Automatic Phonetic Transcription of Laughter and its Application to Laughter Synthesis

Jérôme Urbain Hüseyin Çakmak Thierry Dutoit
Circuit Theory and Signal Processing Lab, Faculty of Engineering, University of Mons
Place du Parc, 20, B-7000 Mons, Belgium
Email: {jerome.urbain,huseyin.cakmak,thierry.dutoit}@umons.ac.be

***Abstract*—In this paper, automatic phonetic transcription of laughter is achieved with the help of Hidden Markov Models (HMMs). The models are evaluated in a speaker-independent way. Several measures to evaluate the quality of the transcriptions are discussed, some focusing on the recognized sequences (without paying attention to the segmentation of the phones), other only taking into account the segmentation boundaries (without involving the phonetic labels). Although the results are far from perfect recognition, it is shown that using this kind of automatic transcriptions does not impair too much the naturalness of laughter synthesis. The paper opens interesting perspectives in automatic laughter analysis as well as in laughter synthesis, as it will enable faster developments of laughter synthesis on large sets of laughter data.**

I. Introduction

Laughter is a significant feature of human communication. Humans are estimated to use this expressive social signal for 7 million years [1]. It is an innate phenomenon that develops around the fourth month of life, way earlier than speech. Nevertheless, laughter has for a long time been neglected in scientific research, most of the work focusing on our primary expression means: speech. In the last decades however, with the emergence of efficient speech processing applications (speech synthesis is intelligible and speech recognition reaches high accuracy in controlled settings), laughter has received a growing interest. Indeed, emotions are generally lacking in current speech processing methods. It is necessary to deal with emotions to make human-machine interactions more natural, and laughter is one of the most important emotional signals.

Several research teams studied laughter and gave descriptions of its features. Ruch and Ekman [1] investigated the whole production mechanism. They reported on the structure of laughter, its respiratory patterns, the involved phonation, facial and body movements. From their analyses, it appears that laughter periodicity is around 200ms and that, due to the facial constraints of smile occurring in joyful laughter, central vowels are the most likely in spontaneous laughter. Bachorowski et al. [2] focused on the acoustic aspects, with similar conclusions. They found that fundamental frequency reaches higher values in laughter than in speech, that unvoiced laughter is almost as frequent as voiced laughter (contrarily to the stereotypical image of voiced laughter we have) and that the "vowels" used in laughter are mainly central vowels. Trouvain [3] summarized the different terminologies used to characterize laughter units. The following structure for voiced laughter emerges: long and intense laughter *episodes* can be composed of several *bouts*, i.e. exhalation parts separated by inhalations. The bouts themselves can include several laughter *syllables* (also called *calls*, *cycles*, or *pulses*), which themselves are generally formed by one consonant (typically h-like) and one vowel. Even if one common conclusion of all these works is the extreme variability of laughter, it can be noticed that all these authors use some kind of phonetic transcription to characterize laughter.

In 1985, Jefferson [4] made the exercise of transcribing laughter and discussed the additional information that can be brought by actually transcribing laughter in conversations instead of only naming that participants are laughing. Glenn [5] uses the same kind of pseudo-phonetic transcriptions, e.g. "*eh h<u>u</u>h huh ↑hah hah*", where _ denotes emphasis and ↑ indicates raising intonation. Chafe [6] uses custom-made symbols to transcribe laughter pulses, distinguishing between inhalation and exhalation, voiced and voiceless, open and closed-mouth syllables. Urbain and Dutoit [7] proposed to annotate laughter phones the same way speech is phonetically transcribed. Bouts and phones (consonant, vowels, and other laughter-related sounds) were annotated.

In this paper, we introduce automatic phonetic transcription of laughs, using the audio signal only. To the best of our knowledge, this is the first work in this way. Applications are obvious in laughter analysis, for clustering similar laughs (with the same acoustic contents, the same global structure, periodicity, etc.), identifying different types of laughter, etc. In addition, one of the main fields where a phonetic transcription is essential is laughter synthesis.

