A Phonetic Analysis of Natural Laughter, for Use in Automatic Laughter Processing Systems

Jérôme Urbain and Thierry Dutoit

Université de Mons - UMONS, Faculté Polytechnique de Mons, TCTS Lab
20 Place du Parc, 7000 Mons, Belgique
{jerome.urbain,thierry.dutoit}@umons.ac.be

Abstract. In this paper, we present the detailed phonetic annotation of the publicly available AVLaughterCycle database, which can readily be used for automatic laughter processing (analysis, classification, browsing, synthesis, etc.). The phonetic annotation is used here to analyze the database, as a first step. Unsurprisingly, we find that h-like phones and central vowels are the most frequent sounds in laughter. However, laughs can contain many other sounds. In particular, nareal fricatives (voiceless friction in the nostrils) are frequent both in inhalation and exhalation phases. We show that the airflow direction (inhaling or exhaling) changes significantly the duration of laughter sounds. Individual differences in the choice of phones and their duration are also examined. The paper is concluded with some perspectives the annotated database opens.

1 Motivation and Related Work

Laughter is an important emotional signal in human communication. During the last decades, it received growing attention from researchers. If we still do not understand exactly *why* we laugh, progress has been made in understanding *what* it brings us (enhanced mood, reduction of stress, and other health outcomes [2, 14]) and in describing *how* we laugh (see [1, 5, 17, 19]). This paper will focus on the last aspect, laughter description, with the aim of improving automatic laughter processing. In particular, we will mainly consider the acoustic aspects.

Bachorowski et al. [1] were the first to extensively report about the acoustic features of human laughter. They classified laughs in three broad groups: song-like, snort-like and grunt-like. They also labeled the syllables constituting these laughs as voiced or unvoiced. They analyzed several features (duration, pitch, formants) over syllables and whole laughs. They found that mainly central vowels are used in laughter and that the fundamental frequency can take extreme values compared to speech. More generally, laughter has been identified as a highly-variable phenomenon. Chafe [5] illustrates a variety of its shapes and sounds with the help of acoustic features (voicing, pitch, energy, etc.).

However, despite the numerous terms used in the literature to describe laughter (see the summary given by Trouvain [21]), there is currently no standard for laughter annotation. Phonetic transcriptions appear in a few laughter-related papers (see [7, 16]) but, to our knowledge, no large laughter database has been annotated that way. For example, the two most used natural laughter databases, the ICSI [9] and AMI [4] Meeting Corpora, do not include detailed laughter annotation (only the presence of laughter in a speech turn is indicated). The ICSI Meeting corpus contains around 72 hours of audio recordings from 75 meetings. The AMI Meeting Corpus consists of 100 hours of audiovisual recordings during meetings. Both databases contain a lot of spontaneous, conversational laughter (108 minutes in the 37 ICSI recordings used in [22]).

With the development of intelligent human-computer interfaces, the need for emotional speech understanding and synthesis has emerged. In consequence, interest for laughter processing increased. Several teams developed automatic laughter recognition systems. In [10, 22], classifiers have been trained to discriminate between laughter and speech, using spectral and prosodic features. Reported Equal Error Rates (EER) were around 10%. The local decision was improved in [11] thanks to long-term features, lowering the EER to a few percent. Recently, Petridis and Pantic [15] combined audio and visual features to separate speech from voiced and unvoiced laughter with 75% of accuracy[1]. No method has been designed to automatically label laughs, classify them in finer categories than simply voiced or unvoiced, or segment long laughter episodes in laughter "bouts" (exhalation phases separated by inhalations).

A few researchers have also investigated laughter synthesis. Sundaram and Narayanan [18] modeled the energy envelope with a mass-spring analogy and synthesized the vowel sounds of laughter using linear prediction. Lasarcyk and Trouvain [13] compared synthesis by diphone concatenation and 3D modeling of the vocal tract. Unfortunately, in neither case the obtained laughs were perceived as natural by naive listeners. A recent online survey [6] confirmed that no laughter synthesis technique currently reaches a high degree of naturalness.

In a previous work, we have developed an avatar able to join in laughing with its conversational partner [24]. However, the laughs produced by the virtual agent were not synthesized but selected from an audiovisual laughter database, using acoustic similarities to the conversational partner's laughs.

