University of Mons

Doctoral School MUSICS

Signal Processing

P H D T H E S I S

to obtain the title of

PhD in Applied Sciences

of University of Mons

Specialty : Signal Processing

Defended by

Jérôme URBAIN

Acoustic Laughter Processing

Thesis Advisor: Thierry DUTOIT

prepared at University of Mons, Faculty of Engineering,
NUMEDIART Institute / TCTS Lab

defended on May 22, 2014

Jury :

Dr. Stéphane Dupont	-	Université de Mons
Pr. Thierry Dutoit	-	Université de Mons
Pr. Joël Hancq	-	Université de Mons
Pr. Marc Pirlot	-	Université de Mons
Dr. Jürgen Trouvain	-	Universität des Saarlandes
Pr. Hugo Van hamme	-	Katholieke Universiteit Leuven

to Joëlle

Il n'y a que les gens qui aiment rire qui sont sérieux. Les autres se prennent au sérieux.[1]

Jean Caplanne

[1] Free translation: "Only those who love to laugh are serious people. The others are taking themselves too seriously."

Abstract

This dissertation relates to acoustic laughter processing. The ultimate objective is the synthesis of natural-sounding laughs, but to achieve this goal we addressed most of the fields related to (engineering) acoustic laughter processing. First, we have tackled the problem of obtaining high quality laughter data, which is obviously critical for any subsequent study. Our own laughter database, the AVLaughterCycle database, has been recorded in the framework of this PhD Thesis and presents unique features. Second, we have investigated the acoustic analysis of laughter and proposed new annotation levels: phonetic transcriptions and arousal signals. Third, we have addressed the field of automatic characterization of laughter episodes, by designing methods to automatically compute phonetic transcriptions and estimate arousal values. Fourth, we have synthesized acoustic laughs and achieved naturalness scores outperforming the state-of-the-art. Finally, the analysis and synthesis methods have been implemented in human-computer applications, which were enhanced by the possibility to detect and express affect through laughter.

Another significant contribution of this dissertation is the extensive state-of-the-art gathered for all these fields, prior to explaining our own developments and suggesting future works. In addition, original analyses of laughter characteristics have been carried out with the help of phonetic transcriptions and arousal annotations. We have for instance demonstrated that laughter phonemes are shorter during laughter exhalation phases than inhalation phases, and that different subjects tend to use different sets of phonemes when laughing.

From a technical side, the biggest achievements have been obtained with the help of Hidden Markov Models (HMMs), by adapting algorithms which had initially been designed for speech processing. HMMs have been used both for producing phonetic transcriptions and for synthesizing acoustic laughs. For the estimation of laughter arousal, Multi-Layer Perceptrons have been trained. Regarding laughter synthesis, several methods are compared and evaluated, among others different vocoders or the training of laughter synthesis via automatically obtained phonetic transcriptions. An important contribution of this dissertation is the development of a method to automatically produce phonetic transcriptions from arousal curves. This method was inspired by concatenative speech synthesis and was proven, through evaluation, to produce acoustic laughs that are as natural as laughs synthesized from lower-level information (phonetic transcriptions). Nevertheless, synthesized laughs are still far from actual human laughs in terms of naturalness, and suggestions of future work are proposed to further improve HMM-based laughter synthesis.

Keywords: laughter processing, laughter analysis, laughter synthesis, laughter databases, Hidden Markov Models.

Remerciements

Je tiens en premier lieu à remercier mon promoteur, le Professeur Thierry Dutoit, sans qui cette thèse n'aurait jamais vu le jour. Il m'a accompagné durant toutes ces années et a orienté mes travaux avec des conseils toujours avisés et des encouragements toujours optimistes.

Je souhaite également remercier les membres de mon comité d'accompagnement, qui ont accepté de suivre ma thèse et de me faire part de leurs avis, remarques et conseils pour définir les directions à explorer aux cours de mes recherches: le Professeur Hugo Van hamme de la Katholieke Universiteit Leuven (KUL), les Professeurs Joël Hancq et Marc Pirlot, ainsi que le Docteur Stéphane Dupont, de l'Université de Mons. Je tiens également à remercier le Docteur Jürgen Trouvain, de la Universität des Saarlandes, qui a rejoint ce comité d'accompagnement pour former le jury de cette thèse. Merci à tous les membres de ce jury pour leur lecture attentive du manuscript, leurs suggestions pour améliorer celui-ci et les discussions intéressantes sur la portée de ce travail et les suites qui peuvent lui être données.

Durant ces dernières années, j'ai eu la chance de travailler sur le projet Européen ILHAIRE qui est en lien direct avec le sujet de cette thèse. Ce projet est coordonné par le Docteur Stéphane Dupont, avec lequel j'ai dès lors beaucoup discuté de mes travaux depuis septembre 2011. Je souhaite vivement le remercier de son aide et de sa compréhension pour alléger un peu ma charge de travail ces derniers mois afin que je puisse rédiger ce manuscript.

Je voudrais évidemment remercier mes parents, sans qui je n'aurais moi-même pas vu le jour... et qui m'ont toujours soutenu et encouragé dans mes démarches. Ce sont eux qui m'ont transmis les valeurs qui collent à ma personnalité et qui m'ont permis de rencontrer toutes les personnes intéressantes citées (ou pas) dans ces remerciements. Je souhaite également remercier tous les proches qui m'ont non seulement soutenu et écouté au fil des ans, mais m'ont aussi permis de me changer les idées et de cultiver le goût de rire. Merci à Marine, Gilles, Anne-Claire, Anne, François, Noémi, Samuel, Catherine, mes beaux-parents, ma famille au sens large. Merci à mes amis de longue date, en particulier Alex, Marie, Charlie, Laurent et bien évidemment la Bande, que je considère un peu comme mon club de rire personnel. Merci à Rose et Alexia, dont je ne suis pas sûr qu'elles m'écoutent vraiment, mais dont l'énergie, les (sou)rires et les mots (mini-claque, tortue ou petit tototame, en vrac) ont un puissant effet sur mon moral et mes zygomatiques; merci à Agathe et Luce également de m'avoir fait redécouvrir la sieste de manière si agréable. Elles ont toutes bien compris que j'étais nettement plus intéressé par les rires que par les pleurs et je leur en suis extrêmement reconnaissant.

Merci à tout le labo de TCTS et aux Professeurs qui le dirigent de m'avoir offert un cadre de travail si plaisant. Un merci particulier à Radhwan, Matthieu, Nicolas R. et Thomas D. pour les discussions animées sur le sport et autres sujets à controverse. Un immense merci à Nicolas D., Alexis, Benjamin, Hüseyin, Onur et Nicolas R. d'avoir relu attentivement des morceaux de cette thèse. Mille mercis à Hüseyin d'avoir

attaqué avec moi la synthèse de rire—c'est tellement plus agréable d'avoir un collègue travaillant sur la même thématique, surtout lorsqu'il est si sympatique, disponible et compétent—et merci à Kévin et Nicolas D. d'avoir rejoint le mouvement! Merci à ce même Nicolas d'avoir spontanément proposé d'encadrer quelques travaux à mes côtés, non seulement pour l'expertise qu'il y apporte mais aussi pour dégager un peu de temps pour finaliser ma thèse. Merci à Alexis, mon Joker, de m'avoir si souvent et si gentiment dépanné sur de nombreux problèmes informatiques (et mentaux, mais n'en parlons pas ici...). Merci également à Christian pour son aide répétée avec MediaCycle. Merci aussi à Maria, Matei, Anderson et Loïc pour les nombreuses sessions de rire collectif. Un immense merci à Nathalie qui s'occupe avec brio de tout ce qui sort de nos recherches proprement dites. Quel confort que tout l'administratif soit toujours réglé si vite et facilement (pour moi).

Merci aux collègues du projet ILHAIRE pour les réflexions intéressantes et les nouvelles perspectives qu'ils apportent à ces travaux. En particulier, merci à Jenny et Radek pour les excellents moments passés ensemble, notamment lors du développement de l'application Laugh Machine. Merci également à Bajibabu Bollepalli et Tuomo Raitio pour leur aide dans la comparaison de vocodeurs pour la synthèse de rire.

Merci à tous ceux qui ont participé aux tests perceptifs indispensables à l'évaluation des méthodes développées dans ce travail et ont sollicité d'autres participants. Je sais ce que vous avez enduré et cela me touche beaucoup que vous ayez consacré ce temps à mes recherches.

Last but not least, merci à Joëlle, mon Buddy chou, qui a rempli pratiquement tous les rôles décrits ci-dessus. Elle m'a apporté un soutien sans faille et une oreille attentive tout au long de cette thèse. Elle m'a procuré des conseils précieux et a été fortement mise à contribution pour relire ce manuscript. Mais en-dehors de ses contributions dans mon travail, c'est évidemment pour ce qu'elle m'apporte au quotidien que je souhaite principalement la remercier...

Contents

List of Figures

List of Tables

List of acronyms

2D	two dimensions ... 31
3D	three dimensions ... 30
AC-PEAK	AutoCorrelation Peak ... 91
ALISP	Automatic Language Independent Speech Processing ... 94
AMI	Augmented Multiparty Interaction ... 22
AUC-ROC	Area Under the Receiver Operator Characteristic (ROC) Curve ... 88
AVEC	2011 Audio/Visual Emotion Challenge ... 24
AVIC	AudioVisual Interest Corpus ... 23
AVLC	AVLaughterCycle ... xv
CGN	Corpus Gesproken Nederlands ... 18
DSM	Deterministic plus Stochastic Model ... 132
EER	Equal Error Rate ... 88
ESN	Echo State Network ... 92
f_0	fundamental frequency ... xv
F1	Frequency of the first formant ... 103
F2	Frequency of the second formant ... 103
F3	Frequency of the third formant
F4	Frequency of the fourth formant ... 50
F5	Frequency of the fifth formant ... 50
FN	False Negatives ... 86
FP	False Positives ... 86
FPR	False Positive Rate ... 87
FPS	Frames Per Second ... 31
GMM	Gaussian Mixture Model ... 89
HCRF	Hidden Conditional Random Field ... 90
HMM	Hidden Markov Model ... 13
HNR	Harmonic to Noise Ratio ... 77
HR	Hit Rate ... 113
HSD	Honestly Significant Difference ... 137
HTK	HMM Toolkit ... 108

Chapter 1

Introduction

Contents

Laughter is universal. There is no reported culture where laughter is absent [Devillers & Vidrascu 2007]. It is estimated to be seven million years old [Ruch & Ekman 2001], used as a communicative and expressive signal before the apparition of speech. Still now, laughter precedes speech: newborns laugh after a few months, later than smiling but long before uttering their first words [Chafe 2007]. But the apparition of speech does not annihilate laughter: that process lies in our genes. Studies have shown that all laughter functions are present and functional at birth. Indeed laughter is easily observable in deaf-blind children [Ruch & Ekman 2001]. It is important to note that even though speech and laughter often co-occur, these are totally separated processes. Mute people do laugh as well. Laughter transmits—voluntarily or not—our feelings and has important social values. It is communicative and generally helps to cheer up our minds. When laughing spontaneously, our awareness state is altered and all our thoughts go to the object of the laugh.

For these reasons and due to the increasing development of natural user interfaces in human-computer interaction, interest for automatic laughter recognition and synthesis devices has grown in recent years. However laughter has long been a neglected research field and many aspects remain unclear regarding its cerebral production

mechanisms or the acoustic patterns enabling humans to identify laughter without any doubt.

But what exactly is laughter? The concept is so obvious for all of us that we would probably have difficulties explaining it without imitating it. The definition given by the Free Dictionary [Free Dictionary 2008] for "to laugh" is:

> *"To express certain emotions, especially mirth or delight, by a series of spontaneous, usually unarticulated sounds often accompanied by corresponding facial and bodily movements".*

This definition is fuzzy, reflecting the large variability of laughter episodes, their miscellaneous meanings and the strong links with emotions. But this vague definition does not embrace the whole phenomenon. Laughter is not always spontaneous—it can be intentionally shaped—, does not all the time reflect positive feelings and can be composed of only one sound[1].

In a nutshell[2], this work aims at improving acoustic laughter processing. We will not consider visual and physiological correlates of laughter. Natural-sounding and easily controllable acoustic laughter synthesis can be considered as the ultimate objective of this PhD Thesis. To achieve this goal, preliminary steps in other fields than synthesis were required. In consequence, this Thesis does not concentrate on laughter synthesis but presents contributions along different aspects of laughter processing: database building and annotation, description, automatic characterization, generation and synthesis. Here is an overview of our developments in these fields:

- Firstly databases are needed to study the phenomenon. We will present our contribution in this field: the recording of a unique audiovisual database of spontaneous laughs from 24 subjects. The database is segmented in laughs and freely available.

- Secondly we investigated convenient ways to characterize laughs, which resulted in phonetic annotations of the laughs from the database as well as their labeling in intensity. These annotations are provided with the database. To the extent of our knowledge, nothing similar exists.

- Thirdly we developed methods to automatically compute laughter phonetic transcriptions and intensity curves. As the annotation of these dimensions was already new, their automatic estimation is obviously something that had not been addressed before.

- Finally laughter synthesis was investigated, with the objective to obtain human-like laughs from a simple symbolic transcription. Methods similar to speech synthesis have been explored, taking a phonetic transcription as input. We

[1] Some people tend to consider laughs with only one sound as smiles, while others, including the author, think that a smile becomes a laugh as soon as there is an audible contribution.

[2] The structure of this document will be recalled in Section 1.5.

have also developed a method to automatically generate such phonetic transcriptions from intensity curves. Hence the synthesis can be driven by these intensity curves. We have also explored the possibility to train laughter synthesis on automatically estimated phonetic transcriptions, so as to avoid the time-consuming task of manual phonetic transcriptions.

The state-of-the-art was advanced in each of these fields, with innovative works. The Thesis is providing the first answers, but also opening research paths in all of these fields. We hope that they will be further investigated in the future.

Before addressing these topics, we would like to continue introducing laughter by presenting the mechanisms of laughter production, without going into acoustic details since they will be described in Chapter 3, as well as some dimensions that will receive no attention in the following chapters but are useful to appreciate the context and importance of laughter processing: laughter social aspects and health outcomes.

1.1 Laughter production

The sounds of laughter are produced by the same organs as speech, the so-called vocal organs or vocal apparatus. The vocal organs include the lungs, the larynx where the vocal folds are located, the pharynx, the jaws, the tongue, the teeth, the lips and the oral and nasal cavities. The vocal apparatus is illustrated on Figure 1.1 (except for the lungs). Chafe [Chafe 2007] states that "*Laughter consists of sudden, spasmodic expulsions of air from the lungs*". These forceful "pulses", caused by jerky contractions of the diaphragm and abdominal muscles [Ruch & Ekman 2001], go up in the larynx, where they can be modulated by the periodic vibration of the vocal folds. This periodic vibration of the vocal folds is called voicing, and makes the difference between the consonants *f* (unvoiced) and *v* (voiced) or *s* (unvoiced) and *z* (voiced) in spoken English. If the vocal folds are relaxed, the pulse is still somehow modified by laryngeal friction [Chafe 2007], but remains unvoiced. Then the pulse is further modified by the tongue and lips, shaping the vowels and consonants of speech. However, unlike in speech, articulation is limited in spontaneous laughter and usually the sounds are produced with the tongue in a resting neutral position [Ruch & Ekman 2001] and the mouth widely open. This generates h-like consonants and central vowels, close to schwa (ə). But the mouth can also be closed, in which case the pulse will escape through the nose, producing an m-like sound if the resonance occurs in the nasal cavity or a grunt-like sound if the turbulence takes place lower, in the oral or laryngeal cavities [Bachorowski *et al.* 2001].

As we have just seen, the organs of the vocal tract (laryngeal cavity, pharynx, oral and nasal cavities, teeth and lips) have a filtering effect on the acoustic vibration emerging from the vocal folds (periodic or not). A model for speech production therefore approximates speech by two contributions: a) the acoustic wave emerging from vocal folds, which is called the *source* or *excitation* and excites b) the *filter* formed by the vocal tract. The model is called *source-filter* model of speech production

and is widely used in speech processing with the assumption of independence between the source and the filter.

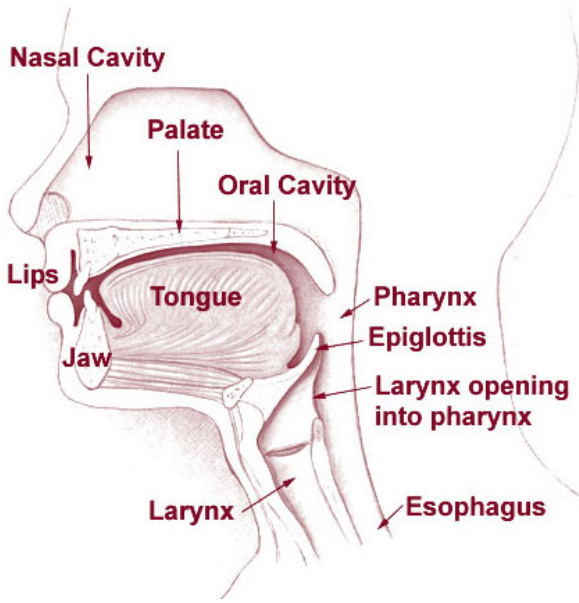

Figure 1.1: Head and neck overview (image produced by the US National Institutes of Health).

Laughter also involves facial movements. It is unusual to laugh naturally without smiling [Chafe 2007]. Exhilaration laughter is accompanied by the "Duchenne" smile [Ruch 1993], referring to the "*joint contraction of the zygomatic major and orbicularis oculi muscles (pulling the lip corners backwards and upwards and raising the cheeks causing eye winkles, respectively)*" [Ruch & Ekman 2001]. All the outcomes of the Duchenne display are however not always present in laughter [Chafe 2007]. For instance, differences have been noticed around the eye region between spontaneous and voluntary laughter [Ruch & Ekman 2001].

Laughter can be accompanied by other body movements. Movements of the trunk and limbs can appear, as well as changes in posture [Ruch 1993]. Ruch also reports vibrations of the trunk and shoulders, due to forced respiration movements of the diaphragm and abdominal muscles.

Laughter is the combination of all these manifestations, as expressed by Fry [Fry 1994]:

> *(...) I believe that we do not laugh merely with our lungs, or chest muscles, or diaphragm, or as a result of a stimulation of our cardiovascular activity. I believe that we laugh with our whole physical being.*

Due to its implications on the whole physical being, laughter is hypothesized by Chafe [Chafe 2007] to prevent the laugher to simultaneously perform any physical or serious mental activity. Chafe considers laughter as a manifestation of a particular emotion which he calls *the feeling of nonseriousness*, which is named "exhilaration" by Ruch [Ruch 1993]. In addition, as the laugh is audible and visible, it is also signaling conversational participants that the laugher is experiencing this feeling, which has strong communicative and social implications. As we will see in the following section, this relationship between social context and laughter is bidirectional: social context also has an influence on laughter.

1.2 Social aspects

Laughter is essential in human communication. It conveys information about someone's emotional state, her/his involvement in the conversation and elicits emotions to its listeners. While laughter is generally associated to positive mood, it can also express negative feelings such as disappointment, embarrassment or stress [Devillers & Vidrascu 2007]. Irony and mockery are other situations where laughter can convey a negative meaning. In consequence, there is no direct mapping between laughter and one particular emotion.

Due to its important communicative aspects, laughter is influenced by the social context. Firstly we are much more likely to laugh when we are not alone [Glenn 2003]. Moreover laughter is communicative—the odds of laughing increase if surrounding people are laughing—and even self-communicative: when we have laughed, we are more likely to laugh again in the near future. This is well-known by humorists: the most difficult part of a one-man show is the beginning; when people have started to laugh it is easier to keep entertaining them. However this communicative property of joyful laughter is not true for gelotophobes [Ruch 1997], i.e. people who are excessively fearful of being laughed at. These subjects have been shown to persistently interpret laughs uttered by others negatively, as rude and offensive acts [Ruch & Proyer 2009].

Secondly the way we laugh depends on whom we laugh with. This was shown by Campbell [Campbell 2007] from an analysis of telephone conversations: a neural network trained with acoustic features of laughs could distinguish the gender (with 62.5 to 67.7% accuracy) and the origin (Chinese or English, with 67.5 to 70% accuracy) of the conversational partner. This confirms everyday behavior: even if we basically feel the same emotions, we will for example not laugh at a business meeting the same way as we do with friends. In addition laughter can be spontaneous or controlled to express a desired meaning [Devillers & Vidrascu 2007].

Furthermore interaction between speech and laughter is interesting. While it is considered as impolite to speak at the same time as a conversational partner, laughing simultaneously is interpreted oppositely. Politeness brings us to laugh when others do. A very basic laughter recognition system in meetings could thus simply count the proportion of people vocalizing simultaneously. Above a certain threshold, it would

be very unlikely that people are talking[3]. It is also surprising to note that laughter does not interrupt speech [Devillers & Vidrascu 2007]. In natural conversations, a lot of overlapping speech and laughs are found, and a given speaker is often doing both simultaneously. The phenomenon is called *speech-laugh* and will be further detailed in Section 3.1.3.

Trouvain noticed that laughter was especially used when conversational partners did not know each other [Trouvain 2001]. We do not have to know people to laugh with them and this might be the start of a conversation. As Victor Borge said, "*laughter is the shortest distance between two people*". People looking happy, which is the case of laughers, are more attractive and gain in self-confidence.

Several researchers have investigated the role of laughter in conversations and realized that laughter is not solely a response to humor. Laughter is used in the case of undesirable or abnormal situations, to mitigate the unpleasantness and help pursuing the conversation in a more enjoyable way [Chafe 2007]. We also simply use laughter when we are nervous or when other people are laughing [Glenn 2003]. Laughter is also important for indicating possibilities to take the turn in conversations.

Glenn [Glenn 2003] summarizes the social impact of laughter as follows (pages 29-31):

> *As this suggests, laughter proves important socially as a means to show affiliation with others. To laugh when someone else has done something humorous, laughed first, or otherwise indicated a nonserious orientation provides a way to display like-mindedness. Similarly, one may laugh first in order to provide the co-participant(s) the opportunity to do the same. (...) Because of its ability to show (and produce) affiliation, laughter proves particularly useful in situations of embarrassment, discomfort, or anxiety. (...) laughter may also contribute to interactional disaffiliation. Laughing at someone may demonstrate a lack of sympathy, consideration, or alignment. Laughter may hurt and may contribute to feelings of hostility or embarrassment. The same laughter that promotes in-group solidarity may be done at the expense of outsiders, thus accomplishing simultaneously both affiliation and disaffiliation (...) More deeply, it may contribute to perpetuating negative attitudes, stereotypes, and temptations to denigrate or dismiss individuals or entire groups of people. These two outcomes viewed together—bonding people while disaffiliating them from others—make laughter a powerful and even subversive social tool.*

Some social contexts can also inhibit laughter. For example, it is not socially acceptable to laugh during funerals, classic concerts or when a Professor is mispronouncing something in a funny way without noticing it[4]. In some cases, not joining

[3] As pointed out by one of the reviewers of this dissertation—but not the author's wife—people vocalizing simultaneously could also be arguing... Politeness is indeed not always present in human communication, but the point is still there.

[4] I cannot be the only one to have suffered from trying to repress giggles in these kinds of situations, can I?

in laughter is the way to show affiliation, for example when someone is telling about her/his troubles or pains [Glenn 2003]. The majority of social contexts however tolerate laughter, or even favor it, as most of the time the underlying feeling of nonseriousness is enjoyable. This supports the creation of humor, which sole purpose is to elicit this feeling. Most of us will agree that laughter is pleasant and in addition it is also thought to have impacts on health, as will be discussed in the next section.

1.3 Laughter and health

1.3.1 Preliminary remarks

A lot is said about the healthy benefits of laughter. In this section we will try to summarize[5] the results and the evidence supporting this belief. Reviews of laughter-related articles from Martin [Martin 2001, Martin & Lefcourt 2004] and Bennett and Lengacher [Bennett & Lengacher 2006a, Bennett & Lengacher 2006b, Bennett & Lengacher 2008, Bennett & Lengacher 2009] show that this idea is largely unfounded, as solid evidence is still lacking. For several aspects, parallel studies have had opposite conclusions.

Martin [Martin 2001] enumerates four mechanisms by which laughter and humor may have a positive impact on health:

1. Physiological changes in various body parameters (blood pressure, heart rate, respiration rate, muscle tone, etc.) could have beneficial effects. If this is true, the consequent benefits are caused by laughter only.

2. Positive emotional states accompanying laughter and humor could be good for health. In this case, the benefits would not be due solely to humor and laughter, but to any method leading to enhanced mood.

3. Health could be indirectly enhanced through the inhibition of stress. Here the crucial variable would probably be the sense of humor (defined in this case as "the tendency to use humor as a coping strategy in daily life" by Martin) rather than laughter.

4. People with a high sense of humor could benefit from a richer social life. Here the important factor would be the capacity to use humor to improve/facilitate social relationships.

The methodology of the vast majority of the experiments linking laughter to impacts of health is imperfect on one or several of the following points [Martin 2001, Bennett & Lengacher 2009]:

[5] Detailed descriptions go way beyond both the scope of the present dissertation and my knowledge, as I have definitely no expertise in the possible roles of body substances with names as strange as norepinephrine, Salivary Immunoglobulin A (SIgA) or suppressor-cytotoxic T cells.

- Being conducted on very small sample sizes. This can partially explain why different studies, done with different groups, lead to contradictory results.

- Failing to include appropriately-designed and diverse control groups to assess on which of the aforementioned four dimensions laughter and/or humor are beneficial compared to other "treatments".

- Lacking experimental measures, such as the perceived funniness, interest or boredom of the stimuli, as wells as the frequency and intensity of laughs. Without these measures, and given the imperfect control groups, it is generally hard to understand what created a difference (or not) between groups: is it really laughter, or is it that the group was distracted, etc.?

For these reasons, evidence of the healthy effects of laughter is still weak. Nevertheless the studies have not proven that laughter does not have a positive impact on health either, nor, worse, that it would have negative effects. This last point must be however nuanced, as Ferner and Aronson's recent review [Ferner & Aronson 2013] is listing both the physical benefits and harms that have been reported for laughter. The list of negative effects contains around 15 items, which in my opinion look quite amusing or specific to a particular population (e.g., smokers, people suffering from asthma, etc.) and/or anecdotal with respect to the positive effects[6]. Troubles that affected one or few laugher(s) are hard to compare to the positive effects enjoyed by the vast majority. In the end "the laughter benefit-harm balance is probably favorable" [Ferner & Aronson 2013]. Nevertheless it is not our intention to hide anything to you, so here is the list of harms that can be provoked by laughter [Ferner & Aronson 2013][7]: weakened resolve and promotion of band preference; inhalation of foreign bodies like gum drops or popcorn; increased dissemination of infection due to the spread exhalation air flows; syncope; arrhythmias; cardiac rupture; stroke; asthma attacks; pneumothorax; interlobular emphysema; cataplexy; headaches; protruding hernia; jaw dislocation; incontinence.

It must also be noted that a lot of studies focus on the importance of the sense of humor, which we will not address here[8]. In addition, and related to the methodological problems identified above, most experiments consist in having people attending some humorous stimulus (e.g., watching a humorous video). It is hypothesized to provoke laughter but generally not measured. The conclusions should often carefully be related to "the effect of a humorous stimulus" rather than "the impact of laughter". Below are the main health consequences that have been related to laughter.

[6] For example the review mentions one case of a woman suffering from long QT syndrome (a heart disorder) who died during intense laughter. This is of course a dramatic consequence, but it occurred to me that people unfortunately die in a range of situations (some die during sleep, would then sleeping be risky?).

[7] But be careful, reading the list could make you laugh, which can be dangerous.

[8] The interested readers will find plenty of information in the mentioned reviews from Martin and Bennett and Lengacher.

1.3.2 Stress reduction

From a psychological point of view, several studies support the idea that humor (and laughter) improves mental health. An experiment conducted by Yovetich et al. [Yovetich *et al.* 1990] showed that humor helps decreasing self-reported anxiety. However humor did not help moderating the physiological correlate of anxiety—increased heart rate. Conclusions are difficult to draw since laughter itself increases heart rate. Berk et al. [Berk *et al.* 1989] found that laughter decreases stress hormones. Laughter could thus be used to lower the effects of stress, which is known to have physiological consequences. Indeed it affects the production and release of growth hormone and insulin among others and disturbs the levels of neurotransmitters and various cells in the immune system [Bennett & Lengacher 2006a]. However Martin [Martin 2001] and Bennett and Lengacher [Bennett & Lengacher 2008] claim that the evidences linking laughter to a reduction of stress hormones are weak and further experiments should be conducted to draw firm conclusions.

Exposure to humorous stimulus was shown to lower self-reported anxiety, although the physiological correlates of anxiety were not modified [Bennett & Lengacher 2009].

1.3.3 Mood changes

Neuhoff et al. [Neuhoff & Schaefer 2002] have found that smiling and laughing did significantly improve mood compared to howling. Furthermore laughing had a larger effect than smiling.

1.3.4 Physiological outcomes

Laughter has also been proven to produce direct physiological outcomes. Fry [Fry 1994] conducted several experiments to measure them. As already mentioned, heart rate is increased when we laugh. Blood circulation and oxygen consumption are increased [Bennett & Lengacher 2008], which reduces the risk of heart diseases [Club de rire Asbl 2008]. In addition laughter causes episodes of deep breathing, rises in blood pressure and modifications of muscle tone, as some muscle groups are activated by laughter. Immediately after laughing, heart rate, muscle activity, blood pressure and respiratory rate drop and a general relaxation state is reached. This relaxation can last up to 45 minutes [Bennett & Lengacher 2008]. Some of these effects—increased heart-rate, respiratory rate and oxygen consumption—are similar to some outcomes of aerobic exercises and this leads to think that laughing could be a good substitute for aerobic exercise. In the same vein, it was reported that 15 minutes of genuine laughter consumed an extra 40 kcal compared to regular activities [Ferner & Aronson 2013]. However the substitution of physical exercise would only be effective if intense laughter could be sustained for long periods, which is hard to achieve. Other experiments reported no change in blood pressure and heat rate after laughter, although a relaxation control group had the expected changes [Martin 2001]. Hence these results do not support the hypothesis that laughter lowers the levels of

heart rate and blood pressure (relaxation period following laughter) and it is difficult to conclude on the physiological impact of laughter without further evidence.

1.3.5 Immune system

There exist several measures to evaluate the strength of the immune system. Two of them have been investigated in humorous experiments: the level of Salivary Immunoglobulin A (SIgA) and the activity of Natural Killer cells.

1.3.5.1 Salivary Immunoglobulin A

In general—although imperfections in the methodology require to take results with caution—it was found that SIgA was increased after exposure to a humorous stimulus [Martin 2001, Martin & Lefcourt 2004, Bennett & Lengacher 2009]. The effect of laughter is unknown as laughter occurrence was not reported. Furthermore the use of SIgA as a measure of immune function is contested [Bennett & Lengacher 2009].

1.3.5.2 Natural Killer cells

Measurement of Natural Killer cell activity in response to humorous stimuli has first lead to contrasted results [Martin 2001, Bennett & Lengacher 2009], and once more the lack of appropriate control groups limits the extent of the results. Nevertheless, one study from Bennett et al. [Bennett *et al.* 2003] has shown that people who laughed out loud during the experiment had a significant increase in immune functions but the long-term effects are still unknown.

1.3.6 Pain reduction

Laughter has the power to raise discomfort threshold [Mahony 2000]. It helps forgetting or lowering pain. Nevertheless it does not seem that the pain reduction effects are a particular property of laughter, as any positive affect introduced (amusement, relaxation) or even negative affects with strong arousal (e.g., watching a tragedy video) yield the same effects [Mahony 2000, Martin 2001]. Even more the effect has been related to smiles rather than laughter (fake laughter even has a counterproductive effect) [Martin & Lefcourt 2004]. To conclude this paragraph about pain reduction, no change was reported on the level of beta-endorphines, contrary to general beliefs [Martin 2001, Bennett & Lengacher 2009].

The above statements explain why laughter is one of the most popular complementary therapies used by patients suffering from serious illnesses (such as cancer) to manage stress, maintain hope and bear pain. Of course, laughter is not the only means to reach these positive effects: relaxation, raised pain tolerance or enhanced mood can also be obtained through meditation, prayer, listening to music, reading, watching a video, relaxation and breathing exercises [Mahony 2000], which are also

used as complementary therapies. But laughter is a pleasant way to obtain these effects.

1.3.7 Miscellanous effects on health disorders

Reviewing recent articles, Ferner and Aronson [Ferner & Aronson 2013] report that laughter lowers the risks of myocardial infection and yields to improved lung functions in patients with chronic obtrusive pulmonary disease. In addition it was shown that the fertility of women undergoing in vitro fecundation was significantly increased if they were entertained by a clown dressed as a chef (compared to no entertainment).

1.3.8 Conclusion on the health benefits of laughter

To conclude this section, we can say that there is only little empirical evidence to support the idea that laughter can improve physical health. Most experiments did not lead to solid conclusions on one side (laughter has healthy effects) or the other. Some possible harms caused by laughter have even been reported, even though they have a limited extent. Large population studies on sense of humor and cheerfulness have even surprisingly revealed that cheerful people die at younger ages than less cheerful people [Martin & Lefcourt 2004]. This might be due to over-optimism in cheerful people, who are less concerned or aware about their health disorders[9].

Laughing clearly has an impact on well-being. It is fun and is at least beneficial psychologically and socially. Individuals with a high sense of humor do not seem to have objectively a better health and do not report more healthy life habits, but they are (subjectively) more satisfied with their health [Martin & Lefcourt 2004]. In consequence, laughter clubs, where people gather to practice laughter together, have been created and their success is growing. International laughter manifestations are organized each year to promote and enjoy laughter. To take advantage of the fun and communicative properties of laughter, a laughter chain has also been launched on Skype [Skype Communications 2009]: people watch their predecessors laughing and join the movement through the world-wide web.

1.4 Motivations of this Thesis

Laughter gives information about its producer's feelings. It is essential in human communication. The first applications of laughter recognition and synthesis are thus to enrich the communication between humans and machines. Recognizing some of the emotions of the speaker through laughter would be beneficial for virtual agents who could react appropriately, for automatic speech recognition in call centers, for people with sensory deficiencies (deaf, blind), etc. Being able to produce human-like

[9] Nevertheless longevity is not the ultimate goal in life, and to cite this famous quote (the original author seems to be unknown): *Life is not measured by the number of breaths we take but by the moments that take our breath away.*

laughs would add a new dimension to speech synthesis and communicative agents, fields which currently see a lot of research to increase the naturalness of the voices and the expression of emotions. Human-computer interaction systems equipped with good laughter recognition/synthesis would seem more natural as well.

Since laughter is communicative, it can also be used to elicit and maintain laughter and in consequence drive the user towards desired emotions. This can be useful to enrich the experience when visiting an artistic installation, to improve laughter therapies or simply for the pleasure of hearing laughs and joining them.

The methods developed in this Thesis aim at enriching human-computer interaction by endowing computers with the capabilities to sense and express affect through laughter. On the one hand, the methods developed to automatically characterize laughs (intensity and phonetic transcriptions) can serve to better describe and understand the behavior and affective states of humans interacting with the system. An appropriate response can thus be triggered. It must be noted that the decisions taken by the machine (e.g., when and how to laugh in response) form a totally separate field (called "dialog management") and are not addressed in this work. On the other hand laughter synthesis gives the computer the possibility to express the required affective states in a richer way (display affiliation, joy, compassion, etc.). In addition parametric laughter synthesis, as developed in this Thesis, provides control over a range of laughter features that can be—and currently are—used to better understand which properties of a laugh influence its perception. For example, we could determine which acoustic features will make the difference between a laugh being perceived as friendly or malicious. The possibility to create laughs with controlled features enables to support psychological experiments on laughter. In particular our methods are used with gelotophobes to investigate whether some laughs could be better accepted by this population.

1.5 Organization of this dissertation

Previous sections have justified the interest of automatic laughter processing. From now on, we will concentrate on the acoustic aspects and explain how laughter can be detected, analyzed and synthesized.

This dissertation is organized as follows. As the Thesis is covering different fields, each chapter is covering one of these fields. Each chapter begins with its own section of related works, including works that have been published after our own developments in the field, and is ending with its own conclusions and suggestions for future work. Chapter 2 is related to databases that have been (or can be) used to study and model laughter. As part of that chapter, the AVLC database, recorded during this work, is detailed. Chapter 3 describes laughter from an audio point of view. We will see that, similarly to speech, laughter can be studied at different levels. The description tracks introduced in this work (phonetic transcriptions and intensity curves) receive particular attention in that chapter. Chapter 4 focuses on automatic laughter recognition and analysis. Recognition was not investigated in the current Thesis, but as

most efforts on automatic laughter processing are recognition tasks, we have decided to cover it in the state-of-the-art section. A few works about laughter classification are then presented, before describing the methods developed for automatic phonetic transcription and intensity estimation of laughter. Chapter 5 focuses on acoustic laughter synthesis. In that chapter we present our developments in laughter synthesis based on Hidden Markov Models (HMMs) and the comparison of several vocoders, as well as an experiment investigating whether laughter synthesis can be trained on phonetic transcription obtained automatically. The method for generating phonetic transcriptions from intensity curves is also detailed in that chapter. Chapter 6 relates to broader applications (human-computer interaction, psychological studies, etc.) in which our methods have been used. Finally Chapter 7 presents the overall conclusions of this work, recalls its main contributions and summarizes the future works suggested in each chapter.

Chapter 2

Laughter databases

Contents

To analyze, characterize and model laughter, samples are mandatory. In this Chapter we will review the databases that can be used to study laughter. First, the objectives and constraints related to gathering laughter data will be described in Section 2.1. Available databases are then presented in Sections 2.2 to 2.4. The AVLaughterCycle (AVLC) database is then described in details, as it is the laughter database which has been recorded in the framework of this PhD Thesis and was used for the developments that will be presented in the remainder of this work.

2.1 Building an emotional database

The most obvious way to build a database of laughter samples is to transpose what has been done with speech for many years: bring people in a lab and ask them to laugh (for speech, participants are generally asked to read out loud some given text). This approach is however already limited for speech, since it does not enable to record all the variations that speech encounters when we are in "real-life", influenced and altered by our emotions, conversational partners' behaviors, surrounding noise, etc. It is very difficult to design an experimental protocol that takes all the targeted situations into account. And even if we were able to do it, how could we be sure that we elicit the same emotions people face in everyday life and that they will react the same way in the lab as in their "natural" environment? Another solution is then to use actors. Actors are able to pretend being in a certain state, but signals are not exactly the same as when they really feel the emotion. This is a major issue in emotion research and, since laughter is an emotional signal, it is affected by the problem. Acted emotions (and laughter) are useful, but never guarantee that real-life signals would exhibit the same patterns. The true relationship between acted and spontaneous feelings is still unknown, but the use of portrayed (i.e., acted) emotions to simulate spontaneous occurrences is contested [Douglas-Cowie *et al.* 2003, Wilting *et al.* 2006, Valstar *et al.* 2007, Ruch & Ekman 2001]. Some people believe that actors, when asked to behave in a given mood, are more portraying a stereotype of how the state is represented in society than its everyday expression [Drahota *et al.* 2008]. On the other hand, capturing the signals in "real-life" without disturbing them is not easy and a lot of post-processing to annotate and isolate the interesting segments is needed. Noise can also be an issue in natural recordings.

Given these considerations, three main techniques have been used to build databases in the fields of emotion and laughter research [Scherer 2003]:

- Natural expressions: data is collected from the real-world, with the subjects free to express themselves and, ideally, not aware that they are being recorded until the end of the data acquisition. A popular setting for emotion recognition is the use of data collected in call-centers [Morrison *et al.* 2007, Devillers & Vidrascu 2007].

- Induced responses: subjects are presented to a stimulus (picture, video, vocal information, etc.) chosen to elicit a target emotion (happiness, fear, etc.). For example, laughter can be induced by presenting a comedy video. Users can be aware that they are being recorded, but everything is done to provoke natural reactions.

- Portrayed emotions: actors, professional or not, are directly asked to portray the emotional state or, in our case, to laugh.

Many different approaches can be found in the literature for obtaining such emotional databases. In the following sections, we will only present a few databases

where special attention was put on laughter. Databases with natural, induced and acted laughter will be respectively presented in Sections 2.2, 2.3 and 2.4. A broad listing of emotional databases can be found on the HUMAINE Network of Excellence website [HUMAINE 2008]. In addition, a growing collection of laughter databases are integrated in the ILHAIRE[1] database [McKeown *et al.* 2012a] and distributed on its website [McKeown 2014]. A summary table of the databases presented in this chapter is available in Appendix A.

2.2 Laughter recorded as part of natural expressions

In this section we will review the databases of natural expressions in which laughter has been annotated. Most of these databases have been initially recorded for other purposes than studying laughter; rather experimenters were interested in regular, natural conversations and, as laughter is an important communicative signal frequently occurring in everyday situations, laughs were captured in the process. It is expected that every corpus of natural human behaviors can include laughs, but only databases for which laughter annotation has been reported are included here. In other words, the following is not an exhaustive list of databases containing laughter.

Note that the boundary between natural expressions and induced emotions can be a bit fuzzy. Some of these databases included here in the "natural expression" group have been recorded in laboratories, but the main objective was not specifically to trigger laughter: experimenters designed a scenario in which the participant would have to talk and react, but not a scenario focusing at the elicitation of laughter. Some of the databases included in this "natural expressions" section could thus arguably be moved to the "induced emotions" section. Reciprocally, parts of the databases in the "induced emotions" section could be considered as natural expressions and therefore be mentioned in the "natural expressions" section too.

Audio-only databases will be presented first in Section 2.2.1, distinguishing between data recorded in the wild ("indubitably natural") and data recorded in labs ("arguably induced"), with a more detailed section on the ICSI Meeting Corpus as it has been used as a reference database to evaluate acoustic discrimination between laughter and other events like speech (see Chapter 4). Then, multimodal databases will be described: data gathered for other purposes than laughter in Sections 2.2.2.1 to 2.2.2.7, then the Belfast conversational dyads (Section 2.2.2.8), which is the only multimodal database in this "natural expressions" section that was specifically recorded to obtain naturally-occurring laughs.

[1] ILHAIRE is a FET European project focusing on the integration of laughter into human-avatar interactions. The author of this Thesis is active in the ILHAIRE project, which helps promoting efforts on laughter processing and studying.

2.2.1 Audio-only databases

2.2.1.1 Recordings in the wild

The Corpus of Spontaneous Japanese [Maekawa *et al.* 2003] contains around 650 hours of spontaneous speech, from over a thousand speakers recorded while giving lectures or presentations. This huge audio database has been transcribed including labels denoting the presence of laughter, but no time boundaries.

Nick Campbell has also been involved in several extensive recordings of spontaneous speech and has shown interest in spotting laughter inside his large corpora [Campbell 2011]. Among others, there are the FreeTalk database (see Section 2.2.2.5) as well as 20 hours of telephone conversations in Japanese between eight pairs (formed from a pool of ten participants) of volunteers accounting for 2001 laughter and 1129 speech-laugh utterances [Campbell 2007].

For his observations of laughs, Wallace Chafe [Chafe 2007] used excerpts of the Santa Barbara Corpus of Spoken American English, which contains recordings of natural conversations in a range of everyday situations (talking about studies, preparing dinner, business conversations, etc.) [Du Bois *et al.* 2000, Du Bois *et al.* 2003, Du Bois & Englebretson 2004, Du Bois & Englebretson 2005, University of South California, Santa Barbara 2011]. Most conversations are face-to-face, but there are also telephone dialogs, radio programs, lectures, story-telling, etc. The corpus contains transcriptions of the audio files, including labels identifying laughter, but one major drawback is that all speakers are recorded on the same channel [Truong & Trouvain 2012a]. Chafe used 60 excerpts of the corpus to study laughter and humor.

The Corpus Gesproken Nederlands (CGN), which translates to "Spoken Dutch Corpus", is a collection of recordings in various socio-situational settings [Oostdijk 2000]. The objective was to build an extensive corpus of spoken Dutch, from both the Netherlands and the Belgian Flanders. The corpus includes scripted (e.g., news, read-out texts), semi-scripted (e.g., lectures, interviews) and unscripted (e.g., business transactions, face-to-face conversations, spontaneous TV commentaries) data. Recordings have been orthographically transcribed, with a specific tag for non-speech events. Truong and van Leeuwen [Truong & van Leeuwen 2007a] have used data from natural face-to-face conversations of the CGN to evaluate algorithms for distinguishing speech and laughter (see Section 4.1.2.1).

Vettin and Todt [Vettin & Todt 2004] were also interested in laughter occurring during natural conversations and recorded naturally occurring conversations: ten participants (six women, four men)—who were acquaintances of the experimenters—agreed to be recorded several months in advance and did not know when the recordings would actually take place. A second set was recorded in a more standardized way, in the lab, with participants discussing with an experimenter without knowing the exact purposes of the recordings.

Devillers and Vidrascu [Devillers & Vidrascu 2007] were interested in the emotions conveyed by laughter and used 20 hours of telephone conversations in a call center

providing medical advices. Verbal and non verbal contents such as laughs or cries have been manually annotated. More than half of the 119 laughter utterances in this corpus have been related to negative emotions.

The COSINE database [Stupakov *et al.* 2012] aims at recording natural speech in the wild. Participants equipped with multiple portable microphones were asked to walk to various noisy locations and talk about anything they like. Conversations included from two to seven people and lasted from 45 minutes to 1.5h. A total of 33 sessions have been recorded (91 unique participants), among which ten have been transcribed (37 unique participants). In the ten hours of transcribed data, locations of 3267 laughter occurrences have been marked [Weninger & Schuller 2012].

The ICSI Meeting Corpus is also audio-only, but as it has been widely used for training and evaluating laughter segmentation methods, we have decided to describe this database more deeply. It receives a dedicated section later (Section 2.2.1.3).

2.2.1.2 Participants involved in scenarios

The HCRC Map Task Corpus [Anderson *et al.* 1991] contains recordings of 64 young adults having to collaborate to reproduce on one participant's map the route that was indicated on the other participant's map. Each participant was involved in four different conversations, in order to vary parameters such as familiarity between the speakers, eye contact, landmark names to mention, etc. The conversations were transcribed. Around one thousand laughter episodes were spotted [Truong & Trouvain 2012a].

The Buckeye Corpus of Conversational Speech [Pitt *et al.* 2007] gathers recordings of 40 participants from Ohio discussing with an experimenter in an interview-style setting. Participants wore a head-mounted microphone. The corpus contains 26 hours of recording [Weninger & Schuller 2012]. Speech has been phonetically transcribed and laughter positions were marked. A total of 1874 laughter occurrences have been found in the data [Weninger & Schuller 2012].

The Diapix Lucid Corpus [Baker & Hazan 2011] consists of microphone recordings of pairs of participants discussing when playing a "spot the difference between two images" game. Twenty pairs of participants took part in the study and each one was involved in three different conversations. The conversations were transcribed and the words were automatically aligned to the sound files (forced alignment). The data contain 582 laughs [Truong & Trouvain 2012a].

For the INTERSPEECH 2013 Social Signals Sub-Challenge of the Computational Paralinguistic Challenge, a corpus containing 2763 audio clips was extracted from 60 phone calls between two participants having to discuss about—and order by importance—a list of items that would be useful to survive in a polar environment [Schuller *et al.* 2013]. Each clip lasts 11 s and contains at least one laughter or filler[2] occurrence. Sixty-three females and 57 males took part in the recordings. The corpus was called SSPNet Vocalization Corpus (SVC).

[2] Fillers are vocalizations like "uhm", "aaah", etc. that indicate attempts to keep the speaking turn.

2.2.1.3 The ICSI Meeting Corpus

In 2000, the International Computer Science Institute of Berkeley (ICSI) launched a project to record a large speech database from meetings, called ICSI Meeting Corpus. The purpose was to obtain speech as natural as possible and they chose to record only meetings that would have occurred anyway [Janin *et al.* 2003]. All the meetings took place in a meeting room of their lab, where they would have taken place even if the ICSI Meeting Corpus project had not existed. The only, but important, unnatural setting the project implied on the meetings was the use of head-mounted microphones, in order to have easier speech activity detection and high quality speech transcriptions, but also to avoid penalizing, with poor acoustic signals, non-acoustic research like dialogue structure analysis [Janin *et al.* 2003]. In consequence, all the subjects knew they were being recorded.

The audio settings were the following: each participant wore a head-mounted microphone and, in addition, four omni-directional PZM microphones and one PDA containing two cheap microphones were placed on the meeting table. The meetings involved three to ten participants, with an average of six [Janin *et al.* 2004]. The recordings ended in 2002, with a total of 75 meetings, corresponding to 72 hours of meetings and 85 hours of speech. The total duration of speech is larger than the duration of meetings due to overlapping speech. The audio signals were downsampled on the fly from 48 kHz to 16 kHz and the channel from each microphone was saved separately. At the beginning or end of the meetings, participants were asked to read digit strings. These sequences of digits may be used to perform research on far-field acoustic issues without having additional complexities introduced by large vocabulary, spontaneity and conversation interactions [Janin *et al.* 2003]. After the meetings, participants could listen to the recordings and remove confidential sections, but it was only requested 19 times, for a total of 2.6 minutes of deleted segments.

Since only naturally occurring meetings at the ICSI lab were recorded, some participants appeared multiple times in the corpus. In total, 53 different speakers were involved in the data collection. Each received a unique identifier mentioning the gender of the participant and whether this participant was a native English speaker or not. Other information such as the date and topic of each meeting or the way the channel signal was transmitted (mostly wireless for head-mounted microphones) is also provided.

Huge efforts were done to annotate the data. For each meeting, there is a full speech transcription, with beginning and ending times of each utterance. More interesting for us, many non-speech sounds were also annotated, in two ways:

- VocalSound: this tag was used for spotting non-speech vocal sounds like laugh, cough, breath, etc. Table 2.1 gathers the most frequently annotated VocalSound events, and the relevant ones for laughter processing [Laskowski & Burger 2007b].

- Comment: various information about the transcribed utterance was encoded under this tag. Among others, modifications of speech due to smile or laughter

are annotated. Table 2.2 presents the most frequent Comment annotations in relation with smile or laughter.

Table 2.1: Top five most frequently occurring VocalSound types in the ICSI Meeting Corpus and all VocalSound related to laughter and smile [Laskowski & Burger 2007b].

Frequency rank	Occurrences	VocalSound Description
1	11515	laugh
2	7091	breath
3	4589	inbreath
4	2223	mouth
5	970	breath-laugh
11	97	laugh-breath
46	6	cough-laugh
63	3	laugh, "hmmph"
69	3	breath while smiling
75	2	very long laugh

Table 2.2: Top five most frequent Comment tags of the ICSI Meeting Corpus in relation with smile and laughter [Laskowski & Burger 2007b].

Frequency rank	Occurrences	Comment Description
2	980	while laughing
16	59	while smiling
44	13	last two words while laughing
125	4	last word while laughing
145	3	vocal gesture, a mock laugh

Laughter processing was not the initial purpose of the ICSI Meeting Corpus and it was not the event that received the most attention. However, thanks to the quality of the database, recorded in a natural environment and presenting various occurrences of spontaneous laughs, it became a standard for laughter processing. Some of the groups using the ICSI Meeting Corpus for laughter processing did additional annotation works to keep only clearly audible laughs [Truong & van Leeuwen 2007a] or localize the boundaries of the laughter segments with more accuracy [Laskowski & Burger 2007b].