Unlike laughter recognition, which has seen a lot of progresses—among others with the works of Kennedy and Ellis [8] or Truong and van Leeuwen [9]—, laughter synthesis is still at its beginning. Sundaram and Narayanan [10] were the first to tackle the problem. They synthesized vowels through Linear Prediction and modeled rhythmic intensity curves with the equations governing a mass-spring system. Lasarcyk and Trouvain [11] compared two methods for laughter synthesis: one based on diphone concatenation (trained on speech data), the other was a 3D-articulatory modeling of the vocal organs. In all cases, synthesized laughs were perceived as significantly less natural than human laughs. Recently, the use of Hidden Markov Models (HMMs) trained on laughter phonetic tran-

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n°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.

978-0-7695-5048-0/13 $26.00 © 2013 IEEE
DOI 10.1109/ACII.2013.32

scriptions was investigated [12]. Using the same evaluation process as Sundaram and Narayanan, it was shown that, although actual human laughs are still out of reach, HMM-based laughter synthesis yields higher naturalness results than the previous methods. This is one of the reasons why we decided to develop automatic phonetic transcription of laughter, with the objective to fasten the development of laughter synthesis (as producing phonetic annotations is time-consuming). In the current work, automatic phonetic transcriptions are used to train laughter synthesis, and the results are compared to the traditional process (HMMs trained on manual transcriptions) to estimate the potential of the method.

The paper is organized as follows. In Section II, the laughter data used for building and evaluating automatic laughter phonetic transcriptions will be introduced. The automatic transcription method is described in Section III. Section IV is devoted to results and measures of the quality of the obtained phonetic transcriptions. Section V focuses on laughter synthesis using automatic phonetic transcriptions. Finally, conclusions and future works are presented in Section VI.

II. Data

The corpus used is the AVLaughterCycle (AVLC) database [13]. It contains recordings of 24 participants watching humorous videos. The database includes video and facial motion capture, but only the audio signals were used in the current work. The corpus contains around 1000 laughter episodes, for a total duration of approximately 1 hour. Audio signals were recorded at 16kHz. In [7], the AVLC laughs were phonetically annotated. Two hundred different phonetic labels were used. For the most part, these labels were taken from the International Phonetic Alphabet (IPA) and standard diacritics [14] to denote particular production modes (e.g., breathy or creaky). In addition, some labels were introduced to characterize laughter sounds that are not covered by the IPA, such as cackles, grunts or hum-like bursts. In addition, inhalation and exhalation phases were annotated in a separate track.

III. Automatic laughter phonetic transcription

Automatic laughter phonetic transcription is a new process. However, numerous methods have already been developed for phonetic segmentation of speech. The most frequently used methods for speech segmentation rely on phonetic Hidden Markov Models (HMMs), trained with spectral features [15]. HMMs are able to model the temporal evolution of signals, which is interesting for characterizing phonemes, as they generally contain several parts: the stable part is surrounded by transition parts with the previous and following phonemes. Nevertheless, it should be noted that automatic speech phonetic segmentation usually relies on existing transcriptions of the utterances, and the objective is to find the best alignment between the acoustic signal and the given phonetic transcription.

In our case, we aim at automatically process any incoming laugh, without any human intervention. No transcription is available for our algorithms. Our approach is actually close to speech recognition, where all the decisions are taken using only the acoustic signal. As HMMs are also widely used in speech recognition, it is this technique we investigated to produce automatic laughter phonetic transcriptions.

The implementation was made with the help of the Hidden Markov Models ToolKit (HTK) [16]. One HMM was built for each phone in the database. HMMs were always trained with a leave-one-subject-out process: the HMMs were trained using all the data from 23 subjects and tested on the laughs of the remaining participant of the AVLC database. The operation was repeated 24 times so as to obtain automatic phonetic transcriptions for all the subjects.