We strongly believe that both automatic laughter recognition/characterization and synthesis would benefit from a detailed phonetic transcription of laughter. On the recognition side, transcriptions can help classifying laughs, on a simple phonetic basis or via features easily computed once the phonetic segmentation is available (syllabic rhythm, exhalation and inhalation phases, acoustic evolution over laughter syllables or bouts, etc.). On the synthesis side, transcription enables approaches similar to those used in speech synthesis: training a system with the individual phonetic units and then synthesizing any consistent phonetic sequence.

In this paper, we present the phonetic annotation of the AVLaughterCycle database [23], which currently is the only large (1 hour of laughs) spontaneous laughter database to include audio, video and phonetic transcriptions. In addition, we use these phonetic transcriptions to study some factors of variability –

[1] Accuracy and Equal Error Rates cannot be directly compared. However, $1 - EER$ is a measure of the accuracy; with no guarantee it is the best the system can achieve.

the airflow direction and personal style –, which received few interest in previous works. The annotation process is explained in Section 2. Section 3 presents the most frequent phones[2] in exhalation and inhalation phases and shows differences in their duration. Section 4 focuses on individual differences in the phones used and in their durations. Finally, conclusions are given in Section 5. They include perspectives we consider with the large phonetically annotated database, which is the groundwork for further developments in the laughter processing field.

2 Annotation Scheme

We used the AVLaughterCycle database, which contains laughs from 24 subjects (9 females and 15 males) [23]. The female and male average ages were respectively 30 (standard deviation: 7.8) and 28 (standard deviation: 7.1). All subjects were participants of the eNTERFACE'09 Workshop in Genova (Italy). They came from various countries: Belgium (8), France (4), Italy (3), Canada (2), UK, Greece, Turkey, Kazakhstan, India, USA and South Korea (1 each). All subjects could speak English. Laughs were elicited with the help of a comedy video. The database consists of audio and video recordings, including facial motion tracking.

Laughs had previously been segmented on the basis of the audiovisual signal. In total, 1021 laughs have been segmented, for a total of 1 hour of spontaneous, hilarious laughs. The database and annotations are freely available on the website of the first author (`http://tcts.fpms.ac.be/~urbain`).

For the present work, one annotator labeled the 1021 laughs in phones in the Praat software [3]. Two annotation tracks have been used (see Figure 1). The first is used to transcribe the "phones"[3], according to the phonetic symbols defined by Ladefoged [12]. Diacritics (symbols added to a letter) have also been used to label voice quality (modal, creaky, breathy) or unusual ways of pronouncing a given phone (e.g. a voiceless vowel or a nasalized plosive), thereby leading to something that looks more like a narrow phonetic transcription of the database. Several sounds encountered in our data could not be found in the extended International Phonetic Alphabet. To describe them, similarly to previous works ([1, 5]), the following labels have been added: hum, cackle, groan, snore, vocal fry and grunt. Examples are available on the website of the first author.

Since the respiratory dynamics are important to process laughter and since the acoustics of laughter are different when inhaling and exhaling, the airflow phases are transcribed on the second annotation track. The airflow phases were segmented using only the audio.

Unsurprisingly, we have noticed that the phones constituting a laugh are often perceived differently when listening to the laugh as a whole than when analyzing each of its phones separately. As a matter of fact, although laughter episodes exhibit no strong semantic contrast (as opposed to words), they still

[2] The phonological notion of "phoneme" is not clearly defined for laughter; we prefer to use the word "phone" for the acoustic units found in our database.

[3] Note that we used only audio for this transcription, while laughter segmentation was done on the basis of both audio and video.