The Corpus is available from the Linguistic Data Consortium (LDC) [Linguistic Data Consortium 2008]. Other annotations were done by the ICSI Meeting Corpus team or other groups to mark the meetings dialogue structure, the parts where speakers were most involved in the meeting, etc. Good overview of these additional annotations and research works carried out with the ICSI Meeting Corpus can

be found in [Janin *et al.* 2004].

2.2.2 Multimodal databases

2.2.2.1 The Belfast Naturalistic Database

The Belfast Naturalistic Database [Douglas-Cowie *et al.* 2003] gathers excerpts of TV programs (chat shows, religious programs, etc.) and studio recordings with a focus on emotional expressions. The criterion to include TV excerpts was that speakers should appear to genuinely experience a given emotion. The studio recordings were one to one interviews over topics that would induce a range of emotional displays. Data were split into clips that include sufficient context to understand the emotional apex and to show how emotions evolve over time. Laughs were identified in 53 out of the 127 clips extracted from TV programs. Due to copyright issues, only five of these clips are disseminated on the ILHAIRE database website [McKeown *et al.* 2012a].

2.2.2.2 The HUMAINE database

The HUMAINE database [Douglas-Cowie *et al.* 2007] is another collection of audiovisual clips from various sources with the aim to cover a broad range of natural affective displays. Some of the clips have been annotated, including labels for paralinguistic events such as laughter. From the 50 clips available, 46 laughter occurrences have been found and included in the ILHAIRE database [McKeown *et al.* 2012a].

2.2.2.3 The AMI Meeting Corpus

The Augmented Multiparty Interaction (AMI) Meeting Corpus [Carletta 2007, AMI project 2011] consists of 100 hours of meetings recorded at the University of Edinburgh (U.K.), Idiap (Switzerland), and the TNO Human Factors Research Institute (The Netherlands). One third are naturally occurring lab meetings. The remaining two thirds were elicited by a role playing game in which participants had to take different roles in a team project. While this is different from the settings of the ICSI Meeting Corpus, it has little influence on laughter naturalness.

All of the 138 role playing meetings involve four participants. Out of the 33 naturally occurring meetings, 25 also involve four participants, five have three conversationalists and the last three have five participants.

The recordings include synchronized audio (individual and far-field microphones), video (individual and room-view cameras), as well as the positions of Logitech pens held by the participants and the illustrations displayed by the meeting participants (PowerPoint slides and whiteboard explanations). At the University of Edinburgh, 24 microphones were used: 16 omni-directional electret microphones were placed on the table, in two groups of 8-microphone circular arrays (10 cm radius). Participants wore two microphones, one head-mounted and one lapel microphone. Audio signals were sampled at 48 kHz, with a resolution of 16 bits and stored in WAV format. For the video signals, six cameras were used: one in front of each participant's chair,

one taking an overhead view from the center of the room and the last one placed in a room corner to obtain a full-scene recording, including the whiteboard and the screen. Data written on the whiteboard and individual notes taken by the participants were stored as x-y coordinates in a XML file format, with the help of dedicated material (Logitech pens, special paper, digital whiteboard). Similar material has been used at Idiap and the TNO Human Factors research Institute, except for slight variations in the position and number of devices (microphones, cameras).

The database is freely available for non-commercial purposes from the AMI website [AMI project 2011]. Signals are provided with a range of annotations. Some recordings include annotations about dialogue acts, emotions, actions, gestures, etc. The vocal activity of each participant in each meeting has also been manually transcribed. These transcriptions include the speech as it has been uttered by the speaker (with grammatical errors, hesitations, etc.) and other non-verbal vocalizations such as laughter and cough. In addition, a phonetic transcription of speech, obtained by forced alignment, is provided.

2.2.2.4 The AVIC database

The AudioVisual Interest Corpus (AVIC) [Schuller *et al.* 2007] contains audiovisual data from conversations in English between an experimenter and a participant. The experimenter is playing the role of a product presenter and presents the product to the participant. The participants and the experimenter are recorded with a lapel microphone and a camera, and an additional far field microphone is placed in the room. In total, 21 participants (ten females, eleven males) have been recorded, for a total of more than ten hours of recording. The spoken content and non-verbal interjections, including laughter, have been annotated.

2.2.2.5 The FreeTalk Database

Petridis et al. [Petridis *et al.* 2013b] refer as "FreeTalk" Database to the laughter data used by Scherer et al. for their experiments on multimodal laughter detection [Scherer *et al.* 2009]. The data were recorded during three multiparty interactions where four or five people sitting around a table discussed freely for around 90 minutes. A 360 degree video camera was used to capture the participants' facial and upper body movements. A centrally-positioned microphone was used to record the audio signal. Speech and laughter segments have been annotated. Scherer et al. [Scherer *et al.* 2009] used 300 laughter segments (average duration of 1.5 s) and 1000 speech segments (average duration of 2 s) for their works. The database is available on the web [Campbell 2014] and possibilities to browse through the data are described in [Campbell 2009].

2.2.2.6 The Green Persuasive Database

The Green Persuasive Database [Douglas-Cowie *et al.* 2007] consists of recordings of conversations between a persuader and a person he tries to convince to adopt more ecological behaviors. Each conversational partner is recorded with a different camera. Close-talk microphones were placed on the participants' clothes. As in any social interaction, there are laughter occurrences in the dataset. Eight interactions between a Professor (the persuader) and different students have been recorded. Each interaction is lasting from 15 to 35 minutes. The database is freely available for research [SSPNET 2014]. In addition, 280 laughter episodes could be identified in the data and these episodes are available from the ILHAIRE database [McKeown *et al.* 2012a].

2.2.2.7 The SEMAINE database

The SEMAINE database [McKeown *et al.* 2012b] consists of audiovisual recordings of participants discussing with either an operator playing the role of a limited automated avatar, an operator-driven avatar or a limited automated avatar, designed to elicit particular emotions. Four different avatars with extreme personalities (constantly angry, happy, gloomy or sensible) have been developed to try to move the participant to the same state of mind. Each file of the database has been labeled in emotional states by six to eight annotators, providing information about the emotional states leading to and following laughter. The database contains audio-visual recordings of the participants with frontal cameras and head-mounted microphones. Laughter has been spotted in 56 out of the 66 interactions (total: 333.7 minutes of recordings) of the corpus. The database is freely available to the research community. A subset of the SEMAINE Corpus has been used for the 2011 Audio/Visual Emotion Challenge (AVEC) and is known as the AVEC Corpus [Schuller *et al.* 2011].

2.2.2.8 The Belfast Conversational Dyads

The Belfast Conversational Dyads data consist in recordings of free conversations between two participants. The objective is to capture the structural particularities in conversations with only two participants: as suggested by Glenn [Glenn 2003], it is socially acceptable for the speaker to laugh first when there is only one listener, but this is no longer true in groups of three or more as it could be interpreted as self-praise. We will further explain that point in Chapter 3. Conversations involving more than two participants are addressed in another corpus with similar settings, the Belfast story-Telling sessions (see Section 2.3.4).

Participants of the Belfast Conversational Dyads were recorded while freely chatting[3] during approximately one hour. Participants wore high-quality close-talk microphones and, as for the Belfast Story-Telling sessions, Kinects and webcams were used to record each participant's movements [Curran & McKeown 2013]. Nine Conversational Dyads have currently been recorded: the full recordings are available on

[3] A random topic could be provided to the participants to start the conversation if they wished.

the ILHAIRE database website [McKeown 2014] and laughter segmentation is under way.

2.3 Induced laughter databases

When laughter is the target of database recordings, inducing it can be an ecological solution as it enables to both control the settings (e.g, close-talk microphones, motion capture equipment, frontal camera views) and increase the number of laughs. Hence, numerous high-quality recordings of laughter episodes can be obtained in a relatively short time. The critical point is to create scenarios that can favor natural laughter, as the objectives of induced laughter databases are to avoid forced laughs (otherwise, one can simply record portrayed laughs).

For their extensive audio analysis of laughs, Bachorowski et al. [Bachorowski *et al.* 2001] enrolled 139 students and let them watch video containing humorous sequences either alone or with a partner. Laughs from 97 individuals (52 females, 45 males) were kept for the acoustic analyses, for a total of 559 female and 465 male laughter bouts (see Section 3.1.1 for a definition of "laughter bout"). It is useful to note that laughs were recorded in a broader setting: participants watched 11 video clips in total, aimed to induce different emotions, and had to vocally express their emotions after viewing each clip. Hence the microphone was needed for the ratings and participants ignored it would also be used to capture their laughs. Kipper and Todt [Kipper & Todt 2007] induced laughter while subjects were reading by putting their own voice in playback with a 200 ms delay.

The AVLC database is using the same induction technique as Bachorowski et al. It was recorded in the framework of this PhD Thesis and will be detailed in Section 2.5. The following sections present other induced laughter databases, which have chronologically been recorded after the AVLC corpus.

2.3.1 MMI Facial Expression database, Part V

Part V of the MMI[4] Facial Expression database includes annotations of voiced and unvoiced laughs [Valstar & Pantic 2010]. As for the AVLC database, participants were left alone in a room to watch video clips. The clips were meant to induce happiness, disgust and surprise. Participants' reactions were recorded with a camera (audiovisual recordings, no close-talk microphone). Nine participants took part in the recordings, providing 109 unvoiced and 55 voiced laughter episodes[5]. The database is available for research [MMI team 2014].

[4] The acronym MMI comes from the first names of the MMI database's authors.

[5] In the MMI Part V annotation conventions, a laugh is considered as voiced if it contains at least one voiced component. The distinction between voiced and unvoiced laughs will be further explained in Chapter 3.

2.3.2 The Belfast Induced Natural Emotion Database (BINED)

The Belfast Induced Natural Emotion Database (BINED) [Sneddon *et al.* 2012] contains camera recordings of the reactions of subjects performing different tasks and watching video clips[6], with the aim to elicit five different emotional states: frustration, surprise, fear, disgust and amusement. Laughter frequently appeared in all these (active or passive) tasks. The corpus was recorded in three different sets with different purposes and sequences of tasks to perform. In particular Set 3 aims at comparing different cultures and includes recordings in Northern Ireland and in Peru. The database (1400 clips in total from 256 participants) is freely available for research [Queen's University Belfast: School of Psychology 2014]. Laughter instances have been extracted from these recordings and included in the ILHAIRE database [McKeown 2014]: 289 laughs from Set 1 and 48 from Set 3 are currently available.

2.3.3 The MANHOB laughter database

The MANHOB laughter database [Petridis *et al.* 2013b] was recorded in 2012 with the aim to provide a benchmark for laughter classification. It is similar to the AVLC database on some points (laughter elicited with humorous videos, participants asked to produce acted laughs) but includes some specific contents related to the objectives of the data recording:

- participants were asked to speak for about 90 s in English, as well as in their mother tongue if it differed from English, so that the database also includes speech from the laughers, which is necessary to develop speech and laughter discrimination;
- participants were also asked to produce posed smiles;
- thermal video recordings are included, to enable thermal images analysis (for example study thermal differences between acted and spontaneous laughter).

The speech sessions consisted in the participant either talking about a selected topic for 90 s, or discussing with a friend or an operator. It is important to note that two operators were in the room together with the participant during the session, which is a significant difference with AVLC, as the social context (and the mere presence of humans) is known to influence people's laughs.

Twenty-five people took part in the recordings. Three of them had to be discarded due to technical failures. Among the 22 remaining participants were ten females (average age: 27; standard deviation (std): 3) and twelve males (average age: 28;

[6] The parts of the database where participants are performing tasks could be considered as "natural settings", yet the objectives of these tasks was to induce specific emotions and the videos were chosen to provoke amusement, which is why it was decided to include the BINED database within the "induced laughter" section.

std: 4). In total, 180 sessions were recorded, distributed as follows: 90 spontaneous laughter recording sessions (watching video clips), 38 speech sessions, 23 acted laughs sessions and 29 posed smiles sessions.

Although only one session (watching humourous clips) was specifically designed to elicit spontaneous laughs, spontaneous laughs actually occurred in all the sessions (speech, posed smiles, acted laughs). For example, the acted laugh session yielded spontaneous laughs resulting from the participant's embarrassment to follow the instructions or their self-perception of inability to produce satisfying examples.

The database was annotated using the ELAN software [Sloetjes & Wittenburg 2008, Max Planck Institute for Psycholinguistics 2014] using several tracks to segment speech, laughter, speech-laughs, acted laughs, posed smiles or other human sounds. Two annotation tracks were actually used to segment all laughter-related events both with and without final inhalation (if present), as both approaches have been used to segment laughs in other databases. In addition, laughs were labeled as voiced or unvoiced by taking the majority class assigned by two human labelers and an automatic decision provided by Praat [Boersma & Weenink 2011] (the laugh was considered as unvoiced if at least 85% of its frames were estimated as unvoiced by Praat). In total, the database contains 563 laughter episodes (318 voiced, 245 unvoiced) for a total duration of 931 s (average laugh duration: 1.65 s).

The MAHNOB laughter database, including the annotations, is freely available on the web [Imperial College London 2014].

2.3.4 The Belfast Story-Telling sessions

With the objective to record laughs related to different affective states, the Belfast Story-Telling sessions [Curran & McKeown 2013] are a replication of a similar experiment that was conducted by the University of Zurich in the framework of the IL-HAIRE project. The experiment, called *The 16 Enjoyable Emotions Induction Task* [Hofmann *et al.* 2011], consisted in recordings of groups of three or four participants where each participant had to tell stories recalling situations in which (s)he experienced each of the given 16 enjoyable affective states [Ekman 2003]. The experiment was first carried out in Zurich, with participants telling stories in Swiss-German. The objective of replicating the experiment in Belfast with English- and Spanish-speaking participants was on the one hand to investigate cross-cultural effects and on the other hand to assess whether participants could identify laughter types without linguistic cues in a large-scale web study (see Section 3.1.7). Such a setup required specific agreement from the participants to broadcast their laughs. Participants wore high-quality close talk microphones and each of them was recorded both with a regular camera and a Kinect (tracking facial action units [Ekman *et al.* 2002], facial landmarks, skeleton, head poses and storing the depth map). All streams were synchronized with the help of the Social Signal Interpretation—formerly Smart Sensor Integration— (SSI) software [Universitat Augsburg 2013, Wagner *et al.* 2011]. Twenty-one participants were

recorded in this setup, making a total of over 25 hours of recordings. The full recordings as well as segmented laughs are available on the ILHAIRE database website [McKeown 2014].

2.3.5 The Belfast and UCL Motion Capture sessions

Still in the framework of the ILHAIRE project, scenarios were developed to record laughter body movements [McKeown *et al.* 2013]. Participants were invited by pairs of friends to participate in a series of tasks aiming at eliciting laughter (e.g., tongue twisters, dancing game, Pictionary, etc.). At least one of the participants was wearing a motion capture system (in the Belfast recordings both participants were equipped with Qualysis markers; in the University College London recording only one participating was wearing an Animazoo IGS-190 inertial motion capture system). Participants also wore close talk microphones, but the atmosphere was in general rather noisy (music, voices of the experimenters, etc.). Eight participants were recorded in Belfast and 18 in London. Motion capture data (both full recordings and stick figures of segmented laughs) have been included in the ILHAIRE database [McKeown 2014].

2.3.6 The MMLI Corpus

Extending from the Belfast and UCL Motion Capture session, the Multimodal Multiperson corpus of Laughter in Interaction (MMLI) [Niewiadomski *et al.* 2013b] aims at capturing full body movements and in particular shoulders, torso and respiration movements. The objective was also to obtain laughs from various contexts of free interactions, hence gathering different types of laughs. To favor laughter and free interactions, participants were recorded in groups of friends performing different tasks together (watching humorous videos, playing simple social games). The database consists of four groups of three friends and two groups of two friends, making a total of 16 subjects, including three females. Their age ranged between 20 and 35. Subjects were equipped with inertial motion capture systems (two XSens MVN Biomech with 17 inertial sensors, one Animazoo IGS 190 with 19 inertial sensors) to record their body movements. In addition, two Kinects were used to record video as well as tracking 100 facial and 20 body points, and detecting six high-level actions such as smiling or frowning. Audio (two close-talk microphones, one ambient microphone) and regular video (six webcams) were also recorded. All modalities were synchronized via the SSI software [Universitat Augsburg 2013]. Laughter episodes were annotated and sum up to 520. The database is available from the ILHAIRE database website [McKeown 2014].

2.3.7 Pinoy Laughter 2

For their developments of acoustic laughter synthesis, Cagampan et al. [Cagampan *et al.* 2013] used audiovisual recordings of four Filipino participants who took part in several tasks like discussing over Skype, watching video clips or asking and

answering questions designed to elicit the six primary emotions (sadness, happiness, anger, fear, disgust, surprise) [Ekman 1992]. The corpus is called Pinoy Laughter 2.

2.3.8 The AV-LASYN database

The AV-LASYN database [Çakmak *et al.* 2014] has been recorded with a similar scheme as the AVLC database, with the specific objective to obtain a larger quantity of audio and facial motion capture data from one single participant, in order to develop audiovisual laughter synthesis. One male participant was recorded while watching humorous video clips. He was wearing reflective markers on the face to track facial movements with the help of the Optitrack system [Natural Point, Inc. 2009] (see Section 2.5.4.2). The database contains 251 laughter episodes, which have been phonetically annotated (similarly to the phonetic annotations of the AVLC database, which will be presented din Section 3.2).

2.4 Portrayed laughter

For identifying acoustic correlates of different emotions in laughter, Szameitat et al. [Szameitat *et al.* 2007, Szameitat *et al.* 2009a, Szameitat *et al.* 2009b], asked eight professional actors (five females, three males) to portray laughs for four different affective states, namely joyous, taunting, tickling and schadenfreude— a German word meaning "pleasure in another's misfortune"—laughs. Actors were asked to get into the emotional states via self-induction techniques (emotional recall, voicing and body movements). They gathered 429 episodes of acted laughs with this process. Laughs were recorded at 48 kHz (16 bits) in a soundproof room.

Suarez et al. [Suarez *et al.* 2012] also recorded two actors portraying laughter related to five different emotions (happiness, giddiness, excitement, embarrassment, hurtfulness). A total of 497 audiovisual laughter recordings were obtained in this way. In addition, three participants were also recorded while watching humorous video clips, with the aim to induce (laughs in) the five emotional states. The corpus is named PinoyLaughter.

John Esling [Esling 2007] used samples from the University of Victoria Larynx Research Project, which includes nasoendoscopic videos of the larynx, to analyze the states of the larynx in acted and spontaneous laughter.

Finally, in the framework of the artistic installation "The world starts every second" [Lafontaine & Todoroff 2007], laughs were recorded, constituting a database of several dozen laughter utterances. The corpus gathers laughs from children and professional singers. Some singers portrayed different states of mind like "lover laugh", "hysteric laugh", "obsessional laugh", etc. The recordings took place in traditional recording rooms or in places with poorer acoustics (echo, etc.). There are also several occurrences of group laughs. From the recordings, 447 laughter episodes have been extracted, totaling 20 minutes of laughs. Some are obviously exaggerated and fake. They do not correspond to spontaneous laughs found in other corpora, but they have

a strong power of eliciting laughter to their listeners. In consequence, they could be used in applications designed to provoke emotional reactions, which was the aim of the artistic installation.

2.5 The AVLaughterCycle database

The AVLaughterCycle (AVLC) project [Urbain *et al.* 2010b], launched during the eNTERFACE'09 workshop[7] held in Genova, aimed at developing an audiovisual laughing machine, capable of recording the laughter of a user and to respond to it with a virtual agent's laughter linked with the input laughter. This goal implied three tasks: laughter detection, laughter analysis/classification (to link the output laugh with the input) and audiovisual laughter (copy-)synthesis. To perform theses tasks, an audiovisual laughter database has been recorded. The aim of this database is to provide a broad corpus for studying the acoustics of laughter, the facial movements involved, and the synchronization between these two signals. During the Workshop, the laughter database has been used to drive the facial movements of a three dimensions (3D) humanoid virtual character, Greta [Niewiadomski *et al.* 2009], simultaneously with the audio laughter signal.

The AVLC database [Urbain *et al.* 2010a], recorded as part of the AVLC project, is meant to be useful for many researches about laughter. To our knowledge, it is the first database of laughter combining both the acoustic signal and facial motion tracking.

2.5.1 Participants

Twenty-four subjects participated in the database recordings: eight (three females, five males) with the ZignTrack [Zign Creations 2009] setting and 16 (six females, ten males) with the OptiTrack [Natural Point, Inc. 2009] setting (see Section 2.5.4). They came from various countries: Belgium, France, Italy, UK, Greece, Turkey, Kazakhstan, India, Canada, USA and South Korea. The female, male and overall average ages were respectively 30 (std: 7.8), 28 (std: 7.1) and 29 (std: 7.3). All the participants gave written consent for their data to be used for research purposes.

2.5.2 Stimuli

Both audio recording and accurate facial motion tracking were desired. Due to the markers required for facial motion tracking, a natural laughter recording was impossible. To push the participants towards spontaneous laughter, a 13-minutes funny movie was created by the concatenation of short videos found on the Internet.

[7] eNTERFACE workshops are one-month summer workshops on Multimodal interfaces. The aim is to meet, learn and work together with researchers from other places. The duration of the event enables to deliver scientific outcomes at the end of the projects.

2.5.3 Database recording protocol

Participants were invited to sit in front of a computer screen, used to display the comedy movie. They wore a headset microphone for audio recording and stimuli listening. A webcam was placed on top of the screen, recording 25 Frames Per Second (FPS) with a 640x480 resolution, stored in RGB 24 bits. The audio sampling frequency was set to 16 kHz, stored in PCM 16 bits. The material for facial motion capture will be presented in Section 2.5.4.

The database was recorded through University of Augsburg's SSI [Wagner *et al.* 2009]. This software enables the synchronization between the different input signals (here microphone and webcam), handles the stimuli display and can process the signals to segment and label interesting parts. SSI was also used for the database annotation (Section 2.5.5).

Participants were asked to relax, watch the video and react freely to it, with two limitations: they should try to 1) keep their head towards the screen, and 2) not put anything between their head and the webcam (e.g., hands) to prevent the facial tracking from failing. All the instructions were displayed on the screen before the experiment. Once the protocol was clear, participants were left alone in the experiment room and started the stimuli playing. For synchronization and data saving reasons, the protocol had to be slightly modified when using OptiTrack (see Section 2.5.4.2).

At the end of the movie, subjects were instructed to perform one acted laughter, pretending they had just seen something hilarious. The main objective of these acted laughs was to provide some material to analyze the differences between spontaneous and acted laughs and to determine whether the subjects, when acting, tend to mimic the spontaneous laughs they had just performed.

2.5.4 Facial motion capture

Since markerless facial motion tracking was not reliable enough to capture the small variations of facial expression during laughter, we turned towards techniques using markers placed on the subject's face. Two systems have been successively used, ZignTrack and OptiTrack.

2.5.4.1 ZignTrack

ZignTrack [Zign Creations 2009] uses one single camera to realize the 3D tracking, which is an extrapolation from a two dimensions (2D) image, using a fixed face template. Facial features are marked with simple stickers or make-up (Figure 2.1). ZignTrack presents the advantages of being cheap and requiring few material, but has several drawbacks: the extrapolation from 2D to 3D causes head distortions, the tracking fails when there are rapid movements and the tracking is unable to recover after an erroneous frame. To obtain the accurate facial motion, a lot of manual corrections are then needed (several hours per recording). For these reasons, we turned

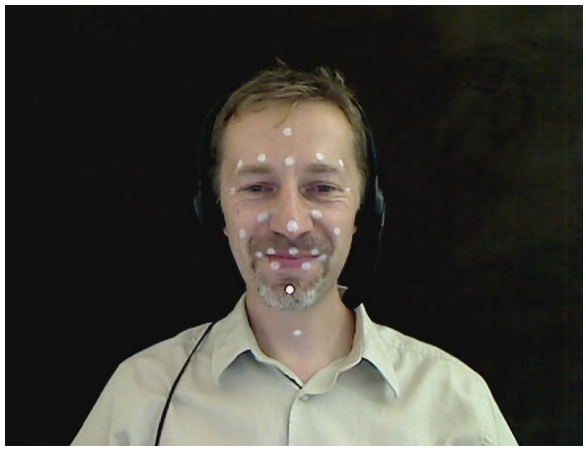

Figure 2.1: Markers drawn for facial motion tracking using ZignTrack.

towards a more professional motion capture system, OptiTrack, after the first eight recordings.

2.5.4.2 OptiTrack

OptiTrack [Natural Point, Inc. 2009] uses seven synchronized infrared cameras: six placed in a semi-circular way for facial motion capture and an additional one for scene audiovisual recording. Each camera contains a grayscale CMOS imager capturing up to 100 FPS. Infrared reflectors need to be stuck on the skin (Figure 2.2). For the recordings performed with the OptiTrack system, the infrared cameras were added to the previous setting (Figure 2.3, with the OptiTrack cameras highlighted by circles). Participants were asked to clap their hands in order to synchronize the facial motion tracking with the audio and webcam signals. OptiTrack provided high quality tracking with few manual corrections required. However, the data acquisition sometimes stopped after around five minutes. To make sure the data of the whole experiment would be usable, it was then decided to shorten the stimuli video to ten minutes and to split it into three parts slightly longer than three minutes. At the beginning of each session, the instructor started the face motion acquisition system, left the room and the subject clapped for synchronization with the other signals. At the end of each session, the subject was again instructed to clap, so that the instructor would enter the room and stop the face motion tracking. The microphone and webcam recorded the experiment from the beginning of the first session to the end of the third session, without interruption.

Figure 2.2: Infrared markers placed for facial motion tracking using OptiTrack.

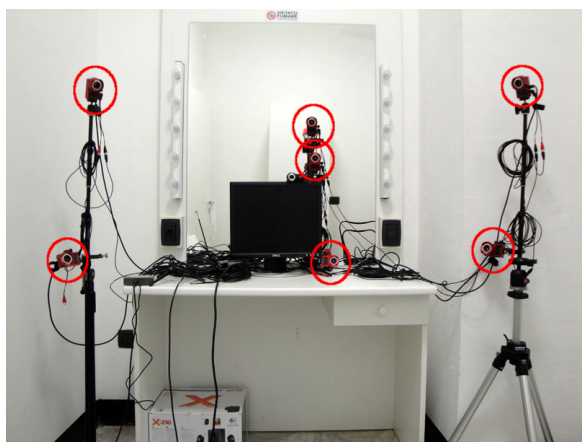

Figure 2.3: Desktop setup for database recording. Optitrack cameras are highlighted by circles.

2.5.5 Database annotation

The recorded data have been labeled by the author, using SSI [Wagner *et al.* 2009]. A hierarchical annotation protocol was designed: segments receive the label of one *main* class (laughter, breath, verbal, clap or trash; silence being the default class) and "sublabels" can be concatenated to give further details about the segment. Laughter sublabels characterize both the temporal structure and the acoustic contents of the laughs.

The laughter temporal structure sublabels follow the three segmentation levels presented in [Trouvain 2003] (see Section 3.1.1). These sublabels indicate whether the laughter utterance contains:

- several *bouts* (i.e., parts separated by inhalations, as will be further explained in Chapter 3), it is then annotated with the sublabel "episode";
- only one syllable, labeled as "monosyllabic", which is not uncommon (e.g., [Edmonson 1987]) but sometimes not considered as a laugh (e.g., [Sudheer *et al.* 2009]);
- several syllables but only one bout - default category (no particular sublabel).

These three temporal structure sublabels are mutually exclusive.

The laughter acoustic contents sublabels describe the type of sound[8]: vowel, breathy (oral exhalation), nasal exhalation, grunt-like, hum-like, hiccup-like, speech-laugh or laughter that is mostly visual (quasi-silent). These sublabels can be combined to reflect content changes during the laughter episode.

For example, a laughter episode composed of several bouts, starting by grunt-like sounds and followed by hiccup-like sounds is annotated: *laugh_episode_grunt_hiccup*.

To cope with exceptional conflicts that might influence the class models when training a classifier (for example when there is a strong noise in the middle of a laughter episode), a "discard" label has been added.

The annotation primarily relies on the audio, but the video is also looked at, to find possible neutral facial expressions at the episode boundaries or annotate silent laughs. In addition, laughs are often concluded by an audible inspiration, sometimes several seconds after the laughter main part. When such an inhalation, obviously due to the preceding laughter, can be found after the main audible part, it is included in the laughter segment.

2.5.6 Database contents

2.5.6.1 Main classes

The number of occurrences of the main classes over the full recordings or only inside the stimuli sessions are presented in Table 2.3. Subjects spent, in average, 21.8% of

[8] The sounds of laughter will be further detailed in Chapter 3. Here, only broad classes are used, following the production modes introduced in Section 1.1.

the stimuli sessions laughing, which is a huge proportion[9]. The number of laughter utterances per participant stands around 42, with extreme values of four and 82, for a total of 1021 episodes inside the dedicated stimuli sessions. The database contains 27 acted laughs, uttered by 22 subjects (two subjects did not produce any acted laughter).

	Occurrences	
Main class	Full database	Stimuli sessions
Laughter	1066	1021
Trash	267	207
Verbal	186	64
Clap	93	1
Breath	41	31
Discard	31	23

Table 2.3: Occurrences of the main classes in the AVLC database annotations.

2.5.6.2 Laughter forms

Table 2.4 presents the occurrences of the laughter sublabels, for the 1021 laughs elicited by the stimuli sessions, considered as spontaneous, as well as the 27 acted laughs. It is important to remember that the acoustic content sublabels can be combined to specify different contents in an episode. This explains why the total number of laughter sublabels is larger than the number of occurrences in the laughter class.

On a structural level, it appears that most laughs contain several syllables forming one single bout. Monosyllabic utterances are relatively frequent (17.5%) when subjects laugh spontaneously, but no subject produced a monosyllabic laugh when asked to pretend he had witnessed something hilarious. Episodes with several bouts separated by inhalations occur from time to time spontaneously and with a larger proportion when acting.

Regarding the acoustic contents, it can be seen that the spontaneous laughs cover a broad variety of sounds: the labels are spread over all the laughter sublabels. One third of the annotations reflect a vowel-like content, and the vowel '*a*' is the most frequent one. Nasal exhalations represent 20% of the annotations. Other categories like breathy (oral exhalation), hum-like, hiccup-like or even silent laughing are also well represented. However, the database contains only 20 speech-laughs [10]. This can be explained by the fact that the subjects were left alone and had nobody to interact with: there is few speech in the stimuli sections, hence few speech-laughs.

[9] For comparison, participants to Bachorowski et al.'s sessions laughed around 4% of the time, laughs account for around 7% of the MAHNOB database, while the call center conversations obtained by Devillers and Vidrascu contain around 2.5% of laughter.

[10] We should even state that in nine out of the 20 cases, speech and laugh do not overlap but follow each other so closely that it is impossible to separate them.

Category	Laughter sublabel	Occurrences	
		Spontaneous	Acted
	TOTAL UTTERANCES	1021	27
Temporal structure	Monosyllabic	179	0
	One bout (several syllables)	677	14
	Several bouts	165	13
	TOTAL	1021	27
Acoustic content	Vowel: a	277	18
	Vowel: e	101	5
	Vowel: i	37	1
	Vowel: o	26	0
	Vowel: u	5	2
	Nasal exhalation	277	1
	Breathy (oral exhalation)	237	2
	Hum-like	169	2
	Hiccup-like	95	5
	Grunt-like	18	1
	Speech-laugh	20	0
	Silent	94	1
	TOTAL	1359	38

Table 2.4: Occurrences of the laughter sublabels for the spontaneous and acted laughs.

The acoustic content sublabel occurrences are different when considering the acted laughs. There, voiced vowels are clearly the most frequent annotations. This might be due to the stereotypical image of laughter ("hahaha").

2.5.6.3 Laughter duration

Excluding laughs involving speech, the average duration of a spontaneous laughter utterance in the database is 3.5 s (std: 5.3 s; median: 2.2 s; min: 0.26 s; max: 82 s). A histogram of the spontaneous laughter duration and its cumulative distribution function are presented in Figure 2.4. The large majority (83%) of the laughter episodes lasts less than 5 s, but longer episodes should not be neglected as they represent 51.4% of the total laughter duration and are the most striking ones. The longest giggle in the database lasts 82 s.

Acted laughs tend to be longer. Their mean duration is 7.7 s (std: 5.94; median: 5.26). A t-test assuming the two samples (spontaneous and acted) come from normal distribution with unknown and unequal variances shows that the difference between the mean duration of spontaneous and acted laughs is highly significant ($p = 0.0012$). Using a t-test might seem daring since the duration distribution is clearly not Gaussian and the number (27) of acted laughs is not sufficient to use the Central Limit Theorem with full confidence. However the outcome of the t-test is strengthened by

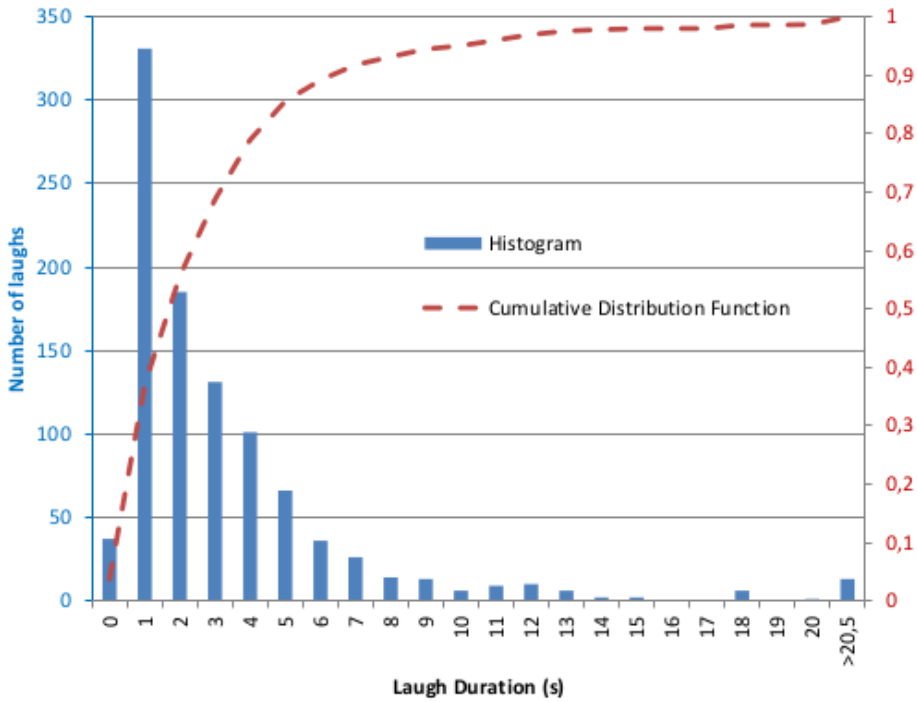

Figure 2.4: Histogram and cumulative distribution function of the laughter durations.

a Kolmogorov-Smirnov test (measuring whether the two samples are likely to belong to the same population, without any assumption on their distribution), which states with high significance ($p = 1.1 \cdot 10^{-8}$) that the spontaneous and acted laughs belong to different distributions.

2.5.7 Limitations and benefits

The biggest limitation of the AVLC database might reside in the absence of active communication provided by the subjects. Unlike popular databases used in laughter processing like the ICSI Meeting Corpus [Janin *et al.* 2003] or the AMI Meeting Corpus [Carletta 2007], participants had no one to interact with. It has been shown that the conversational partners influence the way we laugh [Campbell 2007]. The laughs from the AVLC database, obtained without conversational partners, might be considered as the "base" laughs from our participants, when they are alone watching a movie, and we have no guarantee these people would laugh the same way when they interact with other people. The most dramatic consequence of the absence of interaction is the very small number of speech-laughs, much less than in human conversations (speech-laughs are as numerous as breath-laughs in the ICSI Meeting Corpus). In ad-

dition, people knew they were being recorded, which is the case of many databases but must still be kept in mind. The protocol (stimulus induction, etc.) was meant to elicit reactions as natural as possible given the constraint of wearing markers on the face. The main benefits of the database are: the number of laughs (over 1000), their variety both in duration and acoustic contents, the presence of visual information including motion tracking, and the annotation focusing on laughter. Finally, the database contains some acted laughs, recorded by the same subjects at the end of the experiment. All the recorded signals, annotations and stimuli videos are freely available for research purposes.

2.6 Summary and perspectives

In this chapter we have reviewed the existing laughter databases. We have seen that several recording settings have been used, from the most natural to the most artificial. Different databases also focus on different social contexts, with some corpora trying to suppress social context (participant is alone in a room) while other settings introduce social context (from the mere presence of experimenters in the room to group conversations). Laughter databases also differ in the set and quality of recorded modalities. Trade-offs have been made to favor one modality or the other and have implications on the whole recording settings (scenario, social context, etc.): for example if one wants high quality body motion tracking, (s)he has to equip participants with motion capture suits (which makes totally natural settings impossible) and allow them to move freely (making frontal facial view difficult). It is also to cope with the impossibility to record some signals with sufficient quality in totally natural conditions that several techniques have been developed to induce laughter—watching humorous stimuli being the most frequently employed technique.

Further works on laughter databases are clear. The more data, the better, so researchers should continue to record—voluntarily or not—laughing people. We can identify three main directions, which are all addressed by the ILHAIRE project. First, investigate the multi-cultural aspects of laughter itself (not the sense of humor) to see whether the laughter patterns or their decoding vary from one culture to another. Second, continue the trend to increase the range of recorded signals: in addition to classic cameras and microphones, recent data collections have included new channels such as depth cameras (e.g., Kinect), motion tracking or respiration belts. All these signals are important to better understand, detect and synthesize laughter. Third, laughter synthesis in particular would benefit from large amounts of laughter data from one single laugher, so as to better model one laugher's voice and style. ILHAIRE attempts to tackle this by inviting the same participants over and over again to recording sessions with various scenarios (see Belfast storytelling sessions, Belfast conversational dyads, Belfast motion capture sessions, etc.). The AV-LASYN database has also been recently recorded with the specific purpose to have a large quantity of audiovisual data from a single subject, to train laughter synthesis.

Even though recording devices become more and more portable (size, weight,

memory) and convenient for natural recordings, we think that there is still a need for scenarios to induce spontaneous[11] laughter in controllable settings (light, noise, frontal view, etc.). These scenarios also slightly reduce the post-processing time for segmenting laughs (as laughter can be expected to be more frequent than in everyday recordings), even though we will see in Chapter 4 that laughter can be reasonably well detected in (audio) recordings.

In this PhD Thesis, we have used the AVLC database. We were interested in clean audio data, in as natural laughs as possible and, for collaborative works, facial motion capture was also necessary. At the time the project started, there existed no such database and this is why we have recorded the AVLC corpus. Although other laughter databases with high-quality audio have been recorded since then, AVLC is still one of the databases that contains the cleanest laughs (i.e., without noise caused by other people) from individual participants. This is essential for synthesis, where it is required to have a lot of examples from a single voice. More importantly, we will see in the next chapter that the AVLC database, in addition to the presence of facial motion capture (which in this Thesis was mostly used in collaborative works and applications), is unique in the way laughs have been transcribed and annotated in intensity[12].

[11] "Inducing spontaneous" seems antinomic, but we mean here inducing as spontaneous laughs as possible.

[12] The AV-LASYN database is now close to match the AVLC features, for a single participant, but with a large quantity of data.

Chapter 3

Hierarchical description of acoustic laughter episodes

Contents

In this chapter we will focus on the audio properties of laughter. As for speech, laughter can be studied at different levels, from the acoustic properties of short time frames to the analysis of the meaning of larger utterances in a given context. In [Rajman *et al.* 2007], the following levels of natural language analysis are identified:

- Phonetics: the study of the sounds used in natural language from an acoustic or an articulatory viewpoint, i.e. with a short time perspective.

- Phonology: the study of the sounds as they are perceived and interpreted by native speakers, that is linguistic units capable of distinguishing meanings, which are called "phonemes"[1].

- Suprasegmental phonology: the study of phenomena that bear on units larger than phonemes, such as syllables. It is for instance the syllable that bears features like height, intensity and duration.

- Morphology: the study of the internal properties of natural language words, including word formation rules. It is the level of "morphemes", which are units that carry a meaning, and of the combination of morphemes to create words.

- Semantics: at this level the meaning of individual words and sentences is studied, independently of their context of use.

- Pragmatics: this level focuses on the interpretation of sentences with respect to their context of use and on the relations between sentences.

Laughter is not emerging from the same construction principles as speech. In particular, laughter does not have phonological and morphological rules of which units carry a meaning and may be combined together. Nevertheless several analysis levels can be considered for laughter as well. The chapter will begin with a description of the different levels at which laughter has already been studied. Section 3.1 will be centered on this state-of-the-art and will show that, besides trying to infer the multi-level structure of laughs, there have been laughter description works that can mainly be related to the pragmatic (interpreting social functions of laughter within its context), semantic (classification of entire laughs) and acoustic phonetic (description of instantaneous acoustic properties such as pitch or formant frequencies) aspects. Some contributions also deal with the suprasegmental phonology (duration of syllables and evolution of pitch over syllables) of laughter. Then, we will present the descriptive works that have been carried out in the framework of this PhD Thesis and which cover some fields that have not really been considered previously. Section 3.2 will focus on phonetic transcriptions. Section 3.3 is related to the overall intensity of laughs. Finally Section 3.4 briefly presents our annotation efforts to characterize laughs with frame-level intensity curves.

3.1 State-of-the-art

In this section, the acoustic analyses conducted on laughter are examined. First, works aiming at describing the structure of acoustic laughs are presented in Section 3.1.1. As respiration has a significant role in laughter structure and has audible

[1] The actual acoustic realization of a phoneme is called a "phone": it may differ from the standard or intended way of pronouncing the phoneme while still be properly related to the corresponding phoneme by native speakers.

effects, considerations about laughter respiration are given in Section 3.1.2. A short parenthesis is then made in Section 3.1.3 to briefly present speech-laughs.

After that, the acoustic laughter properties are described, from low-level to high-level perspectives. Section 3.1.4 refers to acoustic phonetic features and especially the fundamental frequency (f_0) and formant values encountered in laughter. Section 3.1.5 focuses on the phonetic level, with descriptions of the evolution of f_0 during a phone[2] as well as efforts to produce phonetic transcriptions of laughter. Section 3.1.6 presents the findings at the syllable and bout levels. Section 3.1.7 considers the classification of laughter episodes in types. Section 3.1.8 refers to characteristics of laughter episodes and their relationship with some laughter types. Finally, Section 3.1.9 summarizes the main conclusions of the presented works with respect to the contributions of this Thesis.

3.1.1 Laughter structure

Laughter sounds have been intriguing scientists for a long time. Describing these sounds is not an easy task, as they are both extremely variable and different from other sounds made by humans.

Laughter sounds are extremely variable: unlike speech, laughter does not seem[3] to follow production rules [Sudheer *et al.* 2009, Sathya *et al.* 2013]. Given this absence of phonological rules, we do not attempt to contrast laughter sounds, as we are not aware of the meaning conveyed by these sounds. In consequence, laughter varies from a subject to another: although it is contested by the small experiment conducted in [Sathya *et al.* 2013], characteristic patterns enable us to recognize people only by the way they laugh [Edmonson 1987, Chafe 2007]. In addition, the laughter signal of a single person may change a lot from one situation to another, influenced by many communicative and social factors, as explained in Section 1.2.

Laughter sounds are also surprisingly different from other sounds humans are making. In 1872, Darwin [Darwin 1872] wrote:

> *The sound of laughter is produced by a deep inspiration followed by short, interrupted, spasmodic contractions of the chest, and especially of the diaphragm. (...) But why the sounds which man utters when he is pleased have the peculiar reiterated character of laughter we do not know. Nevertheless we can see that they would naturally be as different as possible from the screams or cries of distress; and as in the production of the latter, the expirations are prolonged and continuous, with the inspirations short and interrupted, so it might perhaps have been expected with the sounds*

[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 laughter.

[3] If they exist, production rules have at least not been found yet. But in any case the rules of language are explicitly established by humans, while laughter is a spontaneous, universal behavior for which no rules of "how to laugh properly" have been specified.

uttered from joy, that the expirations would have been short and broken with the inspirations prolonged; and this is the case.

Since Darwin, laughter has been described by several research teams (e.g., [Ruch & Ekman 2001, Bachorowski *et al.* 2001]). Efforts have been recently conducted to standardize the laughter terminology. Two laughter categories are broadly accepted (see Section 3.1.7): voiced (involving quasi-periodic vibrations of the vocal folds) and unvoiced laughter. In the voiced laughter set appears the stereotypical melodious laugh (e.g., "*hahahahaha*"), and a hierarchical structure has been proposed by Trouvain [Trouvain 2003], as illustrated in Figure 3.1:

- Episode: The entire laughter utterance is called an "episode".
- Bout: An episode can contain several "bouts", which are exhalation periods, delimited by inhalations[4].
- Syllable: Each bout is further decomposed in laughter syllables, formed by the concatenation of a vowel and a consonant. The consonant is generally a fricative (aspired "h" sound) [Ruch & Ekman 2001].
- Consonants and vowels: The basic components of voiced laughs are alternating vowels (voiced) and consonants (unvoiced, mostly breath-like "h"). Vowels generally present higher energy values than consonants.

Using a broader definition to take unvoiced laughs[5] into account as well, Ruch and Ekman [Ruch & Ekman 2001] and Chafe [Chafe 2007] use the term *pulse* to refer to a burst of expelled air. Bachorowski et al. [Bachorowski *et al.* 2001] and Sundaram and Narayanan [Sundaram & Narayanan 2007] denote these bursts as *calls*. A laugh bout is composed of laugh pulses or laugh calls, and each pulse can be voiced or not. Pulses or calls are separated by pauses with significantly lower energy, called *inter-call* by Bachorowski et al. In the case of voiced laughs, the calls are typically formed by vowels and the inter-calls are formed by h-like consonants, hence the combination of a call and an inter-call forms a syllable. But even in open-mouth unvoiced laughter syllables[6]—that at first sight are close to simple exhalations—, laryngeal friction usually makes laughter exhalations more audible than during normal breathing [Chafe 2007]. Ruch and Ekman [Ruch & Ekman 2001] also use the term *cycle* in place of *bout*.

The presented structure mainly focuses on the stable, rhythmic part of laughs. Several researchers have also considered the beginning and ending phases of laughs, which generally differ from the rhythmic structure.

[4] Chafe [Chafe 2007] calls "phrase" the segment formed by a bout and its successive inhalation.

[5] Unvoiced laughs do not contain voiced segments, hence syllables could not be formed with the definition given above as there is no vowel.

[6] Closed-mouth laughter syllables are generally powerful enough to be audible as well, in particular grunts that imply resonances in the laryngeal cavity, as well as strong exhalations through the nostrils—called nareal fricatives—as will be introduced later in this chapter.

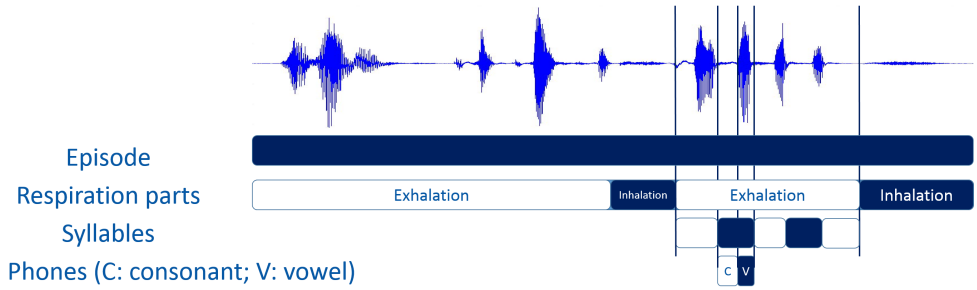

Figure 3.1: Hierarchical structure of laughter. (Exhalation parts = *bouts*)

For acoustic laughter synthesis and considering only melodious laughs, Lasarcyk and Trouvain [Lasarcyk & Trouvain 2007] proposed the following structure[7]:

- An *onset*, caused by a strong exhalation.
- The *main part*, composed by a succession of laughter syllables, each of them containing a voiced (laughter call) and an unvoiced part (inter-call interval).
- A *pause*.
- The *offset*, including at least one audible inhalation.

Analyzing the 178 laughter bouts of their corpus, Kipper and Todt [Kipper & Todt 2007] observed a similar typical structure of laughter episodes, but expressed it differently:

- Initialization by one or two "singular" elements, where "singular" means that at least two features among the maximum f_0, the duration of the element and the interval between two elements varied by more than 50% with respect to the next element.
- A homotype series, composed of similar (i.e., non-singular) elements, thus somehow predictable. This homotype series was present in 95% of the analyzed laughs, and generally formed the major part of the laughter bouts. The average periodicity of elements within this series was 213 ms, with variations between individuals but no difference between men and women.
- Sometimes singular elements again.
- Ending with a sound during inhalation.

[7] This structure is similar to the definitions given by Ruch and Ekman [Ruch & Ekman 2001], who used also visual cues (smile, etc.) to characterize the laughs.

A similar structure can be inferred from the various examples presented by Chafe [Chafe 2007]: an initiating burst is commonly found, followed by "normal" and somehow regularly spaced exhalation bursts, and usually a recovery inhalation. The inhalation is most of the time unvoiced, but can be voiced.

In the remainder of this Thesis, we will use the following terms, as defined earlier and illustrated in Figure 3.1: vowels, consonants, syllables, bouts, episodes. The definition of "syllable" is extended to the unvoiced case by considering that a syllable is formed by a call and an inter-call duration. In addition we will refer to frames, which are short-time (typically around 30 ms) windows used for local acoustic analysis. Each unit (vowel, consonant, syllable, etc.) is a sequence of frames.

3.1.2 Laughter respiration

A short note is worth to be made about laughter respiration. Unlike speech, laughter is disrupting normal breathing [Chafe 2007] and audible inhalations can occur in laughter. Respiration occurs in cycles of inhalations, inhalation pauses, exhalations and exhalation pauses. A resting state typically involves 14 cycles per minute, and the ratio between the durations of inhalations and exhalations is around 60%. The frequency of respiration cycles remains in the same boundaries during laughter, but the proportion of inhalation periods diminishes under 40% of the exhalation duration [Ruch 1993]. Exhalations and inhalations are stronger during laughter, and the respiration amplitude can be up to 2.5 times higher than in normal breathing.

No matter where in its respiration cycle the laugher is, laughter usually begins with a forceful exhalation that expires the tidal volume[8] [Ruch & Ekman 2001]. This can explain the initial explosive exhalation in some laughs, which serves to expel the excessive amount of air in the lungs. Then, the following pulses are initiated around the functional residual capacity[9] and laughs can terminate close to the residual volume[10] [Ruch & Ekman 2001]. This causes an urgency to inhale again, which can lead to the strong, audible inhalations that frequently occur in laughter.

As we have seen in the previous section, in the case of (sustained) laughter, such audible inhalations can take place in the middle of the laugh, clearly delimiting bouts. Inhalations are considered by Jefferson et al. [Jefferson *et al.* 1987] and Glenn [Glenn 2003] as indicating to conversational partners that they can take the speaking turn, in which case the laugh would probably stop[11]. If no-one is speaking following the inhalation, the laugher may (or may not) continue to laugh and produce a new series of forceful exhalations and inhalations. The laugher can then either stop laughing or go on with laughter. Regular breathing is resumed in a few breaths after

[8] The tidal volume is the volume of air usually displaced during normal breathing.