The acoustic signal was segmented in 32ms frames (512 samples at 16kHz) with a 10ms shift. For each frame, 70 acoustic features were extracted:

- spectral centroid, spectral spread, spectral variation, 4 values of spectral flatness, spectral flux, spectral decrease [17];
- 13 Mel-Frequency Cepstral coefficients (MFCCs), their first and second derivatives;
- Zero-Crossing Rate, Root-Mean-Square energy and loudness [17];
- Chirp Group Delay [18] and 4 values for Harmonic to Noise Ratio [19];
- 12 Chroma features [20];
- the fundamental frequency computed with the SRH method [21], as well as the value of the maximum SRH peak;

Initial tests revealed that the HMMs could not deal with the large quantity of different phones. There were numerous confusions, as some phones are really close to each other from an acoustic point of view. Furthermore, many phones only had a few available occurrences for training, which resulted in inaccurate modeling. As it has been shown that good laughter synthesis can be achieved with a limited set of phones [12], we decided to group acoustically close phones in broader phonetic clusters. Several combinations have been experimented. The grouping illustrated in Figure 1 appeared as a good trade-off between precision (keeping a sufficient number of different classes to distinguish between laughter sounds) and efficiency (limiting the number of classes so that it is manageable by automatic transcription algorithms). Some labels introduced to characterize laughter sounds that are not covered by the International Phonetic Alphabet also formed phonetic clusters, namely *cackles*, *grunts* (including pants and chuckles) and *nareal fricatives*—audible voiceless friction sound through the nostrils, that actually have a phonetic symbol: ñ̊. As clicks and glottal stops are very short and subtle phones, which would make it very hard to detect, but do not provide critical information for applications that can be built on automatic laughter phonetic transcription (e.g., clustering, laughter synthesis, etc.), these phones were discarded from our models. The size of the phonetic clusters used in our experiments is given in Table I.

HTK provides control over several parameters to design HMMs. It goes beyond the scope of this paper to present detailed results regarding the optimization of each of these parameters for acoustic laughter transcription. Most of the parameters have been manually tuned and the resulting automatic transcriptions were compared with the manual (reference) phonetic transcriptions. The following parameters have been used in our experiments: all the HMMs have 3 states and

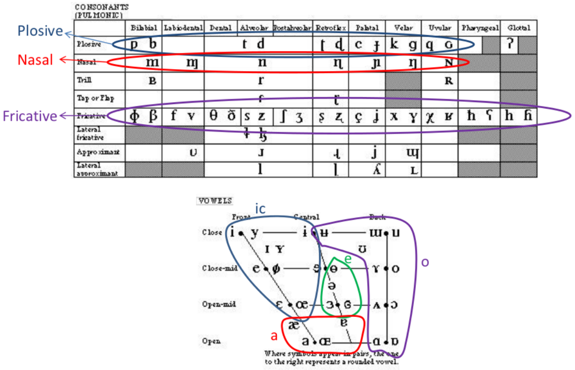

Fig. 1. Grouping of phones to build consistent phonetic clusters. The original IPA chart was retrieved from [14]

TABLE I. PHONETIC CLUSTERS USED FOR HMM-BASED LAUGHTER PHONETIC TRANSCRIPTION, ORDERED BY FREQUENCY OF OCCURRENCE

Inhalation or Exhalation	*Phonetic cluster*	*Occurrences*
e	silence	6612
e	fricative	3261
e	e	1549
e	a	1432
e	ɪ	1203
e	ɦ̊	839
i	fricative	774
e	nasal	717
e	cackle	704
e	plosive	286
e	o	256
i	e	219
i	ɦ̊	166
e	grunt	156
i	ɪ	153
i	plosive	43
i	silence	9

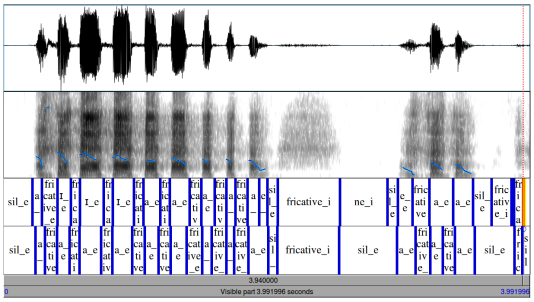

Fig. 2. Example of automatic phonetic laughter transcription. From top to bottom: 1) waveform; 2) spectrogram; 3) automatic (HTK) transcription; 4) reference transcription. In the phonetic transcriptions, the _e and _i suffixes indicate exhalation and inhalation phases, respectively.

transitions between all theses states are allowed; the emission probabilities of each state are modeled with 10 Gaussian distributions; to avoid excessive insertions, the word insertion penalty was set to -20; the grammar consisted in bigram modeling of the succession of phones; the language factor (i.e., the weight of the grammar in the recognition process) was set to 2. An example of the obtained phonetic transcriptions is shown in Figure 2.

