IMITATING CONVERSATIONAL LAUGHTER WITH AN ARTICULATORY SPEECH SYNTHESIZER

Eva Lasarcyk, Jürgen Trouvain

Institute of Phonetics, Saarland University, Germany
{evaly|trouvain}@coli.uni-saarland.de

ABSTRACT

In this study we present initial efforts to model laughter with an articulatory speech synthesizer. We aimed at imitating a real laugh taken from a spontaneous speech database and created several synthetic versions of it using articulatory synthesis and diphone synthesis. In modeling laughter with articulatory synthesis, we also approximated features like breathing noises that do not normally occur in speech.

Evaluation with respect to the perceived degree of naturalness indicated that the laugh stimuli would pass as "laughs" in an appropriate conversational context. In isolation, though, significant differences could be measured with regard to the degree of variation (durational patterning, fundamental frequency, intensity) within each laugh.

Keywords: Laughter synthesis, articulatory synthesis, synthetic laughter evaluation.

1. INTRODUCTION

Enriching synthetic speech with paralinguistic information including non-verbal vocalizations such as laughter is one of the important challenges in current speech synthesis research. The modeling of laughter has been attempted for concatenative synthesis [4, 12] and formant synthesis [10].

We present an initial study to find out whether articulatory synthesis is a viable alternative. To this end, we analyze the articulation of laughter and create three synthetic laughs on the basis of this analysis. The synthetic laughs differ with respect to degree of variation and with respect to the synthesis method used (see also Sec. 1.2).

The second goal of this study is to investigate if the variation of the details of a laugh (e.g. fundamental frequency, intensity, durational patterning) increases the degree of perceived naturalness of the laugh.

We present a perceptual evaluation that tested, firstly, whether our laugh imitations are "good" enough to pass as a laugh in conversation, and, secondly, to find out whether the rating in naturalness improves when we put more variation into the modeling of a laugh.

1.1. Laughter

A laugh as a whole is very rich in variation and very complex. There are, however, attempts (see e.g. [11] for an overview) to categorize different types of laughter. Bachorowski *et al.* [1] for example introduced three types of human laughs: *song-like*, *snort-like*, and *unvoiced grunt-like*. We will concentrate on the song-like type, "consisting primarily of voiced sounds", including "comparatively stereotyped episodes of multiple vowel-like sounds with evident F_0 modulation …" (p. 1583).

Categorizations have to focus on high level descriptions but authors emphasize at the same time that laughter is *not* a stereotypical sequence of laugh sounds [1, 5]. In [5], Kipper and Todt state that acoustic features like fundamental frequency (F_0), intensity, and tempo (durational pattern) as well as their changing nature "seem to be crucial for the identification and evaluation" of a laugh (p. 256).

Regarding re-synthesized human laughs, Kipper and Todt [5] found that stimuli were rated most positively when they contained varying acoustic parameters (p. 267), which in their case were the durational pattern (rhythm) and the fundamental frequency (pitch).

While a laugh event *itself* can be described, one has to take into account that laughter naturally occurs in a phonetic *context*. The preceding stretch of speech of the laughing person *himself/herself* influences the characteristics of the laugh. It is important, for instance, to match the degree of intensity of the laugh with its phonetic context [12]. Otherwise a laugh would be easily perceived as inappropriate. The phonetic context can also be the utterances of the dialog *partner* where a too intense laugh would be equally inappropriate.

1.2. Synthesis

We used two different synthesis programs to synthesize our laugh samples. One of them was an articulatory speech synthesis system [3], the other one was a diphone synthesis system [8] (see Sec. 3.1 and 3.2). However, the main emphasis was put on the use of the articulatory system. Since the diphone system draws its speech material from prerecorded regular speech (excluding laughs etc.), it obviously cannot be as flexible as a synthesizer that simulates the whole production process. It was mainly used here to delineate the possible advantages (or disadvantages) of the articulatory system.