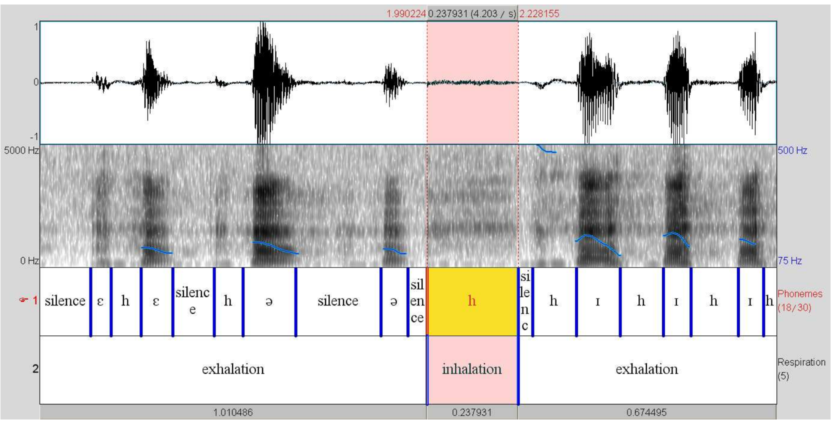

Fig. 1. Laughter annotation in Praat

obey strong phonotactic constraints (e.g. we will have the impression of *hahahaha* when actually listening to *haha-aha* because the first instance is more likely to happen). In addition, psychoacoustic effects are likely to influence our perception of continuous laughter, given its fast succession of sounds that can be highly contrasted in amplitude. In this work, we annotated laughter phones as they had been produced, rather than how they actually sounded, following a long tradition of articulatory phonetic transcription.

3 Laughter Phonetic Description

Out of the initial 1021 laughs, the 20 laughs involving speech and 4 short laughs labeled as only silence (i.e. they only had visual contributions) were discarded from our phonetic analysis, leaving 997 acoustic laughs. Excluding the silences outside acoustic laughs (as the laughs had been segmented with the help of visual cues, most of the times there are silences before the first phone and after the last phone), 17202 phones have been annotated: 15825 in exhalation phases and 1377 in inhalation phases. If we take diacritics into account[4], 196 phonetic labels appear in the database: 142 during exhalations and 54 during inhalations. This reinforces the idea that laughter is extremely variable.

For the sake of simplicity, the diacritics will not be considered in this paper. This reduces the number of labels to 124 (88 during exhalations, 36 during inhalations). The most frequent phonetic labels in exhalation and inhalation phases are respectively listed in Tables 1 and 2, with their average duration.

The outcomes of our annotation are mostly in line with previous findings ([1, 17, 19]). During exhalation phases, if we exclude silences that are extremely

[4] The following diacritics, showed here on the letter e, have been used: ẽ (nasalized), ḛ (creaky), e̤ (breathy), e̊ (voiceless), é (high tone).

Table 1. Most frequent phonetic labels in laughter exhalation phases

Label	Occurrences	Average duration (std)	Label	Occurrences	Average duration (std)
silence	4886	0.308s (0.427s)	ǀ	214	0.031s (0.032s)
h	2723	0.121s (0.068s)	x	176	0.228s (0.170s)
ə	1422	0.073s (0.044s)	ʌ	160	0.094s (0.066s)
ɐ	1373	0.082s (0.047s)	ħ	152	0.175s (0.085s)
ñ̇	839	0.210s (0.134s)	ɦ	135	0.175s (0.114s)
ɪ	741	0.077s (0.039s)	ɵ	109	0.116s (0.058s)
cackle	704	0.034s (0.024s)	k	102	0.048s (0.051s)
hum	639	0.077s (0.042s)	t	81	0.073s (0.035s)
ɛ	370	0.076s (0.035s)	grunt	81	0.126s (0.104s)
ʔ	269	0.027s (0.016s)	ʉ	79	0.093s (0.090s)

Table 2. Most frequent phonetic labels in laughter inhalation phases

Label	Occurrences	Average duration (std)	Label	Occurrences	Average duration (std)
h	640	0.305s (0.133s)	s	38	0.340s (0.141s)
ə	172	0.095s (0.059s)	ħ	24	0.340s (0.154s)
ñ̇	166	0.346s (0.170s)	t	23	0.049s (0.032s)
ɪ	108	0.097s (0.064s)	i	23	0.148s (0.058s)
ɦ	38	0.226s (0.121s)	ɛ	17	0.094s (0.039s)

frequent inside laughs, we obtained a large number of h-like phones (h, x, ɦ, ħ), and voiced parts are mainly central vowels (ə, ɐ, ɵ, ʉ). As stated in [5], but contested in [17], voiced segments can be abruptly ended by a glottal stop (ʔ).

