Automatic acoustic synthesis of human-like laughter[a)]

Shiva Sundaram[b)] and Shrikanth Narayanan[c)]
Speech Analysis and Interpretation Lab (SAIL), Department of Electrical Engineering-Systems, 3740 McClintock Ave, EEB400, University of Southern California, Los Angeles, California 90089[d)]

(Received 3 February 2006; revised 18 October 2006; accepted 18 October 2006)

A technique to synthesize laughter based on time-domain behavior of real instances of human laughter is presented. In the speech synthesis community, interest in improving the expressive quality of synthetic speech has grown considerably. While the focus has been on the linguistic aspects, such as precise control of speech intonation to achieve desired expressiveness, inclusion of nonlinguistic cues could further enhance the expressive quality of synthetic speech. Laughter is one such cue used for communicating, say, a happy or amusing context. It can be generated in many varieties and qualities: from a short exhalation to a long full-blown episode. Laughter is modeled at two levels, the overall episode level and at the local call level. The first attempts to capture the overall temporal behavior in a parametric model based on the equations that govern the simple harmonic motion of a mass-spring system is presented. By changing a set of easily available parameters, the authors are able to synthesize a variety of laughter. At the call level, the authors relied on a standard linear prediction based analysis-synthesis model. Results of subjective tests to assess the acceptability and naturalness of the synthetic laughter relative to real human laughter samples are presented. © *2007 Acoustical Society of America.* [DOI: 10.1121/1.2390679]

PACS number(s): 43.72.Ja [DOS] Pages: 527–535

I. INTRODUCTION

Expressiveness is a unique quality of natural human speech. The ability to adequately convey and control this key aspect of human speech is a crucial challenge faced in rendering machine-generated speech more generalizable and acceptable. While the primary focus of past efforts in speech synthesis has been on improving intelligibility, and to some extent naturalness, recent trends are increasingly targeting on improving the expressive quality of synthetic speech (Hamza *et al.*, 2004; Junichi Yamagishi and Kobayashi, 2003, 2004; Narayanan and Alwan, 2004). For instance, natural expressive speech quality is essential for synthesizing long exchanges of human-machine dialogs and for information relaying monologs. There are many ingredients that play a role in imparting expressive quality to speech. These include variations in speech intonation and timing (Dutoit, 1997), modifications of spectral properties, appropriate choice of words, use of other nonlexical (expressions such as throat clearing, tongue clicks, lip smacks, laughter, etc.) and nonverbal (physical gestures, facial expression, etc.) cues. Emotion is an important underlying expressive quality of natural speech that is communicated by a combination of the aforementioned variations. Inclusion of nonlexical and/or nonverbal cues in emotional speech can also regulate the type and degree of emotion being expressed and also improve the clarity of emotions in speech. For example, in Robson and MackenzieBeck, 1999 (and references therein) it has been determined that speech with labial spreading (a nonverbal cue) is aurally interpreted as "smiled" or happy sounding speech. Another avenue being explored involves addition of nonlexical cues in machine synthesized speech (Sundaram and Narayanan, 2003) that can better express the desired emotion or the state of a human-machine dialog. Nonlexical cues for expressing happy sounding speech is important in this respect. Prior work has shown (Bulut *et al.*, 2002; Trouvain and Schröder, 2004) that synthesizing happy sounding speech is one of the most challenging problems and that one has to look beyond just intonation variation, especially, if speech accommodates laughter. This is because the implicit nature of laughter causes variations in the supporting speech and vice versa (Nwokah *et al.*, 1999). Laughter may have different functions in interpersonal speech communication, expressing an amusing or happy context is key among them. Hence it is an important attribute in this context of expressive, synthesized speech. The focus of the present work is restricted to automatic acoustic synthesis of laughter by machines. It can be used to enhance the expressive quality of the accommodating synthesized speech and/or aid in communicating a happy or amusing context.

Speech, in humans, is a more controlled and a better understood process that is governed by the rules of a language's grammar. Therefore, for machine synthesis of speech, for example, the phrase "*How are you*?," the required sequence of sounds (as phonemes) and the expected intonational variation are fairly well prescribed, even for different situations and context. Text analysis on the given phrase, and existing intonation models are used in the generation of the final waveform. Also, natural speech audio examples are abundantly available for aiding analysis and modeling. Thus the inputs required to generate any word in

[a)]Abstract previously appeared in the J. Acoust. Soc. Am. 116, 2481 (2004). Part of this paper has been published previously as an invited lay language paper for the 148th ASA Meeting, San Diego, California.
[b)]Electronic mail: shiva.sundaram@usc.edu
[c)]Electronic mail: shri@sipi.usc.edu
[d)]URL: http://sail.usc.edu

