EVALUATION OF HMM-BASED LAUGHTER SYNTHESIS

Jérôme Urbain and Hüseyin Çakmak and Thierry Dutoit

TCTS Lab, Faculté Polytechnique, Université de Mons, Place du Parc 20, 7000 Mons, Belgium

ABSTRACT

In this paper we explore the potential of Hidden Markov Models (HMMs) for laughter synthesis. Several versions of HMMs are developed, with varying contextual information and algorithms for estimating the parameters of the source-filter synthesis model. These methods are compared, in a perceptive tests, to the naturalness of actual human laughs and copy-synthesis laughs. The evaluation shows that 1) the addition of contextual information did not increase the naturalness, 2) the proposed method is significantly less natural than human and copy-synthesized laughs, but 3) significantly improves laughter synthesis naturalness compared to the state of the art. The evaluation also demonstrates that the duration of the laughter units can be efficiently learnt by the HMM-based parametric synthesis methods.

Index Terms— Laughter, HMM, synthesis, evaluation

1. INTRODUCTION

Laughter is a significant feature of human interactions. It conveys information about our emotions and fulfills important social functions such as back-channeling. With the progress of speech processing and the development of human-machine interactions, in the last decades laughter received a growing interest as one signal that machines should be able to detect, analyze and produce.

In 2001, Ruch and Ekman [1] published an extensive report on the phonation, respiration, muscular and facial activities of laughter. Laughter is described as an inarticulate utterance, operated on the expiratory reserve volume, with a cycle of around 200ms. The same year, an analysis of the acoustic properties of human laughter was conducted by Bachorowski et al. [2]. They showed that the fundamental frequency of laughter is highly variable and generally takes higher values than speech, and that formant frequencies in laughter correspond to central vowels. In addition, they demonstrated that an important proportion of laughs is unvoiced (40 to 50%). Chafe [3] also describes the mechanical production of laughter and illustrates several profiles of laughter. A common conclusion of these studies is the high variability of the laughter phenomenon, in terms of voicing, fundamental frequency, intensity and more generally, types of sounds (grunts, cackles, pants, snort-like sounds, etc.).

A few years later, systems to automatically distinguish laughter from other sounds like speech started to be developed. Classification typically relies on Gaussian Mixture Models (GMMs) [4], Support Vector Machines (SVMs) [4, 5], Multi-Layer Perceptrons (MLPs) [6] or Hidden-Markov Models (HMMs) [7], trained with traditional spectral and prosodic features (MFCCs, PLP, pitch, energy, etc.). Equal error rates vary between 2 and 15% depending on the data used and classification schemes.

On the other hand, acoustic laughter synthesis is an almost unexplored domain. Sundaram and Narayanan [8] modeled the laughter intensity rhythmic envelope with the equations of an oscillating mass-spring system and synthesized laughter vowels by Linear Prediction. The naturalness of the obtained laughs was assessed in a perceptive study. Participants rated each laugh on a 5-point Likert scale (0-very poor to 4-excellent). Results showed that synthesized laughs are perceived as unnatural (average score of 0.71) and that human laughs do not receive a perfect naturalness score (average score: 3.28). Lasarcyk and Trouvain [9] compared laughs synthesized by a 3D modeling of the vocal tract and diphone concatenation. The articulatory system gave better results, but synthesized laughs were still far from human laughs.

Given the good performance achieved, in speech, by HMM-based approaches, we decided to explore the potential of HMMs for improving laughter synthesis naturalness. This paper presents the developed methods and the results of a perceptive evaluation assessing the naturalness of the synthesized laughs. The paper is organized as follows: the algorithms employed for HMM-based laughter synthesis are presented in Section 2; Section 3 focuses on the laughter data used; the setup of the perceptive evaluation is described in Section 4; the results of the evaluation are presented in Section 5 and discussed in Section 6; finally Section 7 concludes the paper and mentions future works.