[9] The functional residual capacity is the volume of air remaining in the lungs after a normal expiration.

[10] The residual volume is the volume of air remaining in the lungs after a forced, maximal expiration.

[11] A decrease in the intensity of the exhalation pulses was identified as another feature that can indicate a possible termination of the laugh, and consequently the opportunity for conversational partners to talk again.

the end of the laugh [Filippelli *et al.* 2001].

3.1.3 Speech-laughs

Laughter and speech can occur simultaneously. The phenomenon is generally called "speech-laugh" and we will briefly present it here. Although it is contested by Provine [Provine 1993], who claims that laughter and speech scarcely ever co-occur, speech-laughs seem to be frequent in conversations: for example 18.6% of the laughs in Nwokah et al.'s mother-infant interactions were co-occurring with speech [Nwokah *et al.* 1999], speech-laughs represent around 10% of the number of isolated laughs in the ICSI Meeting Corpus (see Section 2.2.1.3) and speech-laughs account for 60%[12] of the laughter labels in the dialogs analyzed by Trouvain [Trouvain 2001]. Comparing speech, speech-laughs and pure laughs during mother-infant interactions, Nwokah et al. found that speech-laughs are generally longer than isolated laughs and have more energy variations than regular speech or isolated laughs. Surprisingly, Nwokah et al. also found that speech-laughs have similar f_0 values as regular speech (while isolated laughs occur at higher frequencies). The latter finding is unexpected as smile usually yields to speech with higher f_0. One explanation could be that mothers tend to speak with higher pitch when talking to their infants. Regarding the places of occurrence of speech-laughs, Nwokah et al. found that they are more frequent during statements than during questions or exclamations. They usually affect two words, but it can be only one word or more than five words.

Speech-laughs acoustic patterns vary from one speaker to the other, but generally laughter is affecting speech by vowel elongation and/or syllabic pulsation[13], and sometimes breathiness and modification of the pitch contour. Although the authors stated that other analyses in different contexts are required to confirm their findings, their conclusions are worth mentioning [Nwokah *et al.* 1999]:

> *A speech-laugh retains the vowels and consonants of the speech, usually adapts to the fundamental frequency of the speech, incorporates the repetitive rhythm and glottal stops/fricatives of laughter, and adopts the amplitude contour of laughter in most cases and (sometimes) the characteristic breathiness of laughter. The duration of the laugh increases, but because the consonants and vowels of speech remain intact, there is sufficient lack of distortion so that the speech is as perceptually intelligible as it would be without laughter. (...) It was also observed that speech appears to be slowed and exaggerated as a result of laughter. (...) (Speech-laughs) only occurred within one breath, and the basic rhythm of laughter was maintained. (...)*

[12] This number is probably an over-estimate as it includes utterances that are perceived rather as speech-smiles by many listeners, but even with a cautious threshold speech-laughs represent at least 10% of the laughter labels.

[13] Syllabic pulsation was defined in [Nwokah *et al.* 1999] as "vocal modulation that usually occurs on vowels producing repetitive-like segments. Each segment is often preceded by a sharp glottal attack or fricative /h/".

> *If the laughter is forceful or intense, it appears to override the speech. However, speech may take increased priority if a mother needs the child to hear her words clearly but is still laughing at the child's antics. This may happen, for example, when the mother is expressing caution. (...)*
>
> *There is no completely predictable result to the reorganization of the two vocal outcomes—each speech-laugh is unique. We have established that the laughter structure appears to have a self-organized timing pattern or laughter signature for each person that is preserved but also adapts with the addition of the rhythmic and phonetic structure of speech.*

Trouvain [Trouvain 2001] analyzed eleven utterances that were confidently scored as speech-laughs by ten naive raters. He also states that speech in speech-laughs remains intelligible. He observed similar acoustic characteristics of speech-laughs as Nwokah et al's: breathiness, stronger aspiration sounds and vibrato, especially in the vowel regions. Speech-laughs were expanded over two syllables most of the times (no matter what the length of the words). The conclusions of Trouvain's analysis are the following [Trouvain 2001]:

> *It is clear that the simultaneous production of speech and laughter is not simply laughter superimposed on articulation. The articulatory configurations for speaking are continuously maintained during speech-laughs. Traces of laugh can be found in increased breathiness and sometimes vibrato on the voiced portions, and a reinforced expiration on phonologically possible locations (e.g., after a plosive release or during an unvoiced segment). A mere superimposing of laughter on speech would probably destroy the temporal relationship between consonant(s) and vowel in a speech syllable, would severely affect the spectral properties of the consonants, and would destroy the local intensity scaling. The sparse data presented here do not allow powerful statements on the acoustics, the frequency and the location of speech-laughs. Nevertheless it became evident that there is indeed no prototypical pattern for speech-laughs.*

Chafe [Chafe 2007] also dedicated one chapter of his book to describe speech-laughs. He shows how laughter modifies regular speech by syllabic pulsation: provoking amplitude modulations (tremolo) as well as inserting aspiration sounds which make the speech more broken. The rhythm of tremolo or laughter-pulse insertions in speech is generally higher than the common rhythms of pure laughter.

Speech-laughs are not considered in this Thesis that focuses on the production of pure laugh, which is different than the production of speech and its modification due to various phenomenons (stress, smiling, laughing, crying, etc.).

3.1.4 Fundamental frequency and formant values

The acoustic parameter of laughter that has been most widely reported is its fundamental frequency (f_0). However, results must be considered with caution as, most

of the time, methods for estimating f_0 in regular speech have been simply used to compute f_0 in laughter, although the two phenomena are quite different. Vocal folds do not close as decisively in laughter as in speech [Chafe 2007], which hinders the automatic estimation of f_0. In addition, laughter f_0 may differ from regular speech frequencies, as will be shown in this section. In other words, methods developed and optimized for computing f_0 in modal speech may not be reliable estimators of f_0 in laughter. For example, Bachorowski et al. [Bachorowski *et al.* 2001] manually checked the outputs of the RAPT f_0 tracking algorithm [Talkin 1995]—which is part of the ESPS Toolbox included in the widely-used Wavesurfer sound analysis software [Sjölander & Beskow 2011]—and realized that the algorithm performed well in only 65% of the cases.

Generally speaking, what is striking is the large variability of f_0 values in laughter. Most researches have also shown that the f_0 range in laughter is wider than in speech and that we are generally laughing with higher f_0 than in regular speech. Bachorowski et al. [Bachorowski *et al.* 2001] obtained an average f_0 of 405 Hz for females and 272 Hz for males, while average f_0s in speech are 220 and 120 Hz, respectively. The maximum f_0 value was 2083 Hz for females and 1245 Hz for males [Bachorowski *et al.* 2001]. Fundamental frequency was also shown to be higher in open-mouth calls than closed-mouth calls. Using laughs produced by actors (see Section 2.4) and only focusing on long (>3 s) episodes, Szameitat et al. [Szameitat *et al.* 2007] obtained fundamental frequencies averaging at 476 Hz for females and 199 Hz for males, with maximum values of 1765 Hz and 595 Hz, respectively. In their corpus of laughs occurring in conversations[14], Vettin and Todt [Vettin & Todt 2004] computed the median of the participants' median average f_0[15] to be 315 Hz for females and 171 Hz for males. These values are lower than what was previously presented, but a) they are still higher than modal speech; b) it is not straightforward to compare means and medians; and c) the considered data are quite different (ten participants in natural conversations, instead of participants watching humorous clips or actors portraying emotions). Vettin and Todt also measured high intra-individual variability in laughter acoustic parameters.

Apart from the fundamental frequency in laughter, several researchers also investigated the position of formants[16]: several studies [Bachorowski *et al.* 2001, Szameitat *et al.* 2007, Tanaka & Campbell 2011] showed that vowels constituting the main part of voiced laughs are mainly unarticulated central vowels. This was expected from the physiological aspects of laughter: the vocal tract is in a relaxed position and raised lips corners and widely opened jaws make articulation difficult

[14] Hence there can be conversational/social laughs in this corpus as well as hilarious laughs.

[15] In other words: they measured the average f_0 of each laugh, then took the median value for each participant, and finally computed the median of the resulting values across all participants.

[16] Formants are the resonances in the spectral contour of speech signals caused by the shape of the vocal tract. Formant positions are commonly used to distinguish vowels. Each vowel is produced with a specific configuration of vocal tract (tongue, lips, jaws, etc.), giving it particular spectral peaks. The frequency of the first formant is denoted F1, the frequency of the second formant F2, and so on.

[Ruch & Ekman 2001]. Edmonson [Edmonson 1987] however used a large spectrum of vowels to transcribe laughs and suggested that deviations from central vowels towards front vowels indicate a higher-degree of self-consciousness and inhibition, while low back vowels (and nasals) are related to higher self-assertiveness. In their corpus of portrayed laughs, Szameitat et al. [Szameitat *et al.* 2007] obtained, for females, voiced sounds mostly falling in the (ʌ) and (a) areas of Hillenbrand et al.'s [Hillenbrand *et al.* 1995] standard vowel-space representation (see Figure 3.2). Males' vowels mostly fell in the same regions and in the (ɜ) range (see Figure 3.3).

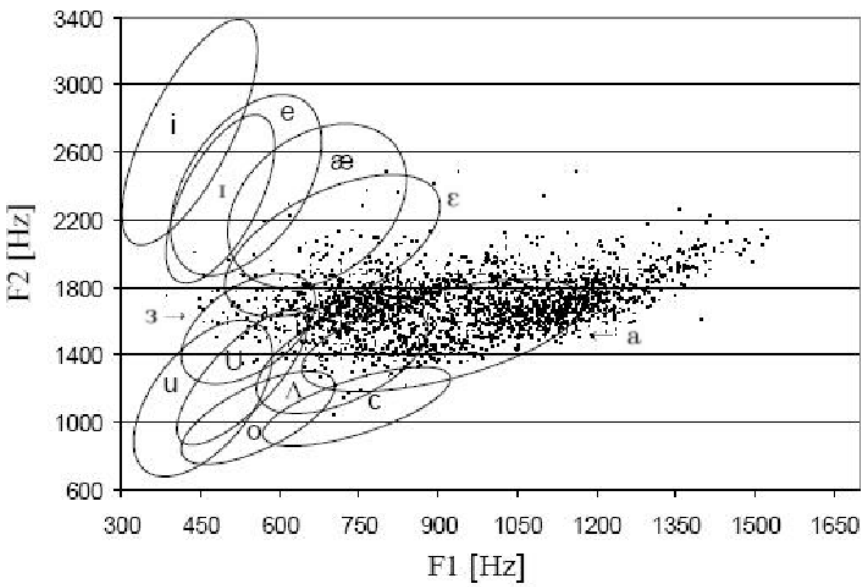

Figure 3.2: F1-F2 plot of female laughter vowels, according to Hillebrand et al.'s vowel representation [Szameitat *et al.* 2007].

As Figures 3.2 and 3.3 illustrate, frequencies of the first formant (F1) of laughter syllables of Szameitat et al.'s corpus can reach very high values [Szameitat *et al.* 2007]: 26% of F1 values are over 1000 Hz, with peaks over 1500 Hz for females and 1300 Hz for males. The average was 924 Hz for females and 728 Hz for males. These large values for F1 can be caused by a wide jaw opening or pharyngeal changes in "pressed" voice [Tanaka & Campbell 2011]. Szameitat et al. indeed noticed that these syllables with large F1 sounded as if they had been produced with a hard, pressed voice. In addition, they obtained formant frequencies higher for females than males, for all the first five formants. Again, these results are somehow contested since Bachorowski et al. [Bachorowski *et al.* 2001] reported different trends, with the Frequency of the fourth formant (F4) in the same range for males and females and the Frequency of the

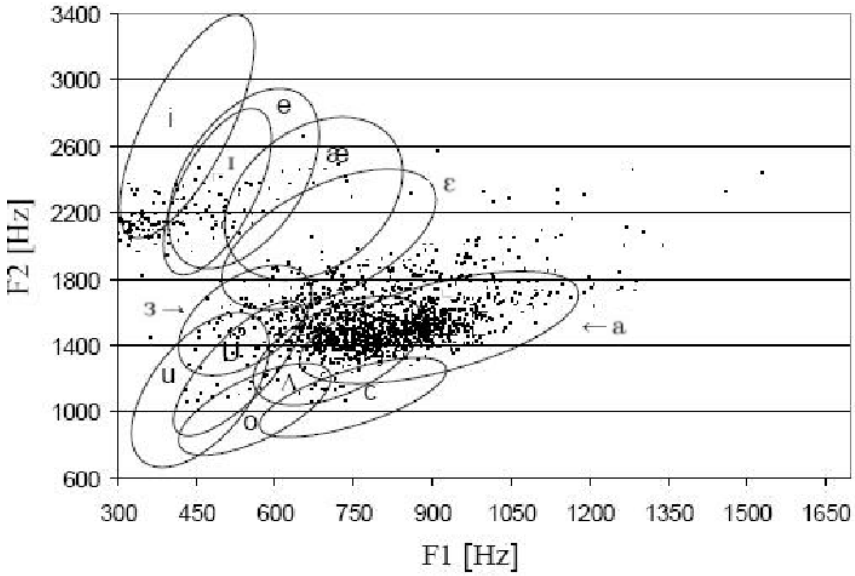

Figure 3.3: F1-F2 plot of male laughter vowels, according to Hillebrand et al.'s vowel representation [Szameitat *et al.* 2007].

fifth formant (F5) even higher for males than females in the case of open-mouth voiced laughs, while only F3 values showed gender differences in the case of closed-mouth laughs. It is important to remind that Szameitat et al. analyzed laughs portrayed by eight professional actors, while Bachorowski et al. [Bachorowski *et al.* 2001] used elicited laughs from 97 subjects. This can explain some differences. Bachorowski et al. obtained much lower F1 values (average: 653 Hz for females; 535 Hz for males). Most of the female calls fell in the ɛ and a regions of Hillenbrand et al.'s vowel space representation, while the most frequent male vowels were ɜ and ʌ.

Both studies report that some calls fell in other vowel regions like ɪ and o. Szameitat et al. also obtained some examples in the æ, ɔ, i, e and ʊ regions.

Finally, calls of a same laughter bout are consistent: one vowel is chosen for the bout, and it generally does not change among calls [Edmonson 1987, Bachorowski *et al.* 2001]. Szameitat and al. even reported that individual laughers tend to use the same set of vowels, which could be one clue to identify people by their laughs. But it is not always the case, and Chafe [Chafe 2007] shows an example of a laugh with a change in vowel quality.

3.1.5 Analysis of call-related features and phonetic transcriptions

3.1.5.1 Evolution of fundamental frequency during calls

Bachorowski et al. [Bachorowski *et al.* 2001] conducted analyses of fundamental frequency (f_0) at the call level (i.e., one phonetic unit). It revealed that f_0 is highly variable within a single call, with average standard deviations of 21.41 Hz and 29.98 Hz for males and females, respectively. The average f_0 range within a call (maximum f_0 minus minimum f_0 of a call) was 59 Hz for males and 86 Hz for females. The average f_0 excursion within a call (initial f_0 minus final f_0 of the call) was 44 Hz for males and 64 Hz for females. There was no dominant f_0 pattern during calls, most f_0 curves were labeled as "flat" (38%), followed by "falling" (29%), "sinusoidal" (19%), "arched" (8%) and "rising" (6%).

3.1.5.2 Phonetic transcriptions

Many terms (snorts, pants, chuckles, cackles, giggles, etc.) have been used by researchers to characterize the sounds of laughter. However only few scholars have actually made the effort to transcribe *all* the sequences of sounds that occur in one (or even better, *any*) laughter episode.

One of such works is due to Gail Jefferson [Jefferson 1985] who analyzed conversations. She is stating that more information about the dynamics of the conversation can be inferred if laughs are actually transcribed (for example "heh-heh-heh") rather than simply reported (i.e., "Mr. X laughed"). In consequence, she proposes a system for transcribing laughter in conversations—which includes transcribing how laughter is modifying regular speech when they co-occur (speech-laughs)—that has been adapted by Glenn [Glenn 2003] for the same purpose of analyzing conversations. The system includes transcriptions of the canonical laughter sounds through standard letters ("h" for the aspired laughter consonants, English vowels and consonant "n" for the voiced parts) as well as a code for breathing sounds that frequently occur in laughter ("hhh") and a symbol to indicate inhalations. Other symbols are sometimes used to provide additional information on pitch (raising, constant, falling), energy (higher or lower volume) and deviations from modal speech (emphasis, pauses, stretched sound, etc.). The full system of symbols can be consulted in Glenn's book (pages xi-xii) [Glenn 2003]. The code also enables to indicate simultaneous vocalizations between the conversational partners—or at least relate the beginning and ending of participants' vocalizations, as the proposed system is not very accurate regarding timing and no duration information is included—which is crucial information for conversation analysis.

Edmonson [Edmonson 1987] gives phonetic transcriptions for a set of laughs. He used phonetic symbols from the International Phonetic Alphabet (IPA) [International Phonetic Association 1999] to characterize laughter aspirates, glottal stops, vowels and nasals. Symbols are also used to give information about the pitch

pattern, the length and the presence of stress. Transcriptions are provided for 83 laughs from different laughers and cultures, to illustrate various phonetic forms laughter can take. Efforts are concentrated on the description of "stable parts" of bouts (fricative-vowels) and sounds outside of the IPA (inhalations, grunts, snorts, etc.) are not transcribed. Furthermore, it is unknown whether the corresponding acoustic samples are available.

To describe the states that the larynx can take during laughter, Esling [Esling 2007] employs phonetic transcriptions and focuses on the diacritics (i.e., symbols added to a letter, like "ẽ" to denote a nasalized "e") to illustrate the changes in vocal quality (breathy, creaky, whispery voice). Phonetic transcriptions are used for theoretical illustrations mostly, as the recorded instances of laughter are not transcribed. Using nasoendoscopic videos to analyze six instances of spontaneous laughter, Esling found that mostly the breathy and modal modes of the larynx are appearing in laughter.

Chafe [Chafe 2007] also proposes a system for transcribing laughter (see page XIII of his book) in conversations (i.e., in the middle of speech), making mainly the distinction between voiced and unvoiced inhalations and exhalations as well as closed-mouth exhalations. Nevertheless, he does not use this system to describe laughs in isolation (i.e., without surrounding speech), but introduces ad hoc terms like "SH", cough, snort, sniff, glottal clicks or creak.

Campbell [Campbell 2007] mentions the "transcription of laughs in Japanese alphabet, wherever possible", for his telephone conversation laughs.

Tanaka and Campbell [Tanaka & Campbell 2011] labeled laughter calls from a single participant with four broad categories (nasal, ingressive, vocal or chuckle).

Two examples of phonetic transcriptions of laughter are provided in [Pompino-Marschall *et al.* 2007]. The total number of laughs transcribed that way is unknown and these transcriptions are related to a small set of acted laughs (as part of a movie).

Finally, the works in [Bachorowski *et al.* 2001, Szameitat *et al.* 2007, Tanaka & Campbell 2011] on the identification of vowels used in laughter (through analysis of formant positions, as we have seen in Section 3.1.4) can be related to phonetic transcriptions, as it concerns identifying and labeling the laughter sounds. Nevertheless these do not form phonetic transcriptions in any way, as no effort was made to transcribe entire laughs (including other sounds than "speech-like vowels") and these works served to identify global trends rather than label individual instances (even the "vowels" have not been labeled along laughter occurrences).

3.1.6 Syllable and bout patterns

In this section we will focus on works that describe laughter properties at the syllable and bout levels. These works have mainly focused on two important quantities: duration and evolution of the fundamental frequency.

3.1.6.1 Syllable rhythm

The rhythm of laughter syllables has been computed by many researchers (see [Ruch & Ekman 2001] for a summary). The laughter syllabic rhythm is similar to speech [Bickley & Hunnicutt 1992]. It was measured between 4 and 6 Hz, with most researchers agreeing on a rhythm slightly below 5 Hz, or a syllable duration slightly higher than 200 ms. Bachorowski et al. [Bachorowski *et al.* 2001] have shown that bouts generally start with one call approximately twice as long as the following calls and that the terminating call of long bouts (over five calls) is generally longer than the middle calls. In addition, the inter-call duration tends to increase during the course of a bout. The syllable rhythm does not vary much however, as the duration of calls themselves tends to slightly decrease over the course of a bout (with the exception of the terminating call mentioned earlier) which roughly compensates for the longer inter-call durations. As illustrated by Chafe [Chafe 2007], series of inhalation pulses can also occur, but in this case the rhythm will be significantly lower than for exhalation pulses: around 1.5 Hz.

Finally, Chafe [Chafe 2007] showed that the energy envelope of laughter pulses is asymmetrical, hence playing laughter backwards does not yield natural laughter sounds.

3.1.6.2 Bout duration

Regarding the duration of bouts, Laskowski and Burger [Laskowski & Burger 2007a] found that the duration of the bouts in the ICSI Meeting Corpus follows a log-Gaussian distribution with a peak around one second. Voiced bouts tend to be slightly longer than unvoiced ones [Laskowski & Burger 2007b]. Bachorowski et al. [Bachorowski *et al.* 2001] report a high variability in laughter bout duration, as the average of 0.87 s is associated to a standard deviation of 0.77 s. We obtained higher values on the AVLaughterCycle database, as bouts lasted 1.69 s on average, with a standard deviation of 1.52 s [Urbain & Dutoit 2011]. The duration of inhalation phases was smaller, with an average of 0.36 s and a standard deviation of 0.15 s. Chafe [Chafe 2007] also indirectly reported about the duration of bouts, indicating that the number of pulses can vary from one to twelve or even more. He insisted on the fact that a bout with only one pulse is not uncommon, something that is also noticed by other researchers (e.g., [Urbain *et al.* 2010a]). Bachorowski et al. [Bachorowski *et al.* 2001] measured the average number of pulses in a bout to 3.61 for males and 3.20 for females.

3.1.6.3 Fundamental frequency inside bouts

Regarding the evolution of fundamental frequency (f_0), as for many other laughter properties, high variability appears to be the main conclusion. Ruch and Ekman [Ruch & Ekman 2001] reported a decrease of pitch and intensity in the course of a bout. Vettin and Todt [Vettin & Todt 2004] obtained opposite results, as they

measured the fundamental frequency (f_0) at the end of a bout to be generally higher than the initial f_0 of the bout: the variation of f_0 is ranging from -68% to +129%, with a median of +12%. The median f_0 excursion within a bout (i.e., the difference between maximum and minimum f_0 within a bout) ranged from 0% to 228% of the minimum f_0, with a median value of 28%.

Bachorowski et al. [Bachorowski *et al.* 2001] stated that f_0 routinely increases and decreases during the course of bouts, without any obvious pattern emerging. Again, the authors insist on the large variability of f_0 within bouts—and even more so in long bouts, which also have higher average f_0 than short bouts. Very large variations of f_0 within a single bout are not uncommon, as 7 males and 13 females (out of their 97 participants, see Section 2.3) produced at least one bout with a variation of 500 Hz or more. There was no evidence of a decrease (or increase) in f_0 during bouts. In accordance with these conclusions, Chafe [Chafe 2007] gave examples of various pitch patterns (rising, declining) for both bouts and syllables.

3.1.6.4 Variability inside bouts

Variability in successive syllables appears to be a key parameter of natural laughter. Vettin and Todt [Vettin & Todt 2004] measured that the median variations in duration and f_0 of successive syllables were 43% and 16%, respectively. These findings confirm two studies conducted by Kipper and Todt.

In a first study [Kipper & Todt 2001], they manipulated a series of seven vowel-like calls from a human laugh, with an average syllabic rhythm of 200 ms, to create three types of rhythmic patterns: a) the original pattern (isolating the seven calls from the laughter onset); b) a laugh with standardized rhythm, created by copying seven times with a fixed interval one of the laughter calls and c) a "reversed" laugh obtained by reversing the order (but not the audio samples themselves) of the seven calls. Speed variations of each rhythmic pattern were also created by multiplying the playing speed of each laugh with factors ranging from 0.4 to 1.8, without modifying the spectral contents of the laughs. Naive participants were asked to tell whether each created laugh was laughter or not. The standardized rhythm laughs received poor evaluations (i.e., were perceived as not being laughter in around 50% of the cases), the reversed rhythm laughs were better rated and the original pattern obtained the best results (around 90% of raters perceived it as laughter). Regarding speeds, it was found that multiplying factors between 0.6 and 1.4 (which roughly correspond to syllabic periods between 140 ms and 300 ms) yielded similar results, while speeds out of this range lead to decreased laughter perception. These results were confirmed on samples from five different male laughers evaluated by different groups of raters. The study was completed with a second experiment in which the fundamental frequency of an eight-call human laugh was modified to create samples with decreasing (from 220 Hz to 110 Hz), constant (165 Hz) and increasing (from 110 Hz to 220 Hz) f_0 patterns. The laughs with decreasing f_0 obtained more positive evaluations—and closer to human laughs—on bipolar scales (happy-sad, pleasant-unpleasant, likeable-

unlikeable, contagious-uncontagious) than the ascending f_0 patterns and finally the constant f_0.

In a second study [Kipper & Todt 2003], Kipper and Todt first investigated more complex rhythmic patterns on six-call bouts and then combined variations of rhythm with variations of pitch patterns. Laughs were evaluated on three bipolar scales as well as through three questions rated on a five-point agreement-disagreement scale ("This series makes me laugh/smile", "The series sounds like laughter", "The series is funny"). Answers were grouped with principal component analysis. Regarding the perception of laughs when only rhythmic patterns vary, the first principal component grouped 5 questionnaire items (pleasant-unpleasant, interested-boring and the three questions presented above). Laughs with constant, predictable (linearly increasing or decreasing) or random rhythmic patterns received less positive scores than laughs with complex yet structured patterns (alternation of long and short calls, or long-short-short patterns, which they called "subphrases"). When varying both the rhythm and the pitch pattern (constant, decreasing or forming two high-medium-low "subphrases"), it was again the "subphrase" pattern that was the most positively evaluated, over samples in which the pitch was linearly modified. The "subphrase" laugh was even perceived as more pleasant and more happy than the corresponding human laugh, but the human laugh was found as sounding more like laughter and making people smile or laugh more. Again, the results were confirmed by similar experiments conducted with laughs from six additional laughers. As the influence of pitch was not evaluated separately from the influence of rhythm in this study, it is impossible to know whether the subphrase pattern for only one of the two features (rhythm, pitch) would have been sufficient to achieve the same effect as the joint modification of the two. In addition, only particular laughs were included in the study (same number of calls, same laughter vowel, etc.) and the majority of the laughs had unnatural settings, so generalization of these results to all laughter patterns should not be done. Nevertheless, these two studies from Kipper and Todt support the idea that (unpredictable[17]) variations of features from one laughter syllable to the other are necessary for the laugh to be perceived as natural.

3.1.7 Laughter types

3.1.7.1 Emotions and laughter

Laughter is highly variable and can occur in a range of social or emotional situations. Some researchers therefore suspected the existence of several types of laughter, related to affective meanings. An experiment conducted by Devillers and Vidrascu [Devillers & Vidrascu 2007] indeed showed that healthy humans are able to discrimi-

[17] For the listener! For the producer, these variations are probably unconsciously shaped. In these experiments, laughs were modified with prepared patterns, so the people who created the laughs can predict these patterns, but they are surprising to the naive listener. Laughter synthesis systems are thus also expected to produce patterns unpredictable for the listeners, even if those patterns are totally deterministic when looking at the mathematical models.

nate between different laughter valences (positive, negative or ambiguous) with agreement rates much higher than chance.

Kori asked an actor to produce 16 laughs covering a range of (hypothesized) types [Kori 1987]. Ten naive listeners rated each laugh on twelve scales (e.g., happy, embarrassed, self-deprecating). He observed that only three of the laughs obtained judgments corresponding to the emotion intended by the actor and concluded that "the content of the laughter is not always unambiguously encoded in its vocal output". Factor analysis performed on the ratings revealed two main lines, which Kori identified as pleasant-unpleasant and superior-inferior. He also investigated several features that contribute to these perceptions and found that pleasant laughs generally had a longer initial burst and a steeper decrease of energy along the bout than unpleasant laughs, while laughs perceived as superior had higher fundamental frequencies, longer durations between vowel onsets and softer decrease of energy than inferior laughs. These findings must however be taken with care as they result from a study with limited data (16 portrayed laughs) and ratings (ten subjects).

Szameitat et al. [Szameitat *et al.* 2009a] investigated the emotional question on portrayed laughter covering four affective categories[18]: joyful, taunting, tickling and schadenfreude (see Section 2.4). Naive participants listened to the laughs and assigned an emotional label (four choices corresponding to the four emotional categories) to each portrayed laugh. The overall recognition rate was 44%, significantly higher than chance level (25%). In a second experiment, using only 160 laughter sequences that were classified above chance level, naive participants were asked to give dimensional ratings to each laugh on a 4-point scale. Four dimensions were assessed: the frequently used valence, arousal and dominance of the speaker, as well as the valence towards the receiver of the laugh (as it was hypothesized that all laughs would have positive arousal and valence for the speaker, but would differ in dominance—with tickling being highly submissive while taunting should be highly dominant—and valence towards the listener—with negative values for schadenfreude and particularly taunting). Results indeed showed that the different emotional categories are associated to different values on the dimensional scales. Nevertheless, this study focused on portrayed laughter, and it is unclear whether these results can generalize to spontaneous laughs.

To verify the existence of laughter types within spontaneous laughs, experiments have been conducted with the Belfast Story-Telling sessions (see Section 2.3.4). The objective was to assess whether participants agree on laughter types when viewing the laugh (sound+video) without any contextual information (i.e., the laugh was segmented so that no linguistic cues before or after the laugh could be used). Two different annotation schemes were experimented [Hofmann *et al.* 2013]:

- UZH scheme: 13 emotions out of the 16 enjoyable emotions which triggered laughter: amusement, relief, contentment, excitement, wonder, visual, schaden-

[18] Tickling has not been proven to be related to emotions—it could simply be a reflex—but for ease of reading the authors included it in the category of emotional laughter.

freude, auditory, fiero, tactile, ecstasy, gustatory and olfactory[19].

- QUB scheme: ten laughter categories: surprised, anxious, backchannel, giggling, happy, hilarious, embarrassed, sad, polite and relieved.

An "other" tag was added to each annotation scheme, allowing annotators to indicate if the laugh did not fall into any of the proposed categories. The experiment was run online using Amazon Mechanical Turk [Amazon.com, Inc. 2014]. A large number of annotations was obtained: 290 people each rated ten laughs with the UZH scheme and 149 participants rated the laughs with the QUB scheme. Results showed poor inter-rater agreement. The best category, amusement, obtained only 30% of agreement. This is contradicting Szameitat et al.'s conclusions [Szameitat *et al.* 2009a] but in agreement with Kori's findings [Kori 1987]. Suarez et al. [Suarez *et al.* 2012] also asked raters to assign categorical and dimensional (on the valence-arousal plane) emotional labels to laughs. It was observed that people do not agree on emotional labels when they only see and hear the laugh. These outcomes tend to indicate that laughter alone does not convey unique emotional meanings, but is an ambiguous signal that is interpreted individually by the listeners and with the help of context[20]. This is in line with Edmonson's opinions [Edmonson 1987] as well as the affect-induction hypothesis of Owren and Bachorowski [Owren & Bachorowski 2003], who concluded that:

> *The critical point of departure lies in proposing that signals are used not to convey information about underlying state, but rather to influence perceived affect and associated behavior. (...) Thus, vocalizer behavior will depend on a combination of affect state, nonconscious goals, and the relationship to the listener. That combination can potentially trigger any of a variety of acoustically differentiated laughs, where the common element is the likely effect that the sounds will have on the other party. We therefore expect humans to produce laughter in virtually any situation where the effect of its affect-inducing features on perceivers can benefit the vocalizer. However, because the sounds are not primarily designed to convey cues to affect and social goals, signaler state and laugh acoustics will not show close, exclusive links.*
>
> *(...) human listeners hearing others laugh are likely able to infer something about the vocalizer's state from a combination of their own affective responses, an evaluation of contextual information, and past experience with laughers.*
>
> *(...) The affect-induction approach to laughter readily produces testable predictions, for instance beginning with the premise that laugh acoustics*

[19] For information the three enjoyable emotions that were not included as they did not conduct to increased proportions of laughter in the initial experiment are: gratitude, naches and elevation.

[20] It has been shown by ILHAIRE partners, although not published by itself as it is rather obvious, that people could identify the 16 emotions with acceptable accuracy when the laughs are presented within context formed by sentences before and after.

> *are not uniquely associated with particular nuances of laugher state, and that listeners are requisitely unlikely to infer those nuances from laugh sounds alone. (...) The ability that listeners have to infer vocalizer state is expected to be critically dependent on contextual cues, including the perceivers' own emotional states at the time, and their experience both with specific vocalizers and laughers more generally. It follows that two listeners hearing the same laughs could derive exactly opposite "meanings" from the sounds and attribute quite different emotions or intentions to the laugher.*

In consequence, it is now strongly suspected that laughter in isolation cannot be reliably classified into emotional types. Rather, context would be necessary to interpret laughter meanings. This would move the emotional analysis of laughter from semantics to the level of pragmatics.

3.1.7.2 Voiced and unvoiced laughter

Despite the inconclusive attempts to classify laughs in emotional types, there exist other ways to characterize laughs. Laughs can be classified based on their articulatory properties. One clear distinction is made between voiced and unvoiced laughs [Laskowski & Burger 2007b]:

- Voiced laughs: The source of energy is indeed a quasi-periodic excitation for at least part of the laugh bursts. In this category, we find song-like laughs (corresponding to Figures 3.1 or 3.4), as well as most chuckles and giggles.
- Unvoiced laughs: The excitation is fricative, there is no voicing. This category groups the open-mouth laughs sounding like panting, and the closed-mouth grunts and nasal snorts. These closed-mouth sounds can appear when trying to retain laughter [De Benedictis 2007], voluntary modifying the laughter sound, often accompanied by a hand movement to cover the mouth. The onset of the laughter episode can in this case be a strong inhalation instead of an exhalation. Some of these unvoiced laughs do not present the rhythmic structure described in section 3.1.1, but are more irregular, as illustrated in Figure 3.5, displaying the waveform of a unvoiced nasal laughter bout. Bursts of energy can still be noticed.

The distinction between voiced and unvoiced laughs was first introduced by Hall and Alliń [Hall & Alliń 1897]. Grammer and Eibl-Eibesfeldt [Grammer & Eibl-Eibesfeldt 1990] showed that these types of laughs have different effects: for instance male subjects were more interested in seeing again a female participant they had just met in the framework of the experiment if she had produced voiced laughter. The different effects of voiced and unvoiced laughs was further highlighted by Bachorowski and Owren [Bachorowski & Owren 2001] who revealed that voiced laughter yields significantly more positive emotional responses in listeners than

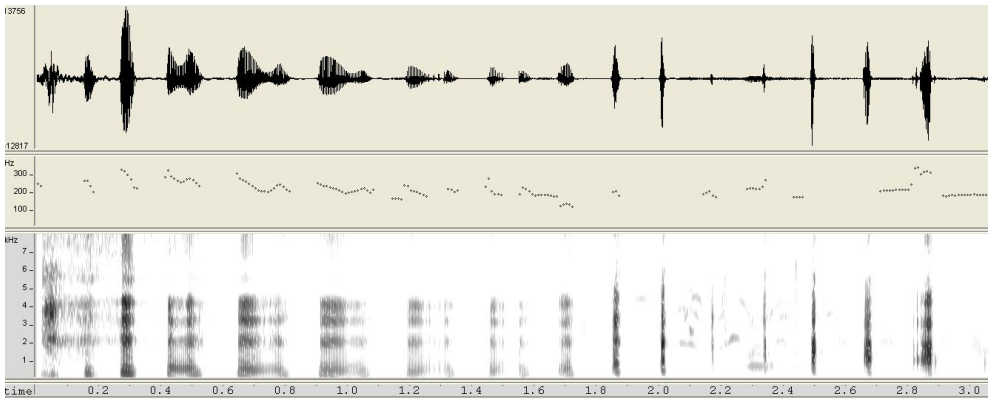

Figure 3.4: A stereotypical voiced laughter episode. Top: waveform; middle: f_0; bottom: spectrogram.

unvoiced laughter. Further splitting the unvoiced category in two, Bachorowski et al. [Bachorowski *et al.* 2001] distinguished the following three categories[21]:

- voiced, song-like laughs: laughs that are mainly composed of voiced syllables, including stereotypical episodes with vowel-like sounds, giggles and chuckles.
- unvoiced, snort-like laughs: laughs that are mainly unvoiced and composed of sounds resulting from turbulences in the nasal cavity.
- unvoiced, grunt-like laughs: mainly unvoiced laughs, with resonances in the oral or laryngeal cavities.

This way of categorizing laughs is appealing, as it is based on articulatory properties of the laughs (objective) rather than perceived emotional states (subjective). Bachorowski et al. have indeed obtained high agreement scores between two annotators classifying laughs among the proposed three categories. As already stated, voiced and unvoiced laughs have also been shown to have different functions in human communication, hence the distinction is not only convenient (relying on objective parameters) but also meaningful. We should however note that the distinction is not as objective as it might look at first sight: where is the boundary between "mainly voiced" and "mainly unvoiced" laughs? This criterion is provided neither by Grammer and Eibl-Eibesfeldt nor by Bachorowski et al., and the limit has been placed at different values by some authors. For example Petridis and colleagues consider laughter to be voiced if it contains at least 15% of voiced frames in [Petridis *et al.* 2013b], but at least 20% of voiced frames in [Petridis & Pantic 2011], while Truong and Trouvain [Truong & Trouvain 2012a] consider a laugh as voiced if it contains at least one voiced

[21] Examples are provided on the web: `http://www.psy.vanderbilt.edu/faculty/bachorowski/laugh.htm`, last consulted on February 18, 2014.

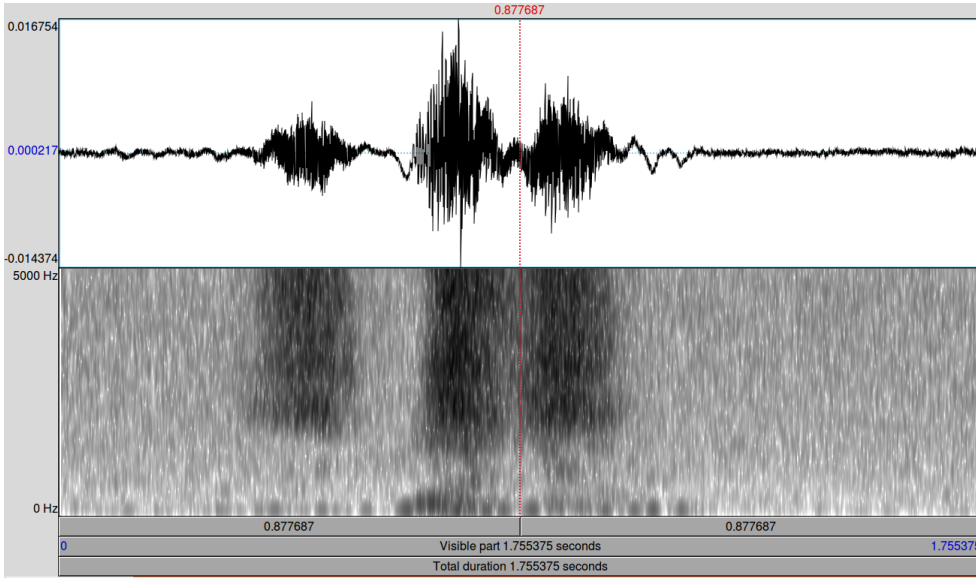

Figure 3.5: Example of an unvoiced laughter bout. Top: waveform; bottom: spectrogram.

frame. Laskoswki and Burger [Laskowski & Burger 2007b] also propose to label as a voiced laugh any laugh that includes at least one voiced call[22] and report quite low agreement rates between voicing annotations in laughs (between 88% and 91% for two annotators), confirming that the distinction between voiced and unvoiced laughter is not objective. One possible explanation for the low agreement is the fact that vocal folds are more abducted in laughter [Bickley & Hunnicutt 1992] and do not close as decisively in laughter as in speech [Chafe 2007], making voicing more breathy and less clear in laughter. This is confirmed by Esling [Esling 2007] who observed the laryngeal states during laughter and concluded that the conditions are met to obtain sounds that are at the same time high-pitched and breathy.

3.1.7.3 Social functions

Another distinction that is made between laughs is related to their functions. In ancient times, lead philosophers Plato and Aristotle considered laughter, while pleasant, as a malicious response to the ignorance of others, and hence educated people should refrain from manifesting such hateful signals [Greenfield 2002]. Laughter is nowadays known to be a signal of amusement, but also a conversational and polite signal to indicate that we are following the discussion. While the former can occur when people are alone, the latter only happens when there is a so-

[22] Without giving subsequent definition of when a call is considered as voiced.

cial context (starting from the mere presence of another person, to group discussions on the other end). In consequence, researchers distinguish between hilarious—as a response to an amusing stimulus—and social/conversational—motivated by social behavior—laughter [Foot & Chapman 1976, Glenn 2003, Vettin & Todt 2004, McKeown *et al.* 2013]. Vettin and Todt [Vettin & Todt 2004] realized that people frequently laugh after their own utterances and suggest that it serves to mitigate the meaning of the preceding utterance, especially if the conversational partner did not react in the expected way. Glenn [Glenn 2003] also noticed this phenomenon, but specifies that it does not happen frequently in groups of three or more, as it could be considered as self-praise, while it is socially acceptable when there is only one conversational partner[23]. Trouvain and Truong [Trouvain & Truong 2013] found that half of the overlapping laughs in dyadic conversations were initiated by the speaker. These laughs are thus inherently social, as are many laughs in natural conversations, that occur after regular statements rather than humorous comments [Provine 1993].

Glenn [Glenn 2003] makes several interesting suggestions about conversational laughs. Among others, his analyses of first laughs suggest that first laugh can be equivocal (similar to a cough or deep breathing), to enable the speaker either to retroactively display it as non-laughter or turn it into more obvious laugh depending on others' reactions (laughing or not). Also, he presents some examples where laughter can be considered as an intermediate answer between total acceptation of the previous comment and resistance towards these comments: the features of the laugh (for example, closed mouth while others are laughing out loud with the mouth open) can indeed indicate some resistance.

Relying on the distinction between hilarious and social laughter, two dimensions for characterizing laughs are currently investigated by the ILHAIRE Consortium [Hofmann *et al.* 2013]: intensity and regulation. The notion of intensity seems so straightforward that many authors (e.g., [Edmonson 1987, Glenn 2003, Martin & Lefcourt 2004, Trouvain & Schröder 2004]) use terms that refer to it without finding necessary to define what this notion implies. This notion is actually quite old. Plato already referred to "violent"[24] laughter (see quote in [Chafe 2007], p. 140), Darwin also refers to this notion [Darwin 1872], with the following statements:

> *A graduated series can be followed from violent to moderate laughter, to a broad smile, to a gentle smile, and to the expression of mere cheerful-*

[23] If you are also wondering in which situations laughing after her/his own statements can be socially accepted (not considered as self-praise), Glenn identified the following situations: after a self-deprecating comment (laughing is then encouraging others to laugh *at* the speaker himself, which is opposite to self-praise), when referring to some laughable comment that was actually made by someone else (retelling a story, reading, etc. then the credit of the laughable comment is attributed to the person who produced it rather than the speaker who laughed first about it), and to mitigate a previous comment, indicating that it was not uttered with a serious aim and in consequence does not require a serious answer.

[24] Although here the meaning of the term "violent" is debatable, as Plato possibly wanted to designate the negative, malicious effect of laughter with this word. Furthermore there may be nuances that were incorrectly translated.

> *ness. During excessive laughter the whole body is often thrown backward and shakes, or is almost convulsed; the respiration is much disturbed; the head and face become gorged with blood, with the veins distended; and the orbicular muscles are spasmodically contracted in order to protect the eyes. Tears are freely shed.*
>
> *(...) Excessive laughter, as before remarked, graduates into moderate laughter. In this latter case the muscles round the eyes are much less contracted, and there is little or no frowning. Between a gentle laugh and a broad smile there is hardly any difference, except that in smiling no reiterated sound is uttered, though a single rather strong expiration, or slight noise—a rudiment of a laugh—may often be heard at the commencement of a smile. (...) We thus see that no abrupt line of demarcation can be drawn between the movement of the features during the most violent laughter and a very faint smile.*

If we try to put words on this obvious notion used by virtually everybody, intensity is used to characterize the "arousal" of the laugher, which is related to a gradual increase of the laughter markers used [Ruch 1993]: starting from a simple smile—Action Unit (AU) 12 (see [Ekman *et al.* 2002] for more information about Action Units)—at low intensity, then involving audio as well (at first with closed mouth), then adding and extending facial activities (cheek raises, AU6, on top of AU12 and AU25, mouth opening) in combination with open-mouth sounds [Ruch & Ekman 2001], finally propagating to the whole body (shoulder shaking, trunk bending, throwing back the head) and lacrimation. Nevertheless, as people have different laughing styles, laughter intensity is encoded differently by different laughers [Edmonson 1987].

Laughter is not an uncontrollable reaction [Glenn 2003]. Regulation indicates whether the laugh is totally spontaneous (resulting from amusement only) or modified (consciously or not) by some social context: in up-regulated laughter, the laugher is trying to appear more amused than (s)he actually is, while it is the opposite in down-regulated laughter [Hofmann *et al.* 2013]. Both behaviors can be motivated by social etiquette. Up-regulated laughs include fake laughs, i.e. the laugher is actually not amused and is laughing to pretend (s)he is (consciously or not). The distinction between fake and spontaneous laugh is indeed well established [Ruch & Ekman 2001], and it has been shown that they are separate processes: some people suffering from brain disorders are not able to voluntary move their abdomen, while they would exhibit large abdominal movements during spontaneous laughter [Bright *et al.* 1986]. It is remarkable to note that the notion of regulation had already been pointed out by Darwin [Darwin 1872], who wrote the following:

> *Laughter is suppressed by the firm contraction of the orbicular muscles of the mouth, which prevents the great zygomatic and other muscles from drawing the lips backwards and upwards. The lower lip is also sometimes held by the teeth (...) Laughter is frequently employed in a forced manner to conceal or mask some other state of mind, even anger.*

The existence of hilarious and social laughs is somehow confirmed by Tanaka and Campbell [Tanaka & Campbell 2011], who proposed five categories to annotate laughs: mirthful, polite, embarrassment, derision and other. Eighteen people were invited to assign one of these labels to 876 laughs uttered by two male speakers. The agreement between raters was fair (Cohen's [Cohen 1960] $\kappa = 37\%$). The authors noticed that embarrassment was hard to acoustically distinguish from polite laughs and grouped these categories. As "derision" and "other" labels had barely been used, this resulted in two remaining categories: mirthful and polite. According to their analysis of the laughs from one of the participants, polite laughs rarely include chuckles and never contain ingressive calls. These findings obviously need to be taken with care until they are confirmed (or not) on a multi-speaker corpus.

Coming back to the notion of intensity, its relevance for characterizing laughter is further stressed by the following two experiments. Firstly, from annotation of laughs on the Mechanical Turk, a high correlation between ratings of amusement (how hilarious the laugh is) and intensity has been observed [Hofmann *et al.* 2013]. Secondly—in what has been called the "ambiguity" experiment [Hofmann *et al.* 2013]—it has been shown that a high intensity laugh in a conversation can be replaced by another high intensity laugh without modifying the naturalness of the sequence, as perceived by naive observers, while this does not hold when low intensity laughs are involved. Intensity is thus, as expected, a dimension that makes sense to laypersons, that is related to other perceptual characteristics (amusement) and that can be used to discriminate laughs having different properties (as high- and low-intensity laughs are not interchangeable).

To avoid confusion with acoustic intensity (amplitude, loudness, etc.), from now on we will use the term *arousal* to refer to the intensity of the emotional state leading to laughter. Arousal is frequently used to characterize affects, for example in the two-dimensional valence-arousal plane.

3.1.7.4 Lexical approach

To conclude this section, it is interesting to look at the lexical study reported in [Hofmann *et al.* 2013] to find and cluster terms that are used to characterize laughter. Large quantities of text (fiction, non-fiction, newspaper articles, transcriptions of spoken language, etc.) in the German language were screened to find utterances of the word "laugh" (or variations of it). Almost 270,000 instances were found, and for each of them all surrounding terms (within ± 5 words of "laugh") were analyzed. Terms that specified the way the person was laughing were included in the laughter description lexicon. A total of 1148 such attributes were found. The words were then grouped in "descriptive" categories. Six major categories of description were identified [Hofmann *et al.* 2013]:

1. Basic parameters: adjectives describing laughter with parameters that are common to other signals, for example its duration, time-course, steepness of the onset, etc.

2. How it sounds.
3. How it looks.
4. How intense the laugh is.
5. Its uniqueness: adjectives relating to how the laugh reflects the identity of the laugher (e.g., "distinctive", "inimitable").
6. Regulation: whether and how the laugh is regulated or modified.

Emotional qualities only formed a minor category, with less terms relating to it than the major six categories. Although the approach is different (analyzing text instead of actual laughter utterances), it is striking that conclusions are similar to previous works on laughter description, with doubts on the appropriateness of emotional categories at the profit of dimensions like regulation and arousal, on top of basic, objective parameters (duration, composing sounds, etc.).

3.1.8 Episode characteristics and relation with laughter types

In this section, we would like to briefly report about characteristics of laughter episodes and some research works that have demonstrated that different categories of laughter have different duration distributions. As could be expected, there is a large variability in laughter duration as laughter can be as short as one single syllable and as long as unstoppable giggles. For example, in the AVLC database, the shortest laugh covers only 260 ms while the longest giggle is lasting 82 s [Urbain *et al.* 2010a]. The vast majority of laughs are however shorter than five seconds (average: 3.5 s; std: 5.3 s; median: 2.2 s), in agreement with Ruch's statement that "single acts of laughter seldomly exceed seven seconds"[25][Ruch 1993].

Devillers and Vidrascu [Devillers & Vidrascu 2007] observed that laughs rated as positive tend to be longer, have higher energy and more voiced frames than negative laughs.

It is also interesting to note that acted laughs tend to be longer than spontaneous laughs: the participants of the AVLaughterCycle database produced acted laughs with an average duration of 7.7 s (median: 5.26 s; std: 5.92 s). The difference between the average duration of spontaneous and acted laughs was found to be highly significant, as explained in Section 2.5.6.3. While the durations of acted laughs are not totally uncommon in spontaneous laughs, the longer durations of acted laughs are betraying the stereotypical idea people have about laughter: when someone is asked to fake laughter, (s)he will most likely try to imitate some long, high-arousal, melodious laugh, rather than short and lower-arousal utterances that are actually more frequent in spontaneous behaviors.

[25] Ruch also considered the visual contributions of laughter, which start before and end after the audio outcomes considered in analyzing the AVLaughterCycle corpus, so it is not surprising that Ruch found slightly higher durations.

In addition, Vettin and Todt [Vettin & Todt 2004] found that conversational laughs are shorter than humorous laughs. For example, in the ICSI Meeting Corpus (which consists of conversation recordings), the average laughter duration is 1.62 s with a standard deviation (std) of 1.24 s [Knox & Mirghafori 2007]. Glenn [Glenn 2003] claims that subtle conversational laughs are frequently neglected in studies describing laughter, who mostly concentrate on intense episodes.