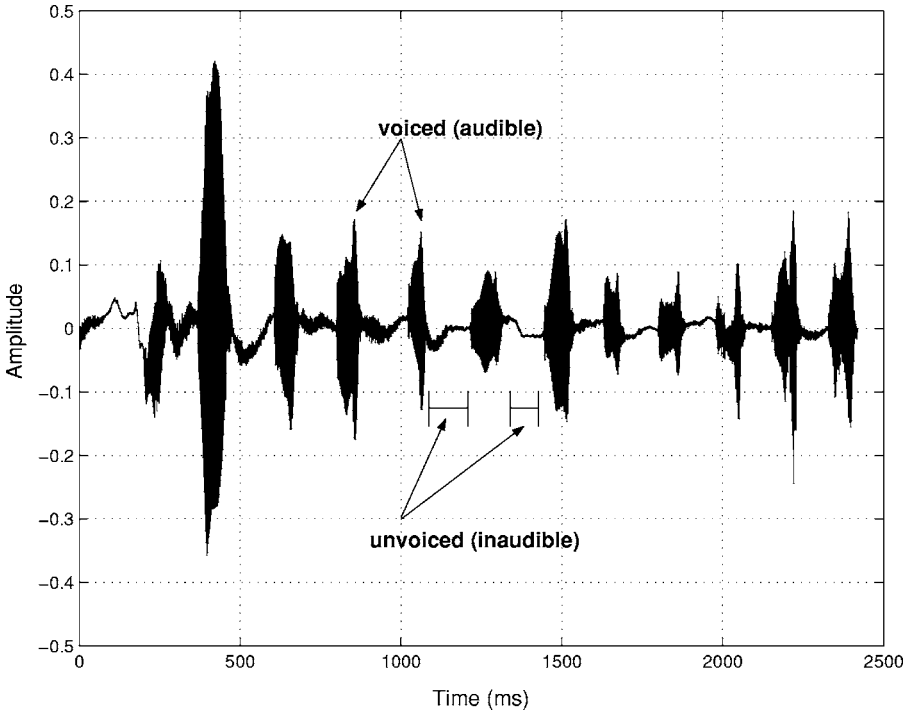

FIG. 1. The alternating phenomenon of a laughter: A laugh cycle with intermittent laugh pulses.

synthesized speech are relatively well defined. Likewise, to synthesize laughter, we require appropriate signal models to generate a particular type of laughter. Unlike spoken language, while there is no guiding grammar for synthesizing laughter, there is a characteristic *texture* to it. Using the simple model proposed in this paper, it is possible to generate laughter using a set of input control parameters, and different types of laughter can be generated by varying these parameters. In this work, we attempt to answer the basic question of *how to synthesize laughter* at the acoustic level; the cognitive/semantic aspects of laughter generation or issues related to including synthesized laughter in conjunction with synthesized speech are beyond the scope of this paper. In the following sections we introduce some common terminologies used to describe laughter and discuss some issues of interest in the synthesis of laughter.

A. Background

We present a brief description of the terms used in this paper to describe the various segments of laughter that have been adopted from acoustic primatology (Bachorowski *et al.*, 2001). In Fig. 1, the waveform of an actual laughter episode is shown. A single instance/episode of laughter, beginning with an inhalation to its end, is known as a laughter *bout*. Each bout comprises alternating voiced (audible) and unvoiced (relatively inaudible) sections. This alternating phenomenon is also termed as a *laugh cycle* with intermittent *laugh pulses* with aspiration sounds in between them (Ruch and Ekman, 2001). The voiced section, illustrated by sections of large amplitude in the figure, is known as a laughter *call* or *laugh pulse* (also referred to as voiced call in this paper). The time interval between two laughter calls is the *inter-call* interval. A laughter call can be a vowel-like sound, for example, in calls such as "ha," or have a grunt-like or snort-like quality (Bachorowski *et al.*, 2001). Such qualities are more evident in a spectrogram of the complete laughter bout. Related details about segmentation of laughter based on its acoustic analysis can be found in Bachorowski *et al.*, 2001; Provine, 2000; and Ruch and Ekman, 2001. Other segmentation schemes are also possible. In Trouvain, 2003 the author discusses syllable and phrase level segments for laughter and its relationship to the terms introduced previously. While it can be useful for studying or categorizing laughter types, these concepts are not directly relevant for the laughter generation model presented here.

B. Variation in laughter and its synthesis

The production of laughter is a highly variable physiological process. Provine (Provine, 2000), describes it to be a very strange expression whose peculiarity is masked by its familiarity. Laughter is an expression with a very distinct pattern. The texture of laughter has variations across gender, and across individuals (Bachorowski *et al.*, 2001; Provine, 2000). Every situation has its own appropriate and inappropriate types of laughter, and even for the same context, an individual can choose to laugh differently at different times. It is used as a vocalized punctuation in a question (Provine, 2000), and it also occurs along with speech (termed as "speech laughs") (Nwokah *et al.*, 1999; Trouvain, 2001). Overall, it can be interpreted as a vocalized expression that bridges the gap between an emotional state of excitement and a neutral emotional state. Laughter differs from smiling because the later is essentially a nonverbal facial expression which, under certain circumstances, may lead to a distinct, audible laughter episode. However, laughter and smiling may share the same facial expression.

While the qualities of the vocalization and issues of duration take their own course during an episode, a limited control by the individual determines the overall duration and number of laugh pulses in a bout or laugh cycle. Thus large

variations in the number of calls per bout and duration of each call is observed in real laughter (Bachorowski *et al.*, 2001). It is known axiomatically that no two instances of laughter are exactly the same, yet they have implicit characteristics that bring out individual traits. Some specific attributes that cause these variations include pitch changes during a bout, pitch changes within a call, duration of the complete bout, duration of a voiced call, the loudness of the calls, and the type of call (vowel-like or grunt-like, etc). Thus, specifications of these components are required to generate an episode of laughter. While the vowel-like sound within a call is a matter of choice, the duration of a bout, duration of each call, and periodicity of the laughter calls are a part of the pattern of the laughter bout. The specifications for the latter are the input control parameters obtained from an appropriately defined generative model for laughter production. From an engineering perspective, a generative model for laughter is challenging because it should meet the following constraints:

- The model should be able to handle a wide range of the variability seen in the physiological process of laughter.
- It should have the provision to generate different types of laughter, e.g., short bursts or a long train of laughter depending on the immediate context. A parametric control over the generated laughter is preferred from an automatic synthesis point of view.
- The model should be convenient to use. It should be able to generate laughter based on simple, easily available information.