IV. AUTOMATIC TRANSCRIPTION RESULTS

There are several ways to evaluate the quality of automatic transcriptions. Most frequently, measures of hit, insertion, substitution and deletion rates are used. These kinds of figures are directly provided by HTK. To compute them HTK only uses the transcription of the file, without paying attention to the temporal segmentation. HTK searches for the best match between the automatic and the reference transcriptions [22] and provides the number of (see Figure 3):

- hits (H): the phones that correspond in the automatic and reference transcriptions;
- substitutions (S): phones that are labeled differently in the automatic and reference transcriptions;
- insertion (I): the number of extra phones in the automatic transcription, where there is no corresponding phone in the reference transcription;
- deletions (D): the number of phones in the reference transcription that have no corresponding phone in the automatic transcription.

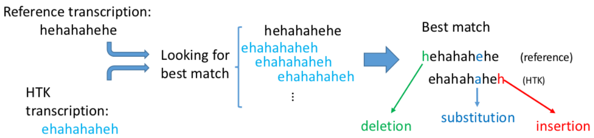

Fig. 3. Basis of the recognition measures outputted by HTK: insertions, substitutions, deletions and hits. Hits are not explicitly represented here, they concern all the matching phones (in black).

Two global measures are also provided [22]:

- the Percentage Correct (PC):

$$PC = \frac{N - D - S}{N} * 100 \qquad [\%] \tag{1}$$

where N is the total number of phones in the reference transcriptions;

- the Percentage Accuracy (PA):

$$PA = \frac{N - D - S - I}{N} * 100 \qquad [\%] \tag{2}$$

While these measures provided by HTK are useful, it must be remembered that these figures do not take the temporal segmentation into account, but only the transcription. The problem of evaluating the quality of a segmentation is discussed in [23]. The proposed methods define a search region around each segmentation boundary of the reference transcription. As illustrated in Figure 4, a *hit* is obtained when there is a detected segmentation boundary inside the search region; a *deletion* occurs when there is no detected boundary inside the search region of a reference boundary; and *insertions* are counted for detected boundaries outside the search regions of the reference boundaries, or when there is more than 1 detected boundary inside the search region of a single reference boundary. Based on these figures, and including the total number of reference boundaries N_{ref} and the total number of detected boundaries N_{det}, the following measures are proposed to evaluate the overall quality of the segmentation [23]:

- Hit Rate (HR), representing the proportion of actual boundaries that have been retrieved:

$$HR = \frac{N_{hit}}{N_{ref}} * 100 \qquad [\%] \tag{3}$$

- Over-Segmentation rate (OS), which is the ratio of supernumerary detected boundaries:

$$OS = \frac{N_{det} - N_{ref}}{N_{ref}} * 100 \qquad [\%] \tag{4}$$

- Precision (PR), which is the ratio of detected boundaries that are correct:

$$PR = \frac{N_{hit}}{N_{det}} * 100 \qquad [\%] \tag{5}$$

- R-distance (R), which is the distance from the optimal functioning point ($HR = 100$ and $OS = 0$):

$$R = 1 - \frac{\sqrt{(100 - HR)^2 - OS^2} + |\frac{-OS+HR-100}{\sqrt{2}}|}{200} \tag{6}$$

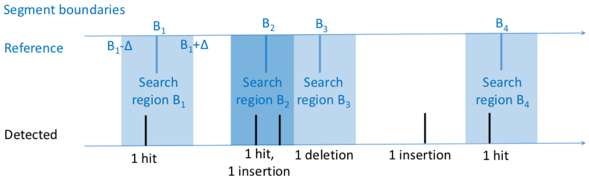

Fig. 4. Basis of the segmentation measures: hits, inserted boundaries and deleted boundaries.