We also found a lot of non-stereotypical laughter sounds. Nareal fricatives (ñ̇) are frequently used, mostly in short laughs with a closed mouth, in which a voiceless airflow going through the nose accompanies a smile. In addition, we have occurrences of non central vowels (ɪ, ɛ, ʌ), which were not found by Bachorowski et al.'s formant frequency analyses [1]. Our data also contains numerous cackles, hum-like sounds (close to vowels, but with a closed mouth), and grunts. More surprising is the presence of a large number of dental clicks (ǀ) and plosives (t, k) that generally take place at the beginning of sudden exhalation phases.

During inhalation phases, the most used phones are similar. Deep breath sounds (h, ñ̇, ɦ) are even more dominant. It can also be noticed that, except for t, the average duration of a phone is longer during inhalation phases than in exhalation phases. Student's *t-tests* show that the average duration in inhalation and exhalation is significantly different at a 99% confidence level ($p < 0.01$) for all the phones that appear in both Tables 1 and 2 (h, ə, ñ̇, ɪ, ɦ and ħ) except for t (no difference) and ɛ ($p = 0.22$). Over the whole database, the average phone duration for exhalation and inhalation phases is respectively $0.165s$ ($std : 0.266s$) and $0.245s$ ($std : 0.159s$). The difference is significant at a 99% confidence level.

Regarding the airflow phases, 1551 exhalation phases and 943 inhalation phases have been annotated. The average duration of exhalation and inhalation phases is respectively $1.69s$ ($std : 1.52s$) and $0.36s$ ($std : 0.15s$). No correlation has been found between the duration of an exhalation phase and the duration of its surrounding inhalations (correlations < 0.1). Table 3 shows the number of laughs presenting a given number of exhalation and inhalation phases.

Table 3. Number of laughs with a given number of exhalation and inhalation phases

N	Number of laughs having N exhalations	Number of laughs having N inhalations
0	1	462
1	733	353
2	156	105
3	54	39
4	26	18
≥ 5	27	20

Most of the laughs have only one "bout" (i.e. exhalation segment separated by inhalations) [21]. The number of inhalation phases is lower than the number of exhalations, meaning that most laughs are not concluded by an audible inhalation. In fact, only 38% of the laughs are ended by an audible inhalation.

4 Interpersonal Differences

We have already stated that the AVLaughterCycle database as a whole contains a wide range of phones, and that these phones have variable durations, influenced by the airflow direction. We will now present some figures corroborating the impression that laughter exhibits individual patterns. We will see that there are more individual differences in the sounds produced than in the duration of the segments. Since the number of subjects and phones are large, we cannot give an exhaustive analysis in this paper and will concentrate on a few examples.

4.1 Phones Used

Subjects used different sets of phones while laughing. The number of phones used per laugher ranges from 2 to 59, with a mean (and median) of 32 (*std* : 14.4). There are large inter-individual differences in the choice of phones. Most laughers are quite consistent from one laugh to another, in accordance to Chafe's statement that users have their "favorite laugh" [5]. Figure 2 displays, for the 5 subjects who laughed the most and the 7 most used exhalation labels (except silence), the individual phone probabilities (i.e. the number of instances of phone X by subject Y, divided by the total number of phones produced by Y). We can see that subject 6 typically uses h and ɐ. His laugh is quite stereotypical. This is not the case for other subjects. Subject 20 produces much more nasal sounds (ň̇ and hum) than others. The choice of the vowel is another difference between subjects: some laughers use up to 3 times more ə than ɐ, others do the opposite.

There are numerous other proofs of individual differences in the produced sounds that do not appear on the graph. For example, subject 14 is the only one