The model described in this paper has two major components. The first component focuses on modeling the behavior of the overall episode (or bout), and is based on a simple second-order mass-spring dynamical systems model, akin to one that describes the simple harmonic motion of an oscillating pendulum (refer to Fig. 2). The second component uses a linear prediction (LP) based analysis-synthesis model, which is widely used in speech processing. Note that for this second component, any other speech synthesis/modification technique such as the time domain pitch synchronous overlap add (TD-PSOLA) (Moulines and Charpentier, 1990) may be used. In this work, we restrict our study to the varieties of laughter that the *spontaneous*, i.e., those that are produced without any restraint. The laughter is assumed to always contain vowel-like voiced calls. The rest of this paper provides details of the model of a mass-spring system and how its equations are used to synthesize a bout of laughter. We also present results of subjective tests performed to assess the perceived naturalness of synthesized laughter against real human laughter. It should be noted, however, that assessment of synthetic speech and laughter is a highly challenging task. It is well known that the rich diversity and variability that make up natural speech also make evaluation of machine-generated speech difficult. The study of perception of everyday natural speech spans a very large domain of problems in speech synthesis and other related sciences. The techniques that are available for speech analysis/synthesis tackle only a subset of the rich possibilities in problems of speech generation (Dutoit, 1994; McAulay and Quatieri, 1986; Moulines and Charpentier, 1990). Many of the challenges to achieve even near natural speech for synthesis are yet to be solved for a fair evaluation of natural versus synthesized speech. For example, in Syrdal *et al.*, 1998 the authors subjectively compared two diphone based speech synthesis techniques with natural speech in terms of intelligibility, naturalness, and pleasantness. It was found that natural speech was consistently perceived to be better than synthetic speech. Still, comparison of real, natural oral gestures to machine synthesized ones provides an assessment of the variables that are useful or lacking in mimicking the gesture under study. We follow a similar approach in evaluating the synthetic laughter samples created in this work.

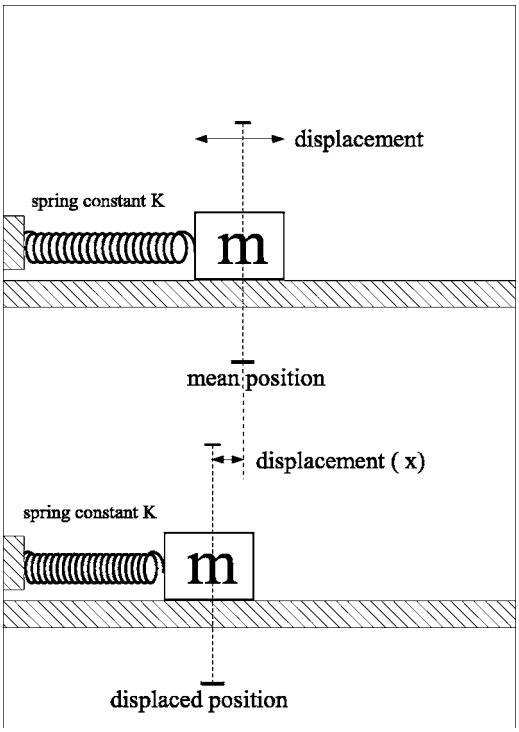

FIG. 2. A mass-spring model.

II. ACOUSTIC MODEL FOR LAUGHTER

An engineering solution to describe an unknown system is to propose a mathematical model based on a set of observations of the system behavior. Figure 1 exemplifies two striking features of a typical laughter bout: alternating segments of audible, voiced section and inaudible unvoiced parts with the envelope of the peaks of the voiced calls falling across the duration of the laughter bout. A laughter starts with a contextual or semantic impulse, that puts the speaker in a *laughing state*. While laughing, there are bursts of air exhalation (along with audible voicing) and aspiration (unvoiced segment) that each last for a short period. This intermittent voicing pattern can be seen as an *oscillatory behavior* that can be observed in most laughter bouts. This pattern has been noted by other researchers as well (Bachorowski *et al.*, 2001; Provine, 2000; Ruch and Ekman, 2001).

We model this oscillatory behavior of alternate voiced and unvoiced segments with equations that describe the simple harmonic motion of a mass attached to the end of a

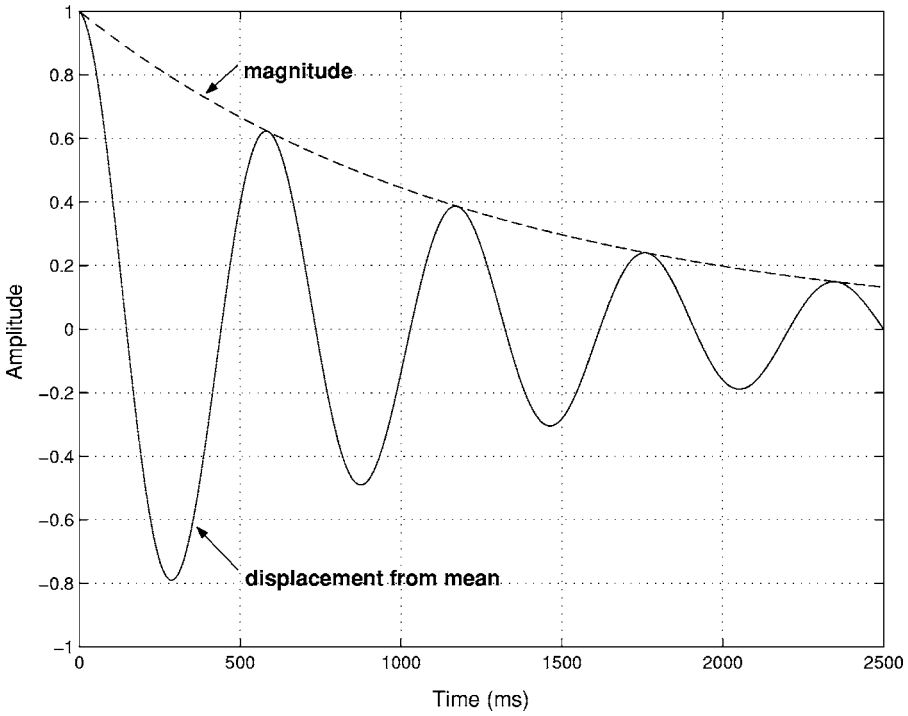

FIG. 3. Damped simple harmonic motion. Amplitude variation over time.