These measures were computed for our automatic transcription methods, using search regions of 40ms ($\Delta = 20ms$) for the segmentation measures. Table II gathers the HTK and segmentation measures for automatic transcriptions obtained with different values of word insertion penalty (WP) and language factor (LF) in HTK. These values illustrate that the values that were empirically determined (WP=-20 and LF=2, row highlighted in bold) indeed form a good compromise between hit rate and over-segmentation, both for the phonetic transcription (HTK measures) and the location of the boundaries (segmentation measures).

TABLE II. MEASURES OF TRANSCRIPTION ACCURACY, FOR DIFFERENT VALUES OF WORD INSERTION PENALTY (WP) AND LANGUAGE FACTOR (LF)

HTK parameters		HTK measures		segmentation measures				
WP	LF	PC	PA	HR	OS	PR	F	R
-20	1	60	43	56	4.4	54	0.55	0.61
	3	63	41	57	11	51	0.54	0.6
	10	58	-204	66	263	18	0.28	-1.4
-20	2	**62**	**45**	**56**	**3.6**	**54**	**0.55**	**0.61**
0		71	2	71	69	42	0.52	0.34
-40		54	47	46	-19	57	0.51	0.62

The obtained results are far from being perfect, reflecting that the automatic phonetic transcriptions do not exactly match the reference ones. However, these transcriptions are already useful for training laughter synthesis, as will be presented in the next section.

V. APPLICATION TO LAUGHTER SYNTHESIS

HMMs were designed for acoustic laughter synthesis, following the same methodology as in [12]. Three-states left-to-right HMMs were trained with the HTS toolkit [24]. The STRAIGHT and DSM methods—which are know for improving speech synthesis and did slightly improve the quality of laughter synthesis [12]—were included. Laughter synthesis was trained on the voice of subject 6 of the AVLC database, with the phonetic transcriptions provided by HTK (see Section III). Sixty-four laughs were available for training. Phones with less than 10 occurrences in the reference transcriptions were mapped to a *garbage* class. The number of occurrences of each phonetic cluster is given in Table III. Laughs were synthesized with a leave-one-out process: HMMs were trained on all the available laughs but one, and the phonetic transcription of the remaining laugh was used as input for synthesis.

TABLE III. NUMBER OF OCCURRENCES IN THE PHONETIC CLUSTERS USED FOR HMM-BASED LAUGHTER SYNTHESIS, FOR THE REFERENCE AND HTK PHONETIC TRANSCRIPTIONS

Inhalation or Exhalation	*Phonetic cluster*	*Occurrences*	
		reference	HTK
e	fricative	439	327
e	a	331	266
e	silence	291	203
e	e	85	101
e	ɪ	50	32
i	fricative	49	99
e	o	45	48
e	cackle	36	85
i	e	11	8

HTS has the possibility to model different contexts for each phoneme. The basic contextual information that can be used to discriminate between contexts is the phonetic labels of the (generally 2) preceding and following phones. But more information can be added. For example, in speech synthesis the position of the phoneme in the syllable, word and utterance is often included as contextual information. An example of contextual features is given in [25]. To be able to use the same kind of contextual features for laughter synthesis, syllables had been manually annotated for subject 6 in [12]. As we desired to use the same kind of information with the automatic transcriptions obtained with HTK, a syllable layer was automatically added. Two-phones syllables were formed when a fricative (or silence) was followed by a vowel, a cackle or a nasal. All the phones that were not included in 2-phones syllables by this process were assigned to a 1-phone syllable.