2. ANALYSIS

2.1. Database

We intended to synthesize a *detailed* laugh and therefore decided to *imitate* natural models of laughter events of spontaneous conversations with overlapping speech. We used a corpus where the two speakers of a dialog were recorded on different audio channels simultaneously [6]. The selected conversation by two male speakers contained 13 situations with one or more laughs and we focused on the *song-like* type of laugh.

2.2. Features of the laugh

Descriptions in [1] concentrate primarily on what we will be calling the *main part* of the laugh (see below). While their definition is plausible for some research questions, we wish to extend the definition of a laugh to include breathing and pausing. Audible breathing can often be observed, framing the *main part* and *pause* of a laugh in the corpus. Since the articulatory synthesizer should be able to generate breath noises, we take this feature into account.

The following structure is thus proposed for the laughs analyzed and imitated in this study:

- an *onset* (an audible forced exhalation [7]),
- a *main part* with laugh syllables, each containing a voiced and an unvoiced portion,
- a *pause*, and
- the *offset*, consisting of at least one audible deep inhalation.

To see a human laugh labeled according to these four phases please refer to image file 1 (top).

In order to re-synthesize the laugh, the following items were specified:

- duration of the *onset*, each laugh syllable in the *main part*, the *pause*, and the *offset*,
- intensity contour of the whole laugh,
- fundamental frequency contour of the laugh,
- vowel quality of the voiced parts.

2.3. Overall results of the analysis

Image file 2 (a) shows a colored screenshot (using the software in [9]) of an oscillogram and a spectrogram of a human laugh from the corpus used here (cf. audio file 1). F_0 and intensity contours are visible in the colored spectrogram (blue and yellow lines.)

The temporal succession of elements can be seen as labels in image file 1: The first element of the laugh (*onset*) is an audible exhalation. This is followed in a *main part* by several laugh syllables of decreasing overall intensity and increasing overall length. Within a laugh syllable, an energy-rich portion (voiced) is followed by a breathy portion (unvoiced), later on with faint sounds in between. The *main part* is followed by a *pause*. The last element of the laugh (*offset*) is a forced inhalation to compensate for the low lung volume.

2.4. Some physiological details

The following physiological and articulatory aspects are important for the control of the articulatory synthesizer.

2.4.1. Subglottal pressure

Luschei et al. [7] state that “laughter generally takes place when the lung volume is low” (p. 442). Nevertheless, the tracheal pressure during laughs can reach peaks of around 1.8 to 3.0 kPa (p. 446), which is higher than the level typical of speech.

2.4.2. Vowel quality

The vowel quality of the voiced portion of a laugh syllable must be defined. Bickley and Hunnicutt [2] found that the formant patterns “do not appear to correspond to a standard … vowel” (p. 929) of the laughers’ mother tongue but do fall into the normal range of speakers’ formant values. Bacharowski *et al.* [1] found that their recorded laughs generally contained “central, unarticulated sounds” (p. 1594).

3. SYNTHESIS

To imitate the human laugh, we used two different synthesis systems both of which have their merits.

3.1. Articulatory synthesis

One system was the articulatory synthesis system described in [3]. The speech output is generated from a gestural score (containing several *tiers*, as can be seen in image file 1) via an aerodynamic-acoustic simulation of airflow through a 3D model of the vocal tract. This allows for a high degree of freedom and control over a number of parameters including subglottal pressure, vocal tract shapes, and different glottal settings. With this type of synthesis it is thus also possible, in principle, to create breathing noise and freely approximate virtually any vowel quality needed.

3.2. Diphone synthesis

The second system was the diphone system MARY [8]. Speech is generated by choosing, manipulating and concatenating appropriate units from a corpus of prerecorded and segmented natural speech. The output is thus based on natural human speech. Since the set of sounds is limited by the corpus, it is not possible to imitate the breathing portions of the laugh, and for the laugh syllables only the predefined phones are available.

3.3. Imitating laughter in different versions

In the following section, we describe the generation of the three different imitations of the human laugh (*version H*) shown in image file 2a.