To conclude this section, let us consider contagion effects on laughter properties. Truong and Trouvain [Truong & Trouvain 2012b] showed that overlapping laughs (i.e., when several people are laughing at the same time) have on average a longer duration, a higher average energy, a higher maximum energy, a higher average f_0 and a higher number of voiced frames than non-shared laughs (when only one person is laughing). Interestingly, these effects are growing with the number of participants laughing (at least partially[26]) simultaneously: the more laughing people, the longer the laughs, the higher the average pitch, etc.

3.1.9 Summary

In this overview of existing works for describing laughter, we have first presented a terminology that can be generalized to any laugh: episodes, bouts and syllables. Then, we have reviewed the analyses of audio laughter properties that have been investigated so far, which mainly concern the duration of suprasegmental units (syllables, bouts, episodes), the fundamental frequency (both isolated values and evolution over the course of syllables or bouts) and the position of formants. The main conclusion is that laughter provokes a large variability of duration and fundamental frequency values and that no general pattern has been identified so far. Finally, we have seen that, perhaps surprisingly, there is no system of laughter categories that receives overall agreement. The existence of emotional categories of laughter has been longly supposed, but recent empirical evidence stems for the opposite. Additional research is needed to understand exactly what type of information isolated laughter can reliably convey, and which contextual information is possibly required to disambiguate between emotional, affective or functional laughter types. The current trend is to come back to simpler systems: voiced-unvoiced (although the boundary between these categories has not been clearly established), conversational-hilarious (again, some laughs cannot be unequivocally placed in these categories) or using dimensional systems like arousal and regulation (the latter dimension enabling to encode social/conversational effects on laughter).

In the remainder of this dissertation, we will focus on acoustic properties of laughter episodes, without considering their context. In other words, we will not try to

[26] Laughs are considered as overlapping as soon as there is one overlapping frame, they do not have to start and end at the same time, or to reach a certain percentage of simultaneous frames. Actually, Truong and Trouvain have found that people join in laughing on average 500 ms after the beginning of the first laugh. This is an interesting finding for building laughing interactive systems, as it means that a few hundred milliseconds are available for computations (laughter detection and decision about joining in laughter).

explain why particular acoustic, phonetic or arousal patterns appear in different situations, nor which effects they can have. We will rather attempt to infer these patterns from laughter episodes (see following sections and Chapter 4) and synthesize laughs that correspond to these patterns (see Chapter 5). As already explained, the interpretation of laughs within their context and the decision to synthesize a laugh with given properties (and expected effects) are, in human-computer interfaces, devoted to Natural Language Processing and Dialog Management [Niewiadomski *et al.* 2013a], which goes beyond the scope of the present work.

3.2 Phonetic transcriptions

We strongly believe that both automatic laughter recognition/characterization and synthesis can benefit from a detailed phonetic transcription of laughter. On the recognition side, transcriptions can help classifying/clustering 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, transcriptions enable approaches similar to those used in speech synthesis: training a system with the individual phonetic units and then synthesizing any consistent phonetic sequence. This is the approach used in this work, as will be presented in Chapter 5.

As we have seen in Section 3.1.5, several researchers (e.g., [Jefferson 1985, Glenn 2003, Pompino-Marschall *et al.* 2007, Tanaka & Campbell 2011]) have approached what we are looking for, but detailed phonetic transcriptions—timely aligned to available acoustic signals—of a large corpus of spontaneous laughs were lacking, as well as a standard system indicating how to do such annotations. In consequence, we decided to propose a way to phonetically transcribe laughs and publicly released the obtained annotations in parallel to the audio signals. The methods are described below (Section 3.2.1), as well as outcomes resulting from analyzing the phonetic transcriptions (Sections 3.2.2 and 3.2.3). These research efforts have been presented in [Urbain & Dutoit 2011].

3.2.1 Transcriptions

The phonetic transcriptions were made on laughs from the AVLaughterCycle database [Urbain *et al.* 2010a] presented in Section 2.5. Laughs had been segmented on the basis of the audiovisual signal. In total, 1021 laughs have been segmented, for a total of one hour of spontaneous, hilarious laughs. Note that only audio was used for the phonetic transcriptions, while laughter segmentation had been done on the basis of both audio and video.

One annotator labeled the 1021 laughs in phones in the Praat software [Boersma & Weenink 2011]. Two annotation tracks have been used (see Figure 3.6). The first is used to transcribe the phones, according to the phonetic symbols defined in the IPA [International Phonetic Association 1999]. Diacritics have also been used

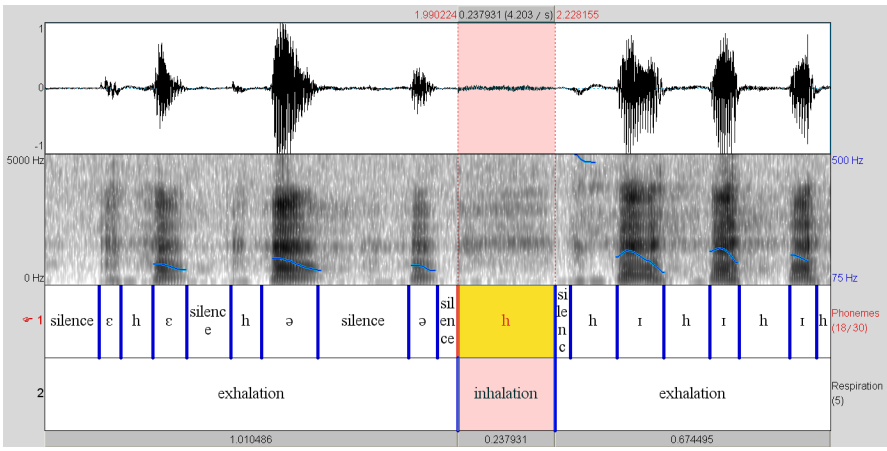

Figure 3.6: Laughter annotation in Praat.

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 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 some previous works ([Bachorowski *et al.* 2001, Chafe 2007]) but unlike others (e.g., [Edmonson 1987]), the following labels have been added: hum, cackle, groan, snore, vocal fry, grunt and nareal fricatives[27]. Nareal fricatives are powerful (hence audible) streams of air traveling through the nostrils and actually have a phonetic symbol in the "extended IPA symbols for disordered speech" [International Clinical Phonetics and Linguistics Association (ICPLA) 2008]: "n̊".

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. In the end, the proposed system is relatively similar to the one used by Jefferson [Jefferson 1985], with an increased range of transcribed sounds/phones, detailed timing information (the phone boundaries are accurately positioned) but no information on the "prosody" of laughter (pitch, energy). This information on prosody is here considered separate from the actual phonetic transcription. Pitch and energy trajectories could however be added as additional tracks. Standard sound analysis programs like Praat [Boersma & Weenink 2011] and Wavesurfer [Sjölander & Beskow 2011] for instance can provide tracks with editable automatic estimations of such features.

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

[27] Examples of the introduced phones are available on `http://www.tcts.fpms.ac.be/~urbain/`.

of its phones separately. As a matter of fact, although laughter episodes exhibit no strong semantic contrast (as opposed to words), they still obey strong phonotactic constraints (e.g. we will have the impression of hearing *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 (it seemed to us that) they had been produced, rather than how they actually sounded, following a long tradition of articulatory phonetic transcription.

3.2.2 Laughter phonetic description

Out of the initial 1021 laughs, 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), 17,202 phones have been annotated: 15,825 in exhalation phases and 1,377 in inhalation phases. If we take diacritics into account[28], 196 phonetic labels appear in the database: 142 during exhalations and 54 during inhalations[29]. This reinforces the idea that laughter is extremely variable.

For the sake of simplicity, the diacritics will not be considered here. 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 3.1 and 3.2, with their average duration.

The outcomes of our annotation are mostly in line with previous findings ([Bachorowski *et al.* 2001, Ruch & Ekman 2001, Szameitat *et al.* 2009b]). During exhalation phases, if we exclude silences that are extremely 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 [Edmonson 1987] and [Chafe 2007], even though it has been contested in [Ruch & Ekman 2001], we found that voiced segments can be abruptly ended by a glottal stop (ʔ).

We also found a lot of non-stereotypical laughter sounds. Nareal fricatives (n̊̃) 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, our database presents occurrences of non central vowels (ɪ, ɛ, ʌ), which were not found in Bachorowski et al.'s formant frequency analyses [Bachorowski *et al.* 2001]. Our data also contain 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. The

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

[29] The same labels can appear in exhalation and inhalation parts. If so, they are here counted twice (one time for inhalation, one time for exhalation) in the total of 196 labels.

Table 3.1: Most frequent phonetic labels in laughter exhalation phases.

Label	Occurrences	Average duration (std)
silence	4886	308 ms (427 ms)
h	2723	121 ms (68 ms)
ə	1422	73 ms (44 ms)
ɐ	1373	82 ms (47 ms)
n̊	839	210 ms (134 ms)
ɪ	741	77 ms (39 ms)
cackle	704	34 ms (24 ms)
hum	639	77 ms (42 ms)
ɛ	370	76 ms (35 ms)
ʔ	269	27 ms (16 ms)
|	214	31 ms (32 ms)
x	176	228 ms (170 ms)
ʌ	160	94 ms (66 ms)
ħ	152	175 ms (85 ms)
ɦ	135	175 ms (114 ms)
ɵ	109	116 ms (58 ms)
k	102	48 ms (51 ms)
t	81	73 ms (35 ms)
grunt	81	126 ms (104 ms)
ʉ	79	93 ms (90 ms)

Table 3.2: Most frequent phonetic labels in laughter inhalation phases.

Label	Occurrences	Average duration (std)
h	640	305 ms (133 ms)
ə	172	95 ms (59 ms)
n̊	166	346 ms (170 ms)
ɪ	108	97 ms (64 ms)
ɦ	38	226 ms (121 ms)
s	38	340 ms (141 ms)
ħ	24	340 ms (154 ms)
t	23	49 ms (32 ms)
i	23	148 ms (58 ms)
ɛ	17	94 ms (39 ms)

Table 3.3: 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

presence of these phones in laughter has been later confirmed in [Wagner *et al.* 2013] (see Section 4.1.2.3).

The duration of exhalation phones varies largely with the nature of these phones: as expected cackles, clicks and plosives (t, k) are really short, while nareal fricatives and fricatives (h, x, ɦ, ħ) are much longer. If we combine the average durations of phones that usually form voiced laughter syllables, we get back to the standard syllable duration of around 210 ms, as "h" has an average duration of 121 ms while "vowels" have average durations in the range [73-116 ms].

During inhalation phases, the most used phones are similar. Deep breath sounds (h, n̥, 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 3.1 and 3.2 (h, ə, n̥, ɪ, ɦ and ħ) except for t (no difference) and ɛ ($p = 0.22$). Over the whole database, the average phone duration during exhalation and inhalation phases is respectively 165 ms (std: 266 ms) and 245 ms (std: 159 ms). 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 1.69 s (std: 1.52 s) and 0.36 s (std: 0.15 s). No correlation has been found between the duration of an exhalation phase and the duration of its surrounding inhalations (correlations < 0.1). Table 3.3 shows the number of laughs presenting a given number of exhalation and inhalation phases.

Most of the laughs have only one bout (i.e., exhalation part separated by inhalations). 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.

3.2.3 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 numbers of subjects and phones are large, we cannot give an exhaustive analysis here and will concentrate on a few examples.

3.2.3.1 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" [Chafe 2007]. Figure 3.7 displays, for the five subjects who laughed the most and the most used seven 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 (ñ̥ and hum) than others. The choice of the vowel is another difference between subjects: some laughers use up to three times more ə than ɐ, others do the opposite.

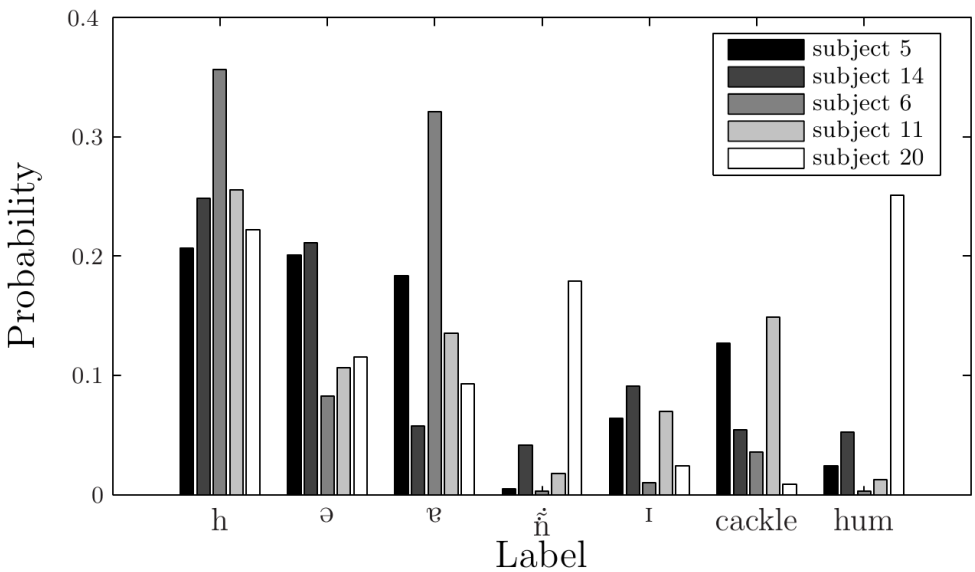

Figure 3.7: Probabilities of the most used phones for the five subjects who laughed the most.

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

3.2.3.2 Phone and airflow phases duration

The average duration of exhalation phones is similar for all subjects: slightly under 100 ms for voiced phones, a bit larger for h-like sounds and nareal fricatives. There is a slightly larger individual variation for inhalation phones. Figure 3.8 exhibits the average duration of the most frequent three inhalation phones for all the subjects, with their corresponding standard deviations (stds). 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.

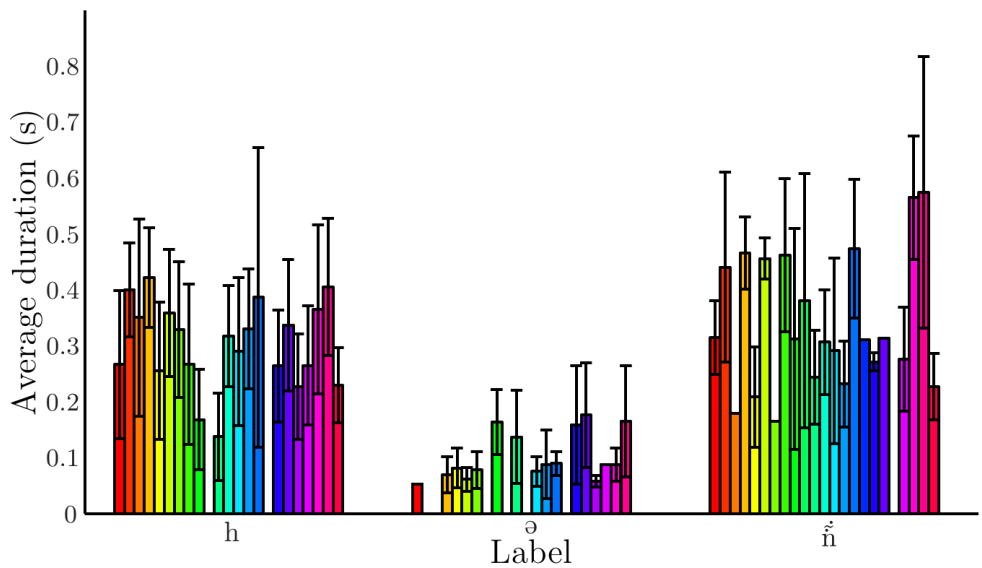

Figure 3.8: Average duration of the most frequent inhalation phones, for all the subjects.

Figure 3.9 shows the average durations (and standard deviations) of exhalation parts 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 [Bachorowski *et al.* 2001].

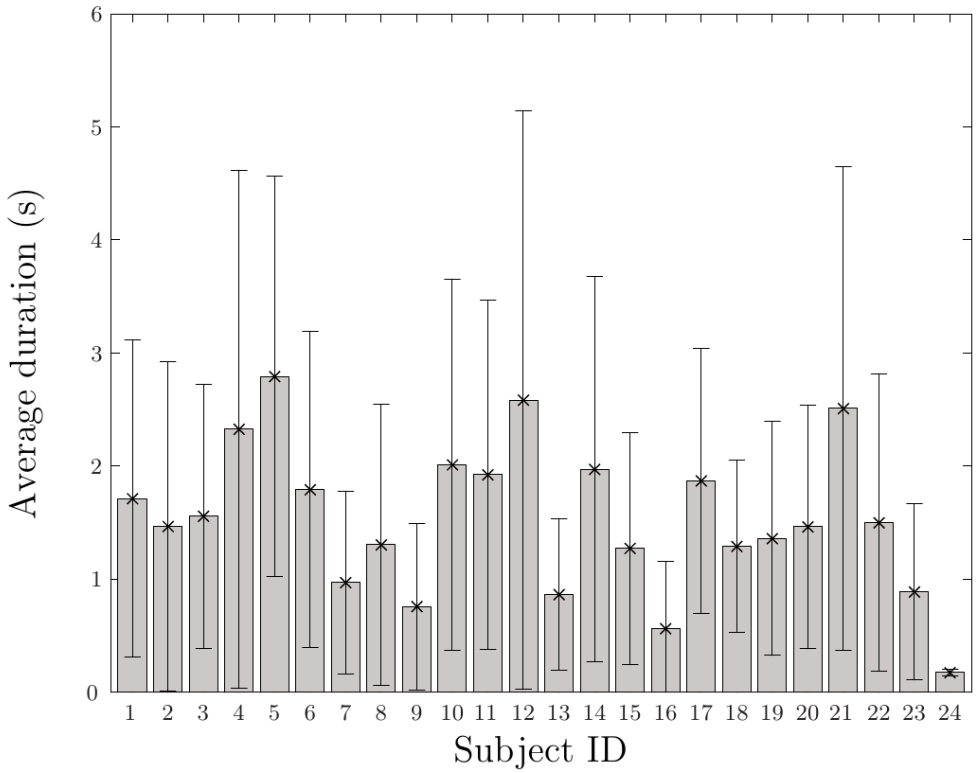

Figure 3.9: Average duration of exhalation phases, for all the subjects.

The laughter "vowels" have been observed to be used mostly in sequences of the same vowel by the AVLC participants. This is illustrated in Table 3.4 which presents the number of occurrences of three-phone sequences beginning with a vowel, followed by a fricative or silence and ending with a vowel. It appears clearly that, although transitions from one vowel to another are allowed, it is much more likely that sequences of constant vowels appear in laughter bouts.

3.3 Overall arousal

As we have seen the notion of laughter arousal reaches consensus among scholars. However, to the best of our knowledge, no laughter database had been annotated in arousal. In this section we will describe the works that we have done to obtain overall arousal values for all the laughs in the AVLaughterCycle database. These efforts were conducted in collaboration with TELECOM-ParisTech. The aims were on one side to obtain arousal annotations for the whole corpus and on the other

Table 3.4: Number of occurrences in the AVLC database of three-phone sequences beginning with a vowel, followed by silence or a fricative and ending with a vowel.

First phone		Third phone					
		a	cackle	ə	ɪ	o	ɛ
	a	802	69	131	43	22	21
	cackle	23	331	59	24	1	16
	ə	94	89	468	75	23	42
	ɪ	29	42	66	339	9	29
	o	17	9	31	6	102	2
	ɛ	22	33	54	41	4	110

side to identify which audio and visual cues contribute to the perception of laughter arousal. We will however only report about the audio conclusions in this section. The whole study, including visual contributions of laughter arousal, was presented in [Niewiadomski *et al.* 2012]. The audio results presented here are nevertheless integrating the latest arousal ratings (which were not included in the paper).

This section is organized as follows. The method for obtaining arousal annotations is described in Section 3.3.1. Acoustic features correlated with arousal are investigated in Section 3.3.2. finally, Section 3.3.2 is related measures that have been taken aside of these arousal annotations to examine acoustic differences between inhalation and exhalations laughter phases.

3.3.1 Arousal annotation

The annotation in arousal was realized through a web application. Participants were invited to the website where they could watch (audio+video) one laugh at a time and rate the arousal of the laugh on a five-point scale (1-low arousal; 5-high arousal). Laughs were randomly presented. Participants could watch one laugh as many times as they wanted before giving it an arousal score. But once they had rated a laugh, they could not come back and change its rating. There was no limit to the test: participants were explained that they could stop annotating whenever they wanted to, but a new laughter sample would always be presented to them after they had rated one laugh.

The experiment was stopped once we had obtained at least six annotations for each laugh of the AVLaughterCycle database. In total, we received 7,272 ratings from 90 participants, including 45 males and 42 females (three participants did not indicate their gender), with an average age of 33.6 (std: 12.6; three participants did not indicate their age). The overall agreement between the raters was fair: Krippendorff's alpha [Krippendorff 2007] was .67. The distribution of ratings is displayed on Figure 3.10. The histogram of the median arousal score of each laugh is shown on Figure 3.11. It can be seen that the distribution of arousal scores is not uniform. Most of the episodes have been rated as produced with low arousal. The maximal arousal

score (5) has only been used 249 times (3% of the ratings) and only 16 laughs have a maximal median arousal value.

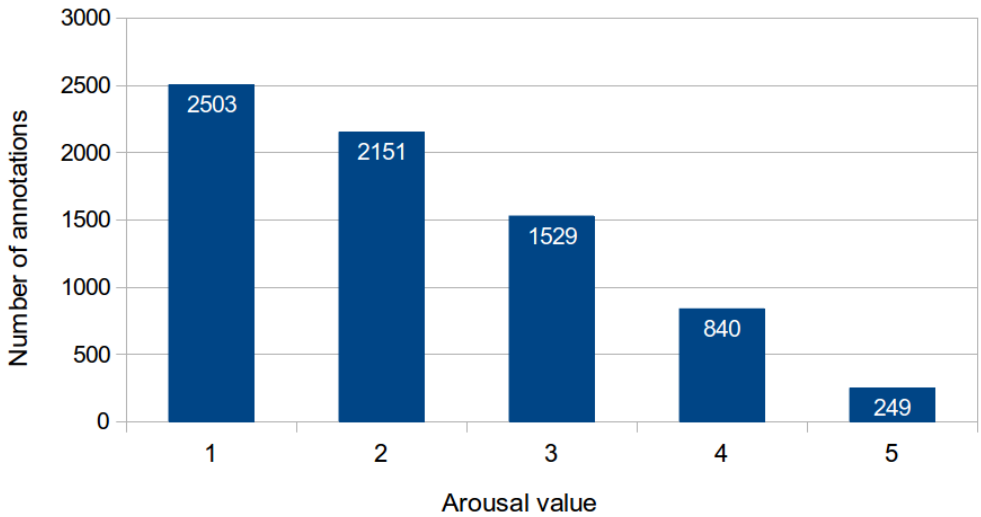

Figure 3.10: Laughter arousal annotations histogram.

It is also interesting to note that after rating laughs, several participants told us that they tended to judge arousal relatively to the laugher's style. As there are 24 laughers in the AVLaughterCycle database, raters who rated a lot of laughs tended to view the same laughers over and over again. They informed us that they got acquainted to individual laughing styles and rated laughter arousal accordingly. This is an important remark for interactive systems: as pointed out by Edmonson [Edmonson 1987], arousal is encoded differently by different laughers and is perceived relatively to the laugher's style. Interactive systems would thus benefit from modeling individual laughing styles and computing arousal in a relative rather than an absolute manner.

3.3.2 Features influencing the perception of laughter arousal

To investigate the audio features that influence the perception of arousal, audio features were extracted from each laugh. The features can be divided into three main categories: spectral low-level descriptors, measures of the noise level and prosody-related low-level descriptors. Spectral low-level descriptors are:

- Twelve Mel-Frequency Cepstral Coefficients (MFCCs), their first (denoted Δ) and second derivatives ($\Delta\Delta$).
- Spectral centroid, spectral spread, spectral variation, spectral flux, spectral decrease [Peeters 2004].

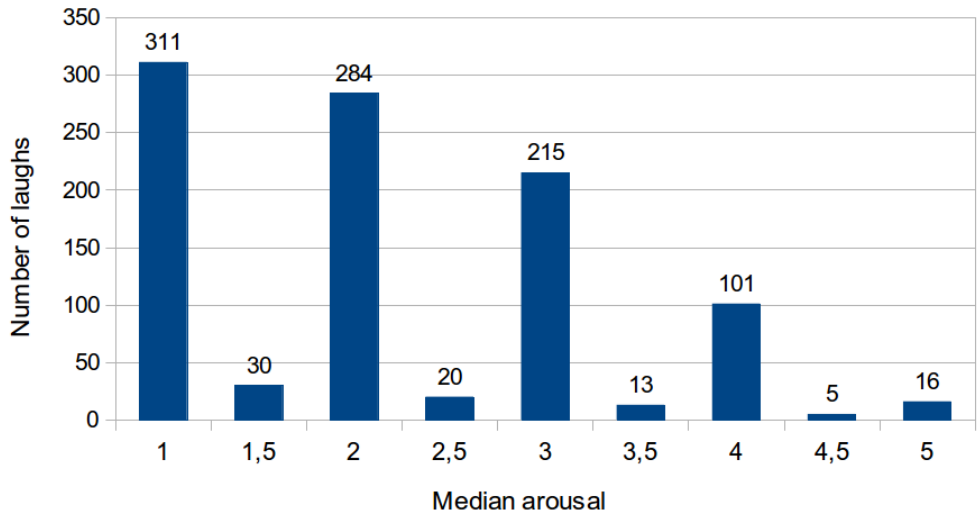

Figure 3.11: Number of laughter episodes for each degree of arousal (median).

- Twelve chroma features [Ellis & Poliner 2007].

The amount of noise is characterized with the following features:

- Chirp group delay [Drugman *et al.* 2011].
- Four values of Harmonic to Noise Ratios (HNRs) corresponding to the frequency bands 0-500 Hz, 0-1500 Hz, 0-2500 Hz and 0-3500 Hz [Drugman *et al.* 2013].
- Zero-Crossing Rate (ZCR).
- Four values of spectral flatness corresponding to the frequency bands 250-500 Hz, 500-1000 Hz, 1000-2000 Hz and 2000-4000 Hz [Peeters 2004].

The prosody-related low-level descriptors include:

- Measures of energy: loudness [Peeters 2004], Root Mean Square (RMS) energy and MFCC0 (with its first and second order derivatives).
- f_0 computed with the Summation of Residual Harmonics (SRH) method [Drugman & Alwan 2011], as well as the value of the maximum SRH peak.
- The four values provided by the Snack ESPS f_0 estimation algorithm—implementing the RAPT method [Talkin 1995]—namely the estimated pitch, probability of voicing, local RMS measurement, and the peak normalized cross-correlation.

- The frequency and bandwidth of the first four formants, computed with Snack [Sjölander 2004].

All these 82 low-level acoustic descriptors were extracted from the 16 kHz audio signals, using frames of 512 samples (32 ms) shifted by 160 samples (10 ms). To get a fixed number of features for each laugh, the frame by frame low-level descriptors (in variable number, depending on the duration of the laugh) are mapped to a fixed-length feature vector with the help of the following nine functionals[30]: minimum over the laugh (abbreviated *min*), maximum (abbreviated *max*), range, mean, standard deviation, skewness (abbreviated *skew.*), kurtosis (abbreviated *kurt.*), percentage of time spent in the upper quartile (%25) and ZCR. Since we had 82 low-level acoustic descriptors, we obtained a feature vector of 738 audio features per laugh. The duration of the laugh was added to the feature vector in order to investigate whether arousal was related to the duration of the laugh. Correlations were computed between all these acoustic features and the median arousal values of the laughs.

Strong correlations between several features and the median arousal annotated for each laugh have been found. Energy features provide the strongest correlations: MFCC0 and its derivatives provide three of the best five correlation coefficients (ρ) with the laughter arousal, while loudness is slightly behind. Figure 3.12 shows the best correlation with the annotated arousal, obtained with Δ MFCC0 range. MFCCs and spectral flatness also provide high correlations. The detailed data for the ten best audio descriptors and pitch are presented in Table 3.5. We can see that the "range" functional is yielding the best correlations for all these low-level descriptors. Energy descriptors ($MFCC0$, $\Delta MFCC0$, $\Delta\Delta MFCC0$ and loudness) are the most correlated with laughter arousal (best correlation coefficients over .8), followed by descriptors of the spectral shape (spectral flatness and MFCCs). Fundamental frequency, extracted through the ESPS method available in Wavesurfer [Sjölander & Beskow 2011], is slightly below with a correlation coefficient of .67.

Interestingly, the overall duration of the laugh is not strongly correlated ($\rho = .48$) with the perceived arousal (Figure 3.13). In other words, an intense laugh does not necessarily last long, and vice-versa.

These results show that some audio features are strongly related to the perceived arousal of laughs. Hence these features are good candidates to predict laughter arousal, as will be investigated in Chapter 4. Yet, (audiovisual) laughter arousal is not only a matter of duration or acoustic energy. Although acoustic energy is strongly correlated with the arousal ratings, there is no one-to-one relationship between these quantities. Other features—or at least temporal patterns that are not captured by our functionals—play a role, for example spectral shapes, pitch, or visual features (jaw opening, etc.) [Niewiadomski *et al.* 2012].

[30] Functionals are descriptors of a sequence of features. For example, the average, minimum and maximum values of a feature during an utterance or a given window are functionals of that feature.

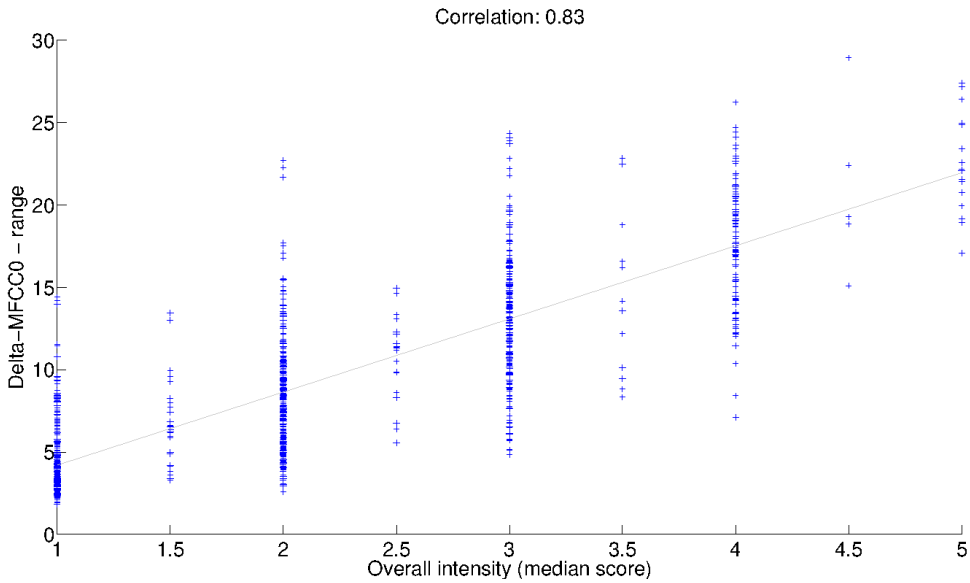

Figure 3.12: Correlation between median arousal and ΔMFCC0 range.

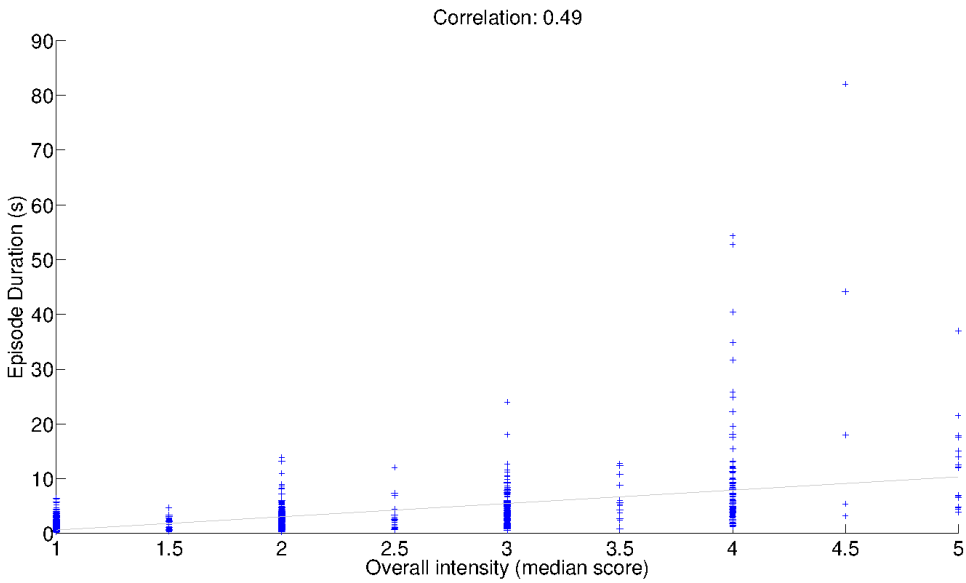

Figure 3.13: Correlation between median arousal and laughter duration.

Table 3.5: Correlation between laughter median arousal and the ten best acoustic descriptors (+ fundamental frequency (f_0)).

	Δ MFCC0	MFCC0	MFCC2	MFCC5	ΔΔ MFCC0	Δ MFCC2	Spectral flatness [500-1000 Hz]	Loudness	Spectral flatness [1000-2000 Hz]	MFCC10	ESPS f_0
min	-0,77	0,25	-0,74	-0,73	-0,74	-0,73	-0,78	0,23	-0,76	-0.67	0.05
max	0,8	0,79	0,3	0,33	0,77	0,73	0,27	0,77	0,26	0,59	0,56
range	**0,83**	**0,81**	**0,81**	**0,80**	**0,79**	**0,79**	**0,78**	**0,77**	**0,77**	**0,77**	**0,67**
mean	-0,1	-0,55	0,27	-0,37	-0,05	0,03	-0,6	-0,57	-0,56	-0,18	0,38
std	0,7	0,69	0,76	0,62	0,68	0,68	0,66	0,67	0,63	0,61	0,54
skew.	0,28	-0,03	-0.39	-0,47	-0,03	-0,29	-0,55	0,26	-0,48	-0.,06	0,12
kurt.	0,41	-0,04	0,11	0,33	0,47	0,5	0,4	0,22	0,35	0,31	0,22
ZCR	-0,29	-0,2	-0,58	-0,43	-0,4	-0,36	-0,6	-0,13	-0,59	-0,47	-0,13
%25	-0,4	-0,23	0,31	0,16	-0,45	-0,3	0,62	-0,4	0,54	-0,22	0,12

3.3.3 Acoustic features and respiration phases

As a parallel study, we investigated whether the acoustic features exhibit different values in inhalation and exhalation phases. Respiration has an important role in the multimodal laughter expression. We expect that information about respiration is crucial to achieve believable audiovisual laughter synthesis: indeed, humans can naturally distinguish these respiration phases when listening or watching to a laugh. The audiovisual signals of the two respiration phases must thus present different patterns.

To investigate this, we extracted the same low-level acoustic features as in the previous section, on a frame-by-frame basis. However functionals were this time computed for all the windows that belong to a same respiration part (exhalation or inhalation), instead of over the whole laugh. Then, for each feature, we compared its distributions in exhalation and inhalation parts.

A Lilliefors test showed that most of the features do not follow a Gaussian distribution; hence a Kolmogorov-Smirnov test was preferred to a t-test to compare the feature distributions over the two classes. The Kolmogorov-Smirnov test resulted in highly significant differences in the distributions of the two classes, for almost all the features[31]. Figures 3.14 and 3.15 present the distributions, for the two classes, of four different features for the two classes. These experiments illustrate that audiovisual features present different patterns in exhalation and inhalation laughter phases, which confirms our expectations since it is easy for humans to distinguish these phases.

It is thus reasonable to assume that these features can help to automatically discriminate between exhalation and inhalation parts in laughter, as will be explored in

[31] The only features that did not reach statistical significance at a 99% confidence value are the percentage spent in the upper quartile of HNR [0-500 Hz] and the minimum values of MFCC10, ESPS voicing probability, ESPS local RMS for f_0 estimation, f_0 estimated by SRH and the maximum SRH peak.

Chapter 4 in the framework of automatically obtaining laughter phonetic transcriptions.

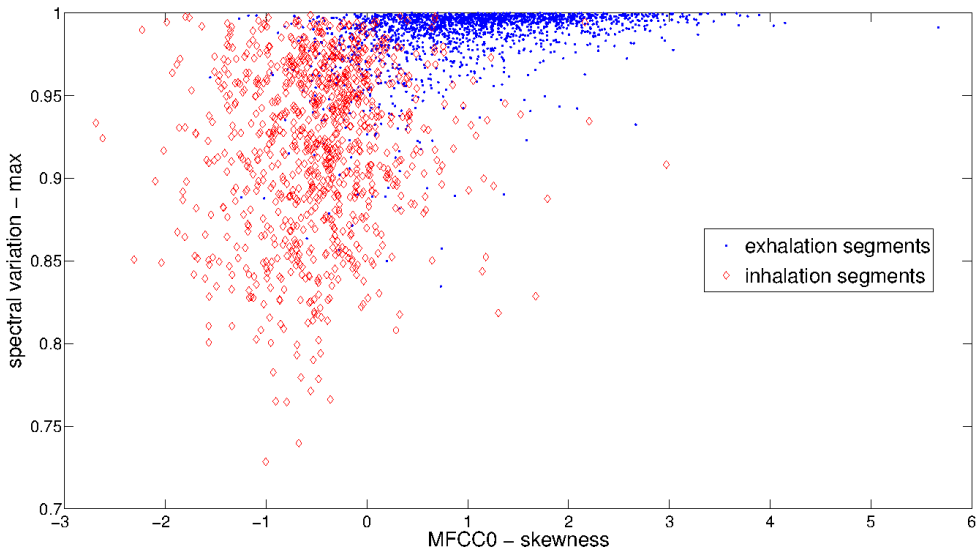

Figure 3.14: Distribution of MFCC0 skewness and maximum spectral variation for exhalation and inhalation laughter parts.

3.4 Arousal curves

As already stated earlier in this chapter, laughter arousal appears as an important dimension to characterize laughter and, consequently, to drive laughter synthesis. It is even interesting to obtain instantaneous arousal signals of laughter to describe their evolution. Instantaneous "arousal" indeed seems to us both convenient to use (it is easy to draw or describe an arousal signal) and highly-correlated with the choice of phones used (for instance low arousal laughs are related to closed-mouth nasal sounds, while higher arousal examples include open vowels [Ruch & Ekman 2001, Ruch *et al.* 2013, Niewiadomski *et al.* 2012]). We investigated how to build laughter arousal signals, with the objective to automatically compute these signals from laughter acoustic features.

To obtain some reference arousal signals against which we could train and evaluate algorithms, the per-frame arousal signal of 49 laughs from 3 subjects of the AVLaughterCycle database were manually annotated by one labeler. The 49 laughs were selected in order to cover the range of overall arousal values (see Section 3.3.1) with, when possible, good agreement between the raters. The 49 annotated laughs

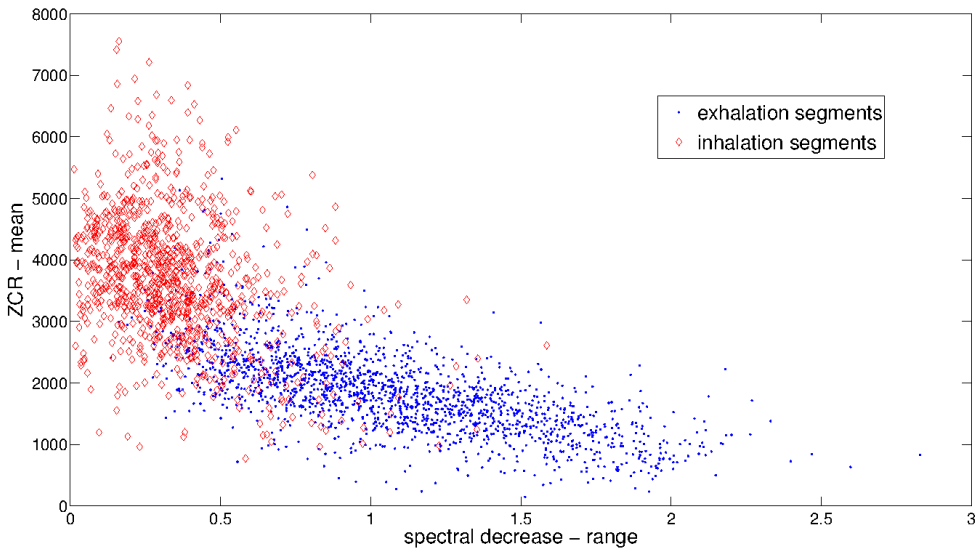

Figure 3.15: Distribution of the range of spectral decrease and average Zero-Crossing Rate for exhalation and inhalation laughter parts.

sum up to 27,693 frames (using windows of 32 ms shifted by 10 ms) labeled in laughter arousal. These frames were used to train automatic arousal estimation, as will be explained in Chapter 4. The arousal signals, along with the audio recordings and the phonetic transcriptions, of all the laughs of the AVLaughterCycle database are freely available to the scientific community. Examples of arousal curves are displayed in Figure 3.16.

3.5 Summary and perspectives

In this chapter we focused on the description of laughter properties at different levels. A first issue is the classification of laughter. We have seen that the existence of emotional laughter types, while highly suspected for a long time, is now contested. This justifies the use of other dimensions to characterize laughs, like their arousal (which is subjective but interpreted similarly by laypersons), their source (hilarious or social) or "objective" acoustic properties (voiced or not, composing sounds, etc.). Parameter analysis mostly concentrated on fundamental frequency and duration of the laughter units. Yet no standard pattern emerged and the main conclusions are that laughter is highly variable and that fundamental frequency can take higher values than in speech.

The introduced annotations (phonetic transcriptions of all the laughs, over-

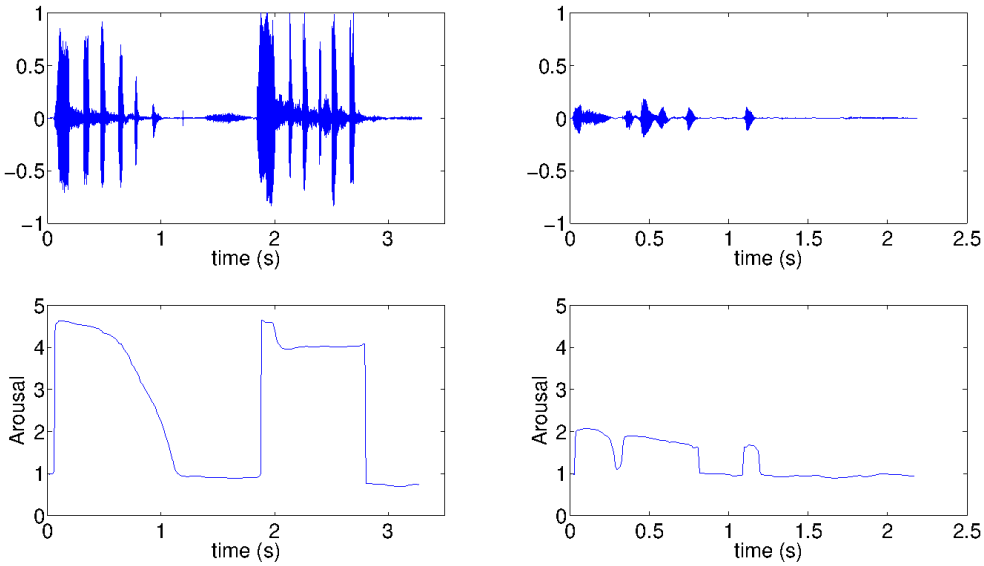

Figure 3.16: Examples of arousal signals. Top: waveform; bottom: corresponding arousal signal.

all arousal for all the laughs, instantaneous arousal for a subset of laughs) make the AVLaughterCycle database absolutely unique for laughter processing, no other database being provided with any of these additional information. In consequence, the AVLaughterCycle database has been the subject of all the developments reported in this Thesis. It is also worth noting that we did not annotate the AVLaughterCycle database on the dimension of regulation (the second dimension in the arousal-regulation characterization mentioned at the beginning of this Chapter) due to the nature of this database: subjects being alone, the social context that could provoke laughter regulation is highly limited and the AVLaughterCycle database is considered to contain only amusement, non-regulated laughs[32].

Perspectives of future works are numerous. First, the problem of describing laughter is far from being solved. New experiments must be conducted to further prove (or not) the existence of some laughter categories or dimensional scales. Cross-cultural experiments are also needed to investigate possible differences in laughter audio patterns across cultures. In addition, the fact that many studies investigating patterns of basic laughter features (f_0, duration, etc.) resulted in inconclusive or contradictory findings might be due to the incorporation of different laughers in the experimental

[32]We cannot however totally reject the possibility that some laughs were regulated, as participants knew they were being recorded, but we hypothesize that it is not the case, or at least that if regulation happened, it was to a small extent. Anyway, all the subsequent analyses would still hold even if regulation had an impact on the data.

sets. Individual styles might exist, possibly combined with contextual factors (social context, underlying emotion, etc.), and reveal some trends. For example, it is possible that certain persons routinely have declining f_0 patterns during laughter, but not everybody. Such individual styles have been revealed in the phonetic analysis conducted within this chapter. Laughter is proving to be an extremely complex signal to analyze, and it might be necessary to take into account interdependencies between several factors to identify "predictable" patterns.

Second, the phonetic analysis presented in this chapter could be extended to study the influence of laughter phones or respiration phases over acoustic features. Here only duration has been investigated. In the same vein, the experiment about laughter arousal could be extended. One interesting development would be to gather new arousal ratings when only audio or video is displayed to the raters and investigate how arousal perception differs when only audio or only visual information is available. Another research path is to experiment new features or functionals that possibly correlate with arousal. We have used here standard audio features covering three types of dimensions (spectral shape, measures of noise, prosody), but numerous other features could be experimented, for example features characterizing the amount of breathiness or the glottal source (e.g., parameters of the Liljencrants-Fant model [Fant *et al.* 1985]), features that are derived from phonetic transcriptions (e.g., the proportion of exhalation and inhalation phases) or new functionals to characterize the evolution of frame-level features over the laugh. For instance, Sathya et al. [Sathya *et al.* 2013] affirmed that the slope of decrease of the fundamental frequency within a call is proportional to the arousal of the laugh and it would be interesting to further investigate this hypothesis.

Third, we have presented here two novel tracks of annotation for laughter databases. It would obviously be beneficial if more databases were annotated the same way, and even several files annotated by different raters in order to estimate to what extent these annotations are "objective" or depend on the rater. All these annotated data are precious for automatic processing. We will see in the next chapters that it is possible to automatize some parts of the description and that the description stages that have been introduced in this chapter (phonetic transcriptions and arousal curves) are useful for laughter synthesis. As we will only use the AVLaughterCycle database for the following developments, our works concentrate on a particular subset of laughs: hilarious (non social), with variable arousal. No clear distinction will be made between voiced and unvoiced laughs, as we are not focusing on the effects or functions of laughter in interactions. Nevertheless this classification is rather implicit given the fact that we will rely on phonetic transcriptions, and that phones can be related to voicing[33].

[33] The relationship between phonemes and voicing is not unique and there are deviations between the acoustic realization and the suspected phonetic class. Nevertheless for most of the laughs it is rather obvious to understand whether they contain voiced parts or not, by simply looking at their phonetic transcriptions. It just requires to spot the presence of modal vowels or nasal consonants. It is easier than finding Waldo.

Chapter 4

Automatic estimation of laughter characteristics

Contents

In this chapter we will present the works that relate to automatic audio processing involving laughter. Most of the automatic processing works related to laughter concern the classification of audio segments (speech, laughter, etc.) or the detection of laughter in continuous audio streams. The state-of-the-art will be reviewed in Section 4.1 and we will see that relatively efficient methods exist for both problems. In consequence, for our own developments, we have made the assumption that we receive segmented laughter as input. Our objectives were to automatically characterize pre-segmented laughs. In Section 4.2, we will present a method for grouping similar laughs, which has been evaluated in its capacity to cluster laughs from the same laugher. The developed methods for automatically obtaining phonetic transcriptions and arousal signals will be described in Sections 4.3 and 4.4, respectively. A very brief introduction to HMMs, which will be repeatedly used in the remaining of this dissertation, is available in Appendix B.

4.1 State-of-the-art

The initial motivation for laughter detection was actually to improve automatic speech recognition systems [Kennedy & Hauptmann 1999]. Laughter was perceived as a noisy signal which could hinder speech recognition, so it was better to spot the laughter segments and discard them. It is only later that interest in laughter itself, as a signal that brings useful information in affective systems, appeared [Schuller *et al.* 2008]. In addition, a few works have targeted the automatic classification of laughs in different types or the identification of the laugher.

All these works will be presented in this section. But before really addressing laughter and speech discrimination, we would like to explain the different measures that are used to evaluate recognition algorithms. Unfortunately all the authors do not use the same measures, and it is sometimes difficult to compare the performance. Section 4.1.1 aims to clarify the results that will be presented in the remaining of the section. Then, in Section 4.1.2 we will begin with audio-only discrimination between speech and laughter (seeing both classification and detection works). This will be followed by an overview of audiovisual discrimination between speech and laughter (Section 4.1.3). These problems are not addressed in this Thesis. However we feel that an extensive state-of-the-art of the methods to automatically distinguish speech and laughter using audio features is important not only to position our works, but also to contribute to promote laughter processing through this dissertation. Finally, we will present the few works on automatic characterization of laughter in Section 4.1.4.

4.1.1 Measures of performance

Let us consider a binary classification problem where the two classes are labeled "+1" and "−1" (usually the positive class is the most interesting class to the researcher; in the case of laughter and speech discrimination, we will consider laughter as positive). To evaluate the classification performance, the classifier predictions are compared to the ground truth. The classification performance for all the instances can be summarized through the confusion matrix, as illustrated in Table 4.1.

Table 4.1: Example confusion matrix

		Ground truth	
		1	-1
Predicted value	1	True Positives (TP)	False Positives (FP)
	-1	False Negatives (FN)	True Negatives (TN)

A first set of evaluation measures can be inferred from the confusion matrix:

- the *Accuracy* is the proportion of correctly classified instances:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{4.1}$$

- the *Recall* or *True Positive Rate (TPR)* is the proportion of actually positive instances that are retrieved by the classifier:

$$Recall = TPR = \frac{TP}{TP + FN} \tag{4.2}$$

- the *Precision* is the proportion of actually positive instances among the objects classified as positive by the classifier:

$$Precision = \frac{TP}{TP + FP} \tag{4.3}$$

- the *False Alarm Rate* or *False Positive Rate (FPR)* is the proportion of actually negative instances erroneously classified as positive by the classifier:

$$FPR = \frac{FP}{TN + FP} \tag{4.4}$$

- the F_1^{score} is a combination of Precision and Recall:

$$F_1^{score} = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4.5}$$

  It reaches its maximum value (i.e., 1) when the classifier is perfect, and is affected by the two types of errors (considering actually positive objects as negative or vice-versa).