spring (illustrated in Fig. 2). In this simple mass-spring system, the stiffness of the spring and the weight of the mass determine the frequency of oscillation of the mass. The initial displacement and the damping factor determine how long the mass would continue to oscillate, and the rate of envelope decay. We next briefly explain the steps in building a mathematical model for an oscillating mass attached to a spring, and also motivate the idea that this model can be used to explain the oscillatory behavior of laughter for its automatic synthesis.

A. Oscillatory behavior of laughter

Let a mass m be displaced from its initial rest position by x (refer to Fig. 2). This will cause the spring to compress in length by an amount x. When the mass is released at some time t, the compressed spring will act on the mass and accelerate the mass in a direction opposite to the initial displacement. This force that the spring exerts on the mass (denoted by F_{spring}) is directly proportional to the compression x, i.e.,

$$F_{\text{spring}} \propto x.$$

By Newton's First Law, the mass m, its acceleration $a = d^2x/dt^2$ and the force F_{spring} are related by the equation

$$m\frac{d^2x}{dt^2} = -kx, \tag{1}$$

where k is the constant of proportionality, also known as the spring constant. The negative sign on the right hand side of Eq. (1) arises because the direction of force generated by the compressed spring is opposite to the direction of the displacement causing the compression.

A solution to this second-order system is given by the expression

$$x = e^{-j(\sqrt{k/m})t}. \tag{2}$$

This is a sinusoid with $(1/2\pi)\sqrt{k/m}$ as its frequency of oscillation. If this system experiences a damping force proportional to its velocity, then Eq. (1) becomes

$$m\frac{d^2x}{dt^2} = -kx - b\frac{dx}{dt}, \tag{3}$$

where b is the damping constant (this case arises with a simplified damping due to a fluid external to the mass m) and the corresponding general solution for damped simple harmonic motion becomes

$$x(t) = Ae^{-Bt}e^{-j(\sqrt{k/m})t}, \tag{4}$$

where $B = b/2m$. The result obtained in Eq. (4) is that of a damped sinusoid, that parametrically describes the motion of a damped simple harmonic motion system.

Figure 3 illustrates the plot of time t versus amplitude x (solid line) of such a damped sinusoid with Ae^{-Bt} (dotted line). By studying the figure it becomes evident that the peak-amplitude envelope decay of the voiced calls in a laughter bout is similar to a damped sinusoid, but where the parameters A, k, m, B are actually $A(t)$, $k(t)$, $m(t)$, $B(t)$, i.e., they are allowed to vary over time. This is illustrated in Fig. 4 where the plot of a damped sinusoid model is superimposed on a real human laughter sample. Here, the parameters $A(t)$, $k(t)$, $m(t)$, $B(t)$ are allowed to vary as a piecewise linear function of time. On visual inspection, even the duration of the positive cycle of the oscillator (which is directly related to the frequency of oscillation) matches with the duration of the intermittent laugh pulses of the laugh cycle. This is also true for the unvoiced segments of the bouts that match with the duration of the negative cycle of the oscillation. This

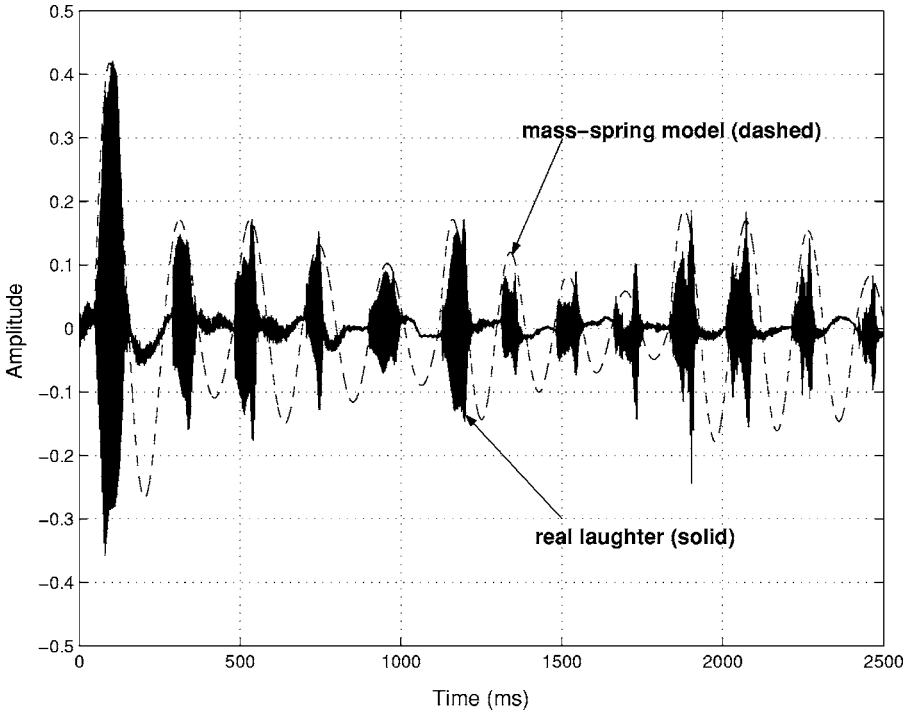

FIG. 4. A mass-spring model trajectory superimposed on a real laughter bout.

aspect is further substantiated by the finding that the call duration and the inter-call intervals are comparable in laughter bouts (Bachorowski *et al.*, 2001).