2. HMM-BASED LAUGTHER SYNTHESIS

In HMM-based speech synthesis, the spectrum, F0 and durations are modeled in a unified framework [10]. From the HMM model, features are predicted by a maximum-

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement $n^\circ 270780$. H. Çakmak receives a Ph.D. grant from the Fonds de la Recherche pour l'Industrie et l'Agriculture (F.R.I.A.), Belgium.

likelihood parameter generation algorithm [11]. Finally, the generated parameters are sent to a parametric synthesizer to generate the waveform. Since this method is known for its flexibility, its use for the synthesis of non-verbal vocalizations, such as laughter in this case, appeared to be relevant. As a first application of this method, we decided to use the canvas provided in the demonstration scripts of HTS (HMM-based Speech Synthesis System) [12] which is a set of tools provided as a patch to HTK (HMM Toolkit) [13] and which allows to perform acoustic speech synthesis based on HMMs.

As an alternative to the feature extraction tools used in HTS demonstration scripts, namely SPTK (Speech Signal Processing Toolkit) for spectrum and Snack for F0, we included the STRAIGHT tools [14], which are known to be efficient in the field of speech processing. Although STRAIGHT provides tools for synthesis and for the extraction of other features, we only used it for extracting spectrum and F0.

To reduce buzziness in the synthesized waveforms for the voiced segments, we used the DSM (Deterministic plus Stochastic Model) method, which has shown improved results in terms of naturalness of synthetic speech [15]. This tool provides an excitation signal, for the voiced segments, which is not simply a pulse train but a waveform closer to the actual human excitation signal.

3. DATA

There exists few laughter data suitable for HMM-based synthesis, which should ideally include a large number of laughs from a single speaker and corresponding phonetic transcriptions. We used the AVLaughterCycle database [16], which includes spontaneous laughs from 24 subjects watching humorous videos, and has been phonetically annotated [17].

The first step regarding the data was to select one voice. The 24 subjects of the AVLaughterCycle database did not laugh the same amount of time and with the same "style". As HMM-based synthesis requires a lot of training examples, we focused on the 5 subjects who had produced the most laughs. Preliminary HMM-based synthesis models have been trained for these 5 voices and the best results (evaluated qualitatively) were obtained with subject number 6. Subject 6 is not the one who laughed the most (only 3 minutes of available laughter data, while subject 5 has 8 minutes), but he tends to use a limited set of phones (see [17]) and hence there are numerous examples of these phones available for training.

The second step was to process the narrow phonetic annotations. Indeed, the phonetic transcriptions for subject 6 contained 45 different phonetic labels, with only 12 of these appearing at least 10 times. It was thus decided to group acoustically close phones into broader phonetic classes, in order to increase the number of training examples available for each label. For example, velar, uvular, pharyngeal and glottal fricatives have been grouped in one "fricative" class. The phones that were barely present and could not be merged with acoustically similar phones to form one representative class with at least 10 occurrences were assigned to a "garbage" class. The resulting phonetic classes for subject 6 are listed in Table 1.

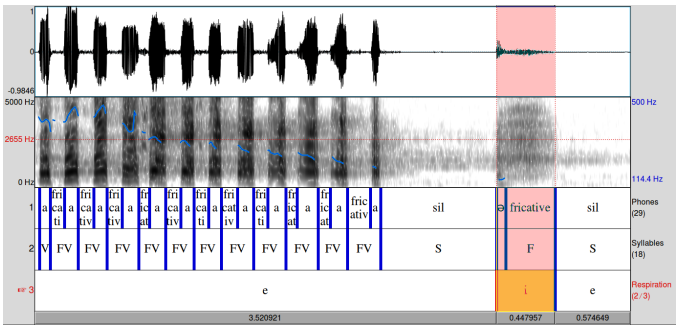

Fig. 1. Example of laughter data used. From top to bottom: waveform; spectrogram and pitch; phone transcription; syllable transcription; inhalation or exhalation transcription.

Table 1. Phonetic classes used for laughter synthesis

Inhalation or Exhalation	*Phonetic class*	*Occurrences*
e	fricative	439
e	a	327
e	silence	296
e	ə	84
i	fricative	49
e	ɛ	40
e	ʌ	39
e	cackle	35
e	ɪ	10
i	ə	9
e	glottal	8

To exploit the capabilities of HTS to model the context of phones, a third data processing step was conducted: laughter syllable annotations have been added to the phonetic transcriptions. Figure 1 shows an example of laughter waveform and spectrogram with the three levels of annotation. Finally, the fundamental frequency of the laughs was manually analyzed and it was found that good estimation of the fundamental frequency could be obtained by setting the boundaries of the estimation algorithm at 100 and 800Hz.

