84**(1), 38–45.

Rao, K. S., and Yegnanarayana, B. (**2006**). "Prosody modification using instants of significant excitation," IEEE Trans. Audio, Speech, Lang. Processi. **14**(3), 972–980.

Robson, J., and MackenzieBeck, J. (**1999**). "Hearing smiles-perceptual, acoustic and production aspects of labial spreading," in *Proceedings of International Conference of the Phonetic Sciences (ICPhS)*, San Francisco, pp. 219–222.

Ruch, W., and Ekman, P. (**2001**). "The expressive pattern of laughter," *Emotions, Qualia and Consciousness*, in Series on Biophysics and Biocybernetics, edited by A. Kaszniak (World Scientific, Singapore), Vol. 10, pp. 426–443.

Sudheer, K., Reddy, M. S. H., Murty, K. S. R., and Yegnanarayana, B. (**2009**). "Analysis of laugh signals for detecting in continuous speech," in *Proceedings of Interspeech*, Brighton, UK, pp. 1591–1594.

Sundaram, S., and Narayanan, S. (**2007**). "Automatic acoustic synthesis of human-like laughter," J. Acoust. Soc. Am. **121**(1), 527–535.

Szameitat, D. P., Darwin, C. J., Szameitat, A. J., Wildgruber, D., and Alter, K. (**2011**). "Formant characteristics of human laughter," J. Voice **25**(1), 32–37.

Trouvain, J. (**2003**). "Segmenting phonetic units in laughter," in *Proceedings of the 15th International Congress of Phonetic Sciences*, Barcelona, Spain, pp. 2793–2796.

Trouvain, J., and Schroeder, M. (**2004**). "How (not) to add laughter to synthetic speech," in *Proceedings of the Workshop on Affective Dialogue Systems (ADS)*, Kloster Irsee, Germany, pp. 229–232.
Urbain, J., Bevacquay, E., Dutoit, T., Moinet, A., Niewiadomskiy, R., Pelachaudy, C., Picart, B., Tilmanne, J., and Wagnerz, J. (**2010**). "The AVLaughterCycle database," in *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, pp. 2996–3001.
Vettin, J., and Todt, D. (**2004**). "Laughter in conversation: Features of occurrence and acoustic structure," J. Non-Verb. Behav. **28**(2), 93–115.