3.3.1. Version V

Of all three synthetic versions, *version V* (image file 2b, audio file 2) contained the highest degree of variation within the laugh in terms of durational patterning, intensity and F_0 contours. The duration of each of the phases and of each laugh syllable within the *main part* was copied from the human laugh sample. Intensity and F_0 movements (yellow and blue lines in the image) were also modeled in a way to match the human ones as closely as possible.

In each laugh syllable in the *main part*, voiced and unvoiced portions alternate. To reflect this basic pattern of vocalization, glottal gestures were placed alternately on the glottal gesture tier in the gestural score (see bottom of image file 1). An “open” gesture corresponds to the unvoiced portion of a laugh syllable, a “close” gesture to the voiced portion (“laugh vowel” [11]). The duration of each gesture was copied from the durational patterning of the human laugh.

To get the appropriate vowel quality in the *main part*, a vowel gesture was placed on the vocalic tier so that when the glottis is ready for phonation, a laugh vowel would be articulated. We approximated the speaker in our sample laugh by using an [ɛ] on the vocalic tier.

In order to model the different levels of intensity within the *main part*, we varied the degree of lung pressure by using different gestures on the pulmonic pressure tier (bottom tier).

The overall (long-term) F_0 contour was modeled with appropriate gestures on the F_0 phrase tier. F_0 accent gestures were used to imitate the (short-term) fundamental frequency contour within one laugh syllable.

Since the kind of laugh imitated here also contains two breathing phases (*onset* and *offset*), we put gestures of generally high lung pressure on the pulmonic tier and gestures of a widely abducted position of the vocal folds (“open”) on the glottal tier. The result was, however, a long way from the original level of intensity. Thus, an additional source of friction was introduced on the consonantal tier (“E:_Pharynx”). This implies a constriction area in the pharynx, and was motivated by introspection, analogous to constrictions in grunt-like laughs [1]. The result was a clearly audible friction noise.

3.3.2. Version S

The second imitation created with the articulatory synthesizer was *version S* (cf. audio file 3, image file 2c). It contained *less* variation in durational patterning, intensity, and fundamental frequency in the *main part*.

The gestural score for this version was constructed by taking *version V* and deleting all the (variation-rich) *main part* gestures except for the gestures of the first laugh syllable. The gap was then filled by repeating the block of gestures for the first laugh syllable until this laugh imitation contained the same number of laugh syllables as the human one and *version V*. *Version S* was thus a more stereotypical imitation than *version V*.

3.3.3. Version D

Due to the inherent phone set restrictions of a diphone synthesis system, the diphone *version D* (audio file 4, image file 2d) was generated without the breathing phases (*onset* and *offset*). As a consequence, the phase containing the pause would become obsolete since no signal followed. The *main part* of *version D* was produced by alternating the phones [ɛ] and [h], which seemed to resemble best the unvoiced and voiced portions of each laugh syllable. The durational pattern was, as in *version V*, adopted from the human laugh.

The fundamental frequency contour was approximated by specifying a target frequency value for each of the [ɛ] and [h] segments. We did not have explicit control over intensity values.

4. PERCEPTUAL EVALUATION

We carried out two perception experiments to get ratings of how natural the laughs would be perceived. In the first experiment, the laughs were integrated in a dialog, whereas in the second experiment, they were presented in isolation.

4.1. Stimuli

For the first experiment, the aim was to keep the verbal interaction presented as natural as possible by placing the synthesized laugh versions at exactly the same location as the original (human) one. Audio file 5 contains the dialog in its original version, 6 to 8 with laugh *versions V*, *S*, and *D*, respectively. The dialog structure of the stimuli was always identical: Person 1 speaks and laughs, directly afterwards, about his own statement; person 2 joins in. In one stimulus, this laugh of person 2 is human (original, version *H*), the other three each contain one of the synthetic laughs.