4. EVALUATION

4.1. Compared Methods

For each laugh, six different methods were compared:

- Method 1: the original human laugh, unmodified
- Method 2: the same laugh, re-synthesized through copy-synthesis (i.e. the parameters are extracted from the laugh and directly used in the source-filter model

to reconstruct the laugh), including STRAIGHT and DSM algorithms

- Method 3: the same laugh, synthesized with the HTS demo process (i.e. using as only input the phonetic transcription of the laugh) with imposed durations (i.e. each synthesized phone is forced to keep the same duration as in the original phonetic transcription), but only using as contextual information the labels of the two preceding and two following phones
- Method 4: same as Method 3, with an extended context including the information available thanks to the syllabic annotation (e.g., position of the phone in the syllable, position of the syllable in the word, etc.)
- Method 5: same as Method 4, with the addition of the STRAIGHT and DSM algorithms
- Method 6: same as Method 5, with the duration of each phone estimated by HTS

Method 1 was included to obtain a reference for naturalness, as it had already been shown that human laughs do not achieve a perfect naturalness score. Method 2 can be seen as the maximum performance achievable with our HMM-based source-filter models. Method 3 is considered as our baseline HMM-based laughter synthesis method, as it is directly available from HTS. Methods 4, 5 and 6 are possible improvements of the baseline method. Our test hypotheses were:

1. H1: Using the full contextual information improves the results (Method 4 better than Method 3)
2. H2: Using STRAIGHT and DSM improves the synthesis quality (Method 5 better than Method 4)
3. H3: HTS can model the durations appropriately (Method 6 is not worse than method 5)

Each of the synthesized laughs (Methods 3-6) was produced with a leave-one-out framework: the models were trained with all the laughs but the one to synthesize. It is important to note that Methods 3-6 used, as only input, the original phonetic transcriptions of the AVLaughterCycle laughs; generation of new phonetic transcriptions is beyond the scope of this paper.

4.2. Design

Sixty-four laughter episodes were available for subject 6 of the AVlaughterCycle database. Thirty-three of these included at least one phone that was present 10 times or less in the available data. These 33 laughs were not included in the evaluation, but were used in the training phase. Each of the remaining 31 laughs was synthesized with the 6 methods presented in Section 4.1. Laughs were presented to the participant in random order and for each laugh, only one of the methods was randomly selected.

The evaluation was performed through a web-based application, on a voluntary basis. Participants were invited by e-mail. The first page of the test asked the participants to provide the following details: their age, sex, whether they would rate the laughs with the help of headphones (which was suggested) or not, and whether they were working either on a) speech synthesis, b) audio processing, c) laughter, d) the ILHAIRE project[1] or e) none of these topics.

The second page explained the task, i.e. rating the naturalness of synthesized laughs on a 5-point Likert scale with the following labels: very poor (score 0), poor (1), average (2), good (3) and excellent (4). As some laughs, were extremely short and/or quiet, participants were also allowed to indicate "I cannot rate the naturalness of this laugh" instead of providing a naturalness value. Participants were also explained that they could listen to each sample as many times as they wanted before moving to the next example, but would then not be able to modify the given answers.

The third page contained 8 examples to familiarize participants with the range of synthesis quality that they would have to rate, with the aim to reduce bias during the evaluation. Laughs 8 (methods 1, 2, 5 and 6) and 20 (methods 1, 3, 5 and 6) were selected to form these examples and were excluded from the evaluation task. In consequence, there were 29 laughs (times 6 methods) remaining for the evaluation.

Finally, the participant was presented one laugh at a time and asked to rate its naturalness. The test was completed after 29 evaluations.

All the text of the evaluation was written both in English and in French (mother tongue of most of our participants).

4.3. Participants

Sixty-six participants completed the study: 37 females (average age: 33.1; std: 10.1) and 29 males (average age: 35.6; std: 13.5). Thirty-eight of these participants used headphones. Regarding the category with respect to laughter-synthesis possible knowledge and expectancies: 45 users selected "none of the above", 12 are involved in the ILHAIRE project, 5 are experts in speech synthesis and 4 are working on laughter.

5. RESULTS

Out of the 1914 received answers, 112 were "I cannot rate the naturalness of this laugh". Table 2 gathers the number of ratings received, the number of "unknown" answers and the average score for each method.

The percentage of naturalness scores obtained by each method is presented in Figure 2.