A web-based evaluation experiment of the synthesized laughs was conducted. As in [10] and [12], naive participants were asked to rate the naturalness of synthesized laughs on a 5-points Likert scale with the following labels: very poor (score 0), poor (1), average (2), good (3), excellent (4). Laughter synthesis trained with the HTK transcriptions was compared to synthesis trained with the manual (reference) transcriptions. For each of these training processes, 2 laughs were synthesized: the first one with imposed duration (HTS had to respect, for each phone, the duration provided in the phonetic transcription), the second one with the duration of each phone estimated by HTS. As the objective was to evaluate whether phones can be accurately modeled for synthesis when they are trained on automatic transcriptions (and segmentation), all laughs were synthesized using the reference transcriptions. For comparison purposes, human laughs were also included in the evaluation. Table IV summarizes the different methods compared in the evaluation experiment. Twenty-three laughs contained at least one occurrence of a *garbage* phone and were not included in the evaluation. Out of the remaining 41 laughs, two laughs ($n°24$ and 29) were shown as examples to the participants prior to the test, so they could familiarize with the range of laughter qualities they would have to rate. These laughs were not included in the evaluation, which included the remaining 39 laughs. Each participant had to rate one laugh at a time. Laughs were presented in random order and for each laugh, only one of the method was randomly selected. The test was completed after 39 evaluations.

TABLE IV. LAUGHTER SYNTHESIS METHODS COMPARED IN THE EVALUATION EXPERIMENT

Method	Training Transcriptions	Duration	Synthesis transcriptions
R1	reference	imposed	reference
R2	reference	estimated by HTS	
A1	automatic (HTK)	imposed	
A2	automatic (HTK)	estimated by HTS	
H	human laughs		

Forty-four participants completed the evaluation. Each method was evaluated between 199 and 255 times. The average naturalness score received by each method is given in Table V and the distribution of received answers for each method is illustrated in Figure 5. A univariate analysis of the variance was conducted and the p-values of the pairwise comparisons of the different methods, using the Tukey Honestly Significant Difference (HSD) adjustment, are presented in Table VI. Statistically significant differences at a 95% confidence level are highlighted in bold. It can be seen that, as expected, human laughs are more natural than synthesized laughs. Among the synthesized laughs, method R2 is the best. As it was already found in [12], the best results are achieved when HTS estimates the phone durations: R2 and A2 are respectively better than R1 and A1, although the difference is significant in neither case. The synthesis methods using automatic phonetic transcriptions (A1 and A2) are less natural than their counterparts using reference phonetic transcriptions (R1 and R2, respectively). Nevertheless, the automatic phonetic transcriptions do not yield to a dramatic drop in naturalness, as method A2 is not significantly less natural than method R1. In comparison with previous works, all our synthesis methods (even using automatic transcriptions for training) received higher naturalness scores than Sundaram and Naryanan's method [10]—which had an average naturalness score of 0.71—, while our reference methods (R1 and R2) obtained similar naturalness scores as in [12], which was expected as we used the same process.

TABLE V. AVERAGE NATURALNESS SCORES RECEIVED BY EACH SYNTHESIS METHOD

	Method				
	R1	**R2**	**A1**	**A2**	**H**
average naturalness score	1.4	1.7	1.0	1.2	3.3
naturalness score std	1.1	1.1	0.9	1.1	0.9

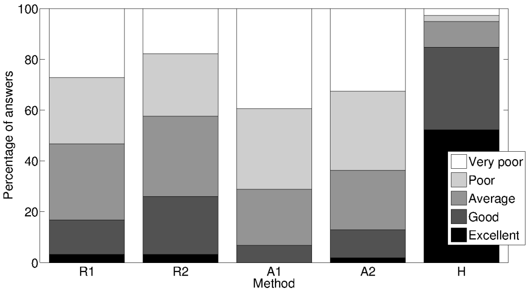

Fig. 5. Naturalness score received by the synthesis methods.