Usually, the decision of the classifier depends on a controllable variable. For example, if the classifier returns the probability $P_+(o)$ of each object o to be positive, the decision can be easily modified using a variable threshold θ, and considering the following decisions:

$$predicted\_value(o) = \begin{cases} 1 & if P_+(o) \geq \theta \\ -1 & otherwise \end{cases} \tag{4.6}$$

If we modify θ, the confusion matrix and all the standard measures defined above will also vary. Hence these values depend on θ. The classifier has different "functioning points": with the same classifier, by simply changing the value of θ, one can lower the false alarm rate if (s)he accepts to miss positive values, or decide that (s)he does not want to miss positive objects, at the cost of more false alarms. Most often, the classification performance is only given for the default value of θ, which is one "functioning point" of the classifier. However, it is interesting to characterize the

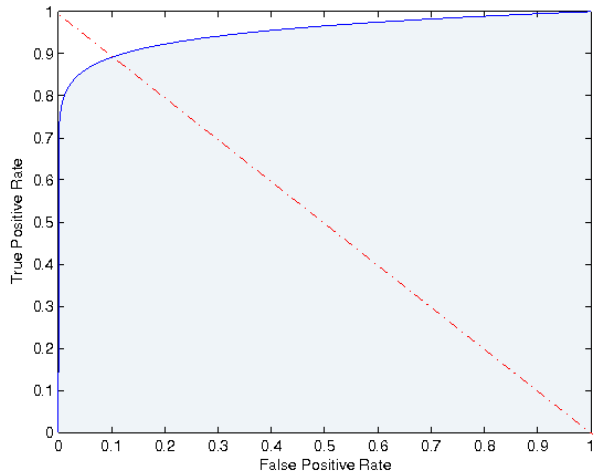

Figure 4.1: An example ROC curve (solid line). The intersection with the (dashed) diagonal is the Equal Error Rate point and the Area Under the ROC Curve is highlighted.

different trade-offs that can be achieved with the classifier. The most explicative way to do it is through the Receiver Operator Characteristic (ROC) curve, which is a graph displaying the evolution of the True Positive Rate with the False Positive Rate (see Figure 4.1).

As it is not always easy to compare graphs, measures have been introduced to characterize the ROC curve:

- the *Area Under the ROC Curve (AUC-ROC)* is the highlighted area on Figure 4.1. The closer it is to 1, the better the classifier, since it means that the classifier can achieve high True Positive Rates with few false alarms.

- the *Equal Error Rate (EER)* is the FPR at the characteristic functioning point where the proportion of FP equals the proportion of FN. This point lies at the intersection of the ROC curve with the dashed diagonal on Figure 4.1. The lower the EER, the better the classification.

To conclude, we can generalize to the multiple classes case (N classes). Accuracy will still be the proportion of correct classification among the N classes. All the other measures presented above can only be given if one class is considered as positive and all the other classes form the negative class.

4.1.2 Audio-only discrimination of laughter versus other events

The works to separate laughter from other sounds can be distinguished depending on the nature of the task. When events are pre-segmented, we talk about *classification*. Classification works will be presented first, in Section 4.1.2.1. On the other hand, when the objective is to detect laughter and locate its boundaries as accurately as possible in audio streams containing other events, we talk about *detection* or *segmentation*. These kinds of works will be presented in Section 4.1.2.2 and prolonged in Section 4.1.2.3, which relates to the specific challenge proposed at the INTERSPEECH'13 conference.

4.1.2.1 Classification of pre-segmented data

The global approach followed up to now for discriminating speech and laughter is to compute usual acoustic features and feed them into typical classifiers: Gaussian Mixture Models (GMMs), Support Vector Machines (SVMs) or Multi-Layer Perceptrons (MLPs) [Haykin 1994]. Several recent studies used the ICSI Meeting Corpus (see Section 2.2.1.3) to train and evaluate the methods, from which speech-laughs were always excluded.

In 2004, Kennedy and Ellis [Kennedy & Ellis 2004] classified one-second segments labeled as laughter when more than one participant was laughing. They used SVMs as classification tool. Features included MFCCs, delta[1] MFCCs as well as modulation spectrum. Modulation spectrum was computed by summing the energy in the 1000-4000Hz range of the spectrogram over a one-second window, computing the Discrete Fourier Transform of this signal and storing the first 20 coefficients. The objective of including modulation spectrum was to capture the repetition of vowels sounds in laughter. However, results showed that MFCCs outperformed modulation spectrum. Kennedy and Ellis achieved 87% of accuracy with six MFCCs.

In 2007, Truong and van Leeuwen [Truong & van Leeuwen 2007a] published a comparative study of several feature sets, classifiers, and fusion of classifiers outputs to classify pre-segmented pure speech or laughter utterances. The ICSI Meeting Corpus, only keeping doubtless laughter episodes, was used to train the methods and evaluate the results and, for a better estimation of the generalization power of the algorithms, the CGN (see Section 2.2.1) was also used for evaluation. Features were split in two categories: *frame-level* features, computed every 16 ms in a 32 ms window, in variable number for a whole utterance—as the number of frames depends on the duration of the utterance—and *utterance-level* features, computed on the whole utterance and generating a constant number of features for each segment. The frame-level features included Perceptual Linear Prediction (PLP) coding features as spectral characteristics, pitch and energy values, as well as the deltas of each feature. A second set of frame-level features consisted of 16 modulation spectrum attributes. As

[1] The term "delta" is commonly used to designate the first derivative. The second derivative is then denoted "delta-delta".

utterance-level features, pitch statistics (mean, standard deviation, excursion, mean absolute slope), the fraction of unvoiced frames and the ratio of non-voiced breaks inside the utterance formed one set of prosodic attributes (called P&V for "Pitch and Voicing"). GMMs and SVMs were used to perform the classification with each set of features.

The best results were achieved with the PLP features, giving clues that spectral features contain useful information to distinguish between speech and laughter. Promising performance was also obtained using only the six P&V features. A discriminant analysis among these six features showed that the mean pitch and the ratio of unvoiced frames have the biggest discriminative power. SVMs performed globally better than GMMs. Then, classifier outputs were fused. The fusion algorithm was a linear combination (with equal weights for each classifier), an SVM or an MLP. This enabled to reduce the classification errors. The best results were obtained by fusing via an MLP the outcomes of SVMs and GMMs taking PLP and P&V features as inputs, with EERs around 3% on the ICSI Meeting Corpus (25 minutes of speech and 21 minutes of laughter to classify) and 7.5% on the CGN, which had balanced proportions of laughter and speech (4 minutes each).

These classification results are very good, but it should not be forgotten that laughter and speech were pre-segmented. Truong and van Leeuwen [Truong & van Leeuwen 2007a] suggested the use of HMMs to capture the temporal variations of features and detect the laughter boundaries in continuous audio files. They addressed the problem of laughter segmentation later on (see Section 4.1.2.2).

Schuller et al. [Schuller *et al.* 2008] discriminated between five unbalanced classes of non-verbal events: breathing (452 instances), laughter (261), consent (325), hesitation (1147) and other (716). They compared features sets based on MFCCs or PLP coefficients. They tested three different classifiers: HMMs, Hidden Conditional Random Fields (HCRFs), which both model the dynamics of features (the evolution of features over a segment), and SVMs, which work with a static vector characterizing the whole segment (by computing functionals of the features over the segments). They also tuned the number of states of the HMMs and of Gaussian mixtures for each state to find the best configuration. The best performance was achieved with HMMs containing nine emitting states and emission probabilities computed with mixtures of eight Gaussians. PLP features yielded slightly better results than MFCCs with an accuracy of 80.7%. HCRFs and SVMs could not beat HMMs performance.

Weninger and Schuller [Weninger & Schuller 2012] addressed the classification of five pre-segmented categories: words; laughter; vocal noise including breathing, sighing or coughing; non-verbal consent ("mhm"); and filled pauses ("um","uh"). Four different databases (AVEC, AVIC, Buckeye and COSINE, see Section 2.2), with available transcriptions of speech and the other targeted events, were used. Data were segmented into word-like units through a forced alignment of the transcription by triphone HMMs. Separate HMMs were then trained for each class, and in the test set each unseen segment was assigned to the most likely class. Authors compared classification performance in two different settings: *intra-corpus* where models are trained

on one part of a database and evaluated on another part of the same database, and *inter-corpus*, where models are trained on one database and evaluated on another one. Authors used the Unweighted Average Recall (UAR) (i.e., the average of the Recall rates obtained for the 5 classes, without weighting these classes) as the performance measure. Performance ranged from 67% UAR to 83% UAR in the intra-corpus setting, but dropped in the inter-corpus classification, with UARs ranging between 30% and 74%. The lower bound could be increased when only leaving one corpus out for testing (i.e., training on the three other corpora), with UARs between 60% and 76%. These results illustrate the difficulties of generalizing models trained on one corpus to data from another corpus and suggest, if generalization to different recording setups is expected, to try to eliminate dependency on the recording conditions by combining several corpora recorded in different environments.

Finally, Tanaka and Campbell [Tanaka & Campbell 2011] used HMMs fed with MFCCs (and their first order derivative) to distinguish between four types of calls produced by one speaker: nasal, ingressive, chuckles or vocal. The overall accuracy was 87%, with most confusions related to the ingressive calls, which were misidentified as chuckles.

4.1.2.2 Laughter segmentation in continuous audio flows

In a second experiment using only ICSI Meeting Corpus files, Truong and van Leeuwen tried to perform laughter segmentation in entire conversation files [Truong & van Leeuwen 2007b]. PLP features were used to take a decision every 16 ms and a Viterbi decoder was used to find the most likely sequence of labels (silence, speech or laughter). They obtained an EER of 11% using this technique, which, as expected, is higher (i.e., worse) than the rate achieved when laughter and speech are pre-segmented (6%). Many of the errors were caused by "noises" such as deep breaths, coughs or external noises, and it seems that unvoiced laughs can easily be confused with these perturbations.

Knox and Mirghafori [Knox & Mirghafori 2007] employed MLPs to segment laughs in the ICSI Meeting Corpus, speech-laughs excluded. They used the 0^{th} and the first 12 MFCCs as spectral features, with their first (delta) and second (delta-delta) order derivatives. In addition, f_0, RMS energy and the AutoCorrelation Peak (AC-PEAK) of each 25 ms frame (with 10 ms shift between consecutive frames), as well as their first and second derivatives were part of the feature vector. They determined the best number of adjacent feature vectors to provide to the MLP to be 75: each frame was classified using its feature vector as well as the attributes of the preceding and following 37 frames. They tried different numbers of hidden units and decided that 200 was a good value. Separate MLPs were trained for each subset of features (MFCCs, f_0, energy, AC-PEAK; and taking into account the feature itself, its delta, its delta-delta, or a combination of the three). The best results were achieved with delta-MFCC features, with an EER under 10%. Fusing the scores provided by MLPs using MFCC and AC-PEAK features, they reached an Equal Error Rate (EER) of

around 8%. This performance is similar to results they had previously obtained with SVMs, but the temporal resolution was then of 25 ms instead of 10 ms.

Knox et al. [Knox *et al.* 2008] improved that work by considering temporal features and using HMMs to process the sequence of laughter likelihoods returned by the MLP. In a first step, they evaluated the combination of seven feature sets to perform laughter and speech discrimination using MLPs. One MLP was trained for each of the feature sets. The features were extracted on windows of 25 ms shifted by 10 ms and were namely:

- 13 delta-MFCCs.
- f_0 and delta-f_0.
- RMS energy and delta-energy.
- AutoCorrelation Peak (AC-PEAK) and delta-AC-PEAK.
- Phones: a phone recognizer trained for speech was used to assign each frame to one of the 46 possible phones.
- Prosody: statistics of jitter, shimmer and long-term average spectrum; these features were extracted over 0.5 s windows.
- Modulation-filtered SpectroGrams (MSGs), computing amplitude modulations in the range 0-16 Hz, which is the frequency range where laughter repetitiveness is expected to appear (see Section 3.1.6). The authors did not specify the size of the window for these features.

Each individual feature set MLP was trained with the features of 101 consecutive windows. The decisions of individual feature set MLPs were fused with another MLP using nine consecutive windows. The output was median-filtered to smooth transitions. The best combination of features included the delta-MFCCs, MSG, energy, autocorrelation and prosody features, reaching an EER of 5.4%. The performance was further improved by including Hidden Markov Models to build a trigram language model with the smoothed posterior probabilities, reaching 78.5% of precision and 85.3% of Recall. In the easier task of classifying a balanced laughter and speech (without other non-speech sounds like deep breathing) test set, the performance increased to an EER of 2.7% without the HMMs and precision and Recall with the HMMs of 99.5% and 88%, respectively. The phone set was useless in this experiment, but no general conclusion can be taken on the use of laughter phones at this point since the speech phone recognizer had to deal with unseen laugh phones, which were mapped to speech phones.

Scherer et al. [Scherer *et al.* 2009] extracted, every 20 ms on 200 ms windows, modulation spectrum features on the FreeTalk database (see Section 2.2.2.5). These features were fed into an Echo State Network (ESN), which, according to the authors, has the advantages to be robust to noisy inputs and to memorize past states and

features, thanks to the introduction of recursive connections. Frame-based detection resulted in an accuracy of 87%. It is important to note that, in the FreeTalk database, only one microphone was used for recording all the participants.

Sudheer et al. [Sudheer *et al.* 2009] proposed a method for extracting the fundamental frequency (f_0) and the strength of excitation (related to the speed of closing of the vocal folds). The method, designed to be more robust than traditional f_0 estimation algorithm to track the rapid variations in laughter, consists in filtering the signal through a zero-frequency resonator with a window length of 3 ms for removing the trend. The positive zero-crossings of the resulting signal are then used to spot the periods of voicing and obtain a first estimation of f_0. The resulting voiced segments are passed through a zero-frequency resonator with a window size adapted to the roughly estimated fundamental frequency of the segment. The positive zero-crossings give the location of the epochs. The pitch period is the interval between two epochs and the strength of the excitation is given by the slope around positive zero-crossings. Features used for laughter detection are the pitch period, the strength of excitation, their slopes and their ratio. A voiced segment was labeled as laughter if, for all its features, at least a given proportion of epochs ("fraction thresholds") fell above fixed thresholds ("value thresholds"). As the authors did not consider monosyllabic laughter as laughter, isolated laughter segments (laughter segments of less than 500 ms having at least 3 seconds of non-laughter either side) were relabeled as non-laughter. The laughter detection method has been evaluated on TV program excerpts, with a Recall of 88.9% and a FPR of 24.1% for the detection of laughter segments, and a Recall of 95.9% and a FPR of 27.4% for the detection of laughter episodes. The performance is relatively satisfying given the simplicity of the method (only one type of features based on f_0, classification via thresholds) but the method is limited to the detection of voiced laughs. It would be interesting to evaluate the contribution of the proposed features when combined to more traditional features to detect all kinds of laughter (voiced and unvoiced) via a trained classifier.

Unlike for speech recognition, and despite the suggestion of Truong and van Leeuwen [Truong & van Leeuwen 2007a], Hidden Markov Models (HMMs) have not been intensively used for laughter detection in continuous streams. We can cite three works using HMMs to distinguish between speech and laughter. Locker and Mueller [Lockerd & Mueller 2002] mention the use of HMMs trained with 70% of a corpus containing 40 laughter and 210 speech segments from a single speaker. Using spectral features, 88% of the test segments were correctly classified. The performance drops to 65% when the data are not pre-segmented. Cai et al. [Cai *et al.* 2003] also used HMMs to spot laughter, cheers and applause events. The features were MFCCs, energy, ZCR, energy in four sub-bands, centroid and spread of the energy spectrum. HMMs of two, four and four states were used respectively for applause, cheers and laughter, and a mixture of four Gaussians modeled the emission probability of each state. The models were evaluated on three types of TV programs with average Recall and precision for laughter of 95% and 86% respectively.

A very interesting approach relying on HMMs has been proposed by Pammi

et al. [Pammi *et al.* 2013]. They used Automatic Language Independent Speech Processing (ALISP), which consists in building models (here HMMs) on a huge, unlabeled data corpus. Hence the models are fully data-driven and can be considered as "universal" sound units if they are trained on a sufficiently diverse database (containing speech, music, laughs, etc.). The data are first segmented into quasi-stationary segments and then acoustically similar segments are clustered to form one single ALISP model. Here the models consisted in HMMs for each of the identified ALISP units and the modeled features are the MFCCs. The models were trained with 240 hours of audio. After building the "universal" ALISP models, Pammi et al. adapted them to laughter on the one hand and non-laughter events on the other hand, this time with the help of labeled data (coming from the AVLaughterCycle, SEMAINE and MAHNOB databases, see Chapter 2). They obtained this way separate ALISP models for laughter and non-laughter units. These models could then be used to decode an unseen audio file: a simple Viterbi decoding was employed to output the most likely ALISP sequence corresponding to the audio file. The resulting sequence was indicative of the presence of laughter in regions where laughter ALISP models were preferred to the non-laughter models. Hence the output ALISP sequence could be converted to a sequence of laughter and non-laughter units (binary signal). To reduce outliers, Pammi et al. proposed to use a median filter on the sequence of laughter and non-laughter units. The ALISP method for laughter detection was compared with more traditional supervised approaches (GMMs, HMMs with several topologies) and was shown to outperform these methods. The best ALISP performance for laughter detection was a Precision of 94.3% and a Recall of 93.9%. It must be noted that some of the used corpora (AVLaughterCycle, MAHNOB) are relatively easy for laughter detection as they mostly contain laughter sounds, but the ALISP methods have the interest to be universal and have shown their potential here for laughter detection.

4.1.2.3 INTERSPEECH 2013 Social Signals Sub-Challenge of the Computational Paralinguistic Challenge

In the framework of the INTERSPEECH 2013 conference, a challenge on the detection of laughter and fillers in speech was proposed. For the challenge, the SSPNet Vocalization Corpus (SVC) (see Section 2.2.1) was released together with frame-based annotations into laughter, filler and garbage (containing anything that is not laughter and fillers, such as speech, cough, etc.) at 100 FPS [Schuller *et al.* 2013]. The corpus was divided into training, development and test sets. Frame-wise features extracted with the OpenSMILE library [Eyben *et al.* 2010] were also available to the challenge participants, including 12 MFCCs, logarithmic energy, voicing probability, HNR, f_0 and ZCR. The first order derivatives of all the features were also included in the feature set, as well as the second order derivatives of MFCCs and logarithmic energy, summing up to a total of 47 features. The feature vector of each frame was extended with the arithmetic mean and standard deviation of the 47 features across 9 frames centered on the considered frame (hence, a window ranging from 4 frames before to

4 frames after the considered frame), for a total of 47*3=141 base features for each frame. Feeding these features into SVMs, the challenge organizers obtained the following results on the test set: AUC-ROC of 82.9% for laughter and 83.6% for fillers, which results in an Unweighted Average Area Under the ROC Curve (UAAUC) of 83.3%. Participants to the challenge were invited to improve these results with the use of other feature sets or classification methods. Here is a summary of the papers that have addressed the challenge.

Krikke and Truong Krikke and Truong [Krikke & Truong 2013] compared several feature sets for laughter detection:

- 13 MFCCs (as well as their first and second order derivatives).
- Pitch and intensity features.
- Position of the first and second formants.
- Voice quality features, based on frequency bands of the long-term averaged spectrum.

The base features were extracted every 10 ms. Functionals were applied to all the features except MFCCs, so that the feature vector of each frame contained the feature value for the considered frame as well as the average, standard deviation and slope of a linear least square fit computed over nine frames centered on the considered frame.

Gaussian Mixture Models were trained for laughter detection against anything else (filler or speech). Results were smoothed through median filtering. The best results were obtained with 128 Gaussian distributions and median filtering based on 51 frames. The MFCCs set outperformed all the other sets (including the combination of all the features) and reached an EER of 9.3% for laughter.

Krikke and Truong explored slightly different feature sets for filler detection and obtained similar results. It is interesting to remark that they have used two features (proposed in [Pruthi & Espy-Wilson 2004]) to characterize the nasality of a frame: one is based on the ratio between the maximum energy below 300 Hz and the maximum energy between 300 and 5500 Hz, the second is the peak frequency below 800 Hz. Although these features did not prove useful for spotting fillers in the challenge, it would be interesting to consider them to discriminate between different laughter sounds.

Oh et al. Oh et al. [Oh *et al.* 2013] investigated features at the syllable level, with a primary focus on laughter. First, they segmented the SVC files into "syllables" by cutting at minimum values of the median-filtered energy envelope. Then, for each syllable they computed features characterizing the energy profile (minimum, maximum, and features measuring the attack and decay), the f_0 pattern (minimum and maximum f_0 values as well as their positions within the syllable), the timbral contour (minimum, maximum and average spectral flux) as well as rhythmic features,

which were estimated over larger windows (e.g., 500 ms) to reflect the periodicity of the energy envelope across all frequencies and also in the 4-6 Hz band which is the band associated to laughter rhythm (as described in Section 3.1.6). The first order derivative as well as the average value of each of the computed features over five syllables (centered on the current syllable) were added in the feature set for each syllable.

Analysis of the discriminant power of the proposed features showed that f_0-related features were not useful for the considered task, while features characterizing the energy profile were the most discriminant, followed by energy modulation inside the 4-6 Hz band. Using these features in combination with the baseline features provided by the challenge organizers, the authors could slightly improve the SVM classification results, which reached an AUC-ROC of 85.9% for laughter and an UAAUC of 85.3%.

An et al. An et al. [An *et al.* 2013] also relied on syllable-based features. They used a pseudo-syllabification algorithm based on the amplitude contour to segment the files into syllable-like regions. They experimented two different methods to improve the baseline detection results.

First, they considered the likelihoods for the three classes (laughter, filler, garbage) over each syllable segment and rescored the frames belonging to this segment if the combined likelihoods reached certain thresholds. It is however surprising that they only rescored frames which have a likelihood below another threshold: they justified this by the will to preserve labels when the classifier was highly confident, yielding a score above the threshold. But on the other hand this goes against their desire to obtain stable labels for several frames (e.g., a syllable) as the events to be detected lasted more than one frame.

Second, they computed features over pseudo-syllables: normalized[2] intensity, normalized pitch, mean spectral tilt (i.e., the average over the syllable segment of the mean slope of the spectrum of the 10 ms frames), duration and duration of the preceding and following pauses. They also included the first derivative of the intensity, pitch and spectral tilt, as well as the position of the syllable (as they noticed that no clip begin with laughter or filler, the position of the syllable can be relevant for that specific corpus). To classify each syllable, they appended in the feature vector the features of the previous and following syllables, making a total of 9 (features per syllable) * 3 (syllables) + 1 (position feature) = 28 features.

In their experiments, they compared the baseline challenge detection (baseline features, SVM) and combinations of the baseline methods with the additional syllable features and/or the proposed rescoring. They found that both additions (rescoring or adding syllable features) individually improved the performance, but that the combination of both yielded to lower results than using the additional features alone (in other words, there was no benefit of rescoring the labels if the detection used their proposed syllable features). The best performance achieved with this method

[2] In this case, "normalized" means that the actual value is divided with respect to the values encountered in the training set.

was AUC-ROCs of 84.64% and 85.06% for laughter and fillers respectively, and an UAAUC of 84.85%.

Janicki Janicki [Janicki 2013] proposed to use in a first step three GMMs to obtain frame-based likelihoods for laughter, filler and speech, respectively, then to rely on a SVM to fuse the decisions. Only MFCCs were used. Several parameters were varied in the experiment: the number of Gaussian components, the length of the centered windows inputted to the GMMs, the impact of adding the first and second order derivatives, the likelihood scores used by the SVM (absolute scores, differences between likelihoods, or combination of both). The best results were achieved when using 128 Gaussian distributions and a window of 60 frames, with the first and second order derivatives of the MFCCs. The method reached the following performance: AUC-ROCs of 90.7% and 89% for laughter and fillers, respectively, UAAUC of 89.8%, EERs of 15.3% and 18.4% for laughter and fillers, respectively. Janicki also considered the features proposed by the challenge organizers, but the addition of these features to the SVM classifier did not improve the overall accuracy[3].

Gupta et al. Gupta et al. [Gupta *et al.* 2013] improved the baseline method by smoothing and masking events probabilities. They used the same features as the challenge organizers, but replaced the SVM by a four-layers Deep Neural Network outputting the likelihoods for each frame to be laughter, filler or anything else (garbage). Then, they filtered the outputs to have smoother predictions over time—as events last several frames, one can avoid oscillations in the likelihood signals. After analyzing the resulting likelihoods, they noticed that when an event (laughter or filler) occurred, there was at least one frame in the event which received a high likelihood for the correct class. Based on this finding, they developed several post-processing steps to weight the likelihoods. The first steps take into account the likelihoods of the neighboring frames: frames that are far from values above a first threshold in the likelihood signals are set to 0, then frames closely surrounded by zero frames are also set to 0, and finally all frames above another threshold are set to 1. The last post-processing step involves a speech recognizer system trained on the data to recognize laughter, filler and garbage. The state occupations obtained when decoding the signals are used to compute the entropy of each frame, which relates to the number of competing states. This number was hypothesized to be higher in laughter (which is highly heterogeneous and variable) than fillers (which are much more stable and can hence be represented with a smaller number of states). Hence different corrections were applied to the laughter and filler likelihoods with respect to the entropy value.

Using this method, Gupta et al. could significantly improve the baseline challenge results, with AUC-ROCs of 93.3% and 89.7% for laughter and speech, respectively, resulting in an UAAUC of 91.5%. Part of the gain can be attributed to the Deep Neural Network, which performs better than the baseline SVM especially for laughter,

[3]The Recall rate was however slightly increased with most of the feature sets.

but the probability smoothing contributes for the biggest part of the performance increase, while the masking steps yield to further small improvements.

Wagner et al. Wagner et al. [Wagner *et al.* 2013] investigated the potential of using phonetic transcriptions to help the detection. They used a standard speech recognizer which enables to output the most likely sequence of phonemes within a file. The speech recognizer included models for 40 English speech phonemes as well as eight "fillers" ("breath", "noise", "cough", "uh", "um", "uhum", "noise" and "garbage"). Wagner et al. used the outputted phonetic sequence for each file and constructed, for each frame, a feature vector with the number of apparitions of each of the 48 "phonemes" within a given window. They noticed that the distribution of "phonemes" differed for each class. Interestingly, the most frequent phonemes in laughter were similar to the ones presented in Section 3.2.2, although they used a recognizer trained for speech recognition.

They fed the SVM with feature vectors containing the baseline challenge features as well as two variations of their phonetic distribution features. Adding the phonetic features improved the detection, in particular when the window for computing the phonetic distributions was larger than 1 second. The best performance achieved with the proposed method was AUC-ROCs of 89.4% and 85.9% for laughter and fillers, respectively, for a global UAAUC of 87.7%.

Kaya et al. Kaya et al. [Kaya *et al.* 2013] employed random forests for classification of laughter, fillers and garbage. They investigated the impact of the random forests parameters: the number of trees and the number of features available (through random selection) for each tree. In addition, they performed feature selection based on mutual information of the features, to limit redundancy. Finally, they explored the effect of Gaussian smoothing on the sequence of frame-wise decisions. They showed that random forests can outperform the baseline SVMs and that feature selection could improve the results. They reported AUC-ROCs of 89.6% for laughter and 87.3% for fillers, corresponding to an UAAUC of 88.4%.

4.1.3 Audiovisual discrimination of laughter versus other events

This section consists in a brief introduction to audiovisual works that attempted to discriminate laughter from speech. As visual analysis is out of the scope of this dissertation, we do not intend to give an exhaustive list of the state-of-the-art here[4], but only to present the main works. In particular, Petridis and Pantic's works are of interest as they are among the few teams who have tried to discriminate between laughter types in addition to distinguish laughter from other acoustic events.

[4]We certainly do not want to claim that the other sections are exhaustive either, in spite of our best efforts, but here we did not even try.

In 2008, Petridis and Pantic [Petridis & Pantic 2008c, Petridis & Pantic 2008a, Petridis & Pantic 2008b] started considering fusion of the audio and video modalities to discriminate pre-segmented laughter episodes from speech segments. They used the AMI Meeting Corpus (see Section 2.2.2.3) and selected 40 laughs (for a total of 58.4 s) and 56 speech segments (total: 118.1 s) from eight speakers. They extracted PLP features, pitch and energy from the audio, and head and facial movements from the video. Twenty facial points were tracked over the course of the video. Facial and global head movements were separated using Principal Component Analysis, which grouped the global head movements along the lower orders of the Principal Components (PCs). The first six PCs were kept for the head movements. PCs #7 and #8 were linked to mouth opening and closing, respectively. PCs #9 and #10 were also included in the facial movements. The authors remarked that the decomposition of head and facial movements did not always work perfectly, and facial variations could erroneously be included in the head features for some samples. They used neural networks to distinguish between speech and laughter and showed that the best features were the facial movements, followed by the spectral features (PLP). They also investigated decision-level fusion of the modalities, which improved the decision compared to audio- or video-only classification. They showed that a simple linear combination fusion between the modalities achieved similar results as a non-linear fusion decision obtained with neural networks. The best results achieved a F_1^{score} of 87.5% when a local decision was taken and 89.3% when considering the temporal evolution of the features.

These efforts led to an application—judging the hilarity of stimuli by laughter analysis—based on feature-level fusion, spectral acoustic features and facial expressions [Petridis & Pantic 2009]. They reported a classification accuracy of 74.7% to distinguish three classes, namely unvoiced laughter, voiced laughter and speech.

Reuderick et al. [Reuderink *et al.* 2008] conducted a similar study. Also using laughs from the AMI Meeting Corpus as well as speech segments (with no smile of the speaker), they performed decision-level fusion between audio and visual classifiers. Audio features were 13 RASTA-PLP features, which add filtering capabilities to PLP features for channel distortion, and their derivatives, extracted every 16 ms over 32 ms windows. Video features were the PCs of the 20 positions of tracked facial points and their derivatives. HMMs and GMMs were investigated for the audio distinction between speech and laughter. Visual classification was performed through SVMs. SVMs and linear combinations were evaluated for the fusion of the decisions taken with the two modalities. For the audio modality, the best results were obtained with GMMs, resulting in an AUC-ROC of 82.5%. Ergodic HMMs (HMMs where transitions are possible from and to any state) were slightly behind with 82.2%, and surprisingly outperformed left-right HMMs (HMMs where transitions are only possible in one direction), which are generally used in speech processing. Video alone had better classification rates, with an AUC-ROC of 91.6%. This figure increased to 93% when fusing both modalities, with similar results for SVM and linear combination fusions.

In 2010, Petridis et al. [Petridis *et al.* 2010] used neural networks to model the

relationship between audio and video features. Four neural networks were trained to predict audio features from video features and vice-versa, for laughter and speech. Then, for each test frame, the neural network giving the best prediction of unseen features was considered as the winner and the frame was labeled accordingly. The system was evaluated on data from the AMI Meeting Corpus as well as from the SEMAINE database. Six MFCCs were used as audio features, while the first four and three PCs of the facial markers were kept as video features for the AMI and SEMAINE data, respectively. The results are hard to compare with previous work, since here the data were evaluated on a database different from the training. Training performed on the AMI Corpus (which is difficult for video features due to large head movements) and testing on the SEMAINE database (easier since head movements are limited) yielded a F_1^{score} of 95%. For comparison a standard neural network, using audio and visual features to discriminate laughter and speech, was trained with the same data and gave the same results. However, when training with the SEMAINE data and testing on the AMI corpus, the modeling of the relationship between audio and visual features gave a better F_1^{score} (76%) than a standard neural network (65%). The authors concluded that modeling the audiovisual relationship for speech and laughter and letting the two models compete has better generalization properties than an usual classifier. Using longer windows (160 ms) and standard neural network classification with feature-level fusion, Petridis and Pantic could improve the classification rates in [Petridis & Pantic 2011]. The problem of generalization from SEMAINE to AMI was still present.

In 2013, Petridis et al. addressed the problem of audiovisual laughter detection in continuous streams [Petridis *et al.* 2013a]. They extracted MFCCs and Facial Action Parameters (FAPs) on the SEMAINE database and trained Time Delay Neural Networks (TDNNs) to recognize laughter and speech frames. One TDNN was trained for each modality (audio or video) and fusion was performed at the decision level (fusion at the feature level was tried too but lead to worst results). A Voice Activity Detector was used to filter out silent portions prior to speech and laughter discrimination. Performance is estimated through ten-fold cross validation. Laughter detection results were poor, with an average Recall of 41.9%, a Precision of 10.4% and a F_1^{score} of 16.4% for the audiovisual detection. Such bad results are partially explained by the unbalanced data, where speech frames largely outnumber laughter frames: in consequence, even if a small proportion of speech frames were misclassified as laughter, the False Positives (FP) would outnumber the True Positives (TP) and yield to a poor Precision score. Results however showed that adding visual information to audio laughter detection helped reducing the misclassification of speech as laughter, hence increasing the Precision score for laughter (and increasing the Recall rate for speech).

Audiovisual laughter detection was also investigated by Scherer et al. [Scherer *et al.* 2009], on the FreeTalk database (see Section 2.2.2.5). Using features characterizing facial and body movements of all the participants fed into ESNs, they obtained an accuracy of around 82%, which is lower than what they achieved using audio features (around 87%), while the combination of both audio and video yielded

improvements, with an accuracy around 91%.

In a second study [Scherer *et al.* 2012], still using the FreeTalk data, Scherer et al. compared different audiovisual feature sets and classifiers for audiovisual laughter classification (pre-segmented laughs) and detection (continuous streams). The compared classifiers were ESNs, HMMs and SVMs. SVMs were fed with GMMs Supervectors, resulting from one step of adaptation of GMMs representing "universal" data (both speech and laughter) towards the inputted features. Audio feature sets were modulation spectrum on one side, extracted over 200 ms windows with 20 ms shift, and PLP coefficients on the other side, extracted every 10 ms on 32 ms windows. Visual features relied on face detection: fore each detected face (i.e., for each of the four participants), two features were computed: one represented the movement within the face, the other the movement of the body (below the face). As one single, centrally-placed, microphone was used to record all the participants, and as visual features also included all the participants, all the feature sets were combinations of all four participants' behaviors. Fusion was performed at the decision level. For the detection experiments, HMMs and SVMs with supervectors operated on fixed-sized windows of 1.2 s. A laugh was considered to be spotted as soon as there was one peak of the estimated laughter likelihood within the boundaries of the actual laugh. In other words, the authors did not aim to locate the boundaries of the laughs, but to count the number of laughs in the data.

For the classification task, SVMs outperformed HMMs, with an accuracy of 96.3% when fusing modulation spectrum and PLP decisions[5]. However, for the detection task, SVMs gave poor results. The authors suggested that this was due to the unbalanced data. HMMs and ESNs, which encode the dynamics of the features, performed better. The best performance with HMMs was achieved with modulation spectrum and PLP features, with an accuracy of 93.5%, a Precision of 64% and a Recall of 80%. ESNs were slightly below: fusion of visual and modulation spectrum yielded an accuracy of 90.9%, a Precision of 52% and a Recall rate of 81%.

4.1.4 Classification of laughs

As already mentioned, Petridis and Pantic [Petridis & Pantic 2009] distinguished between voiced laughter, unvoiced laughter and speech. The overall classification rate was 74.7% and the F_1^{scores} for voiced and unvoiced laughter classification were 70.4% and 66.6%, respectively.

To the best of our knowledge, there exist only three other works which aimed at automatically distinguishing laughs from other laughs using audio features. For each of these works, laughter was pre-segmented, grouped in several classes and the objectives were to predict those classes from acoustic features.

First, Bachorowski et al. [Bachorowski *et al.* 2001] investigated whether laugher sex and identity could be inferred from acoustic features. The analysis was performed

[5] ESNs could not be used in this task, as they require some time to initialize and are hence not suited to process short segments.

at the call level: the objective was to classify calls, not entire laughs. Separate classification was performed for voiced and unvoiced open-mouth calls. Data from a given participant were included in the study only if the participant had produced at least six occurrences of the considered call type. As a result data from 19 males and 13 females were included for voiced calls, and from 11 males and 7 females for unvoiced calls. Features were extracted on call segments. Several feature sets were compared in the task: formants F1-F2-F3, formants F1-F2-F3-F5[6], call duration, vocal tract length (estimated from formants positions) and f_0-related features (mean, std, excursion and change over the call) only for voiced calls. Discriminant analysis was used for classification.

Classification of laugher sex was relatively successful, with accuracy rates of 86.3% for voiced calls and 87.4% for unvoiced ones. The best performance was achieved by combining all available features, although close performance was obtained when using formant values only. On the other hand f_0-related features were not really useful (improvement of chance level by only 22%), as could be expected given the large variability of f_0 in laughter, regardless of laugher sex. Despite the relatively good classification rates, Bachorowski et al. noted that performance is lower than what can be achieved with speech vowels.

Identification of the laugher was performed separately for males and females. Performance was again satisfying, with accuracy rates between 40.7% and 53.2% depending on the cases (voiced or unvoiced, females or males) and corresponding improvements over chance levels (which differ in each case as there are different numbers of participants involved) ranging from 30.8% to 49.3%. As for gender classification, the best classification rates were obtained when combining all the features, closely followed by the formant values while the f_0-related features yielded poor classification. And as for gender classification, although the automatic identification of the laugher was largely above chance levels, the authors noted that the obtained performance was lower than what can be achieved from speech vowels.

Second, as already said in the introduction (Chapter 1), Campbell [Campbell 2007] trained neural networks on laughs from telephone conversations to predict the gender and origin (English or Chinese) of the conversational partner. Speech-laughs were also included. For each laughter or speech-laugh episode, the following features were extracted in Snack [Sjölander 2004]:

- Pitch: minimum, maximum, position of the maximum in the episode and voicing proportion.
- Power: minimum, maximum and position of the maximum in the segment.
- Duration and speaking rate (obtained by dividing the duration by the number of transcribed units in the laugh).

[6] F4 was discarded as previous observations had shown that F4 values did not differ among gender, as explained in Chapter 3.

- Spectral shape: location and energy of the first two harmonics, amplitude of the third formant and the difference in energy between the first harmonic and the third formant as a measure of breathiness.

Principal Component Analysis was applied to these features and the first five PCs were retained for laughter classification. Classifications rates were higher than chance for both dimensions (language and gender of the conversational partner), indicating that the laughing style depends on the interlocutor.

Third, Szameitat et al. [Szameitat *et al.* 2009b] tried to automatically distinguish the emotions (joyful, taunting, schadenfreude and tickling) of their portrayed laughs. For this study, they used 127 laughter sequences which had been assigned by naive raters to their emotional class with success rates above chance level (see Section 3.1.7.1)[7].

Laughs were manually segmented in bouts and calls. A range of features were extracted either on the full laughter episode (number of calls, number of bouts, average bout duration, etc.) or on calls, to characterize their duration, energy pattern (ratio between mean and maximum intensity, relative position of the maximum energy within the call, etc.), fundamental frequency pattern (mean, minimum, maximum f_0, relative position of the maximum f_0 value within the call, etc.), formants (positions of first five formants, bandwidth of first formant, etc.) and vocal parameters such as jitter, shimmer, percentage of voicing, HNR as well as the center of gravity, skewness and kurtosis of the spectrum. To obtain one single value for each acoustic parameter and each laugh, the features extracted over calls were averaged across bouts and then further averaged across episodes. As the first call was found to differ from the following calls and to bring only little additional information about the underlying emotion, the average of calls only involved calls numbered two to eight of each bout.

After analyzing the discriminant power of each feature to distinguish between the emotions, a set of twelve features was built with the attempt to maximize the discriminant information and minimize redundancy between features. The following features were selected: f_0, Frequency of the first formant (F1) and Frequency of the second formant (F2), average call duration, maximum peak frequency, ratio between maximum frequency and average f_0, difference between maximal and minimal intensity, percentage of voicing, average HNR, center of gravity of the spectrum, average bout duration and average number of calls per second. Discriminant analysis using these twelve features could classify the emotions with an accuracy of 76%. The majority of errors concerned schadenfreude laughs, which were correctly identified in only 43% of the cases, while the other three emotions were correctly classified in over 80% of the cases.

Classification performance was thus relatively good, but the data used suffer from two important limitations. First, it is portrayed laughter and, as explained in Chapter 2, the use of acted episodes to study laughter is contested, as it is unclear whether

[7] It is unknown to us why only 127 laughter episodes are mentioned in the automatic classification study while the characterization study presented in Section 3.1.7.1 had 160 laughs classified above chance level, with equal overall classification agreement rates of 63%.

acted and spontaneous occurrences share the same acoustic properties. Second, Szameitat et al. only considered a subset of laughs (only 30% of their corpus) for which humans could correctly identify emotions. It can thus be expected that these laughs do have distinct patterns (that were related to emotions), and it is not surprising that automatic classification is also efficient on this subset. In consequence, despite the good classification results, this experiment does not prove that emotional laughter types exist and can be classified, even in portrayed laughs. Recent experiments indeed suggest that emotional laughter types do not exist, as was explained in Section 3.1.7.1.

4.2 Laugher retrieval

The AVLaughterCycle database containing around a thousand laughs, we were interested in developing methods to efficiently browse through the database[8]. This problem is different, but related, to the classification of laughs that has just been presented in Section 4.1.4, as here we do not aim to classify laughs into a defined number of categories, but to examine the similarity between laughs and group laughs that are acoustically similar to each other.

We will report in this section the work that has been conducted to group laughs according to the timbre of the laughing voice. Timbre is defined as the quality of tone distinctive of a particular singing voice or musical instrument [Merriam-Webster 2009]. It is linked to the relative intensities of a sound harmonics, independently from its pitch and intensity. It has been shown to be related by the spectral envelope of the sound, using a non-linear scale. In consequence MFCCs have been used to characterize the timbre [Dupont *et al.* 2009b]. The timbre of a voice is an individual characteristic, hence we expected that by grouping laughs according to their timbre, we would group laughs uttered by the same laugher.

To organize the database according to the laughter timbre, the following spectral features were extracted on each laughter episode:

- 13 MFCCs, and their first and second derivatives.
- Spectral flatness and spectral crest values, each divided in four analysis frequency bands (250Hz to 500Hz, 500Hz to 1000Hz, 1000Hz to 2000Hz and 2000Hz to 4000Hz).
- Spectral centroid, spread, skewness and kurtosis.
- Loudness, sharpness and spread, computed on the Bark frequency scale.

[8] This was part of the development of a browsing application for large databases, not restricted to laughter but to any audiovisual data, in the framework of the NUMEDIART research program: http://www.numediart.org/, consulted on February 28, 2014. The application was created within the AudioCycle project [Dupont *et al.* 2009b, Urbain *et al.* 2009b] and extended with the MediaCycle [Siebert *et al.* 2009], LaughterCycle [Dupont *et al.* 2009a] and AVLaughterCycle [Urbain *et al.* 2009a] projects.

- Spectral slope, decrease, roll-off and variation.

In addition, two temporal features were included: RMS energy and ZCR. In total, 60 features were extracted for each frame of 340 samples (sampling frequency: 16 kHz), with 75% overlap. The similarity estimation requires comparing laughs of different lengths, so also comparing different numbers of frames. To obtain a constant feature vector size, it was decided to store only the mean and standard deviation of each feature over the whole segment. More complex models could be investigated but this simple transform provides promising results and establishes a baseline, useful to measure future improvements. This simplification had been successfully used in other similarity computation [Dupont *et al.* 2009b] or laughter classification [Petridis & Pantic 2009] contexts and was assumed applicable to laughter timbre characterization. Normalized Euclidean distance between feature vectors is used to compute the similarity between laughter episodes.

The capability of the method to group laughs uttered by the same laugher has been objectively evaluated, as presented in [Urbain *et al.* 2010b]. For these experiments, some laughs were discarded from the AVLC database: 20 laughs involving speech; 19 laughs from subject #1 for which we do not have facial tracking[9]; the laughs from Subject #24, who only uttered 4 short laughs, which is not enough to perform reliable similarity tests. In total, these experiments involved 978 laughs. To evaluate the similarity estimation, we have taken each laugh L of the AVLaughterCycle database and used the similarity algorithm to retrieve its N closest neighbors. Two different measures have been computed.

In the first measure, the retrieval was considered successful if at least one of the N retrieved laughs had been uttered by the same laugher as L. Figure 4.2 gives the individual success rates for $N = 1, 3, 5$ and 10. The gray bar represents the likelihood of a successful search if randomly organizing laughs instead of using the similarity algorithm.

The random success score for speaker i and N random picks equals

$$R_i^N = 1 - \prod_{k=1}^{N} \frac{N_{tot} - N_i - k + 1}{N_{Tot} - k} \tag{4.7}$$

where N_i is the number of laughs from speaker i out of the N_{tot} laughs in the database.

For each value of N, the similarity algorithm performs significantly better than chance, at a 95% confidence level (all p-values are largely lower than 0.05, using one-sided paired t-tests). However, for some individuals (subjects #10, #19 and, to a smaller extent, #8) the similarity algorithm does not outperform chance. This is probably due to the fact that these subjects mainly uttered nasal or breathy laughs, for which it is very hard to discriminate between subjects (there is no perceived timbre). On the other hand, Subject #11, who gets a (nearly) perfect success rate,

[9] Facial tracking was not used in the experiment reported here, but was important for other experiments. See [Urbain *et al.* 2010b] for further information.

produced a large majority of voiced ("vowel") laughs[10].

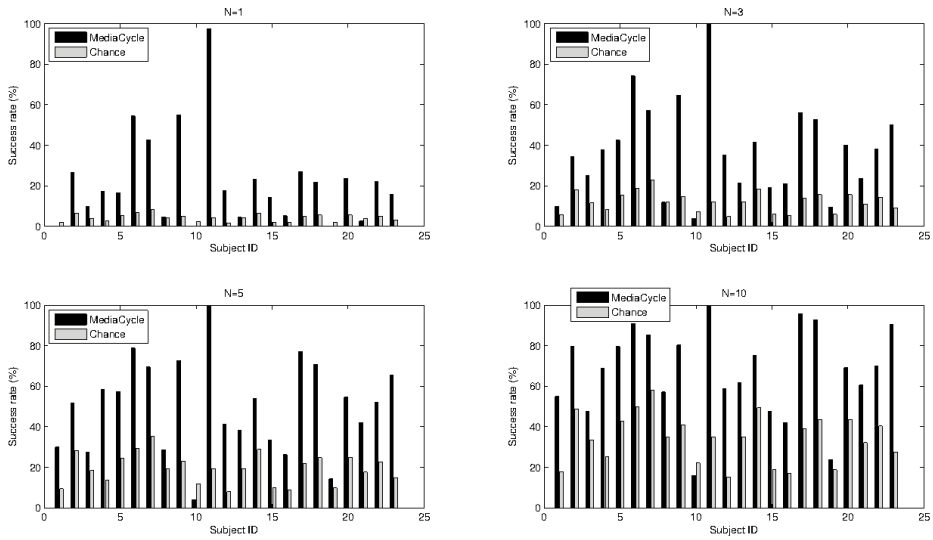

Figure 4.2: Success rates achieved by the similarity browsing application, MediaCycle (black), against chance (gray) for laugher retrieval, using N picks.

To complement this information and illustrate the interest of using the similarity algorithm to organize a laughter database according to the speaker, we have computed, for each laugh, the average number of utterances one needs to pick to find one laugh from the same speaker. Again, the timbre similarity algorithm (utterances ordered by distance to the input laugh in the feature space) was compared against chance. The mean chance score for speaker i equals:

$$C_i = \sum_{u=1}^{N_0+1} u \cdot \frac{N_1 - 1}{N_{tot} - u} \prod_{t=1}^{u-1} \frac{N_0 - t + 1}{N_{tot} - t} \tag{4.8}$$

where N_i is the number of laughs from speaker i out of the N_{tot} laughs in the database and $N_0 = N_{tot} - N_i$ is the number of laughs from other speakers. The results are shown on Figure 4.3, with the standard deviation intervals for the similarity algorithm. The grouping by timbre is undoubtedly better than random search[11], though for five subjects (#3, #7, #14, #20, #21), the $mean + std$ value goes above (i.e., is worse than) the chance performance. The one-sided paired t-test gives a p-value of 7.1×10^{-8}. For unknown reasons, the similarity algorithm was not able to efficiently improve the search for Subject #3, who uttered 40 laughs spread over the laughs types.

[10] I have been informed that (s)he will most likely attend the PhD dissertation, so you could get a chance to hear her/his nice laughs if I succeed in making one joke.

[11] The lower the number of picks, the faster the search.

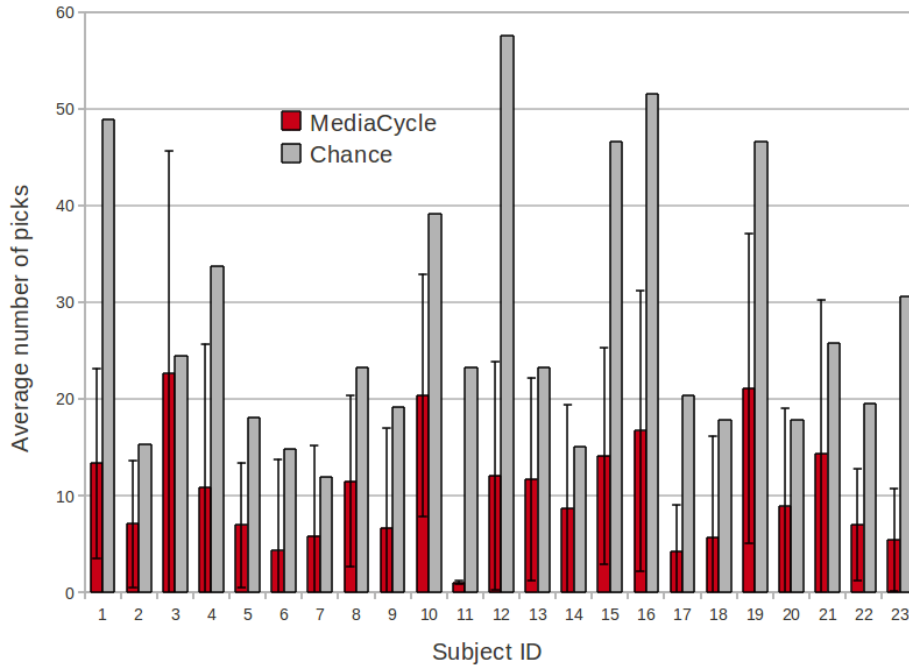

Figure 4.3: Average number of picks needed to find one laughter from the same speaker: similarity algorithm (MediaCycle) against chance (light gray).

These measures, although they are not pure classification experiments, indicate:

1. That laughter has individual traits, hence that we can recognize laughers by hearing their laugh only. This goes in contradiction with the recent experiment conducted by Sathya et al. [Sathya *et al.* 2013], as they found that listeners could not recognize people through their laughs. However, our findings join general beliefs and conclusions drawn by Bachorowski et al. [Bachorowski *et al.* 2001], who could identify laughers above chance levels.

2. That the proposed timbre features can at least partially encode such individual differences. We must however remain prudent on that side, as we cannot certify that the proposed features are characterizing only timbre: some correlates of rhythm or amplitude can also be encoded in the feature set.

This study was part of broader experiments, including evaluation of audiovisual similarity, and the developed similarity algorithm was integrated in an application aiming at answering to participant's laughs with the most similar laugh in the AVLaughterCycle corpus. More details are available in [Urbain *et al.* 2010b].

4.3 Automatic phonetic transcriptions

Given the importance of laughter phonetic transcriptions for laughter synthesis (see Chapter 5) as well as the possible impact for laughter detection—see the works of Knox et al. [Knox *et al.* 2008] and Wagner et al. [Wagner *et al.* 2013] who used speech-trained phonetic transcriptions to discriminate between laughter and speech, or Pammi et al. [Pammi *et al.* 2013] who used ALISP models for the same task)—and characterization (discriminating between different types of laughs based on their phonetic contents), it seemed interesting to us to develop a method for automatically transcribing laughs.