[1] ILHAIRE is a European project centered on laughter, whose participants had already been presented some examples of acoustic laughter synthesis, which could cause a bias in the ratings

Table 2. Received answers for each method

Method	*# ratings*	*# unknown*	*Total*	Av. score (std)
1	284	19	303	2.8 (1.2)
2	317	11	328	2.2 (1.2)
3	294	28	322	1.4 (1.0)
4	320	22	342	1.3 (1.1)
5	285	28	313	1.5 (1.1)
6	302	4	306	1.6 (1.1)
ALL	1802	112	1914	1.8 (1.2)

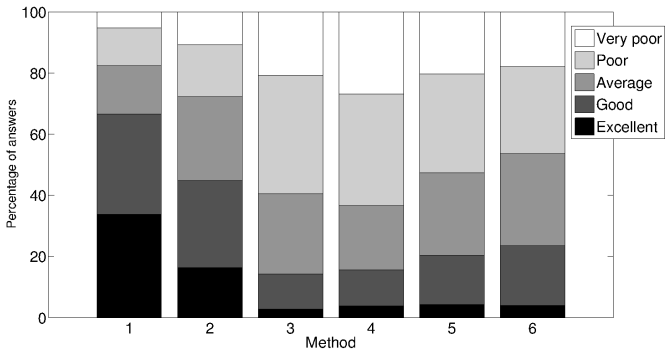

Fig. 2. Naturalness scores obtained by each method

Finally, an univariate analysis of the variance has been conducted, with the naturalness score as dependent variable and the method and laugh as explaining factors. Table 3 presents the p-values of pairwise comparisons between the different methods, using the Bonferroni test. Statistically significant differences at a 95% confidence level are highlighted.

Table 3. Pairwise p-values between the synthesis methods

Method	*1*	*2*	*3*	*4*	*5*	*6*
1	/	**0**	**0**	**0**	**0**	**0**
2	**0**	-	**0**	**0**	**0**	**0**
3	**0**	**0**	-	1	1	**0.022**
4	**0**	**0**	1	-	0.098	**0**
5	**0**	**0**	1	0.098	-	1
6	**0**	**0**	**0.022**	**0**	1	-

6. DISCUSSION

As it has already been found in previous studies, actual human laughs are not rated as perfectly natural by participants: method 1 has an average score of 2.8 out of 4. Even more, the perceived naturalness for human laughs is highly variable from one laugh to the other, as indicated by the large variance. Nevertheless, human laughs received significantly better naturalness scores than copy-synthesized laughs and our 4 HMM-based synthesis methods.

Regarding our hypotheses, the results of the evaluation contradict H1: adding more contextual information does not yield to higher naturalness scores. While this goes against our initial expectations, it can possibly be explained by the limited amount of training data: adding context enables HTS to build contextual subgroups for each phonetic class, which gives better dynamics to the laughs, at the expense of degraded acoustic models, as they have less training examples. This should be verified in the future with a larger laughter database. H2 has not been verified either: although Method 5 is clearly better rated than Method 4, the difference does not reach statistical significance. Finally, H3 has been verified: letting HTS model the duration of the phones does not impair the quality of the synthesis. Method 6 is actually better than Method 5, although the difference does not reach statistical significance. This indicates that the generation step (i.e. producing, from high-level instructions, the phonetic transcription of a laugh to synthesize) does not have to produce duration information along with the sequence of phones.

Among the 4 synthesis methods, method 6 yields the best results. The obtained average score of 1.6 is clearly better than the 0.71 achieved by Sundaram and Naryanan [8]. Objective comparison with Lasarcyk and Trouvain is not possible as they only reported about the rank of their methods. However, qualitative tests tend to favour our method.

7. CONCLUSION AND FUTURE WORK

In this paper we have presented an innovative way of synthesizing acoustic laughter, adapting methods that have proved efficient in speech synthesis. The proposed method yields significant improvement compared to previous work.

The potential of HMM-based laughter synthesis has been demonstrated with limited training data. Larger, single speaker, phonetically-annotated laughter databases would likely help improving the results. Recording and (automatically) annotating such a laughter database is part of our future work and will open new possibilities for laughter synthesis development.