Other possible variations of the damped oscillator described earlier include forced and damped oscillation where the system is either forced periodically or at random instances during the oscillation of the body. Variations in the nature of the damping force can also cause different oscillatory behavior. One such example is when the damping force is a constant frictional force (Marchewka *et al.*, 2004). In essence, any arbitrarily complex oscillatory behavior can be generated using this basic model, and virtually any pattern of oscillation can be generated by controlling the basic parameters such as $A(t)$, $k(t)$, $m(t)$, $B(t)$.

The other model component relates to the voiced-call units. Since these are vowel-like vocalizations, analysis-synthesis techniques used in conventional speech processing can be directly adopted. Different vowel-like laughter calls can be synthesized by changing the user-defined linear prediction (LP) coefficients or the speech data associated with the synthesizer. The procedure is explained briefly. LP based analysis-synthesis assumes a source-filter model of speech production. The LP coefficients can be extracted (the analysis part) from a sample waveform of speech using standard, well known procedures such as the Levinson-Durbin algorithm (many existing speech analysis software tools have inbuilt LP analysis functions). The estimated LP coefficients define an all-pole filter; and when excited with an appropriate input (such as a pulse train), it can generate (the synthesis) a speech sound at the output (for example, a vowel). Since the set of LP coefficients is primarily dependent on the sample waveform at the time of analysis, different vowel-like sounds for laughter calls can be synthesized by changing the speech data during analysis (essentially using a different set of LP coefficients). Further details about the LP analysis-synthesis techniques can be found in Rabiner and Schafer, 1978.

Thus, one could synthesize segments of voiced calls by using the above duration, time-position and peak-amplitude information, and thereby synthesize a complete laughter bout. By changing the input parameters, laughter bouts with different patterns can be generated. For example, if the damping factor of the previously described system is reduced, then the oscillation will last for a longer duration and thus a longer laughter bout can be synthesized. Similarly, if the values of mass or spring constant are changed, then the frequency of the laughter calls in a bout can be changed. Also, by using different waveform synthesis schemes, other snorting or grunt-like qualities can be imparted to the laughter calls.

The main advantages of this model are summarized below:

- For a given set of parameters, the same model directly presents the duration, timing, and peak-amplitude decay of the laughter calls in an episode of laughter simultaneously.
- By an appropriate choice of parameters such as pitch variation, and choice of $A(t)$, $k(t)$, $m(t)$, $B(t)$ functions, any real human laughter can be accurately represented.
- There is a clear, direct, and predictable relation between the control parameters and the generated pattern of laughter.
- There is no restriction on the speech synthesis technique used for synthesizing the calls in the laughter. Linear prediction (LP) analysis-synthesis method has been used in this work due to its ease of implementation. Other speech modification techniques such as the TD-PSOLA can also be used.

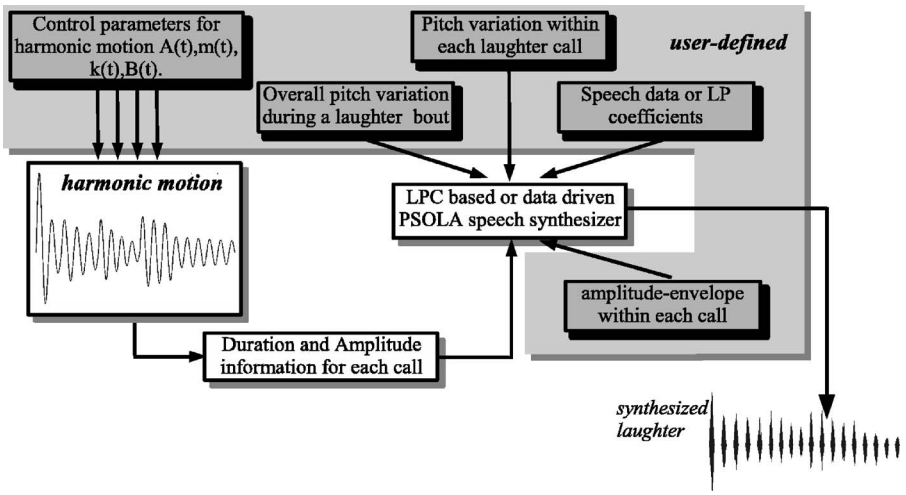

FIG. 5. Steps involved in the synthesis of laughter described in this paper. The shaded boxes depict the inputs required from a user and the unshaded boxes combine to form the laughter synthesizer.

Thus we have a simple *generative* model with user-defined input control parameters such as $A(t)$, $k(t)$, $B(t)$, $m(t)$ which would give us the ability to model and vary the duration of the laughter calls, time position of the calls, and the peak amplitude variation of the calls over the course of a laughter bout. Depending on the synthesizer used, there is also no restriction on the type of laughter call that would be used to generate the complete laughter bout. Figure 5 illustrates the complete laughter synthesis methodology adopted in this paper.