TABLE VI. PAIRWISE P-VALUES BETWEEN SYNTHESIS METHODS

Method	*R1*	*R2*	*A1*	*A2*	*H*
R1		0.11	**0**	0.08	**0**
R2	0.11		**0**	**0**	**0**
A1	**0**	**0**		0.24	**0**
A2	0.08	**0**	0.24		**0**
H	**0**	**0**	**0**	**0**	

To conclude this section, it is interesting to take into account the comments given by some participants after evaluating the naturalness of the laughs. First, several participants informed us that, for many laughs, a large proportion of the laugh was nicely synthesized, but a few phones were strange and definitely not human. Rating the whole laugh was then complicate, even though generally in such cases the ratings were made towards the non-natural end of the scale. This is related to the second main remark from the participants: what exactly is the definition of naturalness? Indeed—even if in this study we purposely did not give any further indication so as to be able to compare our results with previous studies employing the same process—several factors can cause a laugh to be perceived as non-natural: a) the sound quality itself (buzzy, etc.); b) the perception of the laugh to be forced/acted/faked (which can also concern human laughs with perfect sound quality) instead of spontaneous/emotional; c) the laugh being far from what participants expect, some kind of laughter stereotype (again, this can also concern human laughs). These remarks can partially explain the large standard deviations in Table V. The overall interrater agreement for the 5 naturalness values is quite low: Fleiss' kappa [26] is .167. Participants however generally agree on whether the laugh sounds natural (score 3 or 4) or not (score 0, 1 or 2) with a kappa value of 0.41 which is significantly different from 0 ($p = 0$). Another reason for the large standard deviations of the synthesized methods is the variability between laughs: even human laughs are affected by a large standard deviation, indicating that, out of their context, laughs can be perceived as unnatural even when they were actually spontaneous (see points b and c above).

VI. CONCLUSIONS AND FUTURE WORKS

In this paper, a method to automatically obtain phonetic transcriptions of laughter has been presented. The topic in itself is novel and has applications in laughter analysis and laughter synthesis. Promising results have been reported. This initial work opens some discussions and interesting perspectives.

First, a methodical optimization of the parameters of the recognition models could be considered. One of the issues here would be the optimization criterion. We have seen in Section III that different measures can be used to evaluate the quality of a recognized sequence and/or a segmentation. In our case, we could consider a combination between phones recognition (Percentage Accuracy) and segmentation (R-distance). Yet, this would be subject to some arbitrary choices (weighting the two original measures) and could depend upon the purposes of the application (the optimal transcriptions for laughter clustering and laughter synthesis are probably different).

Second, the choice of phones can be further investigated. Again, it is most likely dependent on the targeted application. Furthermore, the set of acoustic features can also be discussed. In this work we used a range of standard acoustic features computed off-line. It would be interesting to see what can be achieved with the features HTK (or other libraries, e.g. openSMILE [27]) can extract in real time. In addition, some laughter-related features could be introduced, for example to characterize the periodicity of laughter, which could improve the results. In the same vein, using larger ($n > 2$) n-gram models should be investigated, as the repetitiveness of laughter syllables would then be encoded.

Third, in this work we have developed a totally automatic, speaker-independent model. The results can most likely be improved with speaker adaptation techniques, frequently used in speech recognition. The cost is that some laughs from the target laugher should then be manually transcribed. This could be evaluated quite rapidly on the AVLC database, as the phonetic transcriptions of all the laughs are available.

Fourth, the automatic phonetic transcription method can be used to cluster laughs according to their phonetic contents. Such an application will soon be launched and evaluated.

Fifth, to deal with the comments made by the evaluators, two modifications of the evaluation process should be investigated in the future: 1) refining the questions: asking several questions and clearly stating which dimension is the focus of each question (e.g. sound quality, what proportion of the laugh is nicely rendered, does the laugh seem fake, is the laugh in line with user's expectations, etc.); 2) identifying which phones are poorly modeled and discard these phones for the synthesis: while it would limit the repertoire of laughs that can be synthesized, it would prevent having some bad parts in the laughs, that affect the overall perception of their quality.

Finally, it has been shown that automatic phonetic transcriptions can be employed to train laughter synthesis. This opens interesting perspectives for laughter synthesis, as new voices can be trained and evaluated rapidly. In the future, we will employ this process to create new laughter voices from even larger quantities of data (recordings are under way), that are not phonetically annotated. There is a small drop in naturalness when comparing laughs trained on automatic transcriptions to laughter synthesis trained with manual phonetic transcriptions. However, the impact of automatic transcriptions can probably be limited by using more training data, which would help reducing the number of badly modeled phones.