Several versions of HMM-based laughter synthesis have been implemented and evaluated. The best model obtained includes STRAIGHT and DSM and duration predicted by HTS. The influence of contextual information must be further investigated, with a more specific evaluation centered on this question, the help of a larger database and/or modified contextual information. For example, we are currently considering the intensity of the phone as a candidate for contextual grouping.

Deeper statistical analysis of our evaluation data will also be performed, to investigate the influence of age, sex, wearing headphones and "expertise" in laughter synthesis on the naturalness scores.

Finally, implementing the generation step is crucial for interactive applications and would enable to produce new laughs from high-level commands. This step will be addressed in the near future within the ILHAIRE project, with the aim to obtain reactive laughter synthesis systems.

8. REFERENCES

[1] W. Ruch and P. Ekman, "The expressive pattern of laughter," in *Emotion, qualia and consciousness*, A. Kaszniak, Ed., pp. 426–443. World Scientific Publishers, Tokyo, 2001.

[2] J.-A. Bachorowski, M. J. Smoski, and M. J. Owren, "The acoustic features of human laughter," *Journal of the Acoustical Society of America*, vol. 110, no. 3, pp. 1581–1597, September 2001.

[3] Wallace Chafe, *The Importance of not being earnest. The feeling behind laughter and humor.*, vol. 3 of *Consciousness & Emotion Book Series*, John Benjamins Publishing Company, Amsterdam, The Nederlands, paperback 2009 edition, 2007.

[4] K. P. Truong and D. A. van Leeuwen, "Automatic discrimination between laughter and speech," *Speech Communication*, vol. 49, pp. 144–158, 2007.

[5] L. Kennedy and D. Ellis, "Laughter detection in meetings," in *NIST ICASSP 2004 Meeting Recognition Workshop*, Montreal, May 2004, pp. 118–121.

[6] M. T. Knox and N. Mirghafori, "Automatic laughter detection using neural networks," in *Proceedings of Interspeech 2007*, Antwerp, Belgium, August 2007, pp. 2973–2976.

[7] R. Cai, L. Lu, H.J. Zhang, and L.H. Cai, "Highlight sound effects detection in audio stream," in *Proceedings of the 2003 IEEE International Conference on Multimedia and Expo (ICME)*, Baltimore, USA, 2003, pp. 37–40.

[8] S. Sundaram and S. Narayanan, "Automatic acoustic synthesis of human-like laughter," *Journal of the Acoustical Society of America*, vol. 121, no. 1, pp. 527–535, January 2007.

[9] E. Lasarcyk and J. Trouvain, "Imitating conversational laughter with an articulatory speech synthesis," in *Proceedings of the Interdisciplinary Workshop on the Phonetics of Laughter*, Saarbrücken, Germany, August 2007, pp. 43–48.

[10] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in hmm-based speech synthesis," in *Proceedings of Eurospeech*, Budapest, Hungary, 1999.

[11] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for hmm-based speech synthesis," in *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*. IEEE, 2000, vol. 3, pp. 1315–1318.

[12] Keiichiro Oura, "HMM-based speech synthesis system (HTS) [computer program webpage]," http://hts.sp.nitech.ac.jp/, consulted on June 22, 2011.

[13] S.J. Young and S. Young, "The HTK Hidden Markov Model toolkit: Design and philosophy," in *Entropic Cambridge Research Laboratory, Ltd*. Citeseer, 1994.

[14] H. Kawahara, "Straight, exploitation of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds," *Acoustical science and technology*, vol. 27, no. 6, pp. 349–353, 2006.

[15] T. Drugman, G. Wilfart, and T. Dutoit, "A deterministic plus stochastic model of the residual signal for improved parametric speech synthesis," in *Proceedings of Interspeech*, 2009, pp. 1779–1782.

[16] Jérôme Urbain, Elisabetta Bevacqua, Thierry Dutoit, Alexis Moinet, Radoslaw Niewiadomski, Catherine Pelachaud, Benjamin Picart, Joëlle Tilmanne, and Johannes Wagner, "The AVLaughterCycle database," in *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 2010.

[17] Jérôme Urbain and Thierry Dutoit, "A phonetic analysis of natural laughter, for use in automatic laughter processing systems," in *Proceedings of the fourth biannual International Conference of the HUMAINE Association on Affective Computing and Intelligent Interaction (ACII2011)*, Memphis, Tennesse, October 2011, pp. 397–406.