B. Laughter synthesis procedure

Referring to Fig. 5, the procedure followed to synthesize laughter is summarized below:

1. For a set of given (user-defined) time varying functions $A(t)$, $k(t)$, $b(t)$, $m(t)$ calculate the duration and peak amplitude and onset time of each positive cycle in the resulting harmonic motion $x(t)$. Let this harmonic motion comprise N_{pos} positive cycles. Thus we will have N_{pos} laughter calls in the synthesized laughter bout.
2. Let $P_{\text{mean}}(i)$, $p_{\text{var}}(t,i)\,\forall\, i \in \{1,2,\ldots,N_{\text{pos}}\}$ and $t \in [0, T_d(i)]$ be the mean pitch and pitch variation within each call, respectively, where $T_d(i)$ is the duration of the ith laughter call. The values for $P_{\text{mean}}(i)$ and $p_{\text{var}}(t,i)$ are defined by the user. Alternatively, these parameters can also be obtained from acoustic analysis of real laughter clips. Note that $P_{\text{mean}}(i)$ is a discrete positive value of a laughter call pitch and the set $p_{\text{var}}(t,i)$ is a set of functions continuous in time that have positive real values $\forall i$. To have meaningful outputs, the order of $P_{\text{mean}}(i)$ and $p_{\text{var}}(t,i)$ is equal to that of measured F0 values of normal speech. Usually, the exact target values are obtained by analysis of clips of real human laughter.
3. Using the peak-amplitude and duration information obtained in Step 1, in addition to the $P_{\text{mean}}(i)$ and $p_{\text{var}}(t,i)$, synthesize each laughter call $\forall i \in \{1,2,\ldots,N_{\text{pos}}\}$.
4. Similar to Step 1, the duration and peak-amplitude information can also be extracted for the negative cycles of the harmonic motion. This can be used to include audible aspiration noise.
5. Finally, arrange the N_{pos} laughter calls in series in time according to the onset time instances obtained in Step 1 and thus construct the overall laughter bout.

The complete laughter synthesis system described in Fig. 5 was implemented in MATLAB (*http://www.mathworks.com*). The user-defined inputs included the overall variation of pitch in a laughter bout, the pitch variation within each laughter, amplitude envelope within each call, and parameters for the call level synthesis. These were provided to the system by the authors using a graphical user interface (GUI) at runtime. The $A(t)$, $k(t)$, $b(t)$, $m(t)$ values were also provided at runtime. The GUI inputs for the amplitude envelope within each call was low-pass filtered with a third-order finite impulse response low-pass filter to smooth the envelope.

It is important to point out that extracting timing and peak-amplitude information for voiced and unvoiced elements of laughter from the positive and negative cycles, respectively, is a matter of practical convenience. It does not bear any direct relevance to the actual physiology of laughter. However, the work we present here alludes to the fact that real human laughter can be interpreted as a form of oscillation. The voiced calls of a real laughter episode are not truly vowel-like sounds. In a real laughter episode, the vocal tract configuration can change rapidly and/or other noise (such as aspiration noise) is always present. Therefore, to get satisfactory synthesis quality, the data for the LP parameters for the call level waveform synthesis were extracted from voiced segments of normal speech. Samples of synthetic laughter can be found at http://sail.usc.edu/emotion.

The next section describes the subjective experiment to assess the perceived naturalness of the synthesized laughter.

III. EXPERIMENT

Subjective evaluation tests were performed on 28 naive volunteers at the Speech Analysis and Interpretation Laboratory (SAIL) at USC. The volunteers were presented with 25 laughter-only clips of which 17 clips were synthesized offline using the technique presented here. The number of calls in the synthesized laughter, its duration, the F0 changes in

TABLE I. A summary of the properties of the 17 synthesized and eight real samples of laughter used in the listening experiments. The eight real laughter samples are marked with an (*).

Sample	mean F0 (Hz)	min F0 (Hz)	max F0 (Hz)	Std F0 (Hz)	No. calls	Duration (s)	Gender
01	203.30	178.0	235.0	17.00	12	2.30	M
02	246.75	175.0	305.0	45.15	9	1.60	M
03	224.60	202.0	259.0	21.05	5	0.84	M
04	276.50	223.0	372.0	53.64	10	1.60	F
05	244.50	165.0	299.0	32.55	18	2.90	F
06	274.00	211.0	350.0	55.94	7	1.00	F
07	190.00	142.0	231.0	26.81	15	2.50	M
08	354.00	286.0	417.0	39.00	8	1.32	M
09	218.66	202.0	235.0	16.50	3	0.43	M
10	365.00	293.0	425.0	15.00	6	0.97	M
11	198.00	171.0	221.0	17.75	8	1.18	M
12	333.40	302.0	362.0	20.48	12	2.45	F
13	267.20	211.0	317.0	46.85	5	0.82	F
14	250.00	192.0	312.0	41.87	6	0.93	M
15	279.35	223.0	419.0	49.92	18	3.02	F
16	225.00	139.0	319.0	51.70	16	2.39	F
17	340.50	304.0	393.0	29.75	11	1.90	F
01*	204.17	85.5	331.5	74.15	11	2.06	M
02*	199.60	111.7	315.5	52.64	17	3.15	M
03*	309.72	234.2	364.5	26.70	10	2.20	F
04*	174.14	81.1	250.4	46.93	10	2.00	F
05*	243.85	142.3	314.4	39.21	17	2.81	F
06*	257.50	161.6	332.7	56.57	18	3.67	F
07*	180.81	111.4	294.2	30.08	12	2.02	M
08*	255.90	192.6	358.9	44.36	14	3.02	M

each sample are given in Table I. The remaining eight were clips of real human laughter. The laughter type in these eight clips matched the 17 clips of synthesized laughter. The 25 clips were randomly played and not grouped in any particular order. The tests were performed in a typical quiet office environment on a computer terminal. Each volunteer had to listen and score each sample for *naturalness* and *acceptability* according to their preference on a scale of 1–5: 1-Very Poor, 2-Poor, 3-Average, 4-Good, 5-Excellent. The samples were presented on an interactive webpage-like GUI. The subject could click on a sample, listen to it and click on the appropriate *naturalness* and *acceptance* score. The samples were played at 22,050 Hz sample rate over a pair of commercially available Sony MDR-XD100 headphones that could be adjusted to snugly fit the listener. The complete evaluation took about 11 min for each subject.