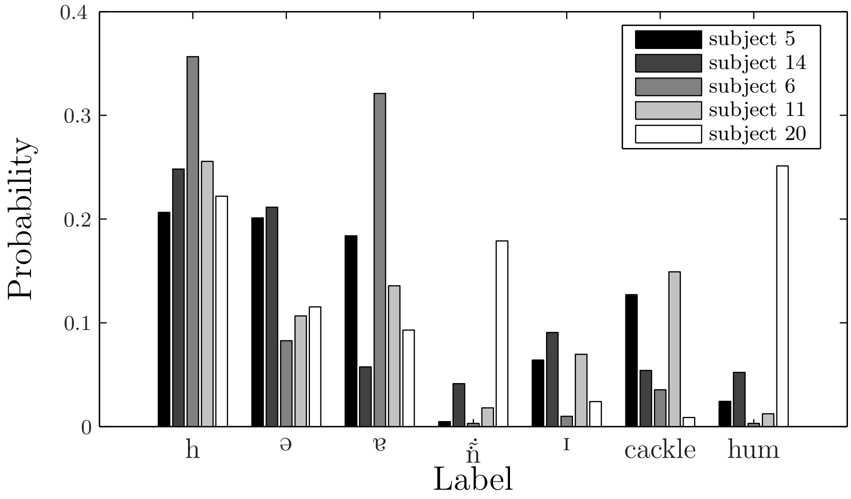

Fig. 2. Probabilities of the most used phones for the 5 subjects who laughed the most

to make a broad use of the phone m, which is present 23 times in her 48 laughs (generally at the end), while there are only 15 other instances of this phone in the database, produced by 11 different subjects. Subject 14 is also responsible for 87 of the 109 instances of the phone ɵ.

4.2 Phone and Airflow Phases Duration

The average duration of exhalation phones is similar for all subjects: slightly under 100ms for voiced phones, a bit larger for h-like sounds and nareal fricatives. There is a slightly larger individual variation for inhalation phones. Figure 3 represents the average duration of the 3 most frequent inhalation phones for all the subjects, with their corresponding standard deviations. No bar means that the subject did not produce the corresponding phone. We can see that there are some extreme values for all three phones, showing some individual influence over the length of inhalation phones.

Figure 4 shows the average durations (and standard deviations) of exhalation phases for all the subjects. We can notice some individual variability, but the large standard deviations prevent us from drawing strong conclusions. The average inhalation durations are similar for all the subjects. The large variability of the laughter phone and bout durations is in line with the findings in [1].

5 Conclusion and Further Work

In this paper, we have presented the phonetic annotation of a large laughter database. The AVLaughterCycle database and these annotations are freely available on the website of the first author (`http://tcts.fpms.ac.be/~urbain`).

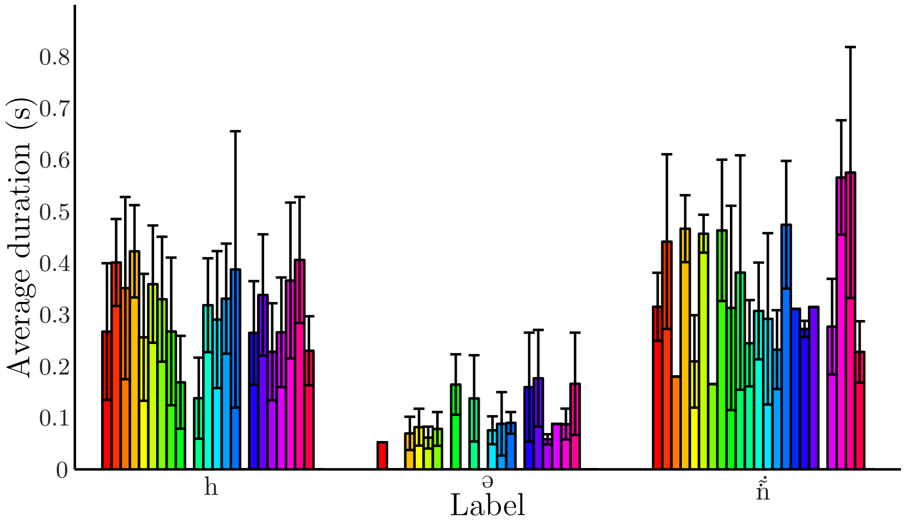

Fig. 3. Average duration of the most frequent inhalation phones, for all the subjects