To the best of our knowledge, only one work could possibly be related to this task: the automatic classification of calls performed by Tanaka and Campbell [Tanaka & Campbell 2011]. It must nevertheless be noted that they used pre-segmented calls from one participant only, and distinguished between only four broad categories (nasal, ingressive, chuckles or vocal). Automatic laughter phonetic transcription in the way we will present it here is thus 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 [Toledano *et al.* 2003]. 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 speaker independent speech recognition, where all the decisions are taken using only the acoustic signal without any a priori knowledge on the speaker identity. As HMMs are also widely used in speech recognition, it is this technique we investigated to produce automatic laughter phonetic transcriptions. It might be useful to recall that a brief introduction to HMMs is given in Appendix B. The method described in this section has served to train laughter synthesis (see Section 5.4) and has been presented in [Urbain *et al.* 2013a].

4.3.1 Hidden Markov Models for automatic laughter phonetic transcriptions

The implementation of our HMM-based laughter transcription was made with the help of the HMM Toolkit (HTK) [Young & Young 1994]. One HMM was built for each phone in the database, relying on the phonetic transcriptions presented in Section 3.2. 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 32 ms frames (512 samples at 16 kHz) with a 10 ms shift. For each frame, 70 acoustic features were extracted:

- Spectral centroid, spectral spread, spectral variation, four values of spectral flatness, spectral flux, spectral decrease [Peeters 2004].
- 13 MFCCs (including $MFCC0$), their first and second derivatives.
- ZCR, RMS energy and loudness [Peeters 2004].
- Chirp group delay [Drugman *et al.* 2011] and four values for HNRs [Drugman *et al.* 2013].
- Twelve chroma features [Ellis & Poliner 2007].
- f_0 computed with the SRH method [Drugman & Alwan 2011], 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 (196 different phonetic labels in the original transcriptions). 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 [Urbain *et al.* 2013b] (we will come back to this in Chapter 5), we decided to group acoustically close phones in broader phonetic clusters. Several combinations have been experimented. The grouping illustrated in Figure 4.4 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). As explained in Section 3.2, some labels introduced to characterize laughter sounds that are not covered by the International Phonetic Alphabet (IPA) [International Phonetic Association 1999] 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: n̊. Clicks and glottal stops were discarded from our models as they are very short and subtle phones—which makes them very hard to detect—but in our opinion do not provide critical information for applications that can be built on automatic laughter phonetic transcription (e.g., clustering, laughter synthesis, etc.)[12]. The proposed grouping reduced the number of phonetic labels from 196 to possibly 22 (three consonant classes, four vowel classes, cackle, grunt, nareal fricative and silence for

[12]It must however be noted that even though glottal stops are not the most prominent laughter sounds, they relate to the degree to which syllables are broken up, which could be an interesting characteristic for some applications.

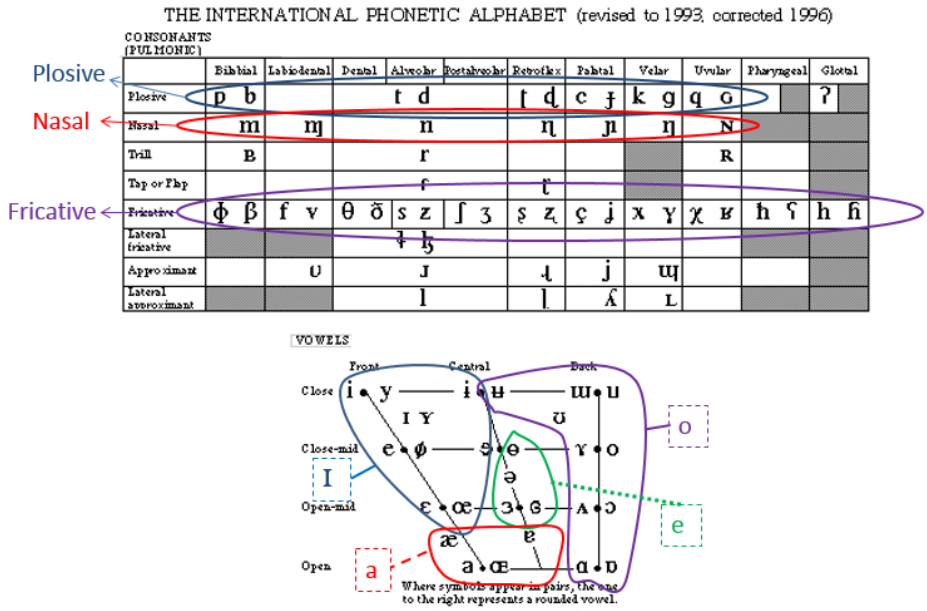

Figure 4.4: Grouping of phones to build consistent phonetic clusters. The original IPA chart can be found in [International Phonetic Association 1999].

exhalation and inhalation parts) and actually 17 as five[13] labels have not been used in inhalation parts. The size of the phonetic clusters used in our experiments is given in Table 4.2.

HTK provides control over several parameters to design HMMs. It goes beyond the scope of this dissertation 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 three states and transitions between all theses states are allowed ("ergodic HMMs").
- The emission probabilities of each state are modeled with ten Gaussian distributions.
- To avoid excessive insertions, the Word insertion Penalty (WP)[14] was set to -20.

[13] Namely "a", "o", nasal, cackle and grunt.

[14] The WP is the cost in likelihood for adding a new "word"—in our case, a new phone—in the transcription. Setting a WP enables to favor longer "words" in comparison to a succession of short words who could have higher local emission probabilities.

Table 4.2: 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	n̊	839
i	fricative	774
e	nasal	717
e	cackle	704
e	plosive	286
e	o	256
i	e	219
i	n̊	166
e	grunt	156
i	ɪ	153
i	plosive	43
i	silence	9

- The grammar consisted in bigram modeling of the succession of phones.
- the Language Factor (LF)[15] was set to 2.

An example of the obtained phonetic transcriptions is shown in Figure 4.5.

4.3.2 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 [Young *et al.* 2006] and provides the number of (see Figure 4.6):

- 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.

[15] The LF represents the weight of the grammar compared to the emission probabilities in computing the likelihood of a transcription sequence.

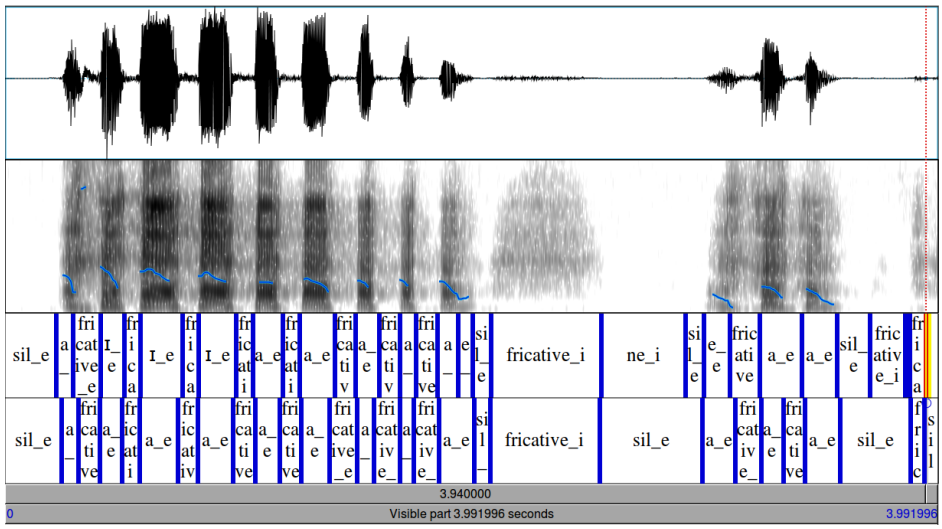

Figure 4.5: Example of automatic phonetic laughter transcription. From top to bottom: 1) waveform; 2) spectrogram; 3) automatic (HTK) transcription; 4) reference transcription. In the phonetic transcriptions, the _e and _i suffixes indicate exhalation and inhalation phases, respectively.

- 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.

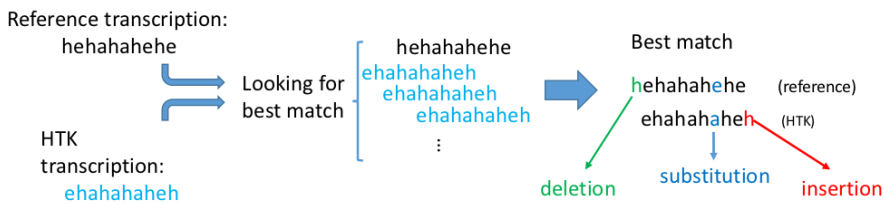

Figure 4.6: Basis of the recognition measures output by HTK: insertions, substitutions, deletions and hits. Hits are not explicitly represented here, they concern all the matching phones (in black).

Two global measures are also provided [Young *et al.* 2006]:

- the Percentage Correct (PCo):

$$PCo = \frac{N - D - S}{N} * 100 \qquad [\%] \tag{4.9}$$

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

- the Percentage Accuracy (PA):

$$PA = \frac{N - D - S - I}{N} * 100 \qquad [\%] \tag{4.10}$$

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 [Räsänen *et al.* 2009]. Räsänen et al. proposed methods that define a search region around each segmentation boundary of the reference transcription. As illustrated in Figure 4.7, 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 [Räsänen *et al.* 2009]:

- Hit Rate (HR), representing the proportion of actual boundaries that have been retrieved (hence it corresponds to what we have previously called *Recall*):

$$HR = \frac{N_{hit}}{N_{ref}} * 100 \qquad [\%] \tag{4.11}$$

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

$$OS = \frac{N_{det} - N_{ref}}{N_{ref}} * 100 \qquad [\%] \tag{4.12}$$

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

$$Pr = \frac{N_{hit}}{N_{det}} * 100 \qquad [\%] \tag{4.13}$$

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

$$R_{dist} = 1 - \frac{\sqrt{(100 - HR)^2 - OS^2} + |\frac{-OS + HR - 100}{\sqrt{2}}|}{200} \tag{4.14}$$

These measures were computed for our automatic transcription method, using search regions of 40 ms ($\Delta = 20$ ms) for the segmentation measures. Table 4.3 gathers the HTK and segmentation measures for automatic transcriptions obtained with

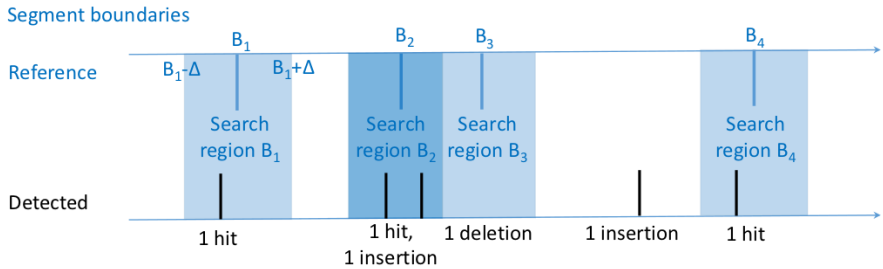

Figure 4.7: Basis of the segmentation measures: hits, inserted boundaries and deleted boundaries.

different values of WP and 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 4.3: Measures of automatic transcription performance, for different values of Word insertion Penalty (WP) and Language Factor (LF)

HTK parameters		HTK measures		Segmentation measures				
WP	LF	PCo	PA	HR	OS	PR	F_1^{score}	R_{dist}
-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 the laughter synthesizer, as will be presented in Chapter 5.

We will finish this section with a short discussion on the evaluation of phonetic transcriptions. The measures presented above (accuracy, etc.) can be considered as figures to optimize. As for classification performance, choices must nevertheless be made on which measure to focus on: accuracy of a transcription is one option, but one must decide whether insertions are taken into account (Percentage Accuracy) or not (Percentage Correct), as well as decide whether only the transcription is evaluated or segment boundaries are also considered. Trade-offs can be looked for using a range of measures, as we have done in this section. Nevertheless, depending on the application, different trade-offs and evaluation measures can be imagined. For

instance, two applications that are considered here will put different constraints on the transcriptions:

- If phonetic transcriptions are used to train the laughter synthesizer, the most important will be the accuracy of the transcriptions. The boundaries between the phones only have a secondary impact, as usually phone boundaries are re-estimated within the training process of the synthesis voice.

- If phonetic transcriptions are used to cluster or classify laughs, boundary locations could have more influence to compute features like the proportion of inhalation parts, the syllabic rhythm, etc. Different errors could also receive different weights, for example one can tolerate confusion between two vowels but not between inhalation and exhalation phases or errors related to more particular sounds like snorts or grunts which are really characteristic of certain laughs.

As we can see, the objectives and architecture of the application will influence what is considered as optimal automatic transcriptions. The measures presented above might not tell everything, and the best can be to evaluate several configurations of automatic transcriptions within the application to elect the one achieving the highest performance. In our case, the main objective of automatic phonetic transcriptions was to train laughter synthesis. We have relied on performance measures as well as visual comparison between the manual and automatic transcriptions to move towards optimal settings. These transcriptions will be evaluated in Chapter 5, but as we will not compare several settings, we cannot guarantee that these transcriptions are the best for laughter synthesis.

4.4 Predicting arousal curves from acoustic data

As explained in Chapter 3, arousal is an important laughter dimension, frequently and naturally used to describe laughs. In consequence, it appeared as an important feature to drive laughter synthesis. It is also a convenient layer in interactive systems to separate the processes of deciding to laugh (with a target intensity), which is independent from the laughter synthesis voice and style, and synthesizing the corresponding laugh, which obviously depends on the modeled individual traits.

Instantaneous arousal seems to us both convenient to use (it is easy to draw or describe an arousal signal) and highly-correlated with the choice of phones used (for instance low arousal laughs are related to closed-mouth nasal sounds, while higher arousal examples include open vowels [Ruch & Ekman 2001, Ruch *et al.* 2013, Niewiadomski *et al.* 2012], while it is suggested in [Edmonson 1987] that consonants are more glottalized at low intensity). In consequence, we investigated how to estimate laughter arousal signals. As for phonetic transcriptions, we aimed here to develop speaker-independent models, that can be included in applications used by

unknown users (as the ones we will see in Chapter 6) rather than models that are tailored to a particular laugher.

Instantaneous arousal, at the frame level, will be referred to as *(per-frame) arousal signal* in the next paragraphs, in contrast to *per-laugh* or *overall arousal* (presented in Section 3.3.1). As explained in Section 3.4, the per-frame arousal signal of 49 laughs (among which 19 from subject #6, who is the laugher we have modeled for laughter synthesis) were manually annotated by one labeler. This makes a total of 27693 labeled frames. An MLP was trained to predict the per-frame arousal signal from acoustic features, using the Weka software [Hall *et al.* 2009]. The acoustic signal was segmented in 32 ms frames (512 samples at 16 kHz) with a 10 ms overlap. For each frame, 82 acoustic features—already used in previous works related to cough analysis [Drugman *et al.* 2013] and already presented in Section 3.3.1—were extracted:

- Spectral centroid, spectral spread, spectral variation, four values of spectral flatness, spectral flux, spectral decrease [Peeters 2004].
- 13 Mel-Frequency Cepstral Coefficients (MFCCs) (including $MFCC0$), their first and second derivatives.
- ZCR, Root Mean Square (RMS) energy and loudness [Peeters 2004].
- Chirp group delay [Drugman *et al.* 2011] and four values for Harmonic to Noise Ratios (HNRs) [Drugman *et al.* 2013].
- Twelve chroma features [Ellis & Poliner 2007].
- f_0 computed with the SRH method [Drugman & Alwan 2011], as well as the value of the maximum SRH peak.
- The four values provided by the Snack [Sjölander 2004] ESPS pitch estimation algorithm—implementing the RAPT method [Talkin 1995]—namely the estimated pitch, probability of voicing, local RMS measurement, and the peak normalized cross-correlation.
- the frequency and bandwidth of the first four formants, computed with Snack.

The MLP was evaluated with a leave-one-subject-out process. Good matching between the predicted and reference curves could be observed. An example of reference and computed arousal signals is given in Figure 4.8 and a histogram of the reference and predicted per-frame arousal values is shown in Figure 4.9. The average absolute error was 0.65 (std: 0.68). Feature selection was performed under Weka, using the correlation-based feature selection algorithm presented in [Hall 1998]. The best features appear to be the MFCCs and the spectral flatness values. Results with this subset of 17 features were close to the performance achieved using the full training set, as the average absolute error was .71. For the work presented in this dissertation, the most important is to have per-frame arousal curves as accurate as possible. It was thus decided to use the full feature

set (82 features) to estimate the per-frame arousal. Computed arousal curves for all the laughs of the AVLaughterCycle database are available on the author's page (http://www.tcts.fpms.ac.be/~urbain/arousal_driven_synthesis).

To further evaluate the quality of the obtained per-frame arousal signals, the per-laugh arousal of each laugh was predicted from its estimated per-frame arousal signal: a second MLP was trained on functionals (mean, standard deviation, minimum, maximum) of the per-frame arousal signal to predict the per-laugh arousal. A leave-one-subject-out method was used for evaluation, using only data from the 21 subjects not involved in training the per-frame MLP: the data from 20 participants were used to train a per-laugh MLP and to predict the per-laugh arousal values of the laughs of the remaining subject. The process was repeated 21 times in order to obtain per-laugh arousal predictions for all the subjects. The correlation between the reference and predicted per-laugh arousals was found to be over .7 for 19 out of the 21 subjects not involved in training the per-frame MLP. A histogram of the reference and predicted per-laugh arousal values for the 820 laughs from the 21 considered subjects is displayed on Figure 4.10, showing good correspondence between the manual and computed values.

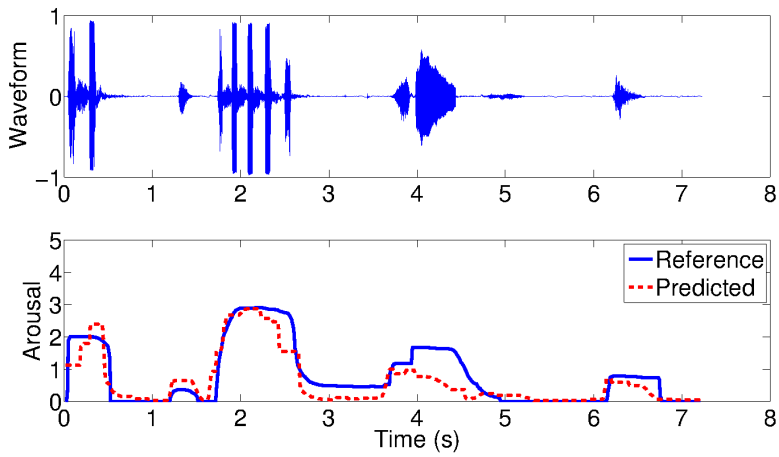

Figure 4.8: Laughter waveform (top) and its reference and computed per-frame arousal signals (bottom).

4.5 Summary and perspectives

In this chapter we have reviewed works on automatic laughter analysis. We have seen that, as expected, the performance of laughter detection in continuous streams is slightly lower than the classification performance, but still satisfying (error rates around 10% depending on the methods and databases). We have also described the

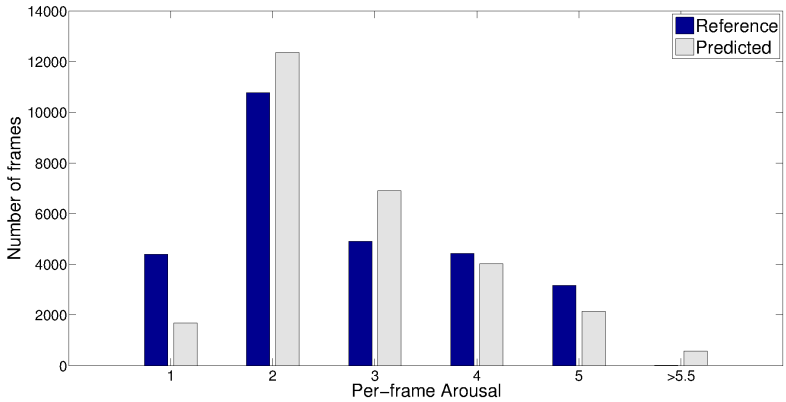

Figure 4.9: Histogram of the reference and predicted per-frame arousal values. Bins are one unit wide, except the rightmost one.

few works that focused on the discrimination between different laughs. Classification largely above chance levels was reported to identify the laugher or her/his gender or the gender and origin of the conversational partner. The identification of portrayed laughter emotions has also been mentioned and discussed, in line with the most recent experiments which suggest that emotional categories do not exist for spontaneous laughter taken out of its context.

After that, we moved to our own developments in laughter characterization, with the assumption that the input is already segmented into laughter episodes. In a first study we could show that individual laughter traits are encoded in spectral features and that similarity measures of those spectral features can help grouping laughs from the same speaker. Finally, we have presented methods to automatically estimate the two layers of annotations introduced in the previous chapter: phonetic transcriptions and laughter arousal (both instantaneous and overall). The performance figures reported in this chapter are however a bit difficult to evaluate given the absence of reference points, as these developments are totally innovative, but we will see in the next chapter that the designed methods can be used to produce state-of-the-art laughter synthesis.

An important direction for future works is to move towards real-time laughter detection and characterization. High reactivity is needed in interactive applications (such as the ones that will be presented in Chapter 6) and this implies spotting laughter as soon as it begins. Some of the methods for laughter detection work on limited time windows (e.g., 1.2 s in Scherer et al.'s audiovisual detection [Scherer *et al.* 2012]) and could be applied to real-time detection. Results are however expected to slightly decrease if the size of the windows is reduced (e.g., if detection is desired within 200 ms). For our laughter characterization algorithms, we could also explore real-time implementations. Performance for phonetic transcriptions is expected to decrease, as

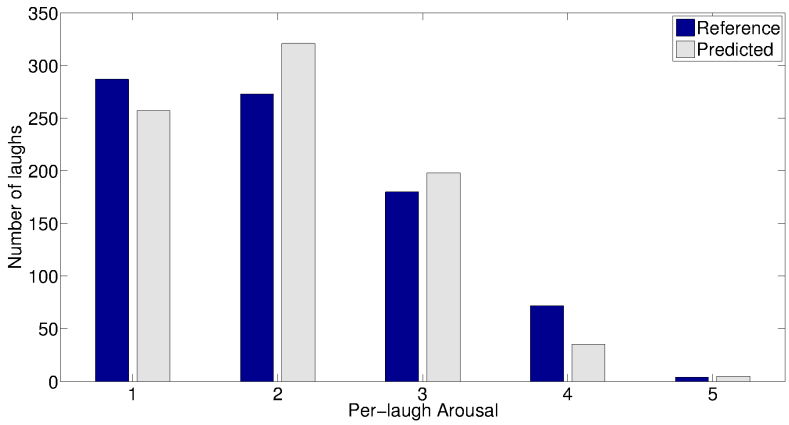

Figure 4.10: Histogram of the reference and predicted per-laugh arousal values. Bins are one unit wide.

for the moment the most likely sequence of phones is evaluated on the whole laughter episode. Instantaneous arousal can already be estimated in real-time as the proposed method is operating on short time frames. A real-time implementation has indeed been integrated in the "Laugh When You're Winning" application (see Chapter 6).

Other areas of improvement concern the feature sets. As we have seen, most of the times usual speech features are extracted (MFCCs, PLP coefficients, f_0, voicing rate, formant positions, etc.). Several researchers have introduced features that are more targeted to laughter characteristics: modulation spectrum, nasality (e.g., [Krikke & Truong 2013]), breathiness (e.g., [Campbell 2007]) or the estimations of fundamental frequency and strength of excitation proposed by Sudheer et al. [Sudheer *et al.* 2009]. These features could be included in our algorithms to assess whether they bring additional information for phonetic transcriptions or estimating arousal. Furthermore, Sathya et al. [Sathya *et al.* 2013] affirmed that the slope of decreasing of the fundamental frequency within a call is proportional to the arousal of the laugh. It would be interesting to investigate this parameter when computing laughter arousal. Contextual features (features computed on neighboring frames or at least derivatives of all the features) could also be included in the arousal estimation algorithm. This would introduce delays in the estimations (hence go against real-time), but could be useful in some offline applications where the quality of the estimation is the only objective and nor the processing time, nor its delay (or latency) do matter a lot. For laughter detection algorithms also, a thorough comparison of all the proposed features (ideally on several corpora) would be useful to assess the importance of each feature, including those that are more closely related to laughter particularities.

As already explained in Section 4.3, laughter phonetic transcriptions could also be used as a feature for laughter detection in continuous streams. Several researchers

have relied on phonetic transcriptions produced by speech recognizers to discriminate between laughter and other audio events. These methods could be extended with the integration of our algorithm focusing on laughter phones.

Finally, speaker adaptation could be explored. Speech recognition is known to perform better when the models are trained on the speaker's voice. Commercial systems frequently propose a training stage to their users, to adapt the speaker-independent models towards the target speaker. Here, laugher-independent models have been developed but a similar adaptation stage could be investigated for applications where the user is known. For estimating arousal, speaker-dependent models could also improve the estimations, as arousal is probably encoded differently by individuals and it should be remembered that people who rated overall arousal explained that they scored arousal relatively to the previous laughs they had observed for a particular laugher.

Chapter 5

Acoustic Laughter Synthesis

Contents

This Chapter addresses acoustic laughter synthesis. First, Section 5.1 will present the state-of-the-art in the field, which is rather scarce. The following sections will

focus on our developments of acoustic laughter synthesis with Hidden Markov Models (HMMs). Section 5.2 is related to the base synthesis scheme, using similar processes as for speech synthesis. Section 5.3 compares different vocoders in their capabilities for laughter synthesis. Section 5.4 explores the possibility of building laughter voices fully automatically (no human intervention needed after data recording, in other words no need for human phonetic transcriptions). Section 5.5 addresses the generation problem, which aims at producing phonetic transcriptions for synthesis from higher-level instructions, as is done in text-to-speech, where text is first converted in phonetic transcriptions before being synthesized. Finally, a summary of this chapter and future works related to HMM-based laughter synthesis are presented in Section 5.6.

5.1 State-of-the-art

As stated by Sundaram and Narayanan [Sundaram & Narayanan 2007], building a good model for laughter synthesis is very complex since:

- The model should be able to generate a broad range of laughter episodes, varying in durations or sounds, as people do in real-life.
- The model needs to produce human-like variations of characteristic parameters, inside a laughter bout. If acoustic features are kept constant, the laughter episode does not seem natural and is rated as less positive [Kipper & Todt 2001, Kipper & Todt 2003, Lasarcyk & Trouvain 2007] (see Section 3.1.6.4).
- The model must remain user-friendly. Laughter must be synthesized providing simple information.

In this section we will present the few reported works on acoustic laughter synthesis. Attempts to synthesize laughs have so far concentrated on voiced laughs. The first efforts related to laughter synthesis are Kipper and Todt's modifications of the pitch of human laughs through Linear Prediction (LP) analysis and re-synthesis [Kipper & Todt 2001, Kipper & Todt 2003] (see Section 3.1.6.4). This is rather modification than actual synthesis of laughter, but it can be seen as a pioneering work to create new laughter signals. The following sections describe recent attempts to (copy-)synthesize[1] laughter from descriptions of its underlying units.

5.1.1 Trouvain and Schröder's diphone concatenation

Trouvain and Schröder [Trouvain & Schröder 2004] conducted a pilot study to investigate how and when laughter should be inserted in synthetic speech, a topic

[1] Copy-synthesis consists in extracting features from a human sample and synthesize a laugh directly from these features. It is used to assess whether the proposed model for speech production (e.g., source-filter) enables to synthesize high-quality utterances, as the features fed into the model are assumed to be perfect. In actual synthesis the trajectories of the features are predicted from a separate modeling stage.

that is actually not covered in this Thesis. To achieve that, they synthesized three versions of a female two-syllable laugh with the MARY diphone speech synthesizer [Schröder & Trouvain 2003]. The duration and f_0 patterns of natural laughter were used to query the diphone database. The versions differed by the vocal quality of the diphones that were selected: soft, modal or loud. These laughs were inserted in short synthesized dialogs. Two natural laughs from the same female voice were also included in the study for comparison purposes, as well as one natural laugh from another female speaker, each with different arousals.

Fourteen naive listeners were asked to listen to the dialogs containing the different laughs and, after each dialog, to indicate how much the speakers liked each other and how well the laugh fitted in the dialog on a 6-point scale. Results indicated that the arousal of the inserted laugh had an effect on the ratings, and that the synthesized laughs could not yield as positive effects as adequate human laughs. The best synthesized laugh (the modal one) obtained a score of 2.9 out of 6 on the appropriateness scale. The moderate-arousal human laughs reached scores of 4.7 (laugh uttered by the speaker) and 5.4 (laugh uttered by a different person than the speaker inserted in place of the speaker's laugh).

The main objectives of this pilot study were not to develop laughter synthesis and in consequence no particular effort was made to improve it or to synthesize a larger amount of laughs. However, the technique of using diphone concatenation, even trained on speech, is interesting and has been further investigated by other research teams, as will appear in the following sections.

5.1.2 Lasarcyk and Trouvain's articulatory system

Lasarcyk and Trouvain [Lasarcyk & Trouvain 2007] compared two systems for generating voiced laughs, inspired by works done in speech synthesis. They relied on the structure described in Section 3.1.1, which states that a laugh is composed of an onset, a main part, a pause and an offset.

The first synthesis system is a 3D simulation of the airflow in the vocal tract, yielding a source-filter decomposition of the laughter-production process. This allows for accurate control of the parameters, both on the source side and on the filter side. This system also allows to produce breathing sounds often encountered in laughs, for example the initial exhalation and final inhalation.

The second system is the diphone speech synthesizer MARY already used by Trouvain and Schröder. It was built on a regular speech database (laughs excluded). The production of sounds is thus limited to the diphones inside the database, and it was not possible to render breathing sounds. Only the main part of laughs could be synthesized with this second system.

The input parameters for generating a laughter episode were the duration of each part (onset, calls, pause, offset), the intensity contour of the episode, its fundamental frequency (f_0) contour and the vowel qualities of the calls. These parameters were copied from an actual human laugh. Three synthetic versions of the human laugh

were created. The first two, versions V and S, were obtained with the articulatory model. Version V contained a lot of variations in glottal gestures (inspired by the human laugh) in all parts of the laugh (onset, main, offset). Version S was a more stereotyped version, as identical glottal gestures were applied in all calls of the main part. Finally, version D was obtained through diphone concatenation with the f_0 and duration patterns of the human calls as targets.

A perceptual evaluation was conducted to evaluate the naturalness of the synthesized laughs. Two experiments were conducted. In the first one, human conversations containing laughter were presented to 14 naive listeners. The laugh in the dialog was either the actual human laugh or a synthesized version of it. Conversations were presented via loudspeakers and the listeners were asked to rate the naturalness of the dialog on a 4-point scale labeled "natural", "less natural", "rather unnatural" and "unnatural". Lasarcyk and Trouvain did not report the overall scores received by each version of the dialog but their average rank. The dialogs with synthesized laughs were not judged much more unnatural than the human ones, except for the diphone concatenation (version D), with the full articulatory system (version V) yielding the best performance.

In the second experiment, only laughs (human and synthetic) were presented to the listeners, and human laughs were evaluated significantly more natural than synthesized ones. Again, version V performed better than the others. This reinforces the conviction that synthesized laughs must present variations in the acoustic characteristics to sound natural and that breathy sounds play an integral role in laughter.

As discussed by the authors, the good performance achieved by placing the laughs inside a conversation might have been due to the fact that the synthesized laughter is the laughter of a secondary speaker, that we perceive in the background of the main speaker's laughter. Furthermore, the listeners were asked to judge the naturalness of the whole dialog. The listeners' attention might not have been focused on this background sound, not noticing the possibly poor quality of the synthesis. This masking effect is removed when only the synthesized laughter is presented. Another limitation of the proposed method is that it has been used only to copy-synthesize one voiced laugh. To fully assess the performance of the method, it would be needed to use it to produce a wider range of laughs. Nevertheless, it remains an interesting approach to laughter synthesis, yielding acceptable laughter bouts including breathing sounds that are often discarded from synthesis.

5.1.3 Sundaram and Narayanan's mass-spring analogy

Sundaram and Narayanan [Sundaram & Narayanan 2007] tackled the laughter synthesis problem under a different angle. Noticing that most of voiced bouts exhibit an oscillatory behavior (as described in Section 3.1.1), they compared the envelope of the laughter waveform to the evolution of the position of a mass attached to the end of a spring. The evolution of the position x of such a mass is given by the following

equation:

$$x(t) = Ae^{-Bt}cos\sqrt{k/m}t, \tag{5.1}$$

where B is the damping factor (symbolizing energy dissipation), m is the mass, k is the spring constant and A a scaling factor.

If we keep only the positive values of this temporal evolution (setting the negative values to zero) and use the resulting signal as the energy envelope of a wave, we obtain the graph showed in Figure 5.1, where the hypothetical oscillatory signal is displayed together with its envelope, corresponding to the positive part of a mass-spring trajectory. The parallel between this waveform and a stereotypical voiced laughter bout is striking. The approximation of a real voiced laughter episode is even better if the parameters of the mass-spring system are allowed to vary over time: *k(t)* or *m(t)* to modify the periodicity of the envelope, *B(t)* to change the damping and *A(t)* to reach the desired amplitude. Figure 5.2 shows the approximation of the envelope of a real laughter episode Sundaram and Narayanan were able to obtain with the mass-spring analogy, modifying the parameters of the mass-spring system to match the waveform.

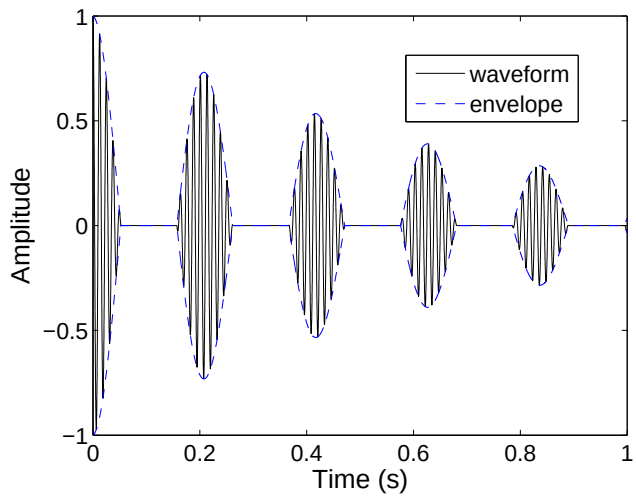

Figure 5.1: Signal modulated by the position of a mass-spring system.

This method enables to model the envelope (i.e., the duration and peak amplitude of each call) of the laughter bouts. One then needs to synthesize the calls. Sundaram and Narayanan restricted themselves to synthesize vowel-like calls via LP. Different vowel-like sounds can be obtained by changing the coefficients of the all-pole filter of the LP model. The LP model was built on vowels from regular speech. The f_0 trajectory of the laugh to synthesize must also be specified (through a Graphical User Interface).

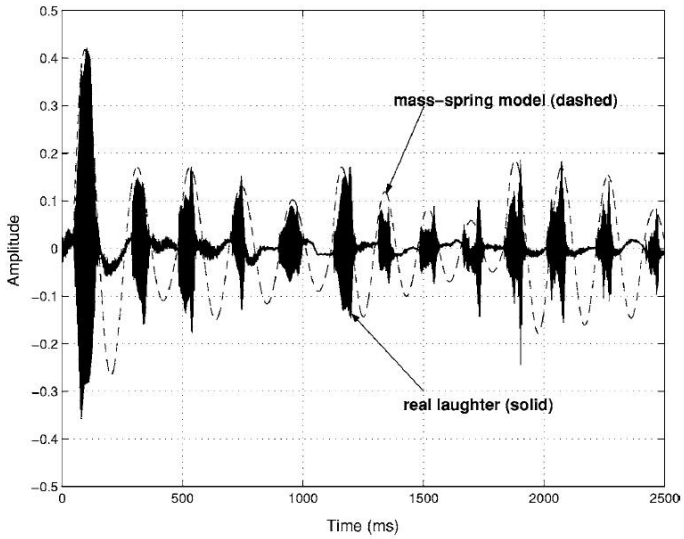

Figure 5.2: A mass-spring model trajectory superimposed on a real laughter bout. Source: Sundaram and Narayanan [Sundaram & Narayanan 2007].

Perceptive tests were conducted to evaluate the naturalness and acceptability of 17 synthesized laughs (with parameters inspired from real episodes), presented randomly with eight acted laughs. A 5-point scale was used. The 28 listeners massively judged the synthesized laughter bouts as non-natural (average naturalness score of 1.71 out of 5) and non-acceptable, while the real utterances received good—but not perfect—scores (average naturalness of 4.28). Possible reasons for the non-natural perception of the synthesis obtained with this approach can be the limited complexity of the features (for example simple vowel-sounds were synthesized) or the absence of breathing sounds. We would also like to add that, while the concepts underlying this synthesis method are rather simple, specifying the mass-spring parameters for modeling a given laughter episode is not straightforward. The authors also insist on the fact that the two proposed synthesis steps can be separated and that one may use the envelope model together with any other "speech" synthesizer (i.e., it is not restricted to LP synthesis trained on speech vowels) to synthesize the calls. Depending on the synthesizer possibilities, snort- or grunt-like calls could also be inserted.

5.1.4 Beller's unit selection and parametric modification

Beller [Beller 2009] proposes an original approach to laughter synthesis, as voiced laughter is synthesized from a neutral speech sentence. The utterance was automatically transcribed in phonemes, and three phones were selected. The first selected

phoneme was the occlusive exhibiting the maximum positive loudness slope, to form the *attack* of the laugh. The second selected phoneme was another occlusive (as the automatic transcription methods mislabeled laughter intercalls as occlusives) with a minimal absolute loudness slope, to form the *inter-call* intervals. Finally, the central vowel with minimal f_0 slope and minimal voicing coefficient was selected. The initial laughter syllable was formed by the *attack* occlusive and the vowel, while the other syllables consisted in a concatenation of the *inter-call* occlusive and the vowel.

The syllables were then individually modified to respect the instructed laughter prosody parameters: syllables could be time-stretched, transposed, scaled and frequency warped to control respectively the rhythm, f_0 pattern, energy pattern and the degree of jaw opening during the laugh.

No evaluation of the obtained laughs has been performed. By listening to the released examples, we found that the synthesized sequences definitely sound like laughter, but in our opinion are lacking naturalness due to the repetition of the same syllable (which is only slightly modified).

5.1.5 Sathya et al.'s modification of excitation characteristics

Sathya et al. [Sathya *et al.* 2013] synthesized voiced laughter bouts by controlling several excitation parameters of laughter vowels: pitch period, strength of excitation and amount of frication. After analyzing these features on a range of human laughs, Sathya et al. concluded that the pitch contour and the strength of excitation of laughter calls can be approximated by quadratic functions, while the amount of frication tends to decrease within and across calls.

To synthesize a call, their method uses a segment of natural or synthetic vowel corresponding to the desired duration of the call as input. The fundamental frequency and strength of excitation are extracted with the method proposed in [Sudheer *et al.* 2009] (see Section 4.1.2.2). Pitch-synchronous Linear Prediction Analysis is then performed to obtain the LP coefficients (approximating the filtering effect of the vocal tract) and the LP residues (representing the excitation signal). These parameters are available for each epoch (instant of significant excitation of the vocal tract system), hence for each pitch period. The pitch patterns of the synthesized calls are approximated by a quadratic curve, determined by the desired minimum and maximum pitch periods within each call. The quadratic approximation gives the positions of the epochs in the synthetic signal. The LP parameters at these new locations are copied from the nearest epochs of the original vowel. The LP residuals are then modified according to the desired (synthetic) pattern of the strength of excitation, which is also modeled with a quadratic approximation. Frication (or breathiness) is then added by band limited Gaussian noise to the signal. The noise signal consists of white Gaussian noise of the length of the call, scaled to correspond to a desired fraction (typically 80 to 120%) of the energy of the LP residuals within the call and then filtered by a band pass filter with a resonance at 2500 Hz and a bandwidth of 500 Hz. Finally, the noise signal is weighted so that the amount of breathiness is

linearly decreasing within the call. The noise signal is added to the residual signal to form the excitation signal filtered with the LP coefficients to obtain the synthesized call. Low amplitude band limited random noise (about 1000 times smaller than the call energy) is used to fill the intercalls.

To synthesize a laugh, the following parameters must be specified: number of bouts, number of calls, intensity ratio from first to last call, as well as the parameters required to synthesize calls (duration of each call and intercall, minimum and maximum pitch period of each call and amount of frication at the beginning of each call). Although technically all parameters can be specified independently, the authors noticed that there are some interdependencies to take into account if one wants to avoid poor quality synthesized laughs. For instance, calls are longer when they are fewer, long bouts tend to have higher f_0 values and intercall duration depends on the position within the bout.

Sixty laughs were synthesized with the proposed method. Each laugh uses one of the following six input vowels: natural /i/ and /a/ from a female voice, natural /i/ and /a/ from a male voice, synthetic /i/ and /a/ from a male voice. For each of these six input vowels, ten different laughs, containing one bout of four to eight calls, are synthesized with different combinations of some synthesis parameters (ratio between first and last call, ratio between the duration of calls and intercalls, breathiness). These laughs were played in random order to 20 subjects who were asked to rate their quality of synthesis and acceptability on a 5-point scale ranging from very poor to excellent.

The synthesized laughs reached relatively high scores in perceived quality and acceptability, with values around 3 on a scale ranging from 1 to 5. The variations of the synthesis parameters had little influence on the scores. However, laughs synthesized from an original synthetic vowel received significantly lower scores for the quality of synthesis (2.36 and 2.73 out of 5) than laughs using human vowels as input (which all have scores over 3.15). It must also be noted that no human laugh was included in the evaluation, which might have had a positive influence on the scores obtained by the synthesized laughs (as there is no "perfect" reference to compare with in the evaluation). To conclude, laughs synthesized with Sathya et al.'s method received relatively high scores compared to Sundaram and Narayanan's, even if the results are still far from perfect naturalness, especially when using synthetic vowels as input. The method is currently limited to synthesize single-bout voiced laughs, as it cannot synthesize inhalations nor unvoiced calls.

5.1.6 Cagampan et al.'s diphone concatenation

Cagampan et al. [Cagampan *et al.* 2013] synthesized laughs by concatenating syllables from the Pinoy Laughter 2 database (see Section 2.3.7). Laughs from the database were segmented into syllables with different labels denoting laughter vowels ('ha','he','hi','ho','hu'), grunt- and snort-like syllables as well as laughter onset and offset. These units were then combined to form laughs with four syllables in the

apex, plus possibly an onset and an offset. Different laughter bouts were synthesized by varying the vowel sounds, the time interval between them (through syllable overlap or dynamic time warping to match the syllabic rhythm of a human laughter sound), the amount of fade in and fade out effects for each syllable and the presence (or not) of an onset and an offset. The naturalness of the laughs was evaluated by naive listeners on a scale from 1 to 5. Two different experiments have been conducted.

In the first experiment, only 'ha' syllables were included, possibly with an onset. It revealed that laughs are perceived as more natural when they include an onset segment (the influence of offset was not investigated) and that dynamic time warping is beneficial, as it introduces some variability in the laugh (by copying human patterns).

In the second experiment, all laughs contained an onset, four syllables (with varied vowels) and an offset. The results somehow contradict those of the first experiment, as dynamic time warping led to improved naturalness in few cases only. However, it generally seems that fade in between concatenated syllables has a positive effect (as it smooths transitions between syllables) and that variability introduced by changes in the vowel quality over the laughter bout increase the perceived naturalness. Laughs with solely the syllable 'he' were also better rated than laughs containing only the stereotypical syllable 'ha'. The authors suggest that it can be related to the expectations that listeners have about the stereotypical laughs. However, we think that the results must be taken with caution: the study does not include sufficient variations in vowels to draw conclusions about which vowel is perceived as more natural than the others, as all laughs contained exactly four syllables and are highly-dependent on the example for each vowel that was selected for concatenation. Similarly, we are quite surprised by the overall low naturalness scores received by human laughs (five out of the six humans laughs received average naturalness scores below 3.4 on the 1-5 scale) as well as by the authors' findings that a laugh with an 'ho-hi-ha' apex was rated as the most natural, as such a combination does not seem natural at all, as explained in Chapter 3. There might be some evaluation bias explaining these results, possibly some cultural effects (laughers and listeners were Filipinos) or some misinterpretation of the scale (funniness/pleasantness instead of naturalness).

Nevertheless, some laughs synthesized with this syllable concatenation technique received high naturalness scores (the best obtaining an naturalness score of 3.6) and were perceived as more natural than five of the six human laughs included in the evaluation.

5.1.7 Oh and Wang: LOLOL

A real-time laughing instrument has been developed by Oh and Wang [Oh & Wang 2013b]. Their main objectives were expressivity and control, rather than quality of synthesis or laughter naturalness. They synthesized vowels by formant synthesis (source-filter decomposition). Real-time control was provided over several parameters: rhythm, pitch, falling pitch at the end of the vowel or not, voicedness/harmonicity of the vowels, formant positions (F1-F2 vowel space), inhalation or

exhalation and glottal waveform (to control voice quality and introduce effects of harshness). As the system was mainly developed as an instrument, for performance purposes, it has not been evaluated methodically, but a video demonstration is available on the web[2].

To the best of our knowledge, this is the first paper on real-time laughter synthesis system[3]. As could be expected, the resulting laughs do not sound natural (inhalations are not well synthesized, only vowel sounds can be rendered, etc.) but the expressivity objective is clearly reached and they are able to convey different "meanings" through these laughter-like signals. It must also be noted that the temporal evolution of laughter syllables over an episode is not modeled at all: the system is relying on the capabilities of the performer to control sound parameters (pitch, rhythm, evolution of formants) in an expressive way, which is clearly achieved by the trained performer in the video. It is however also possible to script the evolution of parameters.

5.1.8 Oh and Wang: modulation of speech

To conclude this state-of-the-art section, it seems interesting to us to make a short parenthesis with the recent works from Oh and Wang [Oh & Wang 2013a] to modulate speech and make it sound like speech-laugh, as opposed to all previous attempts on pure laughter. The method takes speech as input and segments it into syllables, based on the energy envelope. Then they provide control over several parameters of the syllables that can be affected by laughter: intensity contour, maximum pitch value, tempo regularity (the degree to which segmented speech syllables are fetched to an isochronous tempo), rhythm (the periodicity of syllables between 4 and 6 Hz). Examples can be listened to on the authors' page[4]. Modifications of speech are clearly audible, although some details are probably missing to make the sentences sound like natural speech-laugh, for example the insertion of aspiration sounds described in Section 3.1.3. Nevertheless this is an interesting first approach to the problem, and it also accounts for the current gain in interest for acoustic laughter synthesis, which has witnessed five additional papers (including ours) in 2013, a similar quantity as in all the preceding years together.

5.1.9 Summary of the state-of-the-art

As we have seen, there have been a few attempts to synthesize laughter and it seems that the topic is currently gaining interest. Laughs synthesized with the described methods are generally lacking naturalness. Several authors have stressed the complexity of producing natural-sounding laughs, as it is a highly variable signal including

[2] http://vimeo.com/58348046, last consulted on February 20, 2014.

[3] Real-time laughter synthesis had already been informally demonstrated by Nicolas D'Alessandro using HandSketch [D'Alessandro & Dutoit 2007] and briefly showcased during concerts (e.g., [D'Alessandro *et al.* 2013]), but had never formed the core of a paper.

[4] https://ccrma.stanford.edu/~jieun5/research/IS2013/show_and_tell/laughter-modulation.mov, last consulted on March 17, 2014

sounds that are usually not modeled, e.g. inhalation sounds. Interestingly, human laughs do not receive a perfect naturalness score either when they are evaluated by naive listeners: the lack of context is probably making some laughs—although totally human and properly recorded—surprising. In the following sections, we will describe our developments of audio laughter synthesis, using Hidden Markov Models (HMMs), which are widely used in speech synthesis but had not yet been experimented for laughter synthesis.

5.2 Hidden Markov Models for acoustic laughter synthesis

5.2.1 Hidden Markov Models implementation scheme

HMM-based speech synthesis roughly uses the same models as those presented in Appendix B. One HMM is trained for each phoneme, to model the evolution of features. Then, for synthesizing an utterance, its phonetic transcription is used to retrieve the corresponding HMMs, put them in a sequence, and generate the sequence of observations that is most likely according to the observation probabilities of each state. As the observation probabilities between two states are modeled independently (in particular between states of different HMMs that are modeled on different pieces of data but can still follow each other for synthesis), first and second order derivatives of the features are usually included in the feature set, in order to obtain smooth synthesized trajectories. Without the introduction of derivatives in the observation probabilities, the most likely sequence of observations would be the sequence of the most likely observation for each state taken independently (hence, the mean vector associated to each state, when considering a Gaussian distribution). The presence of derivatives pushes the system to find the most likely *trajectory* of features, which can deviate from the local optimums.

In HMM-based parametric speech synthesis, the spectrum, fundamental frequency and duration of phonemes are modeled in a unified framework [Yoshimura *et al.* 1999]. Based on the resulting HMM, a maximum-likelihood parameter generation algorithm is used to predict the source/filter features [Tokuda *et al.* 2000], which are then sent to a parametric synthesizer to produce the waveform. HMM modeling being known for its flexibility, its use for the synthesis of non-verbal vocalizations, such as laughter in this case, appeared to be relevant. To develop laughter synthesis we used as a baseline the canvas provided in the demonstration scripts of the HMM-based speech synthesis system (HTS) [Oura 2011]. HTS is a set of functions designed for HMM-based acoustic speech synthesis and provided as a patch to the HMM Toolkit (HTK) [Young & Young 1994].

As an alternative to the feature extraction tools used in the HTS demonstration scripts, namely the Speech Signal Processing Toolkit (SPTK) [SPTK online 2013] for spectrum and Snack for f_0 [Sjölander 2004] (which implements the RAPT method

[Talkin 1995]), we included the Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum (STRAIGHT) tools [Kawahara 2006], which are known in the field of speech processing to provide efficient source and filter representations. Although STRAIGHT provides tools for synthesis and for the extraction of other features than spectrum and f_0, we only used it for extracting spectrum and f_0 in our baseline method.

The traditional excitation used by HMM-based speech synthesizers—denoted as MCEP in Section 5.3—is either pulse train or white noise, during voiced and unvoiced segments respectively. To reduce buzziness in the synthesized waveforms of vocalized segments, we used the Deterministic plus Stochastic Model (DSM) source model, which has been shown to improve the naturalness of synthesized speech [Drugman & Dutoit 2012]. This tool provides an excitation signal which is closer to the actual human excitation signal than the original pulse train. More details on the MCEP, STRAIGHT and DSM methods will be given in Section 5.3.

In addition, the fundamental frequency of the laughs was studied and it was found that a good estimation of f_0 could be obtained by setting the boundaries of the estimation algorithm at 100 and 800 Hz for the considered voices (see below). The number of states per HMM was left to five, as qualitative experiments showed few differences when using three or four states per HMM. We have chosen MFCCs as descriptors of the spectrum and the state probabilities were modeled with one Gaussian distribution per state.