The eight clips of real isolated laughter were collected from two sources: four were extracted from a compact disk and downsampled to 22,050 Hz (Junkins, 2005). These were from tracks of recorded laughter intended for laughter therapy and the remaining four were obtained from a database of laughter episodes that we recorded independently. This database was created by recording volunteer subjects who were simply asked to laugh impromptu for a laughter synthesis project. For each of the subjects the first few laughter instances seemed to be forced and were rejected and the later episodes that seemed more natural to us were kept in the database. The eight clips were selected based on how well they aurally matched the synthesized laughter used in listening tests. Many candidate clips were rejected because of speaker's movement while laughing, unidentified noises picked up by the microphone, and change of laughter call type during the bout. To make all the tracks similar, and reduce any extraneous bias in listener assessment, noise extracted from the silent parts of the compact disc tracks were extracted, downsampled to 22,050 Hz sample rate, and added to the other synthesized and recorded clips.

IV. RESULTS

At the time of analysis of the results the evaluations of the volunteers were grouped into Group I and Group II according to their language background. Group I comprised four female and five male subjects whose first language was American English and Group II comprised seven female and twelve male subjects whose second or third language was English. For the analysis of the evaluations, we make the assumption that each laughter clip is an independent encounter by an individual subject. Thus, for $N=28$ subjects and 17 synthesized samples, we have a total of $28\times17=476$ samples and for the eight real laughter clips, we have $28\times8=224$ samples.

The mean and variance of the evaluation scores are listed in Table II. The evaluation results are summarized below:

Mean evaluation scores: A t Test (with unequal variance, and degrees of freedom (df)$=440$) was performed to compare the evaluation scores of real and synthesized laughter clips. For the given experiment it was found that at $\alpha=10^{-4}$, there is a significant difference in the mean natural-

TABLE II. Mean evaluation scores for natural and synthesized clips

Group	Synthesized clips *mean, variance*	Real clips *mean, variance*
Groups I evaluations	naturalness: 1.49, 0.044	naturalness: 4.36, 0.014
	acceptability: 1.66, 0.053	acceptability: 4.34, 0.079
Groups II evaluations	naturalness: 1.81, 0.084	naturalness: 4.38, 0.034
	acceptability: 2.21, 0.077	acceptability: 4.36, 0.014
Groups I & II evaluations	naturalness: 1.71, 0.600	naturalness: 4.28, 0.59
(overall)	acceptability: 2.03, 1.020	acceptability: 4.35, 0.73

ness scores between the real and synthesized laughter clips. It was also found that at $\alpha=10^{-4}$, there is a significant difference in the mean acceptability scores of real and synthesized laughter clips.

Sample-wise test: A parametric single-factor analysis of variance (ANOVA) was performed to determine differences in the evaluation among the synthesized clips. For a total of $N=28$ evaluations for each clip, it was found that at ($\alpha=0.05$; $df=16/459$) there was a significant difference in the mean naturalness scores among the synthesized laughter clips. However, at $\alpha=10^{-4}$, there was no significant difference in the mean naturalness score. A single-factor ANOVA performed on the acceptability scores among the synthesized clips indicated that for ($\alpha=10^{-4}$; $df=16/459$) there was no significant difference in the mean scores among the synthesized laughter clips. A single-factor ANOVA of the mean naturalness score ($\alpha=10^{-4}df=7/216$) for the real laughter clips showed no significant differences among the evaluations of the real laughter clips. The single-factor ANOVA of the mean acceptability scores ($\alpha=10^{-4}$, $df=7/216$) also indicated no significant differences among the evaluations of the real-laughter clips.

Group-wise test: A parametric single-factor ANOVA test of the naturalness scores of real laughter clips between Group I and II indicated a significant difference in the scores at ($\alpha=0.05$; $df=1/223$). However, at ($\alpha=10^{-4}$; $df=1/223$), the test indicated no significant differences in the mean evaluation scores between the groups. This same trend was observed when the mean evaluation acceptability scores for synthesized clips were compared for Group I and II. The same ANOVA test ($\alpha=10^{-4}$; $df=1/223$) performed on the mean acceptability scores indicated no significant difference in the mean scores between the two groups for the synthesized clips. However, for synthesized laughter clips it indicated significant difference between Group I and II at ($\alpha=0.05$; $df=1/223$) and at ($\alpha=10^{-4}$; $df=1/223$).

V. DISCUSSION, CONCLUSION AND FUTURE WORK

In this paper we have presented a two-level parametric model for human laughter. The first level of the model captures the overall temporal behavior of a laughter episode. At the next level, we model the audible calls with conventional LP coefficients based analysis-synthesis and/or TD-PSOLA speech modification technique that are widely used in speech processing. The model presented is based on the idea that laughter in human beings can be interpreted as an oscillation, where exhalation alternates with inaudible segments. We also presented properties of laughter episodes that can be captured by the model parameters. Motivated by the need for computer synthesis of laughter for emotional speech synthesis, we applied this idea to synthesize different varieties of laughter and evaluate them in terms of two subjective measures: *naturalness* and *acceptability*. We also compared this evaluation with evaluation of real laughter clips.