For the second experiment, each of these four laughs (one human, three synthetic) was prepared to be presented in isolation by cutting it out of the conversational context. The aim of presenting them in this isolated way was to allow for a more direct focus on the laugh itself in order to asses its intrinsic naturalness. The human laugh (audio file 1) obviously contained the highest degree of variation, *version V* a mid-high degree of variation, and *versions S* and *D* contained less variation (regarding durational patterns, intensity and fundamental frequency).

4.2. Experimental setup and participants

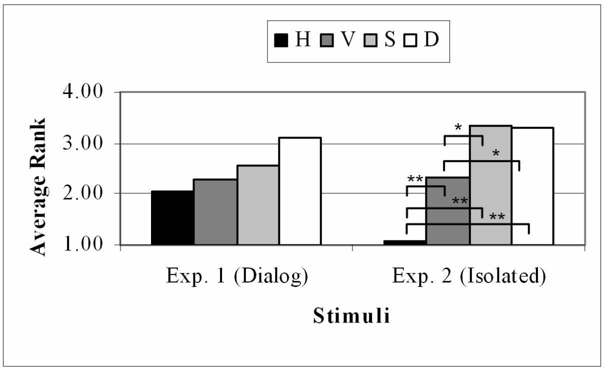
The experiments were conducted together, one immediately after the other. All participants (14 in total, 8 female, 6 male, with an average age of 25 years) participated in both sessions. The audio material was presented to each person individually via loudspeakers in a separate randomized order for each participant to minimize order effects. The participants were asked to rate each stimulus with respect to naturalness on a scale of 1 to 4: 1 “natural”, 2 “less natural”, 3 “rather unnatural”, and 4 “unnatural”. Thus, in experiment 1, they were asked to give their overall impression of how natural they found the dialog in total. In experiment 2, they were asked to rate the naturalness of the laugh stimulus by itself.

For both experiments, we calculated the average ranks of each stimulus (dialog or laugh). A non-parametric Friedman test (significance threshold 5 %) was applied to ascertain significant effects of laugh type within an experiment.

For experiment 1, the null hypothesis was: There is no dependency between the rating of a dialog and the laugh stimulus placed in the dialog. The alternative hypothesis was: The rating of the dialogs depends on which laugh stimulus is placed into them.

For experiment 2, the null hypothesis was: There is no dependency between the rating of an isolated laugh and its degree of internal variation. The alternative hypothesis was: The rating of an isolated laugh depends on how rich its internal variation is.

Fig. 1 Average ranks regarding naturalness in experiments 1 and 2. Bars between pairs mark significant differences of $p < 0.0083$ (*) and $p < 0.001$ (**). Properties of the stimuli *H*, *V*, *S*, and *D* are explained in Sec. 3.3.



4.3. Results

Fig. 1 shows the average ranks of the ratings for experiment 1 (left) and 2 (right).

The dialog stimuli of experiment 1 were ranked in the following order: *H* (average rank of 2.07), *V* (2.29), *S* (2.54), and *D* (3.11). It has to be added that the ratings for this experiment did *not* differ significantly.

For experiment 2, the order of the stimuli was similar, only the last two were reversed: *H* (1.07), *V* (2.32), *D* (3.29), and *S* (3.32). In this experiment, the ratings differed significantly. Thus, we conducted post-hoc pair-wise comparison tests (Wilcoxon) to determine which versions differed significantly from one another. The 5 % significance threshold was corrected to 0.83 % since we had 6 pairs to compare.

We found a significant difference between all the pairs except between stimuli *S* and *D*. *H* was ranked as significantly more natural than *V*, *S*, and *D* ($p < 0.001$). *V* was ranked as significantly more natural than *S* ($p = 0.002$) and *D* ($p = 0.008$). The more natural rating of *D* with respect to *S* was *not* significant ($p = 0.688$).

5. DISCUSSION

5.1. Experiment 1

The outcome of experiment 1 might indicate that all synthetic laughs are "good enough" to pass as laughter in the dialog.