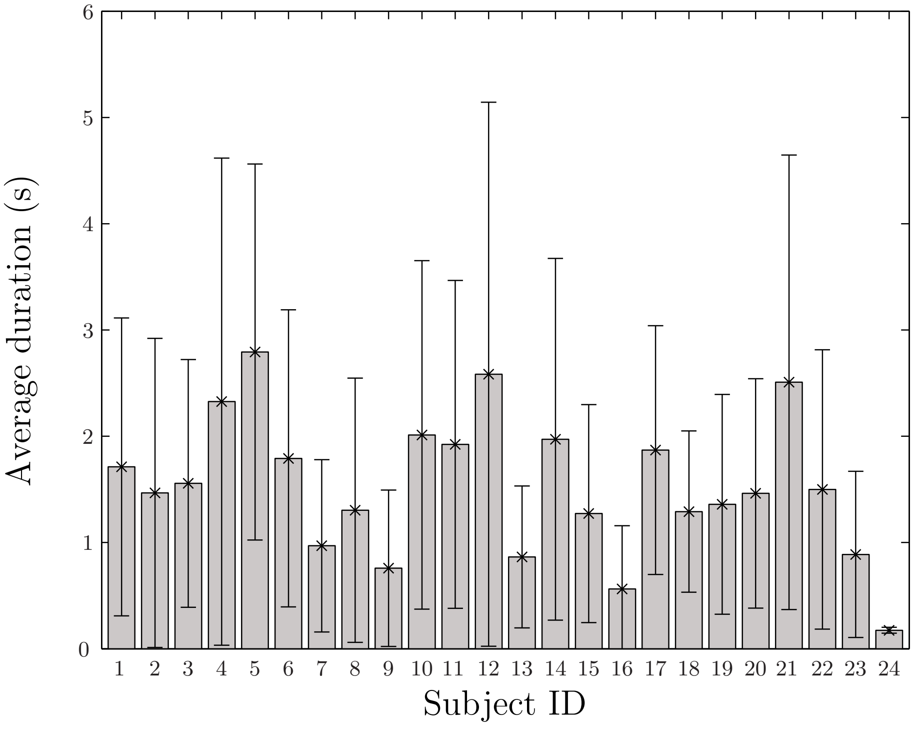

Fig. 4. Average duration of exhalation phases, for all the subjects

This large, phonetically annotated database can be used for a broad range of purposes. First, it can serve to study and describe laughter, its variability, and factors responsible for these variations. We have started this type of analyses in this paper, showing that 1) the airflow direction influences the phone duration; 2) individuals have their own favorite subset of phones for laughing; 3) the duration of laughter units (phones and airflow phases) can also vary with individuals. More acoustic features (fundamental frequency, formants, etc.) could be extracted and compared over phones or individuals. We are currently working on robust fundamental frequency estimation for laughter.

Second, since there is currently no standard of annotating laughter, we hope that this paper will be an important step toward this type of agreement. Among the other available laughter databases (for example the ICSI [9] and AMI [4] Meeting corpora), the AVLaughterCycle is unique given its audiovisual data – including facial motion tracking – and annotation.

Manual phonetic annotation is extremely time-consuming. One of our objectives is to develop automatic laughter phonetic transcription, going beyond current laughter recognition systems that consider at most two categories [15].

Such a phonetic transcription is crucial to natural laughter synthesis, for which a phonetic description of laughter will make it possible to use efficient speech synthesis approaches (e.g. unit selection [8] or parametric synthesis [20]) to develop text-to-laughter (or more accurately, labels-to-laughter) synthesis.

Combining these approaches, we aim to improve our AVLaughterCycle application [24], which consists in enabling a virtual agent to detect its conversational partner's laugh and answer with an appropriate, human-like laugh.

All these aspects will be addressed within the European FP7 FET project ILHAIRE starting in September 2011. In this project, not only the computing aspects of how to recognize, characterize, generate and synthesize laughter will be studied, but also the psychological foundations of this important signal (to avoid inappropriate laughs sounding rude to the user) as well as cultural differences.

References

1. Bachorowski, J.A., Smoski, M.J., Owren, M.J.: The acoustic features of human laughter. Journal of the Acoustical Society of America 110, 1581–1597 (2007)
2. Bennett, M.P., Lengacher, C.: Humour and laughter may influence health. III. Laughter and Health Outcomes. Evidence-based Complementary and Alternative Medicine 5(1), 37–40 (2008)
3. Boersma, P., Weenink, D.: Praat: doing phonetics by computer (version 5.2.11) [computer program]. `www.praat.org` (Retrieved on January 20, 2011)
4. Carletta, J.: Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus. Language Resources and Evaluation Journal 41(2), 181–190 (2007)
5. Chafe, W.: The Importance of not being earnest. The feeling behind laughter and humor., Consciousness & Emotion Book Series, vol. 3. John Benjamins Publishing Company, Amsterdam, The Nederlands, paperback 2009 edn. (2007)