5.2.2 Adaptation of laughter data to HMM-based synthesis

Several post-processing stages were required in order to adapt the AVLC data to HMM-based synthesis. First, a specific voice had to be selected. The 24 subjects of the AVLaughterCycle database laughed with variable (total) durations and acoustic contents. As HMM-based synthesis requires a lot of training examples, we focused on the five subjects who produced most laughs. Preliminary HMM-based synthesis models have been trained for each of these five voices. Informal evaluation revealed that subject #6 provided the best acoustic (laughter) quality. Only three minutes of laughter data are available for subject #6, but he tended to use a limited set of phones (see [Urbain & Dutoit 2011] or Section 3.2.3) and hence there are numerous examples of these phones available for training. In addition, subject #6 is highly expressive on the acoustic side (while other subjects tend to have more silent periods in their laughs).

The second step was to decide how to handle the available narrow phonetic annotations. Indeed, the phonetic transcriptions of the 64 laughs from subject #6 contained 45 different phonetic labels, with only twelve of these appearing at least ten times. As for automatic phonetic transcriptions, it was decided to gather acoustically close phones into broader phonetic clusters, in order to increase the number of training examples available for each label. For example, velar, uvular, pharyngeal and glottal fricatives were grouped in one "fricative" cluster. The phones that were barely present

and could not be merged with acoustically similar phones to form one representative cluster with at least ten occurrences were assigned to an *unknown* set. The resulting phonetic clusters for subject #6 are listed in Table 5.1. The way vowels were grouped is illustrated in Figure 5.3. The grouping is slightly different than the one used for automatic phonetic transcriptions, as here front vowels are divided in three clusters instead of two. We do not think it makes a big difference. The grouping for automatic transcriptions was actually realized after the first attempts for laughter synthesis presented here, and was downsized to two clusters for front vowels after we observed confusions between laughter phones by the HMMs. The grouping for automatic transcriptions has also been used for our latest laughter synthesis experiments, which will be presented in Sections 5.3 and 5.4.

Table 5.1: Phonetic clusters used for HMM-based laughter synthesis, ordered by frequency of occurrence. (Note: Average arousal will be used in Section 5.5).

Respiration part	*Phonetic cluster*	*Occurrences*	*Av. arousal (std)*
Exhalation	fricative	439	3.4 (1.0)
Exhalation	a	327	3.5 (1.0)
Exhalation	silence	296	2.0 (0.9)
Exhalation	ə	84	2.4 (0.8)
Inhalation	fricative	49	1.9 (0.8)
Exhalation	ɛ	40	2.3 (0.6)
Exhalation	o	39	3.9 (0.6)
Exhalation	cackle	35	1.5 (0.5)
Exhalation	ɪ	10	2.3 (0.5)

Finally, the HTS framework enables to build context-dependent HMMs, where the contexts that yield different acoustic realizations of one phoneme are identified with the help of contextual features and decision trees [Zen *et al.* 2007b]. Different emission probabilities can be modeled for the same phoneme HMM and the distribution that is used at each step is determined by the context in which the phoneme lies.

To exploit the capabilities of HTS to model phonetic context, a third post-processing step was conducted: laughter syllable annotations were added to phonetic transcriptions. Syllables generally contain two phones (typically a fricative "h", and a vowel "a") and their syllabic label is the sequence of the involved phonetic classes (hence, a syllable containing a fricative and a vowel is labeled FV).

The basic information that can be provided to HTS for contextual modeling of one phoneme is the labels of the (generally two) preceding and following phonemes. This is immediately transposable to laughter synthesis and our baseline method includes the labels of the preceding and following two phones. However, additional contextual information is frequently used to improve speech synthesis, with contextual features like the position of the phoneme in the syllable, the position of the syllable in the word, etc. (for an example, see [Zen 2006]). Using the phonetic, syllabic and respiration transcriptions, similar contextual features have been computed for our laughs. The

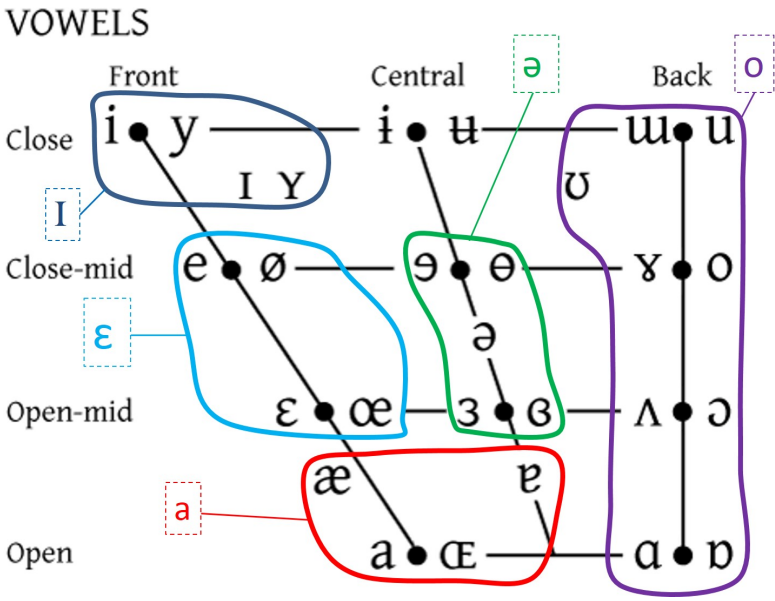

Figure 5.3: Grouping of the vowel phones in phonetic clusters. The original vowel chart can be found in [International Phonetic Association 1999].

following 16 contextual features have been added:

- Position of the current phone in the current syllable (forward and backward counting, two features).
- Number of phones in the previous, current and following syllables (three features).
- Position of the current syllable in the current respiration phase (forward and backward counting, two features).
- Position of the current syllable in the whole laughter episode (forward and backward counting, two features).
- Number of syllables in the previous, current and following respiration phases (three features).
- Position of the current respiration phase in the whole laughter episode (forward and backward counting, two features).
- Total number of syllables in the laughter episode.
- Total number of respiration phases in the laughter episode.

Given all that information, HMMs could be trained with the audio and transcription files from subject #6 of the AVLaughterCycle database, and acoustic laughs could then be synthesized. The outcome of the synthesis has been evaluated, and the results will be presented in the next Section.

5.2.3 Evaluation of HMM-based laughter synthesis

5.2.3.1 Compared Synthesis Methods

In search for the best algorithms to synthesize laughs, six different methods for each laugh were included in the evaluation test for comparison purposes:

- Method H: the original human laugh, unmodified.
- Method CS: the same laugh, re-synthesized through copy-synthesis (i.e., f_0 and spectral parameters are extracted from the laugh and directly used in the source-filter model to reconstruct the laugh), including the STRAIGHT and DSM algorithms.
- Method $S1$: the same laugh, synthesized with the HMM-based synthesis 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 preceding two and following two phones.
- Method $S2$: same as Method $S1$, with an extended context including the contextual information available from syllabic annotation, as explained in Section 5.2.2.
- Method $S3$: same as Method $S2$, with the addition of the STRAIGHT and DSM algorithms.
- Method $S4$: same as Method $S3$, with the duration of each phone estimated from the HMMs by HTS.

Method H 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 CS can be seen as the maximum performance achievable with our HMM-based source-filter models. Method $S1$ is considered as our baseline HMM-based laughter synthesis method, as it is directly available from HTS. Methods $S2$, $S3$ and $S4$ are possible improvements over the baseline method. Our test hypotheses were the following:

- H1: Using full contextual information improves the results (Method $S2$ better than Method $S1$).
- H2: Using STRAIGHT and DSM improves synthesis quality (Method $S3$ better than Method $S2$).

- H3: HTS can model the duration of laughter phones appropriately (Method $S4$ is not worse than method $S3$).

Each of the synthesized laughs was obtained with a leave-one-out method, to ensure that we were not simply able to reproduce learned trajectories. The laugh to synthesize was not included in the training phase of the HMMs.

5.2.3.2 Experimental set-up

Sixty-four laughter episodes were available for subject #6 of the AVLaughterCycle database. Thirty-three of these included phones that were present fewer than eleven times 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 methods presented in Section 5.2.3.1. For the evaluation, laughs were presented to participants in random order and, for each laugh, only one of the methods was randomly selected for each participant. In consequence, all participants were assigned a different set of laughs, varying both in the ordering of the laughs and in the method to evaluate for each laugh.

The evaluation was performed through a web-based application. The first web-page of the test asked participants to provide the following details: their age, gender, 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[5] or e) none of these topics.

Once this information was filled, participants were presented with an instruction page explaining the task, i.e. rating the naturalness of synthesized laughs on a 5-point scale with the following labels: very poor (score 1), poor (2), average (3), good (4) and excellent (5). As some laughs, with some of the methods, were extremely short and/or quiet, participants were also allowed to indicate "I cannot rate the naturalness of this laugh"—which will be referred to as "unknown" naturalness rating in the following sections—instead of providing a naturalness value, but asked to use this option only if they could not hear the laugh. Participants were also explained that they could listen to each sample as many times as they wanted before moving to the next example.

The third page contained eight laughter examples to familiarize participants with the range of synthesis quality that they would later have to rate, with the aim to reduce interpersonal variability during the actual evaluation. Two laughs (each presented with four different methods evaluated in the experiment) were selected to form these examples and were excluded from the evaluation task. In consequence, there were 29 laughs remaining for evaluation. Participants were presented one laugh at a time and asked to rate its naturalness. The test was completed after 29 evaluations.

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

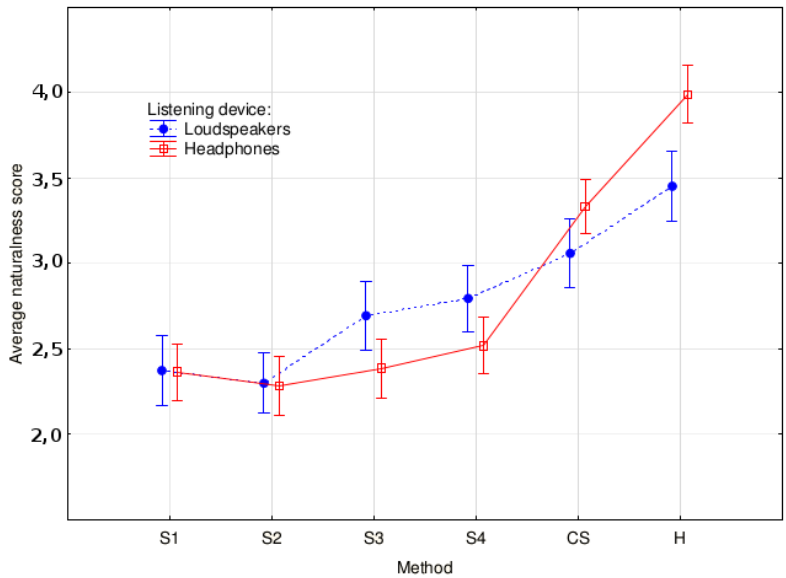

Figure 5.4: Average naturalness scores obtained by each method, depending on the listening device.

5.2.3.3 Results

The focus of the evaluation was to compare synthesis methods to each other. Nevertheless, analyses of variance with two independent factors have been conducted to investigate the influence of gender, of using headphones or not, and of possible experience in the laughter synthesis field, over perceived naturalness. The synthesis method was always one of the two independent factors. The Tukey Honestly Significant Difference (HSD) adjustment has been used to compute all p-values. Generally speaking, no statistically significant differences related to the participants' profiles were found. Although using headphones or not did not yield to statistically significant differences in the evaluation, synthesized laughs tended to be slightly better evaluated when listening to them via loudspeakers (see Figure 5.4). We will hence here only report on the results related to people wearing headphones.

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. Their profiles are summarized in Table 5.2. Out of the 1102 received answers from people using headphones, 53 were "I cannot rate the naturalness of this laugh". Table 5.3 gathers the number of ratings received, the number of "unknown" answers and the average score for each method. As the naturalness score differed for each human laugh, it was decided to compute naturalness scores relatively to the reference human laugh: the relative rating (RR_x) of the naturalness score x is the

difference between the average score received by the corresponding human laugh (H_x) and x:

$$RR_x = H_x - x \tag{5.2}$$

RR_x can be seen as the distance between a synthesized laugh and its human counterpart: the lower RR_x, the more natural the laugh sounds.

Table 5.3 also presents p-values resulting from a univariate analysis of the variance, with the relative naturalness score as dependent variable and the method as explaining factor. Only the p-values between successive synthesis methods (corresponding to our hypotheses) are presented in Table 5.3. The Tukey HSD correction was used. Although the average naturalness scores of our synthesis methods differ (S4 received better naturalness scores than S3, for example), no statistically significant difference is obtained.

Table 5.2: Participant profiles.

Participant category	Loudspeakers		Headphones		Total
	Females	Males	Females	Males	
Non-expert	17	5	13	10	45
Synthesis expert			1	4	5
Laughter researcher	2			2	4
ILHAIRE participant	1	3	3	5	12
Total	20	8	17	21	66

Table 5.3: Received answers for each synthesis method, only including participants using headphones. Note: p-val denote the p-values of the pairwise comparisons between one method and the method from the line above.

Method	*# ratings*	*# unk.*	Av. score (std)	Av. RR score (std)	p-val
S1	176	11	2.4 (1.0)	1.7 (1.2)	-
S2	166	11	2.3 (1.1)	1.7 (1.2)	1
S3	164	12	2.4 (1.1)	1.6 (1.2)	.81
S4	176	2	2.5 (1.1)	1.4 (1.0)	.94
CS	196	8	3.3 (1.2)	0.7 (1.2)	0
H	171	9	4.0 (1.2)	0 (1.0)	0
ALL	1049	53			

5.2.3.4 Discussion

As it has already been found in previous studies, actual human laughs are not rated as perfectly natural by participants: method H has an average score of 4 out of 5 (see Table 5.3). Even more, the perceived naturalness for human laughs is highly variable

from one laugh to the other, as indicated by the large variance. This is why relative ratings have been used. Nevertheless, human laughs received significantly better naturalness scores than copy-synthesized laughs and than our HMM-based synthesis methods. It should however be noticed that being around 1 point below human utterances on a 5-point naturalness scale is comparable to the performance achieved by the best speech synthesis methods (for example, see the results in [Yamagishi *et al.* 2009]).

Regarding our hypotheses about synthesis methods, 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 an even larger laughter database (e.g., the AV-LASYN database which has been recorded recently, see Section 2.3.8). Our second hypothesis, H2, was not verified either: using the STRAIGHT and DSM algorithms did not improve the quality of the synthesized laughs sufficiently to 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 S4 is actually better than Method S3, 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, as the duration can be properly modeled by the synthesizer itself. We will come back to this in Section 5.5.

Among the four synthesis methods, method S4 yields the best results, with an average naturalness score of 2.5 (std: 1.1; minimum: 1.64; maximum: 3.82) out of 5. The obtained average score of 2.5 is clearly better than the 1.71 achieved by Sundaram and Narayanan [Sundaram & Narayanan 2007] and similar to the scores obtained by Cagampan et al. [Cagampan *et al.* 2013] and by Sathya et al. [Sathya *et al.* 2013] when using synthetic vowels. One of the advantages of our method is the range of sounds that can be synthesized, simply from a phonetic transcription: contrarily to previous attempts, we did not limit ourselves to a fixed number of syllables, episodes without inhalation/breathing sounds, solely voiced laughs nor copied acoustic parameters from human laughs.

5.3 Comparison of vocoders in HMM-based laughter synthesis

We have presented our first developments of HMM-based laughter synthesis in the previous section. For this first implementation, we had already integrated the DSM vocoder and results have shown that including the STRAIGHT (for feature extraction) and DSM methods (for creating the excitation signal), only slightly improved the laughter quality compared to the baseline method (the difference does not reach

statistical significance). Other vocoders exist, however. Most of them have been developed for speech synthesis. In this section, we will compare the performance of several vocoders for laughter synthesis. These developments were made in collaboration with researchers from KTH (Stockohlm, Sweden) and Aalto University (Espoo, Finland) and have been published in [Bollepalli *et al.* 2014].

5.3.1 Vocoders

The following vocoders were chosen for comparison:

- Impulse train excited Mel-CEPstrum based vocoder (MCEP) (the vocoder of our baseline experiments).
- Deterministic plus Stochastic Model (DSM) [Drugman & Dutoit 2012], also already used in Section 5.2.
- STRAIGHT [Kawahara *et al.* 1999, Kawahara *et al.* 2001] using mixed excitation.
- GlottHMM vocoder [Raitio *et al.* 2011].

All these vocoders rely on the source-filter model of speech production. The vocoders can differ on the extraction of the spectrum (to build the filter) and on the model for the excitation signal. The vocoders and their parameters are listed in Table 5.4 and described in more detail in the following sections.

Table 5.4: Tested vocoders and their parameters and excitation type.

System	Parameters	Excitation
MCEP	MFCC: 35 + f_0: 1	Impulse + noise
STRAIGHT	MFCC: 35 + f_0: 1 + band aperiodicity: 21	Mixed excitation + noise
DSM	MFCC: 35 + f_0: 1	DSM + noise
GlottHMM	f_0: 1 + Energy: 1 + HNR: 5 + source LSF: 10 + vocal tract LSF: 30	Stored glottal flow pulse + noise

5.3.1.1 Impulse train excited mel-cepstral vocoder

The impulse train excited mel-cepstrum based vocoder is the baseline vocoder used in Section 5.2. In this vocoder, speech is described by only two features: f_0 (for excitation) and spectrum (the filter). Here, spectrum is represented by MFCCs as

they provide a good perceptual representation of the speech spectrum. Spectrum is computed with the Speech Signal Processing Toolkit (SPTK) and f_0 with Snack.

For synthesis, the excitation is either a pulse train at the desired f_0 for voiced speech, or white noise for unvoiced parts. This simple excitation method causes the rendered speech signal to frequently sound buzzy.

5.3.1.2 STRAIGHT

The STRAIGHT method proposed by Kawahara [Kawahara *et al.* 1999, Kawahara *et al.* 2001] decomposes speech into three components: a) the spectrum, again represented with mel-cepstrum coefficients, but extracted using pitch-adapted spectral smoothing; b) the fundamental frequency (f_0) extracted using instantaneous-frequency-based estimation; c) band-aperiodicity parameters which represent the proportions of voiced and unvoiced contributions in spectral sub-bands.

For voiced speech excitation, STRAIGHT uses a mix between an impulse train and noise according to the values of the band-aperiodicity parameters [Yoshimura *et al.* 2001]. White Gaussian noise is used for unvoiced excitation. The filter corresponds to the MFCCs. STRAIGHT is widely used in speech synthesis as it is robust and produces speech of good quality [Zen *et al.* 2007a].

5.3.1.3 Deterministic plus Stochastic Model (DSM)

The Deterministic plus Stochastic Model (DSM) of the residual signal [Drugman & Dutoit 2012] first estimates the speech spectrum, and uses the inverse of the filter to reveal the speech residual. Glottal closure instant (GCI) detection is used to extract individual GCI-centered residual waveforms, which are further resampled to fixed duration. The residual waveforms are then decomposed into the deterministic and stochastic parts in frequency domain, separated by the *maximum voiced frequency* F_m fixed at 4 kHz. The deterministic part is computed as the first Principal Component (PC) of a codebook of residual frames centered on glottal closure instants and having a duration of two pitch periods. The stochastic part consists of a white Gaussian noise filtered with the LP model of the average high-pass filtered residual signal, and time-modulated according to the average Hilbert envelope of the stochastic part of the residual. White Gaussian noise is used as excitation for unvoiced speech. The DSM vocoder has been shown to reduce buzziness and to achieve comparable speech synthesis quality as that of STRAIGHT [Drugman & Dutoit 2012]. Here, STRAIGHT is used to extract f_0 and MFCCs for the DSM analysis. For synthesis, the excitation is built with the DSM model.

5.3.1.4 GlottHMM

The GlottHMM vocoder [Raitio *et al.* 2011] relies on glottal inverse filtering to separate the excitation and filter contributions. Inverse filtering is actually performed in an iterative way, according to the Iterative Adaptive Inverse Filtering (IAIF)

method presented in [Alku 1992]. The filter is represented by LP coefficients. As opposed to usual methods, the spectral contributions of the excitation are here not included in the filter model which is excited by white noise or a pulse train, but are included in the excitation signal which is allowed to exhibit decaying spectral envelopes [Raitio *et al.* 2011]. In the IAIF method, the glottal spectrum and vocal tract contributions are iteratively estimated and removed from the original speech signal in order to compute the next estimations. For the sake of completeness, we should mention that the resulting LP coefficients are then converted to Line Spectral Frequencies (LSF) [Soong & Juang 1984], for a better parameterization for the statistical modeling. Besides the coefficients of the all-pole filters representing the spectral contributions of the glottal source and the vocal tract, f_0 (estimated through autocorrelation) and HNRs in four sub-bands ([0-2],[2-4],[4-6],[6-8] kHz) are extracted from the glottal source signal. All these features, as well as the energy of the frames, are modeled by the HMMs. In the case of unvoiced segments, f_0 and HNR values are set to 0, and only the glottal source estimates of voiced segments are used for synthesis.

For synthesis, one example of glottal pulse from a human voice producing a sustained vowel is used to compose the excitation of voiced frames and processed to match the target parameters outputted by the HMMs. First, the human glottal pulse is interpolated in the time domain to correspond to the target f_0 and scaled to reach the desired energy. Noise is then added to the pulse in the spectral domain (Fast Fourier Transform coefficients are modified with a random factor) according to the target HNRs. Finally, the glottal pulse is filtered to match the target glottal spectrum. Unvoiced excitation consists of white noise.

A perceptive evaluation of the GlottHMM vocoder was performed on speech synthesis signals in [Raitio *et al.* 2011]: GlottHMM was preferred to STRAIGHT and MCEP and proven more intelligible than STRAIGHT.

5.3.2 Evaluation

To compare the performance of the four vocoders for laughter synthesis, a perceptive evaluation was conducted. For each vocoder, two types of laughs were synthesized: a) copy-synthesis laughs, where the necessary features for synthesis were extracted from a human laugh and synthesis was directly performed using those human features; b) HMM-based synthesis, where laughter models were trained on a laughter database and laughs were then synthesized, using the models, from the phonetic transcription of a human laugh.

Copy-synthesis can be seen as the theoretically best synthesis that can be obtained with a particular vocoder, while HMM-based synthesis shows the current performance that can be achieved when synthesizing new laughs. Human laughs were also included in the evaluation for reference.

It is worth noting that, contrarily to all other perceptive evaluations reported in this dissertation, the laughs synthesized for comparing the vocoders have not been

obtained with a leave-one-out process. In other words, the synthesized laughs belonged to the training set. Time constraints are the only reason for this, but it is not expected to affect the comparison between vocoders, as they have all been used in the same conditions.

5.3.2.1 Training data

For the purpose of this work, two voices from the AVLaughterCycle database [Urbain *et al.* 2010a] were selected: a female voice (subject #5, 54 laughs) and a male voice (subject #6, the same voice as in the other sections, 64 laughs). Phones were grouped in phonetic clusters according to the grouping illustrated in Figure 4.4. For each voice, the phonetic clusters that did not have at least eleven occurrences were assigned to an *unknown* class.

For the test, five laughs lasting at least 3.5 seconds were randomly selected for each voice. For each vocoder, these laughs were synthesized from their phonetic transcriptions (HMM synthesis with HTS) as well as re-synthesized directly from their extracted parameters (copy-synthesis). The five original laughs were also included in the evaluation. This makes a total of 5 (original laughs) + 5 × 2 (HMM and copy-synthesis) × 4 (number of vocoders) = 45 laughs in the evaluation set for each voice.

5.3.3 Results

The perceptive test was run on the web. Participants were asked to listen to all the laughs, presented one by one in random order, and to rate the naturalness of the synthesis on a 5-point scale. Participants were suggested to wear headphones. Eighteen participants evaluated the male voice while 15 evaluated the female one. All listeners were between 25–35 years of age, and some of them were speech experts.

The average results received by each vocoder, for each voice and each setting (synthesis or copy-synthesis) are displayed in Figure 5.5, together with the 95% confidence intervals over the average. The pairwise p-values between the methods (using the conservative Bonferroni adjustment) are displayed in Tables 5.5 and 5.6.

As for the initial experiments presented in Section 5.2, we can see that: a) human laughs are perceived as more natural than synthesized and copy-synthesized laughs; b) copy-synthesis laughs received better scores than synthesized laughs. However, the main aim of this study is the comparison between vocoders.

GlottHMM was rated as less natural than all other vocoders. STRAIGHT and DSM seemed to be equivalent for copy-synthesis, but DSM received slightly better scores (not reaching statistical significance) for HMM synthesis. Regarding MCEP, despite its simplicity and known buzziness, it reached high naturalness values for the male voice in the copy-synthesis mode as well as for the female voice in HMM synthesis, where it was the vocoder receiving the best scores (although the difference is not statistically significant with DSM in both cases).

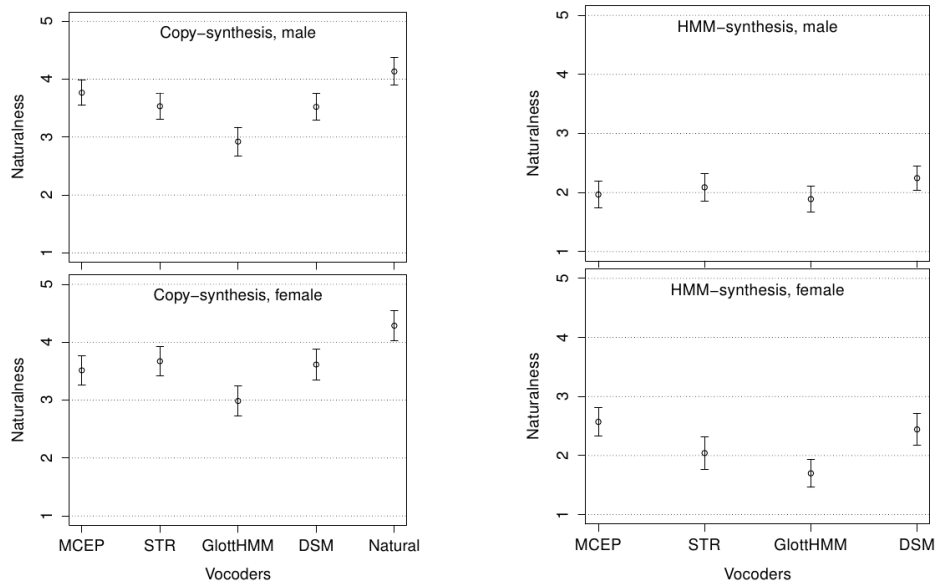

Figure 5.5: Average naturalness scores obtained by each vocoder, for copy-synthesis (left) and HMM synthesis (right), for the male (top) and female (bottom) voices.

5.3.4 Discussion

According to the results, DSM and MCEP appear like the best (of the evaluated) choices for HMM-based laughter synthesis. Both vocoders rely on the same features (f_0 and MFCCs, although for DSM these were extracted with the STRAIGHT

Table 5.5: Pairwise p-values between the vocoders copy-synthesis and natural laughs. Statistically significant results are marked in bold.

Female	System	DSM	Glott	MCEP	STRAIGHT	Natural
	DSM	–	**0.006**	1	1	**0**
	Glott	**0.006**	–	**0.04**	**0.002**	**0**
	MCEP	1	**0.04**	–	1	**0**
	STRAIGHT	1	**0.002**	1	–	**0**
	Natural	**0**	**0**	**0**	**0**	–
Male	System	DSM	Glott	MCEP	STRAIGHT	Natural
	DSM	–	**0.003**	1	1	**0**
	Glott	**0.003**	–	**0**	**0.002**	**0**
	MCEP	1	**0**	–	1	**0.027**
	STRAIGHT	1	**0.002**	1	–	**0**
	Natural	**0**	**0**	**0.027**	**0**	–

method) and mainly differ on the shape of the excitation signal for voiced segments. The pulse train used by MCEP can cause buzzy signals, but it seems that this effect did not appear or was not prominent compared to other possible defects. In particular, the female voice used fewer voiced segments than the male one, which can explain why MCEP was relatively well rated for female HMM synthesis.

STRAIGHT performed well in copy-synthesis but could not maintain the good performance in the synthesis settings. The unstable estimation of aperiodicity features can be one explanation to this, as laughter signals are challenging for pitch estimation. Furthermore, STRAIGHT pitch estimation is known to be unreliable for non-modal voices [Raitio *et al.* 2013], and laughter does not correspond to modal speech. Although possible pitch estimation errors themselves affected DSM as well, they could have caused deviations in the STRAIGHT aperiodicity features, resulting in poor models for the mixed excitation signal.

Regarding GlottHMM, it models the voiced sounds in a totally different way than other vocoders. It could thus suffer even more from bad estimations of voicing and f_0. Indeed, GlottHMM is known to suffer from pitch estimation errors in the case of challenging signals, to which laughter clearly belongs. GlottHMM may also be penalized by the use of a glottal pulse retrieved from a sustained vowel. Laughter is a highly dynamic process and it might be difficult to represent the evolution of laughter glottal pulses by simply altering (interpolating, scaling, filtering) one human sample.

To conclude, these findings justify the use of the DSM vocoder, at least for our male voice, as it performs slightly better than the other studied vocoders.

Table 5.6: Pairwise p-values between HMM synthesis of different vocoders. Statistically significant results are marked in bold.

Female	System	DSM	Glott	MCEP	STRAIGHT
	DSM	–	**0.003**	1	0.16
	Glott	**0.003**	–	**0**	0.34
	MCEP	1	**0**	–	**0.02**
	STRAIGHT	0.16	0.34	**0.02**	–

Male	System	DSM	Glott	MCEP	STRAIGHT
	DSM	–	0.14	0.46	1
	Glott	0.14	–	1	1
	MCEP	0.46	1	–	1
	STRAIGHT	1	1	1	–

5.4 Use of automatic phonetic transcriptions for HMM laughter synthesis

In this section, we will investigate whether good quality HMM-based speech synthesis can be obtained by training the models with automatic phonetic transcriptions. If so, new laughing voices will be available without needing to manually transcribe the training laughs, which is time-consuming, subjective and error-prone.

To do so, HMMs were trained using the same methods as described in Section 5.2. The STRAIGHT and DSM methods 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 4.3). The HTK models were trained using all the AVLaughterCycle data, except the files from subject #6, then used to estimate the phonetic transcriptions of subject #6's laughs. Hence, these transcriptions were obtained in a speaker-independent way.

For synthesis, sixty-four laughs were available for training. Phones with less than eleven occurrences in the reference transcriptions were mapped to an *unknown* class. The number of occurrences of each phonetic cluster is given in Table 5.7. 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 5.7: 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

As explained in Section 5.2, a syllable layer of annotations had been manually added on the transcription files in order to compute contextual features for HMM-based synthesis. 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 Section 5.2, naive participants were asked to rate the naturalness of synthesized laughs on a 5-point scale with the following labels: very poor (score 1), poor (2), average (3), good (4), excellent (5). Laughter synthesis trained with the HTK transcriptions was compared to synthesis trained with the manual (reference) transcriptions. For each of these training processes, two 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 5.8 summarizes the different methods compared in the evaluation experiment. Twenty-three laughs contained at least one occurrence of an *unknown* phone and were not included in the evaluation. Out of the remaining 41 laughs, two laughs were shown as examples to the participants prior to the test, so they that 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 methods was randomly selected. The test was completed after 39 evaluations.

Table 5.8: 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 5.9 and the distribution of received answers for each method is illustrated in Figure 5.6. A univariate analysis of the variance was conducted and the p-values of the pairwise comparisons of the different methods, using the Tukey HSD adjustment, are presented in Table 5.10. Statistically significant differences at a 95% confidence level are highlighted in bold. It can be seen that, as expected, human laughs sound more natural than synthesized laughs. Among the synthesized laughs, method R2 performed the best. As it was already found (see Section 5.2), 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). This is due to errors in the transcriptions (insertions, deletions, substitutions) which degrade the quality of the estimated models. 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 Narayanan's method [Sundaram & Narayanan 2007]—which had an average naturalness score of 1.71—, while our reference methods (R1 and R2) obtained similar naturalness scores as in Section 5.2, which was expected as we used the same process.

Table 5.9: Average naturalness score received by each synthesis method.

	Method				
	R1	**R2**	**A1**	**A2**	**H**
average naturalness score	2.4	2.7	2.0	2.2	4.3
naturalness score std	1.1	1.1	0.9	1.1	0.9

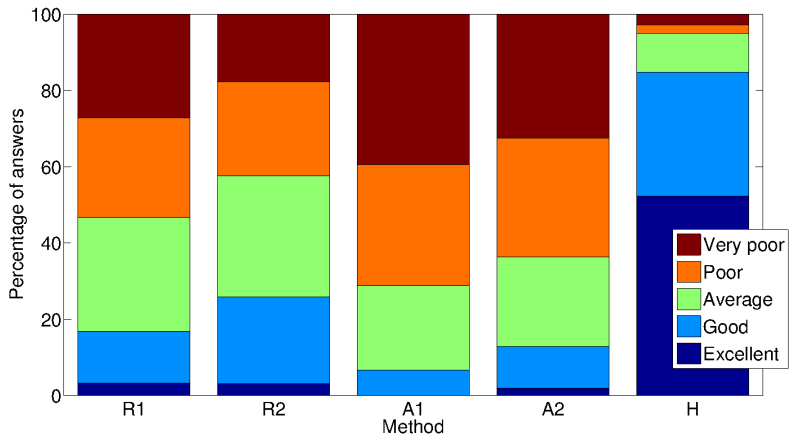

Figure 5.6: Distribution of naturalness scores received by the synthesis methods.

Table 5.10: 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 complicated, 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 (stds) in Table 5.9. The overall inter-participant agreement for the five naturalness values is quite low: Fleiss' kappa [Fleiss *et al.* 1981] is .167. Participants however generally agree on whether the laugh sounds natural (score 4 or 5) or not (score 1, 2 or 3) with a kappa value of .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).

5.5 Arousal-driven generation of laughter phonetic transcriptions

In this section, we will describe our algorithm developed to generate laughter phonetic transcriptions from arousal signals. The objective is to simplify the overall synthesis process by enabling users to drive it from higher-level instructions (e.g., arousal signal) than phonetic transcriptions. The generation algorithm is first presented in Sections 5.5.1 and 5.5.2, before evaluating it with standard perceptive tests in Section 5.5.3. The generation algorithm, together with evaluation of the base synthesis (see Section 5.2), has been published in [Urbain *et al.* 2014].

5.5.1 Generation of transcriptions by unit selection

As explained in Sections 3.4 and 4.4, the per-frame arousal of 49 laughs of the AVLaughterCycle database has been manually annotated and neural networks have been trained to automatically estimate the per-frame arousal of laughs.

To develop the algorithm to generate phonetic transcriptions from arousal signals, the per-frame arousal of all the laughs from our selected voice (subject #6) were needed. As the objective here was to design the generation algorithm in the best possible way, rather than evaluate the quality of the speaker-independent automatic

per-frame arousal estimation, the neural networks were here trained only with the 19 manually transcribed per-frame arousal signals from subject #6. The obtained neural network was then used to compute the per-frame arousal signals of all the 64 laughs from subject #6.

For each laugh, our database contained the per-frame arousal signal and the associated phonetic transcription. Our aim was to generate the phonetic transcription (with syllable and respiration information) from the arousal signal only. To do so, we developed a method inspired by unit selection for concatenative speech synthesis [Hunt & Black 1996]. All the data from subject #6 (see Table 5.11) were gathered and segmented into syllables. Target cost was set to the cumulated distance between the target per-frame arousal and the arousal signal of each syllable (each presenting one value every 10 ms), divided by the syllable length as a normalization factor. Concatenation cost was obtained as the inverse of the n-gram likelihoods trained on the syllable sequences of subject #6 via the MIT Language Modeling Toolkit [Hsu & Glass 2008]. Given the synthesis purposes of this work, it is interesting to use high-order n-grams to encode long-term dependency effects (for example, inhalations are more likely after a higher number of exhalation syllables, which is better encoded with high-order n-grams than low order ones). We used 6-grams in our experiments, taking advantage of the back-off values [Katz 1987] to compute the likelihoods from lower orders if they did not appear in the 6-grams.

Table 5.11: Number of available units (from subject #6).

	# phones	# syllables	# parts	# laughs
exhalation	1311	781	79	64
inhalation	64	58	50	

One of our long-term objectives is to achieve real-time laughter synthesis. Recently a lot of progress has been made in real-time HMM-based speech synthesis [Astrinaki *et al.* 2013]. To be able to exploit these advances towards real-time synthesis, the generation algorithm was implemented with a real-time approach: the transcription was computed step by step, from left to right. As the vast majority of laughter syllables last less than 500 ms (see Figure 5.7), we decided to use this duration as look-ahead value. Let $S_n = \{s_1, s_2, \ldots, s_n\}$ be the sequence of syllables selected at step n, D_n the total duration of S_n, I the target arousal signal with duration D_I, L the library of available syllables, $T(i, x)$ the target cost between target arousal fragment i and the candidate syllable x (i.e., the normalized cumulated distance between the corresponding arousal signals) and $C(S_n, x)$ the concatenation cost of candidate syllable x given S_n (obtained via n-gram likelihoods). The algorithm is implemented as follows:

1. Initialization: $n = 0$, $S_0 = \emptyset$, $D_0 = 0$.

2. While $(D_n < D_I)$:

(a) Increment n.

(b) Cut the target arousal fragment:
$I_n = I[D_{n-1}, D_{n-1} + 0.5s]$.

(c) Build the subset $L_n \subset L$ of the 20 syllables[6] l with the lowest target cost $T(I_n, l)$.

(d) Select the syllable s_n with minimal overall cost:

$$s_n = \underset{l \varepsilon L_n}{\arg\min}(T(I_n, l) + C(S_{n-1}, l)). \tag{5.3}$$

(e) Add the selected syllable to the sequence:
$S_n = \{S_{n-1}, s_n\}$.

3. Build the phonetic transcription P by concatenating the phonetic transcriptions of the syllables in S_n.

This generation process is illustrated in Figure 5.8. Iteration #4 is displayed. The current target arousal curve lasts for 500 ms after the already processed values. The 20 syllables from the syllable library that have the lowest target scores are preselected. n-grams are used to compute the overall cost and the best syllable is added to the sequence of validated syllables. Iteration #5 will consider the 500 ms that follow the end of syllable #4.

The generated laughter transcriptions can then be synthesized and evaluated through the HMM-based laughter synthesis model presented in Section 5.2.

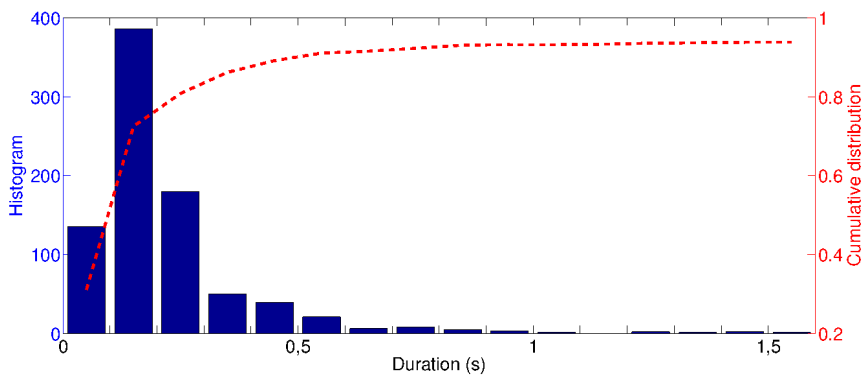

Figure 5.7: Histogram and cumulative distribution function of the laughter syllable durations (subject #6 of the AVLC database).

[6] To ensure a good correspondence between the target arousal and the selected syllables, it was decided to only consider the q syllables that have the lowest target cost. q was empirically set to 20 in our experiments.

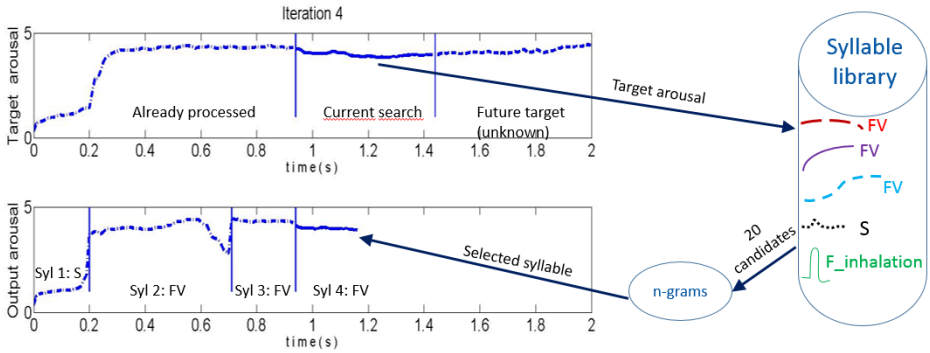

Figure 5.8: Algorithm for generating laughter phonetic sequences from arousal signals.

5.5.2 Refinements of the method

Using the aforementioned algorithm sometimes resulted in unusual phonetic sequences: *1)* mostly short syllables were selected, resulting in an unnatural succession of very short sounds; *2)* inhalation syllables were barely selected, which is obviously unnatural given the limited human pulmonary capacity; *3)* as the selection algorithm was only looking at syllable transcriptions and "ha" and "ho" have the same syllabic transcription (i.e., FV), some transcriptions had unnatural oscillations between these these vowels (e.g., "hahahohahoha"). The generation algorithm was therefore modified as follows.

Improvement 1 In order to favor longer syllables, the overall cost (Equation 5.3) was divided by the candidate syllable duration. Note that this is not normalization, as the target cost was already normalized with respect to the syllable duration.

Improvement 2 To include inhalation syllables at appropriate times, the duration and area under the arousal signal—which can be related to the pulmonary effort—of the bouts have been studied. Figure 5.9 displays the cumulative distributions of the duration and area under the arousal signal of the first bout of each laugh of subject #6. It can be noticed that almost all bouts last at least 1 s and have an area under the arousal signal greater than 2. Starting from these values, the likelihood to enter in an inhalation phase increases almost linearly (from 1 to 3 s on the duration graph; from 2 to 8 for the area under the curve). These observations were integrated in our generation algorithm. If the last selected syllables (from indexes m to n) in S_n are exhalation syllables, with a cumulated duration D_{mn} and a cumulated area under the arousal signal A_{mn}, the overall cost of the 20 candidate syllables is modified as follows:

- If the candidate syllable is an exhalation syllable, no modification.

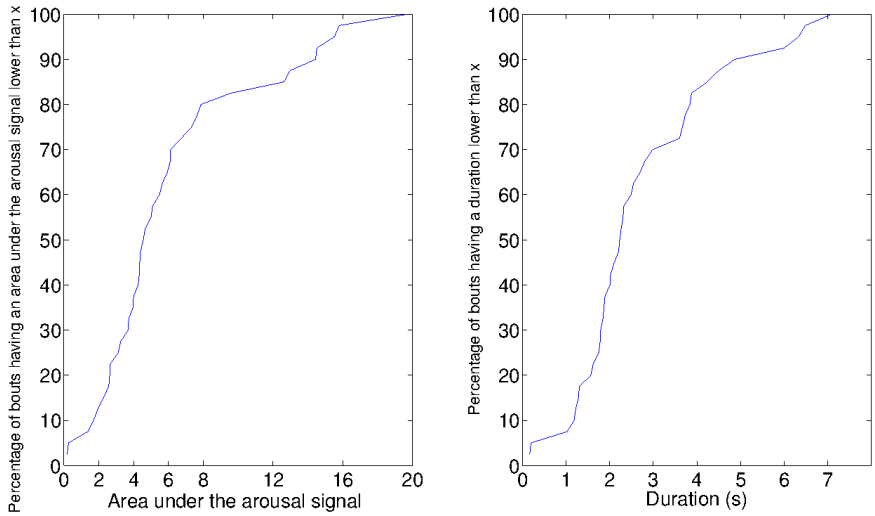

Figure 5.9: Cumulative distribution functions of the area under the arousal signal (left) and duration (right) of the bouts.

- If the candidate syllable is an inhalation syllable:
  - divide its overall cost by $2 \times D_{mn}$ if $D_{mn} \geq 1s$.
  - divide its overall cost by A_{mn} if $A_{mn} \geq 2$.

These divisions lower the cost of inhalation syllables compared to exhalation syllables, and will in consequence favor inhalation syllables in the selection algorithm.

Modification 3 Finally, to prevent disturbing oscillations between the vowels "a" and "o", the "o"s in the generated transcription were replaced by "a"s, as for subject #6 "a" is much more frequent (as was shown in Table 5.1).

5.5.3 Evaluation

The evaluation was conducted with the same process as already presented in previous sections. The only change lies in the evaluated signals.

Here, we aimed at comparing laughs synthesized in a classical way (using phonetic transcriptions as input) with laughs synthesized from arousal signals (using the generation algorithm to produce the phonetic transcriptions that are then synthesized by HTS). The same HTS models have been used in all cases, with STRAIGHT and DSM.

The classical synthesis methods were methods $S3$ and $S4$ already presented in Section 5.2:

- Method $S3$: the laugh is synthesized using the HMM-based synthesis process (using as only input the phonetic transcription of the laugh), using the STRAIGHT algorithms for spectrum and f_0 extraction and the DSM vocoder, with the phone durations imposed by the phonetic transcriptions.

- Method $S4$: same as Method $S3$, with the duration of each phone estimated by HTS.

The following generation methods were included in the experiment:

- Method $G1$: the basic generation method, as explained in Section 5.5.1, with imposed duration for synthesis.

- Method $G2$: the generation algorithm was modified to favor long syllables and inhalation phases (improvements #1 and #2 explained in Section 5.5.2), with imposed durations for synthesis.

- Method $G3$: same as Method $G2$, with the addition of modification #3 described in Section 5.5.2.

- Method $G4$: same as Method $G3$, with phone durations estimated by HTS during the synthesis step.

Human laughs (method H) were also included in the evaluation.
The test hypotheses were the following:

- H4: the refinements of the generation algorithm (see Section 5.5.2) improve the naturalness of synthesized laughs (method $G3$ is better than method $G2$ which is better than method $G1$).

- H5: the generation module is efficient in producing natural laughter transcriptions (methods $G3$ and $G4$ have comparable results to methods $S3$ and $S4$, respectively).

Each of the produced laughs was obtained with a leave-one-out method, to ensure that we were not simply able to reproduce learned trajectories. For the synthesis laughs (methods $S3-4$), the laugh to synthesize was not included in the training phase of the HMMs. For the phonetically generated laughs (methods $G1-4$), the syllables of the laugh to generate (and synthesize) were withdrawn from the library of available syllables. The arousal signal of the corresponding laugh was the only input used to obtain the phonetically generated laughs. The types of data used for each of the methods are summarized in Figure 5.10.

Table 5.12 presents the average durations of the evaluated laughter units for each method, as well as the average f_0 and its standard deviation. It can be observed that the average f_0 values are similar in all the methods. Regarding the durations, it appears that method G1 tends to include short syllables and that the modification introduced to overcome this problem (see Improvement 1, Section 5.5.2) indeed brings

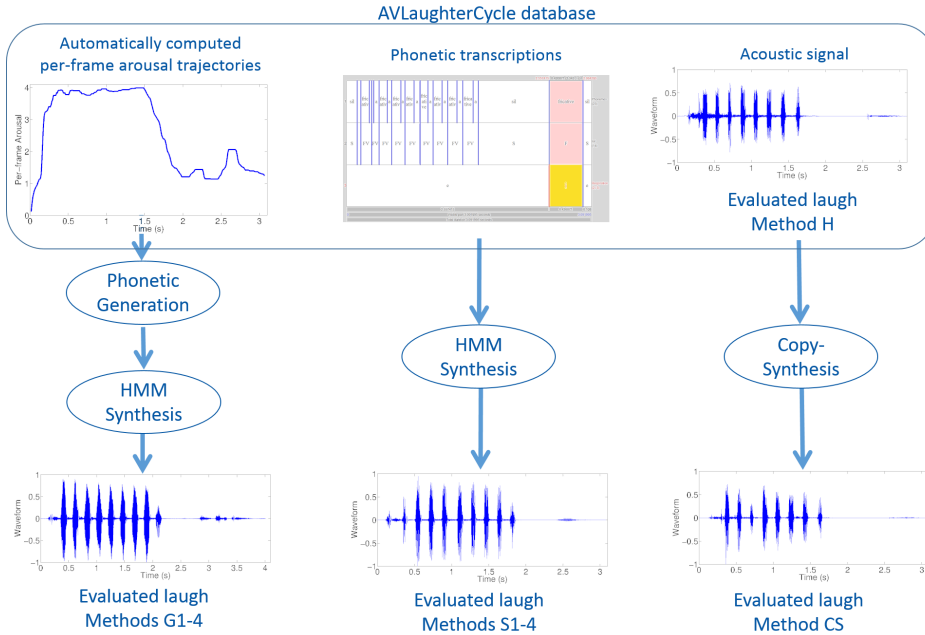

Figure 5.10: Types of data used for the different methods compared in the evaluation experiment. Note: Methods S1-S2 and CS were not used in this evaluation, but have been placed on the graph for clarity and completeness, as they have been used in previous experiments.

the syllable durations back to more common values (around 200 ms, as in the original human laughs). Examples of evaluated laughs are available on `http://www.tcts.fpms.ac.be/~urbain/arousal_driven_synthesis`.

The web-based evaluation application was exactly the same as the one described in Section 5.2. Results are presented in the next sections.

5.5.4 Results

Fifty-three participants completed the study: 16 females (average age: 30.2; std: 15.4) and 37 males (average age: 24.5; std: 7.3). Their profiles are summarized in Table 5.13. As in Section 5.2, only the results from participants wearing headphones will be reported here. Out of the 841 received answers from participants using headphones, 55 were "I cannot rate the naturalness of this laugh". Table 5.14 gathers the number of ratings received, the number of "unknown" answers and the average score for each method. The distribution of naturalness scores received by each method is displayed in Figure 5.11.

A univariate analysis of the variance was conducted. Table 5.15 presents the p-values of pairwise comparisons between the relative naturalness scores of the different

Table 5.12: f_0 averages, standard deviations and average durations of the evaluated laughter units. Note: Exh. and Inh. stand for Exhalation and Inhalation parts, respectively.

Method	f_0 (s)		Average duration (ms)			
	Mean	std	Exh.	Inh.	Syl.	Phones
S3	199	86	1522	384	207	126
S4	197	84	1523	345	205	124
G1	218	80	822	147	104	72
G2	214	91	1082	220	199	118
G3	211	80	1082	220	199	118
G4	216	86	1446	376	256	152
H	193	88	1522	384	207	126

Table 5.13: Participant profiles. Note: the "none" or "laughter" category is related to the experience of the participant in laughter synthesis.

Category	Loudspeakers		Headphones		Total
	Females	Males	Females	Males	
None	8	15	8	20	51
Laughter	0	1	0	1	2
Total	8	16	8	21	53

methods, using the Tukey HSD correction. Statistically significant differences at a 95% confidence level are highlighted in bold.

5.5.5 Discussion

As the results indicate, the modifications proposed in Section 5.5.2 slightly increased the performance, but not enough to reach statistical significance. H4 is thus not verified. Again, the best results are achieved when HTS estimates the duration of the phones, confirming that the generation step only has to produce phonetic sequences (phone durations are not important at this stage, they are better modeled by the synthesizer). One explanation for the good performance achieved when durations are estimated by HTS rather than replicating durations of actual human laughs is that the models are less constrained and can thus produce better feature trajectories.