This is especially noteworthy with respect to the laugh *version D*: It was created with a diphone synthesis system that can assemble a laugh only from regular speech sounds. This may in a way support the indication Bickley and Hunnicutt found in their study [2] that "in some ways laughter is speech-like" (p. 930) since they found *similar* measurements of the temporal and spectral characteristics of their *laughs* to what they found in *speech*. It may also indicate that the natural (human) origin of the diphones to a certain extent counterbalances the purely synthetic voice of the articulatory system. Sounding more natural *per se* may be advantageous; another issue is the degree of flexibility (discussed below in Sec. 5.2).

It can be argued, though, that the context chosen here was *masking* the target laugh too much, and that the major part of the dialog was made up of unprocessed natural speech/laughing. This can be seen as an "advantage", yielding relatively high values of naturalness. Nevertheless, it was a real-life context, and joint laughter of two speakers is presumably not uncommon [12]. Still the question arises: What other context would be better suited to the test?

Another point of discussion is the fact that, in experiment 1, the participants were asked to rate the naturalness of the dialog *as a whole*. Our initial intention had been to compare a laugh *within a dialog* with a laugh *in isolation*. In order to do this, it might have been possible to address the laugh item in the dialog directly, when giving the instructions, and in this way create a bias in the expectation of the listener. However, we did not want to influence the participants before they heard the dialog by saying that it contained laughter. Thus, we could not compare the ratings directly with those of experiment 2.

5.2. Experiment 2

The results of experiment 2 indicate, firstly, that all synthetic versions are perceived as much less natural than the natural version. This can be expected, since natural speech introduces an extremely high standard and laughs in particular can be very complex. Furthermore, the synthetic stimuli created here were an initial approach to modeling laughter.

Secondly, while *all* synthetic stimuli in our experiments seemed "good enough" to pass as laughter in speaker-overlapping context, presenting them in isolation brought to light that there are differences in perceived naturalness with regard to the *variation* within a laugh. The significantly better (i.e. more natural) ranking of *version V* suggests that, in principle, it should be possible to improve perceived naturalness by putting more details and variations into a laugh stimulus. This result may be seen as confirmation of previous findings; see e.g. the overview and study in [5] which concludes that variation within a laugh is important for its evaluation.

It can be argued, though, that the *version D* and *version S* laughs sound rather simple and in consequence, the better rating of *version V* should not come as a surprise. The stimuli *D* and *S* were meant to be reasonable initial imitations of the human laugh, though with less variation than *V*. Some features were impossible to model in the diphone synthesis system, such as the breathing noise, the selection of the "laugh vowel", or the lack of intensity control. Other features were deliberately generated in a less varied way (such as the durational pattern, fundamental frequency, and intensity in *version S*). Maybe a more fine-grained scale of variation could be designed and implemented in laugh stimuli synthesis in the future.

Another dimension in the discussion is whether articulatory synthesis provides any advantages when imitating laughter. In general, synthesizing laughter "from scratch" in an articulatorily transparent way seems quite promising, the reason being that with the different gestures one could model the articulation processes quite directly – we have to note, though, that the gestural solutions used here do not necessarily mirror correctly what humans do when producing laughter. The results of experiment 1 might only indicate that this is *one* way of doing it.

Apart from the advantage of modeling gestures directly, we also noted limitations to the current articulatory approach. The first is of a more technical nature. E.g. the current limit of 1 kPa to the pulmonic pressure is appropriate for speech but seemingly not high enough for laughter. In this case we compensated by introducing the *ad hoc* constriction in the pharynx in order to achieve the desired level of friction noise. This choice might not reflect accurately what really happens during laughter.

The second kind of limitation stems from our limited knowledge of some aspects of laughing. We need to know exactly what sort of excitation there is at the glottis. When modeling singing, we add tremolo to the voice; what could be the adequate or necessary additions to the regular source signal when modeling laughter?