References

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

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

[4] G. Jefferson, "An exercise in the transcription and analysis of laughter," in *Handbook of discourse analysis*, ser. Discourse and Dialogue, T. V. Dijk, Ed. London, UK: Academic Press, 1985, vol. 3, pp. 25–34.

[5] P. J. Glenn, *Laughter in interaction*. Cambridge University Press, Cambridge, 2003.

[6] W. Chafe, *The Importance of not being earnest. The feeling behind laughter and humor.*, ser. Consciousness & Emotion Book Series. Amsterdam, The Nederlands: John Benjamins Pub. Comp., 2007, vol. 3.

[7] J. Urbain and T. Dutoit, "A phonetic analysis of natural laughter, for use in automatic laughter processing systems," in *Proceedings of ACII 2011*, Memphis, Tennesse, October 2011, pp. 397–406.

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

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

[10] 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.

[11] 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, 2007.

[12] J. Urbain, H. Cakmak, and T. Dutoit, "Evaluation of hmm-based laughter synthesis," in *Proceedings of ICASSP'13*, Vancouver, Canada, 2013.

[13] J. Urbain, E. Bevacqua, T. Dutoit, A. Moinet, R. Niewiadomski, C. Pelachaud, B. Picart, J. Tilmanne, and J. Wagner, "The AVLaughterCycle database," in *Proceedings of LREC'10*, Valletta, Malta, 2010.

[14] P. Ladefoged, "A course in phonetics," Online: http://hctv.humnet.ucla.edu/departments/linguistics/VowelsandConsonants/course/chapter1/chapter1.html, Consulted on January 20, 2011.

[15] D. T. Toledano, L. H. Gómez, and L. V. Grande, "Automatic phonetic segmentation," *Speech and Audio Processing, IEEE Transactions on*, vol. 11, no. 6, pp. 617–625, 2003.

[16] S. Young and S. Young, "The htk hidden markov model toolkit: Design and philosophy," in *Entropic Cambridge Research Laboratory, Ltd*. Citeseer, 1994.

[17] G. Peeters, "A large set of audio features for sound description (similarity and classification) in the CUIDADO project," Institut de Recherche et Coordination Acoustique/Musique (IRCAM), Tech. Rep., 2004.

[18] T. Drugman, T. Dubuisson, and T. Dutoit, "Phase-based information for voice pathology detection," in *Proceedings of ICASSP'11*. IEEE, 2011, pp. 4612–4615.

[19] T. Drugman, J. Urbain, N. Bauwens, R. Chessini, C. Valderrama, P. Lebecque, and T. Dutoit, "Objective study of sensor relevance for automatic cough detection," *IEEE Transactions on Information Technology in BioMedicine*, 2013.

[20] D. P. Ellis and G. E. Poliner, "Identifying cover songs' with chroma features and dynamic programming beat tracking," in *Proceedings of ICASSP'07'*, vol. 4. IEEE, 2007, pp. IV–1429.

[21] T. Drugman and A. Alwan, "Joint robust voicing detection and pitch estimation based on residual harmonics," in *Proceedings of Interspeech 2011*, Firenze, Italy, August 2011.

[22] S. J. Young, G. Evermann, M. Gales, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, "The htk book version 3.4," 2006.

[23] O. J. Räsänen, U. K. Laine, and T. Altosaar, "An improved speech segmentation quality measure: the r-value," in *Proceedings of Interspeech'09*, 2009.

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

[25] H. Zen, "An example of context-dependent label format for hmm-based speech synthesis in english," *The HTS CMUARCTIC demo*, 2006.

[26] J. L. Fleiss, B. Levin, and M. C. Paik, "The measurement of interrater agreement," *Statistical methods for rates and proportions*, vol. 2, pp. 212–236, 1981.

[27] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of ACM*, Florence, Italy, 2010, pp. 1459–1462.