6. Cox, T.: Laughter's secrets: faking it – the results. New Scientist (27 July 2010), http://www.newscientist.com/article/dn19227-laughters-secrets-faking-it--the-results.html
7. Esling, J.H.: States of the larynx in laughter. In: Proc. of the Interdisciplinary Workshop on the Phonetics of Laughter. pp. 15–20. Saarbrücken, Germany (2007)
8. Hunt, A., Black, A.: Unit selection in a concatenative speech synthesis system using a large speech database. In: icassp. pp. 373–376. IEEE (1996)
9. Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., et al.: The ICSI meeting corpus. In: Proc. of ICASSP'03. vol. 1, pp. I–364. IEEE, Hong-Kong (2003)
10. Kennedy, L., Ellis, D.: Laughter detection in meetings. In: NIST ICASSP 2004 Meeting Recognition Workshop. pp. 118–121. Montreal (2004)
11. Knox, M.T., Morgan, N., Mirghafori, N.: Getting the last laugh: automatic laughter segmentation in meetings. In: INTERSPEECH 2008. Brisbane, Australia (2008)
12. Ladefoged, P.: A course in phonetics. http://hctv.humnet.ucla.edu/departments/linguistics/VowelsandConsonants/course/chapter1/chapter1.html (Consulted on January 20, 2011)
13. Lasarcyk, E., Trouvain, J.: Imitating conversational laughter with an articulatory speech synthesis. In: Proc. of the Interdisciplinary Workshop on the Phonetics of Laughter. pp. 43–48. Saarbrücken, Germany (2007)
14. Mahony, D.L.: Is laughter the best medicine or any medicine at all? Eye on Psi Chi 4(3), 18–21 (Spring 2000)
15. Petridis, S., Pantic, M.: Is this joke really funny? Judging the mirth by audiovisual laughter analysis. In: Proc. of ICME'09. pp. 1444–1447. New York, USA (2009)
16. Pompino-Marschall, B., Kowal, S., O'Connell, D.C.: Some phonetic notes on emotion: laughter, interjections and weeping. In: Proc. of the Interdisciplinary Workshop on the Phonetics of Laughter. pp. 41–42. Saarbrücken, Germany (2007)
17. Ruch, W., Ekman, P.: The expressive pattern of laughter. In: Kaszniak, A. (ed.) Emotion, qualia and consciousness. World Scientific Publishers, Tokyo (2001)
18. Sundaram, S., Narayanan, S.: Automatic acoustic synthesis of human-like laughter. Journal of the Acoustical Society of America 121(1), 527–535 (January 2007)
19. Szameitat, D.P., Alter, K., Szameitat, A.J., Wildgruber, D., Sterr, A., Darwin, C.J.: Acoustic profiles of distinct emotional expressions in laughter. The Journal of the Acoustical Society of America 126(1), 354–366 (2009)
20. Tokuda, K., Zen, H., Black, A.: An HMM-based speech synthesis system applied to english. In: 2002 IEEE TTS Workshop. Santa Monica, California (2002)
21. Trouvain, J.: Segmenting phonetic units in laughter. In: Proc. of the 15th International Congress of Phonetic Sciences. pp. 2793–2796. Barcelona, Spain (2003)
22. Truong, K.P., van Leeuwen, D.A.: Evaluating automatic laughter segmentation in meetings using acoustic and acoustic-phonetic features. In: Proc. of the Interdisciplinary Workshop on the Phonetics of Laughter. Saarbrücken, Germany (2007)
23. Urbain, J., Bevacqua, E., Dutoit, T., Moinet, A., Niewiadomski, R., Pelachaud, C., Picart, B., Tilmanne, J., Wagner, J.: The AVLaughterCycle database. In: Proc. of LREC'10. Valletta, Malta (2010)
24. Urbain, J., Niewiadomski, R., Bevacqua, E., Dutoit, T., Moinet, A., Pelachaud, C., Picart, B., Tilmanne, J., Wagner, J.: AVLaughterCycle: Enabling a virtual agent to join in laughing with a conversational partner using a similarity-driven audiovisual laughter animation. JMUI 4(1), 47–58 (2010), special Issue: eNTERFACE'09