6. CONCLUSIONS

Imitating human laughter in conversation proves to be a challenging field when it comes down to modeling the articulatory aspects of laughter, not all of which are known in full detail yet. The general approach seems promising and the perceptual tests conducted suggest that the articulatory synthesizer used for our stimuli is indeed capable of producing purely synthetic laugh-like sounds with varying degrees of variation.

It therefore presents a viable alternative to other forms of parametric laughter synthesis like formant synthesis [10]. In contrast to concatenative synthesis, more room for improvement and fine-tuning exists.

In concatenative systems, the continuum of possible variation is limited. A *regular* diphone synthesis system, for example, relies on speech sounds only. Thus, only (stylized) “haha” laughs are possible, restricting the set of possible variations to fundamental frequency, duration, and the phone choice of the laugh vowels.

In a further approach, whole prerecorded laughs are inserted into concatenative speech, either in combination with diphone speech [12] or as autonomous units in unit-selection synthesis [4]. However, the laughs must either be selected according to yet unknown criteria or they must be manipulated again in ways with unclear phonetic results for the listener. It is easy to sound ridiculous with the wrong laugh.

Further work could include the generation of laugh stimuli with articulatory synthesis that allow for more detailed testing of different features, varied with respect to intensity, fundamental frequency, breathing noise, friction sources, one or more laugh vowels etc. Several goals could be pursued: The set of articulatory gestures that work best for imitating particular laughs could be investigated, or articulatory synthesis could be used to build systematically varying laugh stimuli to test the impact that particular features have on the listener.

Another aspect associated with laughter is the question of *speech laughs*, i.e., where laughing occurs simultaneously with speech. It would be a highly challenging task to undertake with the articulatory synthesizer.

7. REFERENCES

[1] Bachorowski, J.-A., Smoski, M.J., Owren, M.J. 2001. The acoustic features of human laughter. *Journal of the Acoustical Society of America* 110, 1581-1597.

[2] Bickley, C., Hunnicutt, S. 1992. Acoustic analysis of laughter. Proc. *2nd International Conference on Spoken Language Processing,* Banff (2), 927-930.

[3] Birkholz, P. 2006. *3D-Artikulatorische Sprachsynthese.* Logos, Berlin. (PhD thesis).

[4] Campbell, N. 2006. Conversational speech synthesis and the need for some laughter. *IEEE Transactions on Audio, Speech, and Language Processing* 14, 1171-1178.

[5] Kipper, S., Todt, D. 2003. The role of rhythm and pitch in the evaluation of human laughter. *Journal of Nonverbal Behavior* 27, 255-272.

[6] Kohler, K. J., Peters, B., Scheffers, M. 2006. *The Kiel Corpus of Spontaneous Speech Vol. IV. German: Video Task Scenario (Kiel-DVD #1).* http://www.ipds.uni-kiel.de/kjk/pub_exx/kk2006_6/InfoDVD1.pdf visited 29-Mar-07

[7] Luschei, E.S., Ramig, L.O., Finnegan, E.M., Baker, K.K., Smith, M.E. 2006. Patterns of laryngeal electromyography and the activity of the respiratory system during spontaneous laughter. *J Neurophysiology* 96, 442-450.

[8] *MARY* Text-to-Speech Synthesis System. http://mary.dfki.de/online-demos/speech_synthesis visited 28-Mar-07

[9] *Praat*: Doing Phonetics by Computer. *www.praat.org* (version: 4.5.14) visited 05-Feb-2007

[10] Sundaram, S., Narayanan, S. 2007. Automatic acoustic synthesis of human-like laughter. *Journal of the Acoustical Society of America* 121 (1), 527-535.

[11] Trouvain, J. 2003. Segmenting phonetic units in laughter. *Proc. 15th. International Conference of the Phonetic Sciences*, Barcelona, 2793-2796.

[12] Trouvain, J., Schröder, M. 2004. How (not) to add laughter to synthetic speech. Proc. of the *Workshop on Affective Dialogue Systems*, Kloster Irsee, 229-232.