The results obtained are similar to the results obtained by Syrdal *et al.*, 1998 where different speech synthesis techniques were evaluated against natural speech. Subjective assessment of real, human, natural expressions is consistently better than synthesized ones. Two main factors that can be attributed to this dichotomous result are the limitations in the variety of features that are included in synthesized laughter and the inherent artifacts present during the final waveform synthesis. For example, unlike natural laughter bouts, we synthesize bouts with relatively simple vowel-like voiced calls. Also, the perceivable artifacts during waveform synthesis are caused due to issues with precise generation of natural sounding pitch contours, and obtaining smooth frame to frame spectral variations. These artifacts give negative cues to the listener that result in unnatural perception of synthesized clips. Another issue deals with the underlying quality that is being evaluated: perceived naturalness. While naturalness is a loose term, a very stringent set of standards is followed to label perceived speech as natural. What is truly regarded as natural and/or acceptable is already encoded in the listener. This is because everyday human speech communication is perceived as highly natural speech and it is abundant with a wide range of qualities that are not imparted in synthesized speech. For the particular case of laughter, due to its high degree of variability, the evaluation in terms of perceived naturalness becomes a bigger issue. It is also difficult to define a quantitative measure or a quantitative set of parameters to define an *acceptable* form of laughter. This is because, such a measure covers a gamut of social, acoustical, and perceptual metrics. Even in the case of real human laughter, for example, if the bout is spontaneous and placed appropriately in a dialog, it is *more* natural and acceptable than when its forced and/or inappropriate. Thus it is difficult to make raw comparisons. The results of the experiments also indicate that synthesized laughter is interpreted differently by different individuals. The evaluation experiments presented here are very limited in scope: they evaluate results of isolated laughter episodes without context or accompanying speech. In attempting to answer the question "*What makes laughter laughter*?" the research presented in this pa-

per sheds a different light on this question by generating laughter than the ones addressed by other researchers through acoustic analysis. The experiments have been designed to evaluate only the synthesis aspects of laughter and its perception.

Computer synthesis of laughter is primarily for expressive speech synthesis, a challenge currently being addressed in the speech synthesis our simple approach appears promising, much remains to be done in integrating laughter within an overall synthesis system. We would like to extend this work to include laughter in synthesized happy speech. To merge laughter and speech requires appropriate prosodic and intonational modifications to the accompanying speech and appropriate choice of words and context tracking. This is a harder problem and part of our future goals. The proposed model can also be incorporated with audio-visual synthesis such as with computer generated *avatars* and other virtual agent technologies. Such an effort would entail combining the acoustic aspects of synthesis with visual gestures such as movement of the lips, face and head. These efforts are topics of our ongoing and future work.

ACKNOWLEDGMENT

The work reported in this paper was supported in part by grants from the NSF and the U.S. Army.

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

Bulut, M., Narayanan, S., and Syrdal, A. (**2002**). "Expressive speech synthesis using a concatenative synthesizer," in *Proceedings of the Seventh International Conference on Speech and Language Processing (ICSLP)*, Denver, pp. 1265–1268.

Dutoit, T. (**1994**). "High quality text-to-speech synthesis: A comparison of four candidate algorithms," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1, pp. 565–568.

Dutoit, T. (**1997**). "An Introduction to text-to-speech Synthesis" (Kluwer, Dordrecht).

Hamza, W., Bakis, R., Eide, E. M., Picheny, M. A., and Pitrelli, J. F. (**2004**). "The IBM expressive speech synthesis system," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Jeju, South Korea, pp. 2577–2580.

Junichi Yamagishi, T. M., and Kobayashi, T. (**2003**). "Modeling of various speaking styles and emotions for HMM-based speech synthesis," in *Proceedings of Eurospeech, Geneva, Switzerland*, pp. 2461–2464.

Junichi Yamagishi, T. M., and Kobayashi, T. (**2004**). "HMM-based expressive speech synthesis-towards TTS with arbitrary speaking styles and emotions," Special workshop in Maui (SWIM).

Junkins, E. (**2005**). "Lots of laughter," http://www.laughtertherapy.com, 3200 N. MacArthur Blvd., Ste. 106, Irving, TX 75062. Last accessed 12/6/06.

Marchewka, A., Abbot, D. S., and Beichner, R. J. (**2004**). "Oscillator damped by a constant-magnitude friction force," Am. J. Phys. **74**(4), 477–483.

McAulay, R. J., and Quatieri, T. F. (**1986**). "Speech analysis synthesis based on sinusoidal representation," IEEE Trans. Acoust., Speech, Signal Process. **34**, 744–754.

Moulines, E., and Charpentier, F. (**1990**). "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," Speech Commun. **9**, 453–467.

Narayanan, S., and Alwan, A. (**2004**). *Text-To-Speech Synthesis: New Paradigms and Advances* (Prentice–Hall, Englewood Cliffs, NJ).

Nwokah, E. E., Hsu, H.-C., Davies, P., and Fogel, A. (**1999**). "The integration of laughter and speech in vocal communication: A dynamic systems perspective," J. Speech Lang. Hear. Res. **42**, 880–894.

Provine, R. R. (**2000**). *Laughter: A Scientific Investigation* (Viking, New York).

Rabiner, L. R., and Schafer, R. W. (**1978**). *Digital processing of speech signals*, Prentice-Hall Signal Processing Series (Prentice–Hall, Englewood Cliffs, NJ).

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

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

Sundaram, S., and Narayanan, S. (**2003**). "An empirical text transformation method for spontaneous speech synthesizers," in *Proceedings of EUROSPEECH*, Geneva, Switzerland, pp. 1221–1224.

Syrdal, A., Stylianou, Y., Garrison, L., Conkie, A., and Schroeter, J. (**1998**). "TD-PSOLA versus harmonic plus noise model (HNM) in diphone based speech synthesis," *IEEE International Conference on Acoustics, Speech and Signal Processing*, Seattle, WA, pp. 273–276.

Trouvain, J. (**2001**). "Phonetic Aspects of 'Speech-Laughs'," in *Proceedings of Conference on Orality and Gestuality (ORAGE)*, Aix-en-Provence, France, pp. 634–639.

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

Trouvain, J., and Schröder, M. (**2004**). "How (not) to add laughter to synthetic speech," in *Proceedings of the Workshop on Affective Dialogue Systems*, Kloster Irsee, Germany, pp. 229–232.

Copyright of Journal of the Acoustical Society of America is the property of American Institute of Physics and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.