It is interesting to note that H5 is verified: laughs generated with our best phonetic generation method ($G3$ and $G4$) achieve the same naturalness scores as laughs synthesized directly from existing transcriptions ($S3$ and $S4$). The generation step is thus efficient, and it also supports the use of arousal signals, which indeed carry sufficient information to generate hilarious laughs. However, it must be further investigated whether the generation algorithm tends to over-privilege the most likely syllables, resulting in laughs that would be excessively similar to each other. This

Table 5.14: Received answers for each method, only including participants using headphones.

Method	# *ratings*	# *unknown*	Av. score (std)	Av. RR score[a] (std)
G1	110	10	2.3 (1.1)	1.4 (1.1)
G2	105	6	2.3 (1.1)	1.5 (1.2)
G3	103	9	2.4 (1.2)	1.5 (1.2)
G4	115	10	2.4 (1.2)	1.2 (1.2)
S3	123	11	2.4 (1.1)	1.5 (1.3)
S4	125	6	2.4 (1.1)	1.4 (1.2)
H	105	3	3.9 (1.1)	0 (0.7)
ALL	786	55		

[a]One laugh could not be included in the relative scores, as the corresponding human laugh had not received any evaluation.

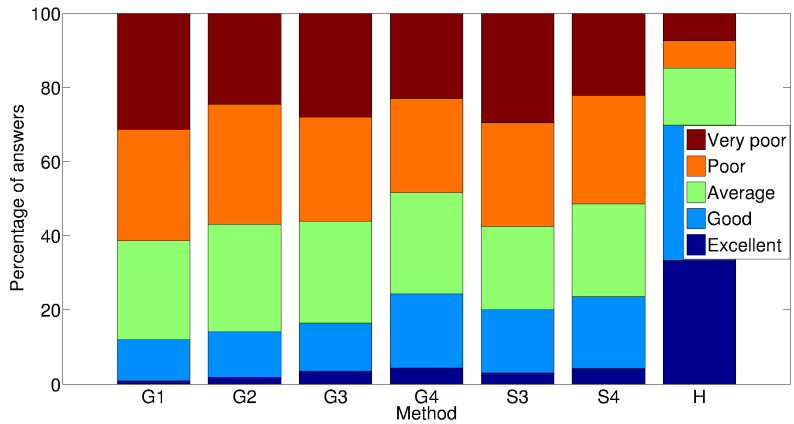

Figure 5.11: Distribution of naturalness scores received by each method.

phenomenon could have negative effects on a listener involved in a human-machine interaction who would perceive that the machine is always laughing the same way.

5.6 Summary and perspectives

In this chapter we have firstly reviewed the state-of-the-art in acoustic laughter synthesis. Then, we have presented our developments using HMMs, which achieve state-of-the-art naturalness scores. The proposed method is largely inspired by speech synthesis and fulfills the required conditions for a laughter synthesis system formulated in [Sundaram & Narayanan 2007]: a large variety of laughs (with different phones, number of syllables, number of bouts) can be synthesized and it relies on very simple

Table 5.15: Pairwise p-values between the generation methods.

Method	*G1*	*G2*	*G3*	*G4*	*S3*	*S4*	*H*
G1	-	0.99	1	0.96	0.98	1	**0**
G2	0.99	-	1	0.62	1	0.96	**0**
G3	1	1	-	0.86	1	1	**0**
G4	0.96	0.62	0.86	-	0.51	0.99	**0**
S3	0.98	1	1	0.51	-	0.92	**0**
S4	1	0.97	1	0.99	0.92	-	**0**
H	**0**	**0**	**0**	**0**	**0**	**0**	-

inputs (a phonetic transcription or even an arousal curve). Contrarily to previous works on laughter synthesis, the presented algorithms relying on HMMs do not focus on one particular type of laughter or phones (voiced laughter). We have to admit that most of the laughs included in our evaluations were also voiced laughs, as it is the type of laughs that was mostly used by the subjects who laughed the most in the AVLaughterCycle database. But this is a consequence of the available data rather than a choice to only address voiced laughs. In addition, some unvoiced syllables have been synthesized (cackles or syllables containing only a fricative[7]), as well as inhalation sounds, which enable to synthesize episodes with several bouts, something that was discarded by all preceding studies.

The HMM-based laughter synthesis process is not limited to one voice, either. Although almost all the results presented here relate to the same male voice, we have also synthesized female laughs with the same methods, as we will see in Chapter 6. It is expected that any laughter voice with sufficient training data (starting from a few minutes) could be modeled with the proposed pipeline. New voices and forms of laughter must nevertheless be developed and evaluated in the future. Laughing styles that are particularly well suited or not to the presented methods could then be identified.

The different experiments conducted within this chapter open several other perspectives for future works. First, regarding classic HMM-based laughter synthesis, some developments did not conduct to improved laughter quality. The addition of contextual information was not helpful so far. One hypothesis is that it is due to the limited data available for training (three minutes of laughter, which is relatively high when we consider spontaneous laughter uttered by a single participant, but is really low compared to the large databases used for speech synthesis). This has to be verified in the future on an even larger laughter database: does more training data enable to improve laughter synthesis quality, and is contextual information useful? The preparation of such large data can hopefully be fastened thanks to the automatic estimation of laughter phonetic transcriptions. Furthermore, new contextual features

[7] There are several of these in the evaluated laughs, even if syllables containing only a fricative are generally of much less intensity than voiced syllables, so their presence is not striking when listening to the laughs.

could be investigated, for example arousal values could be added in the contextual features.

Second, the HTS algorithms include a range of parameters (boundaries for f_0 estimation, number of states in the HMMs, models for computing the derivatives, number of MFCCs, etc.), to which we can add the phonetic grouping for our phonetic transcriptions. Although these parameters have been empirically tuned in our developments, a thorough optimization would be welcome. Some of the parameters (e.g., boundaries for f_0 estimation) should also probably be adjusted to the individual. The parameters of the generation algorithm could also be further studied and it would be desirable to include vowel information directly into the syllable labels to avoid having to post-process the generated phonetic transcriptions.

Third, the results of the comparison of vocoders suggest that the robustness of parameter estimation is crucial for laughter synthesis. Efforts on identifying robust algorithms for laughter parameters extraction should thus be increased. Among others, in line with the conclusions of Chapter 4, it would be interesting to investigate the robustness of the algorithm presented by Sudheer et al. for f_0 extraction in laughter—which was actually used by Sathya et al. for laughter synthesis—for HMM-based synthesis.

Fourth, deep analysis of which synthesis errors create the perception of unnaturalness would be interesting. This is somehow related to the comments received by the participants who rated the naturalness of the laughs. Identifying accurately what is going wrong (is it a problem of dynamics of the laugh, badly estimated durations, some phones that are poorly modeled and constantly poorly synthesized?) would help better understanding the current limits of the methods and consequently improving them. New perceptive evaluations, with refined questions for deeper analysis than the fuzzy notion of "naturalness", could be necessary to efficiently address these issues.

Fifth, the algorithm to generate laughs from arousal curves opens new perspectives for the integration of laughter synthesis in human-computer interactions. Laughter synthesis can now be driven by high-level input instead of low-level descriptors (e.g., phonetic transcriptions or acoustic parameters used in other laughter synthesis works), which are hard to produce in a consistent way. Further experiments with the generation algorithm should however be conducted to investigate whether it can be generalized to new arousal curves and new voices.

Finally, real-time laughter synthesis can be investigated. Recent developments of real-time HMM-based speech synthesis are promising. Adaptation of the current models to real-time (or as close to real-time as possible, by investigating how much of the future has to be known to maintain an acceptable synthesis quality) and coupling with a real-time implementation of the generation algorithm would constitute an important step forward for a range of applications (psychological studies on the perception of laughter, human-computer interactions, artistic performances including synthesized laughter, etc.). Some of these applications relying on acoustic laughter synthesis will be presented in the next chapter.

Chapter 6

Applications

Contents

In this chapter, we will focus on applications in which our laughter synthesis methods have already been integrated. We will briefly present three applications we have been involved in. The first two are related to human-computer interaction: the eNTERFACE'12[1] Laugh Machine project (Section 6.2) and the eNTERFACE'13 Laugh When You're Winning project (Section 6.3). The last application concerns laughter modifications for psychological studies (Section 6.4). As for other chapters, we will start with a section presenting related works, centered on human-computer interaction involving laughter. Experiments on laughter manipulations for perceptive experiments have already been presented in Section 3.1.6.4 (see Kipper and Todt's experiments).

6.1 Related works

To the best of our knowledge, there exist only few—and recent—applications specifically integrating laughter in human-machine interactions. None of them actually relied on synthesized laughter, but rather include recorded human laughter.

In 2007, Melder et al. designed the Adaptive Affective Mirror, an application aiming at pushing participants towards positive emotions and laughter [Melder *et al.* 2007]. The application consists in mirroring and distorting the face of the participant, in order to create funny images. Acoustic laughter detection (based on the works of Truong and van Leeuwen [Truong & van Leeuwen 2007a]), visual smile detection, affective keyword spotting and hand gesture recognition (through accelerometers) are performed in real-time to assess the emotional state of the participant, and in particular if (s)he is smiling or laughing. The mirror feedback is a

[1] To recall what was introduced in Section 2.5, eNTERFACE is a one-month Summer Workshop taking place every year and enabling researchers to work together and deliver scientific outcomes.

distorted image of the participant's face, with increased modifications when the participant is laughing or smiling. The objective was to create an interaction loop, i.e. that the first (slightly) distorted images would push the participant to smile, which would provoke increased distortion, leading the participant to smile more or laugh, and so on.

In 2009, we proposed AVLaughterCycle[2], similar to Adaptive Affective Mirror in its objectives but using a laughing agent [Urbain *et al.* 2010b]. The application aimed at creating laughter loops by triggering agent laughter when the participant was laughing. It is in the framework of that project that the AVLaughterCycle database was recorded. The recorded laughs formed a laughter library containing two modalities: audio as well as facial movements (obtained through motion capture, and synchronized with the audio). When vocal activity was detected (there was no actual laughter detection in the application, it was assumed that any input would be a laugh), features of the laugh were computed and the most similar laugh in the library was identified, using the similarity algorithm presented in Section 4.2. The selected laugh was then played by the virtual agent, Greta [Niewiadomski *et al.* 2009]. As for Adaptive Affective Mirror, it was expected to create laughter loops, where the first laugh could be forced to trigger a first reaction from the agent, which would trigger more spontaneous laughter, provoking another agent laugh, etc.

Becker-Asano et al. [Becker-Asano *et al.* 2009] studied the impact of auditory and behavioral signals of laughter in different social robots. They discovered that the social effect of laughter depends on the situational context including the type of task executed by the robot, and on verbal and nonverbal behaviors (other than laughing) that accompany the laughing act [Becker-Asano & Ishiguro 2009].

Finally, Fukushima et al. [Fukushima *et al.* 2010] used toy robots that were shaking their heads and playing recorded laughs when the system detected that the user was laughing. The objective was again to push participants to laugh. The system was evaluated and results showed that the application indeed enhanced the participants' laughter activity.

6.2 Laugh Machine

The Laugh Machine project has been implemented in the framework of eNTERFACE'12 and of the ILHAIRE project. The objectives were a) to build an interactive virtual agent able to laugh appropriately (right time, right arousal) when watching a humorous video together with a human being and b) to evaluate the impact of such a laughing agent on the participant's experience. We will only give an overview of the system here. More detailed descriptions can be found in [Urbain *et al.* 2012] and [Niewiadomski *et al.* 2013a].

[2]We have included this work in the Related Works section of this chapter as it does not include acoustic laughter synthesis. Parts or the AVLaughterCycle project have however already been presented in this dissertation: the database in Section 2.5 and the algorithm for evaluating similarities between laughs in Section 4.2.

The interactive laughing agent must be able to exhibit reactions to both the stimulus film and the participant's behavior (in this case, her/his laughter). To estimate the funniness of the film, 13 researchers provided a continuous rating of the film funniness. The average and standard deviation of the scores given at each moment formed the funniness information used in the application.

A range of sensors were included to monitor the participant's behavior: a Kinect, a webcam and a respiration belt. Laughter detection was performed on audio only with Support Vector Machines (SVMs), which returned the participant's laughter likelihood every 200 ms. The other modalities were nevertheless included in the experiments in order to record useful data to develop multimodal laughter detection in the future.

A decision component was trained—with recordings of human dyads watching the same movie—to decide when and how (duration and arousal) to laugh. The inputs of the decision component were the two streams of information (sent every 200 ms): the funniness of the film and the other participant's behavior (laughter likelihood and arousal). A simplified bloc diagram of the application is displayed in Figure 6.1. The Greta agent was then used to display a laugh corresponding to the instructed duration and arousal. The audio was synthesized using the HMM-based method presented in the previous chapter, trained on a female voice (subject #5 of the AVLC database). Laughs were synthesized from the phonetic transcriptions of subject #5. The visual display was derived from the video of the selected laugh in the AVLC database, to animate the agent's facial action parameters. The set of available laughs from subject #5 formed a laughter library inside which the decision component could select an episode corresponding to the desired duration and arousal. Although decisions were taken every 200 ms, laughs could not be interrupted: once a laugh had be chosen to be played, it had to be played until the end.

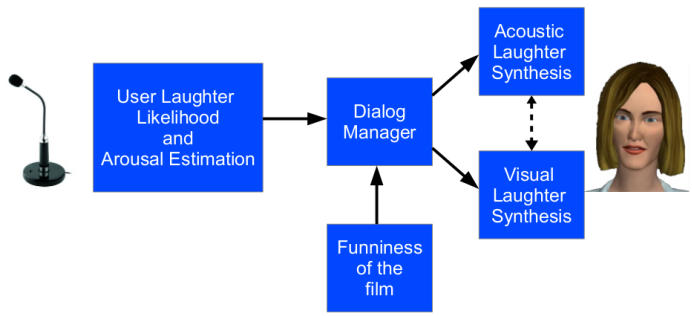

Figure 6.1: Bloc diagram of the Laugh Machine Application.

To evaluate the application, three experimental conditions were designed. One of the conditions is the interactive agent as presented above, i.e. reacting to both the humorous video and the participant's video. Two control conditions were also

implemented: fixed laughter and fixed speech. In the fixed laughter condition, the agent was laughing at eight predefined time points of the 8-minute humorous movie, where there was a peak in the funniness ratings. The participant's behavior had no influence at all on the agent's behavior. In the fixed speech condition, the agent was expressing verbal amusement (e.g., "Oh, that is funny", "I liked that one") at the same fixed time points as in the fixed laughter condition.

Twenty participants took part in a pilot study to evaluate the system, presented in [Niewiadomski *et al.* 2013a]. Eleven were assigned to the interactive condition, four to the fixed speech condition and five to the fixed laughter condition. Participants watched the humorous video in company with the avatar and where then asked to fill in questionnaires about their experience. Due to the small sample sizes, the two fixed conditions were grouped for analyzing the results, as the focus of the work was the impact of the agent's responsiveness.

Results indicate that the interactive agent lead to statistically significant higher felt amusement and higher perceived emotional contagion than the fixed conditions. The feeling of social presence was also higher with the interactive agent, but the difference failed to reach statistical significance.

After building the full processing chain (from laughter detection to synthesis) and following the positive results obtained with the pilot study—which indicated that an interactive laughing agent indeed has an effect on amusement and emotional contagion—, a broader study with 60 new participants was conducted. The results of this study, with deeper analyses (including psychological outcomes) of the impact of the interactive agent should be published soon. The main conclusions are that the interactive laughing agent is able to enhance the participant's mood, in particular for participants who are in a bad mood prior to watching the movie, and is perceived as more natural, believable and human-like than the fixed agents.

One of the limitations of the Laugh Machine application was that laughter was detected from the audio channel only, which prevented from detecting most low-arousal laughs. This problem was addressed in the following application, Laugh When You're Winning.

6.3 Laugh When You're Winning

The eNTERFACE'13 Laugh When You're Winning project was built upon the Laugh Machine project, but with a different scenario. Here, participants were invited to play a game with the virtual agent, who had an active role. The game was played with two participants at a time, in order to explore social behaviors that could not be addressed with the Laugh Machine scenario: cooperation, competition, etc. Laughter can be related to these behaviors (affiliation or malicious laughs). To amplify these social behaviors through laughter, mimicry of one participant's behavior was developed. The aim was to investigate whether mimicry has an effect on the perception of the agent by both participants.

The game was a simple yes/no game in which one participant must avoid saying the words "yes" or "no". Each game involved two human participants as well as the virtual agent. One of the two participants and the virtual agent could say whatever they wanted and were asking questions to the other participant, who had to answer them without saying "yes" or "no". The choice for such a simple yes/no game was motivated by the low game development needs (as the game itself was not the focus of the project, and we had few resources to implement it) while offering a complex scenario in which diverse emotional reactions (enjoyment, frustration, etc.) and social behaviors (cooperation, competition, etc.) could emerge along with laughter. Again, only an overview of the project will be presented here. Extended descriptions can be found in [Mancini *et al.* 2014].

The behavior of each participant was recorded with a head-mounted microphone, a Kinect and a webcam. Green markers were placed on the participants' shoulders to track their movements. A scheme of the application setup is displayed in Figure 6.2. For each participant, information from the different modalities (shoulder tracking with the webcam, facial features from Kinect, audio features from the microphone) was fused to estimate the likelihoods that the participant was smiling, laughing, speaking or silent. One-second frames were used to compute these probabilities, which were updated every 500 ms. Laughter arousal was also estimated from the audio channel, using the algorithm presented in Section 4.4. In addition, the dominant frequency of shoulder oscillations during laughter was computed.

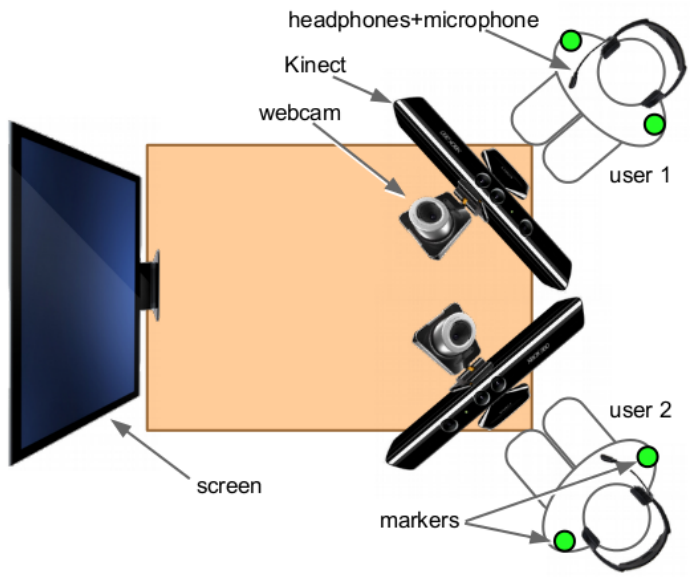

Figure 6.2: Setup of the Laugh When You're Winning game (the agent was displayed on the screen and acting as a third participant).

The decision component used the speaking, smiling and laughing likelihoods from both participants to decide whether the agent should laugh, ask a question or remain silent. Simple politeness rules were implemented: the agent should not speak if a participant is speaking, but it should laugh if someone is laughing. If a period of silence was detected, the agent had to ask a question to try to make the participant fail. From observations of humans playing the game, we realized that streams of questions on the same topic (e.g., "What is your name?", "Is that your full name?", "Really?") were a common strategy to make the participant fail, and such streams of questions were implemented in the agent's list. The agent was asking questions randomly from the list, and was following a stream of questions as long as no speech was detected from its partner (the human participant who could say anything). When laughter was instructed, the decision engine also had to specify the duration and arousal of the laugh. Synthesis laughs were more intense and longer when the cumulative arousal of the participants' laughs was higher. As for the Laugh Machine project, synthesis laughter could not be interrupted. Nevertheless, to increase the number of short laughs in the laughter library, laughs containing several bouts were split in individual bouts by cutting the phonetic transcriptions after inhalation parts. Shorter synthesis laughs artificially made the system more reactive (as it was "frozen" for a shorter duration).

As already explained, one of the objectives of the Laugh When You're Winning project was to explore the effects of mimicry of the agent: would the participants notice if the agent is mimicking the behavior of one participant, would the agent be better perceived by that participant or the other participant? To start investigating this question, it was decided to copy the laughter rhythm of one of the two participants. As already explained, the frequency of shoulder oscillations from the participants was computed for each laugh. One participant was selected to be mimicked and her/his average shoulder frequency was updated each time a laugh had been detected. When laughing, the agent had to match the mimicked participant's rhythm, in addition to display a laugh with the instructed arousal and duration. The rhythm of the synthesis laughs was estimated as the average duration of fricative-vowel syllables in exhalation parts. To increase the range of available rhythms for synthesis laughs, we modified the durations in the phonetic transcriptions: the duration of all the laughter phones was multiplied by a factor X, so as to obtain slower ($X > 1$) or faster ($X < 1$) rhythms. All synthesis modalities (audio, facial movements, shoulder oscillations) were driven by the phonetic transcription, and hence synchronized to the desired rhythm.

Three experimental conditions have been designed to evaluate the impact of the laughing agent as well as the impact of mimicry. In the first condition, the agent is not laughing and can only ask questions. In the second condition, the agent can speak and laugh, but there is no mimicry (the rhythm of the laughs is not modified and corresponds to the modeled human voice). In the last condition, mimicry is added. To the best of our knowledge, this is the first time laughter is altered to match a participant's features (here rhythm) in a human-computer interaction. The game

application, albeit simple, is also far more complex than previously explored scenarios, as it introduces social dimensions that could not be explored with a laughing virtual agent in the past.

Participants were recruited in pairs to play the game together with the agent. Participants went through the three conditions in random order. For each condition, two games were played as participants took turns in answering questions. Each game lasted a maximum of one minute. If the answering participant failed within one minute, the questioning participant was asked to click on a mouse and the game was stopped. In that case of a lost game, the agent was acknowledging it either verbally, in the non-laughing condition, or by laughing in the other two conditions, as it was noticed from humans playing the game that laughter was consistently appearing when a game was lost. If the answering participant could hold one minute without saying "yes" or "no" (or the questioning participant did not notice it), the agent verbally congratulated the answering participant for having won the game.

Prior to the first game and after each condition, participants were asked to fill in mood questionnaires and questionnaires about the previous set of games.

Nine pairs of participants played the game in the pilot experiment. Results and participants' comments revealed some technical flaws that hindered the expected effects of the laughing agent. For instance, the non-laughing agent was perceived as a more competent game player than the laughing agent. This was likely caused by false laughter detection alarms due to the open setting: the agent's laughter was rendered through loudspeakers, got back to the participants' microphones, laughter was then detected and the agent was instructed to laugh again. This created laughter loops where the agent was laughing over and over again without apparent reason for the participants, which was disturbing and going against any game strategy. Other flaws were identified and are currently corrected for further experiments. Data is also collected to train a dialog manager on this specific scenario, and replace the implemented rule-based decision component. Nevertheless, the laughing agent was generally perceived as more natural than the non-laughing agent (see Figure 6.3) and the non-verbal behavior of the mimicry agent was statistically significantly better rated than that of the non-laughing agent (see Figure 6.4). Mimicry itself had limited effects in this pilot study, partly because of interaction flaws, but probably also due to the short experiment durations, where mimicry would at best occur for two minutes.

The biggest limitation of the scenario related to laughter synthesis is the impossibility to control the synthesis in real-time. This restricts both reactivity of the agent and mimicry options. This is hence the major development to address in the future.

6.4 Laughter variations for perceptive experiments

As introduced in the previous section, our HMM-based process enables us to easily modify some properties of the synthesized laughs (e.g., rhythm). This opens the possibility to investigate which dimensions influence laughter perception, in a similar way

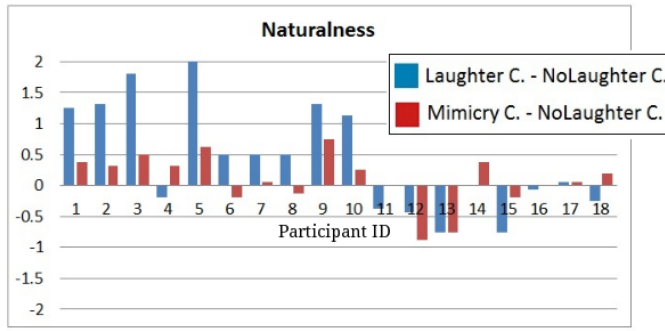

Figure 6.3: Difference in the naturalness ratings for the three conditions. For each participant (numbered #1 to #18), the differences in naturalness ratings between the laughing and non-laughing conditions is represented in blue (left bar) and between the mimicry and the non-laughing conditions is shown in red (right bar).

to the experiments on laughter rhythm and pitch conducted a decade ago by Kipper and Todt (see Section 3.1.6.4), with concatenative synthesis and a phase vocoder.

The objectives are to understand which acoustic properties can make laugh be perceived as friendly or malicious. This is of general interest for human-computer interaction, as presented in the previous sections, in order to know which kind of laugh to display to enhance participant's mood, show affiliation, etc. But it is particularly important for a better understanding of gelotophobia, also known as "the fear of being laughed at". Gelotophobes have the tendency to misperceive any laugh as malicious laugh (laughing at them) [Ruch & Proyer 2008]. Identifying which features (if any) could lead to laughter being (better) accepted by gelotophobes would help treating the disease.

Four dimensions—related to the features investigated by Kori [Kori 1987] and Kipper and Todt [Kipper & Todt 2001, Kipper & Todt 2003], see Section 3.1—have been selected: rhythm, f_0, energy and number of syllables. Rhythm was modified in the same way as explained above, by changing the durations of the phonetic transcriptions prior to synthesis. This method was preferred to changing the playing speed of synthesized laughs (with a phase vocoder to preserve the frequencies) in order to keep the HMM modeling of the excitation and filter trajectories.

To modify f_0, we used the pitch trajectories generated by our HMMs. The target f_0 pattern can simply be replaced before building the excitation signal and attacking the filter. Here, we have decided to amplify the deviations from the average f_0 of the synthesized laugh. The modified f_0 value at time t, $f_0^m(t)$, is obtained from the initial f_0 value for synthesis $f_0(t)$, the average f_0 value $\bar{f}_0 = \frac{1}{T}\sum_{t=1}^{T} f_0(t)$ and the multiplying factor f by:

$$f_0^m(t) = \bar{f}_0 + f * (f_0(t) - \bar{f}_0) \tag{6.1}$$

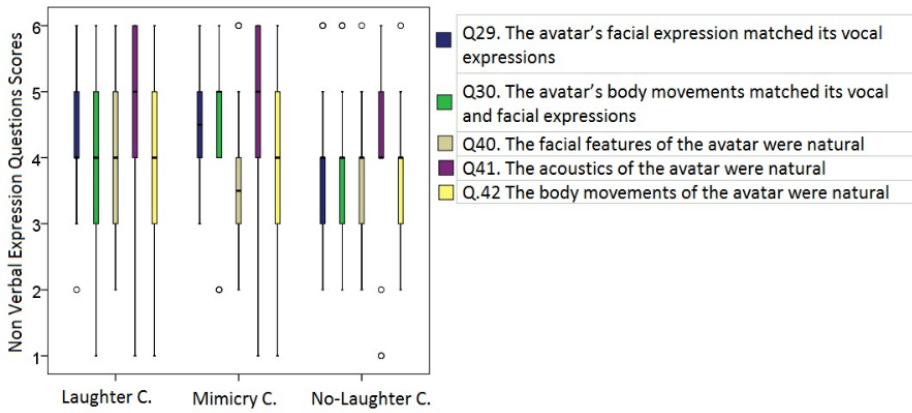

Figure 6.4: Box plots of the ratings received in the three experimental conditions for the questions related to the non-verbal behavior of the agent.

If $f = 0$, we obtain a totally flat f_0 profile, if $0 < f < 1$ we attenuate the f_0 variations, while we amplify them if $f > 1$.

The energy profile is modified by linearly weighting the amplitude of the synthesized laughs. The weighting factor is either decreasing or increasing linearly over the laugh episode, with a maximum value of 1 (at the beginning or end of the laugh, respectively) and a minimal value of I (<1) (at the end or beginning of the laugh, respectively). This enables to attenuate the attack of the laugh (increasing patterns) or obtain a more gently fading out of the laugh (decreasing patterns).

To vary the number of syllables, we looked for the longest series of "fricative-vowels" syllables (FV series) in the human laugh transcription. To obtain laughs with smaller numbers of syllables, we deleted syllables in the series in a uniform way: if only 1 syllable has to be deleted, we select the middle syllable from the human FV series; if 2 syllables have to be dropped, we take them at $1/3$ and $2/3$ of the human FV series; if 3 syllables are removed, we take them around $1/4$, $1/2$ and $2/3$ of the human FV series. To obtain laughs with higher numbers of syllables, we added syllables in the FV series in a uniform way. The duration of the inserted syllable is the average between the durations of the preceding and following syllables, while the vowel of the inserted syllable is copied from the preceding syllable.

Examples of modifications of the number of syllables, rhythm and f_0 are available from `http://www.tcts.fpms.ac.be/~urbain/laugh_synthesis_examples/`, for a female and a male voice, with or without corresponding animations of a virtual agent. All four modifications are currently used in psychological experiments with gelotophobes. Some of the laughs are easily perceived as fake, for example when the rhythm is very slow, and can be considered as malicious by certain people.

The variations introduced so far (rhythm, pitch, number of syllables and

energy profile) were also accessible to Kipper and Todt [Kipper & Todt 2001, Kipper & Todt 2003]. Our aim is to push evaluation and modifications forward (e.g., they always used the same number of syllables) as well as to explore new ways of manipulating these dimensions (e.g., amplifying the f_0 trajectories modeled by our HMMs). Possibilities are numerous and will be refined when evaluation results become available. The evaluation and identification of laughter features that create particular perception for gelotophobes and control subjects is indeed only at its beginning, but will be carried out in the future in collaboration with psychologists from the University of Zürich.

6.5 Conclusions of the applications using laughter synthesis

In this chapter, we have browsed a few applications integrating laughter, and in particular some projects where our HMM-based audio laughter synthesis methods have been used. Although the results are still limited due to the small sample sizes reported here, it has generally been showed that a laughing agent has an impact on the participants' experience. The development of such laughing agents should hence continue, to further investigate what they can bring in a range of situations (companions, game players, coaches, etc.) and understand when and how laughter can be beneficial in human-computer interactions. We have also seen that acoustic parameters of laughs can easily be modified, which will help to better understand the impact of laughter features, either in isolation or within interactions.

The Laugh When You're Winning application is the first human-computer interaction in which we altered the laughter synthesis in order to match the participant's behavior. Here we only played with the laughter rhythm to create an effect of mimicry. However laughs are so far not interruptible: a laugh is selected for synthesis and must be played until its end. Laughter-enabled human-computer interactions would clearly benefit from the possibility to modify laughs on the fly, in order to be more reactive, as humans do: laughter is amplified if the funny stimulus is persisting, while we would try to suppress it if we realize it is inappropriate or the situation suddenly becomes serious.

Another application we are currently working on involves a revised version of the laughter similarity algorithm, including features derived from the automatic phonetic transcriptions (rhythm, number of bouts, types of phones involved), for organizing laughter databases and easily browse through them by acoustic similarities.

Chapter 7

Conclusion

This document has presented several innovative research areas in acoustic laughter processing. For each of these areas, a significant contribution to the state-of-the-art has been brought during this PhD Thesis.

The AVLaughterCycle database itself, with its phonetic annotations, is the first contribution. Even though it does not really involve signal processing, we believe this database (and the proposed annotations) is bringing new insights for laughter analysis and processing. The database has already been used for several studies—most of which have been reported on in this dissertation, but there have also been analyses of facial movements, which were beyond the scope of the present work. In addition, the AVLC database was also a starting point for the new generation of laughter databases, trying to overcome the limitations of existing corpora. For instance, the MAHNOB laughter database is largely inspired by our database, but includes speech segments and (limited) social interactions, which were both absent in the AVLaughterCycle recordings. Recordings done in Belfast in the framework of the ILHAIRE project also aim at increasing the amount of laughter data available from single speakers, which is another limitation of the AVLC database: although state-of-the-art laughter synthesis could be developed with our data, it is believed that more data for a single voice would further improve the synthesis quality. This was also the motivation for recording the AV-LASYN database, which is specifically targeting audiovisual laughter synthesis.

Second, annotations of laughter arousal, both at the frame and episode levels, have been performed for the first time. Arousal is probably the characterizing dimension of laughter that receives the biggest consensus, as attempts to identify laughter types have so far failed to gather convincing proofs and reach agreement. We are thus expecting that annotations and analyses of arousal will gain interest in the near future.

Third, the proposed phonetic annotations of the AVLC database yielded to original laughter analysis, since no one before could report on phone variations, for instance. Differences in phone durations have been investigated. Individual differences in the choice of phones have also been highlighted, to confirm one of the key messages of laughter-related studies: laughter is a highly variable signal, hence difficult to analyze and model.

Fourth, we have proposed methods to automatically obtain the two annotations tracks introduced in this PhD Thesis. For estimating arousal at the frame level, we have used neural networks fed with acoustic features. Good correlation with manual annotations could be obtained, in a speaker-independent scheme. Arousal at the

laughter episode level can be computed from simple statistics of the per-frame arousal trajectory: another neural network achieved good correlation with the reference per-laugh arousal values, again in a speaker-independent scheme. Laughter phonetic transcriptions were automatically obtained with the help of Hidden Markov Models. Acceptable matching with the manual transcriptions could be achieved, which was later confirmed by the satisfying quality of laughter synthesized with models trained on the automatic phonetic transcriptions.

Fifth, we have adapted HMM-based speech synthesis to laughter. Our method achieves state-of-the-art results, even though competing methods—targeted to specific voiced laughter structures—have been simultaneously developed. Besides offering the possibility to synthesize any laughter sounds (inhalations, nareal fricatives, etc.) as long as there is sufficient training data, the proposed HMM-based laughter synthesis seems to us relatively convenient to use: either from phonetic transcriptions (without the need to also specify the duration of each phone, which can be efficiently estimated by the HMMs) or from high-level arousal information, using our phonetic generation method.

Finally, we have already implemented several other applications, like laughter-enabled virtual agents or manipulation of laughter features, that can interest a broader community. Laughing virtual agents can be used in various applications like games, educational companions or health coaches. We have also seen that unique psychological experiments can be conducted thanks to laughter synthesis. Laughter is also of interest for some artistic installations, as it is naturally conveying affect. All these applications will benefit from every improvement in the points addressed above. Let us recall some of the directions for future research in the next paragraphs, in correspondence to the major contributions summarized above.

First, more laughter data would obviously be beneficial. We have seen that there already exists a reasonable number of corpora containing laughter. In addition, the efforts are currently growing and tackling limitations of current databases. We can thus expect that large amounts of high-quality laughter data will be available in the next years.

Second, new annotations are required, not only for using the (new) laughter data, but also to assess the validity of proposed dimensions to characterize the laughs. Arousal seems to have been universally used for centuries to describe laughs, but it would be really interesting to know whether people agree on annotating per-frame arousal. Annotations on different modalities (audio only, video only, both audio and video, only face, face plus body movements, etc.) would also help to identify which information is contributing to arousal. For phonetic transcriptions also, measures of inter-rater agreements would be useful to validate the set of phones used, for example. Furthermore, it would be nice to have annotations to study cross-cultural variations (both in the laughs and in the way to annotate it) or how people adapt their annotations once they know the laugher (e.g., rating arousal relatively to laughs already seen from the same laugher).

Third, coming back to the phonetic annotations, we have proposed a two-step

process here: first annotating with a lot of details, then grouping acoustically close labels. The advantage of the first step is that it enables different groupings without needing to re-annotate the whole dataset. Nevertheless, the groupings we have used in this dissertation are tentative ones, resulting from our own observations and needs. Assessing which laughter phones actually make a perceptive difference—or at least which grouping is truly optimal for the different objectives of describing and synthesizing laughs—is another body of work that could emerge from this dissertation and feed back to all our developments.

Fourth, there is obviously room for improving our methods for automatic phonetic transcriptions and arousal estimation. One path is to investigate new features, like the slope of f_0 for arousal, descriptors of breathiness or nasality for phonetic transcriptions, etc. Another option would be to include contextual information to take decisions that are less local, especially for arousal estimation. Phonetic transcriptions are currently computed using the full laughter episode, but each frame is characterized by local features only, and only bigrams have been implemented to constrain the sequence of phones. Exploring higher-order n-gram models is another research direction. As already discussed, adaptation to the speaker is also expected to be beneficial, both for estimating arousal and phonetic transcriptions, as laughter is exhibiting high variability between subjects.

Fifth, the whole synthesis process could be studied point by point and tuned to laughter. It should not be forgotten that the used methods have originally been designed for speech and have been experimentally adapted to laughter in this dissertation. Optimal parameters should be looked for, and one must keep in mind that some of them (like f_0 boundaries) should probably be tuned to the individual laugher. Further analyses of the vocoders could also help, and the development of a vocoder that would account for laughter specificities (breathiness, vocal folds not closing as decisively as in speech, etc.) would be an interesting PhD topic.

Furthermore—although we already consider the generation algorithm as an important step forward for promoting the use of laughter synthesis in a range of environments (as it is extremely easy to drive laughter synthesis from arousal curves)—the generation method should be further evaluated to assess whether it can work for different voices and arousal patterns.

Finally, applications integrating laughter—including human-computer interactions, artistic installations, etc.—would definitely benefit from real-time implementations of laughter characterization and synthesis processes. Unfortunately, all algorithms cannot be applied to real-time in a straightforward way or without degrading the performance. Implementing real-time versions of all our algorithms that are not real-time yet (in particular automatic phonetic transcriptions and laughter synthesis) would be a first step towards increased reactivity of applications. It is however expected that the actual research efforts will be to optimize the algorithms so that they do not suffer too much from real-time constraints (limited context, restricted computational time, etc.).

To conclude this PhD, we would like to recall the importance of laughter in hu-

man communication. Laughter is an essential social signal, generally transmitting "socially-positive" information (politeness, affiliation, amusement). It can also convey negative feelings, either intentionally (malicious laughs, bullying, etc.) or as a result of misinterpretation (and not only by gelotophobes). This is why we were motivated in advancing the state-of-the-art in laughter processing, to enable the integration of laughter in human-computer interactions and increase their naturalness, but also as a means of studying the complex role of laughter in human communication. Numerous fields are concerned by laughter (medicine, psychology, communication, engineering, etc.). It makes the challenging tasks of laughter processing even more fun and interesting to deal with. We hope this dissertation will contribute to promote these ideas.

Bibliography

[Alku 1992] Paavo Alku. *Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering.* Speech Communication, vol. 11, no. 2–3, pages 109–118, 1992. 142

[Amazon.com, Inc. 2014] Amazon.com, Inc. *Amazon Mechanical Turk.* https://www.mturk.com/mturk/welcome, Consulted on February 26, 2014. 58

[AMI project 2011] AMI project. *The AMI Meeting Corpus.* http://corpus.amiproject.org/, Consulted on June 1, 2011. 22, 23

[An *et al.* 2013] Gouzhen An, David Guy Brizan and Andrew Rosenberg. *Detecting Laughter and Filled Pauses Using Syllable-based Features.* In Proceedings of the 14^{th} Annual Conference of the International Speech Communication Association (INTERSPEECH), pages 178–181, Lyon, France, 2013. 96

[Anderson *et al.* 1991] Anne H. Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller*et al.* *The HCRC map task corpus.* Language and speech, vol. 34, no. 4, pages 351–366, 1991. 19

[Astrinaki *et al.* 2013] Maria Astrinaki, Nicolas D'Alessandro, Loic Reboursière, Alexis Moinet and Thierry Dutoit. *Mage 2.0: New features and its application in the development of a talking guitar.* In Proceedings of the 13^{th} Conference on New Interfaces for Musical Expression (NIME'13), Daejon and Seoul, Korea Republic, 2013. IEEE. 150

[Bachorowski & Owren 2001] Jo-Anne Bachorowski and Michael J. Owren. *Not all laughs are alike: Voiced but not unvoiced laughter readily elicits positive affect.* Psychological Science, vol. 12, no. 3, pages 252–257, 2001. 59

[Bachorowski *et al.* 2001] Jo-Anne Bachorowski, Moria J. Smoski and Michael J. Owren. *The acoustic features of human laughter.* Journal of the Acoustical Society of America, vol. 110, no. 3, pages 1581–1597, September 2001. 3, 25, 44, 49, 50, 51, 52, 53, 54, 55, 60, 68, 69, 73, 101, 107

[Baker & Hazan 2011] Rachel Baker and Valerie Hazan. *DiapixUK: task materials for the elicitation of multiple spontaneous speech dialogs.* Behavior research methods, vol. 43, no. 3, pages 761–770, 2011. 19

[Becker-Asano & Ishiguro 2009] Christian Becker-Asano and Hiroshi Ishiguro. *Laughter in Social Robotics - no laughing matter.* In Proceedings of the International Workshop on Social Intelligence Design (SID2009), pages 287–300, 2009. 162

[Becker-Asano *et al.* 2009] Christian Becker-Asano, Takayuki Kanda, Carlos Ishi and Hiroshi Ishiguro. *How about laughter? Perceived naturalness of two laughing humanoid robots.* In Proceedings of Affective Computing and Intelligent Interaction, pages 49–54, 2009. 162

[Beller 2009] Grégory Beller. *Analysis and Generative Model for Expressivity. Applied to Speech and Musical Performance.* PhD thesis, June 2009. 126

[Bennett & Lengacher 2006a] Mary P. Bennett and Cecile Lengacher. *Humour and Laughter may Influence Health. I. History and Background.* Evidence-based Complementary and Alternative Medicine, vol. 3, no. 1, pages 61–63, 2006. 7, 9

[Bennett & Lengacher 2006b] Mary P. Bennett and Cecile Lengacher. *Humour and Laughter may Influence Health. II. Complementary Therapies and Humor in a Clinical Population.* Evidence-based Complementary and Alternative Medicine, vol. 3, no. 2, pages 187–190, 2006. 7

[Bennett & Lengacher 2008] Mary P. Bennett and Cecile Lengacher. *Humour and Laughter may Influence Health. III. Laughter and Health Outcomes.* Evidence-based Complementary and Alternative Medicine, vol. 5, no. 1, pages 37–40, 2008. 7, 9

[Bennett & Lengacher 2009] Mary P. Bennett and Cecile Lengacher. *Humor and laughter may influence health IV. humor and immune function.* Evidence-Based Complementary and Alternative Medicine, vol. 6, no. 2, pages 159–164, 2009. 7, 9, 10

[Bennett *et al.* 2003] Mary P. Bennett, Janice M. Zeller, Lisa Rosenberg and Judith McCann. *The effect of mirthful laughter on stress and natural killer cell activity.* Nursing Faculty Publications, page 9, 2003. 10

[Berk *et al.* 1989] Lee S. Berk, Stanley A. Tan, Barbara J. Napier and William C. Eby. *Eustress of mirthful laughter modifies natural killer cell activity.* Clinical Research, vol. 37, no. 1, page 115A, 1989. 9

[Bickley & Hunnicutt 1992] Corine A. Bickley and Sheri Hunnicutt. *Acoustic analysis of laughter.* In Proceedings of the 2^{nd} International Conference on Spoken Language Processing (ICSLP), pages 927–930, 1992. 54, 61

[Boersma & Weenink 2011] Paul Boersma and David Weenink. *Praat: doing phonetics by computer (Version 5.2.11) [computer program].* `http://www.praat.org`, Retrieved on January 20, 2011. 27, 67, 68

[Bollepalli *et al.* 2014] Bajibabu Bollepalli, Jérôme Urbain, Tuomo Raitio, Joakim Gustafson and Hüseyin Çakmak. *A Comparative Evaluation of Vocoding Techniques for HMM-based Laughter Synthesis.* In Proceedings of the International

Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, May 2014. 140

[Bright *et al.* 1986] KE Bright, TJ Hixon and JD Hoit. *Respiration as a laughing matter*. WHIMSY IV, pages 147–148, 1986. 63

[Brown *et al.* 1992] Peter F. Brown, Peter V. Desouza, Robert L. Mercer, Vincent J. Della Pietra and Jenifer C. Lai. *Class-based n-gram models of natural language*. Computational linguistics, vol. 18, no. 4, pages 467–479, 1992. 206

[Cagampan *et al.* 2013] Bernadyn Cagampan, Henry Ng, Kevin Panuelos, Kyrstyn Uy, Jocelynn Cu and Merlin Suarez. *An Exploratory Study on Naturalistic Laughter Synthesis*. In Proceedings of the 4^{th} International Workshop on Empathic Computing (IWEC'13), Beijing, China, 2013. 28, 128, 139

[Cai *et al.* 2003] Rui Cai, Lie Lu, Hong-Jiang Zhang and Lian-Hong Cai. *Highlight sound effects detection in audio stream*. In Proceedings of the 2003 IEEE International Conference on Multimedia and Expo (ICME), pages 37–40, Baltimore, USA, 2003. 93

[Campbell 2007] Nick Campbell. *Whom we laugh with affects how we laugh*. In Proceedings of the Interdisciplinary Workshop on the Phonetics of Laughter, pages 61–65, Saarbrücken, Germany, August 2007. 5, 18, 37, 53, 102, 119

[Campbell 2009] Nick Campbell. *Tools and resources for visualising conversational-speech interaction*. In Multimodal corpora, pages 176–188. Springer, 2009. 23

[Campbell 2011] Nick Campbell. *Nick's Data website*. http://www.speech-data.jp/, Consulted on June 5, 2011. 18

[Campbell 2014] Nick Campbell. *Conversation Data in English*. http://www.speech-data.jp/taba/nov07/index.html, consulted on February 17, 2014. 23

[Carletta 2007] Jean Carletta. *Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus*. Language Resources and Evaluation Journal, vol. 41, no. 2, pages 181–190, 2007. 22, 37

[Çakmak *et al.* 2014] Hüseyin Çakmak, Jérôme Urbain, Joëlle Tilmanne and Thierry Dutoit. *The AV-LASYN Database: A synchronous corpus of audio and 3D facial marker data for audio-visual laughter synthesis (accepted for publication)*. In Proceedings of the 9^{th} conference on International Language Resources and Evaluation (LREC), Reykjavik, Iceland, May 2014. 29

[Chafe 2007] Wallace Chafe. The importance of not being earnest. The feeling behind laughter and humor., volume 3 of *Consciousness & Emotion Book Series*.

John Benjamins Publishing Company, Amsterdam, The Nederlands, paperback 2009 edition, 2007. 1, 3, 4, 5, 6, 18, 43, 44, 46, 48, 49, 51, 53, 54, 55, 61, 62, 68, 69, 72

[Club de rire Asbl 2008] Club de rire Asbl. *The official website of laughter clubs in Belgium.* `http://www.clubderire.be/`, Consulted on September 10, 2008. 9

[Cohen 1960] Jacob Cohen. *A coefficient of agreement for nominal scales.* Educational and Psychological Measurement, vol. 20, pages 37–46, 1960. 64

[Curran & McKeown 2013] Will Curran and Gary McKeown. *ILHAIRE project, Deliverable 1.4: Mono-Cultural database of Conversational Laughter*, 2013. 24, 27

[D'Alessandro & Dutoit 2007] Nicolas D'Alessandro and Thierry Dutoit. *HandSketch Bi-Manual Controller: Investigation on Expressive Control Issues of an Augmented Tablet.* In Proceedings of the International Conference on New Interfaces for Musical Expression (NIME), pages 78–81, New York City, USA, June 2007. 130

[D'Alessandro *et al.* 2013] Nicolas D'Alessandro, Christophe d'Alessandro, Lional Feugère, Maria Astrinaki, Johnty Wang and Olivier Perrotin. *Vox Tactum Meets Chorus Digitalis: Seven Years of Singing Surfaces.* In Proceedings of the International Conference on New Interfaces for Musical Expression (NIME) (to appear), Daejon/Seoul, South Korea, May 2013. 130

[Darwin 1872] Charles Darwin. *Chapter 8: Joy, High Spirits, Love, Tender Feelings, Devotion.* In The expression of the emotions in man and animals, pages 196–219. New York: D. Appleton & Company, 1872. 43, 62, 63

[De Benedictis 2007] Marianna De Benedictis. *Psychological and cross-cultural effects on laughter sound production.* In Proceedings of the Interdisciplinary Workshop on the Phonetics of Laughter, pages 31–35, Saarbrücken, Germany, August 2007. 59

[Devillers & Vidrascu 2007] Laurence Devillers and Laurence Vidrascu. *Ensemble methods for spoken emotion recognition in call-centres.* In Proceedings of the Interdisciplinary Workshop on the Phonetics of Laughter, pages 37–40, Saarbrücken, Germany, August 2007. 1, 5, 6, 16, 18, 56, 65

[Douglas-Cowie *et al.* 2003] Ellen Douglas-Cowie, Nick Campbell, Roddy Cowie and Peter Roach. *Emotional Speech: Towards a new generation of databases.* Speech Communication, vol. 40, no. 1, pages 33–60, 2003. 16, 22

[Douglas-Cowie *et al.* 2007] Ellen Douglas-Cowie, Roddy Cowie, Ian Sneddon, Cate Cox, Orla Lowry, Margaret Mcrorie, Jean-Claude Martin, Laurence Devillers, Sarkis Abrilian, Anton Batliner*et al. The HUMAINE database: addressing the*

collection and annotation of naturalistic and induced emotional data. In Proceedings of the 2^{nd} International Conference on Affective Computing and Intelligent Interaction (ACII), pages 488–500, Lisbon, Portugal, 2007. Springer. 22, 24

[Drahota *et al.* 2008] Amy Drahota, Alan Costall and Vasudevi Reddy. *The vocal communication of different kinds of smile.* Speech Communication, vol. 50, pages 278–287, 2008. 16

[Drugman & Alwan 2011] Thomas Drugman and Abeer Alwan. *Joint Robust Voicing Detection and Pitch Estimation Based on Residual Harmonics.* In Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Firenze, Italy, August 2011. 77, 109, 116

[Drugman & Dutoit 2012] Thomas Drugman and Thierry Dutoit. *The Deterministic plus Stochastic Model of the Residual Signal and its Applications.* IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, pages 968–981, 2012. 132, 140, 141

[Drugman *et al.* 2011] Thomas Drugman, Thomas Dubuisson and Thierry Dutoit. *Phase-based information for voice pathology detection.* In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pages 4612–4615. IEEE, 2011. 77, 109, 116

[Drugman *et al.* 2013] Thomas Drugman, Jérôme Urbain, Nathalie Bauwens, Ricardo Chessini, Carlos Valderrama, Patrick Lebecque and Thierry Dutoit. *Objective Study of Sensor Relevance for Automatic Cough Detection.* IEEE Transactions on Information Technology in BioMedicine, 2013. 77, 109, 116

[Du Bois & Englebretson 2004] John W. Du Bois and Robert Englebretson. *Santa Barbara Corpus of Spoken American English, Part 3.* University of Pennsylvania: Linguistic Data Consortium, 2004. 18

[Du Bois & Englebretson 2005] John W. Du Bois and Robert Englebretson. *Santa Barbara Corpus of Spoken American English, Part 4.* University of Pennsylvania: Linguistic Data Consortium, 2005. 18

[Du Bois *et al.* 2000] John W. Du Bois, Wallace L. Chafe, Charles Meyer and Sandra A. Thompson. *Santa Barbara Corpus of Spoken American English, Part 1.* University of Pennsylvania: Linguistic Data Consortium, 2000. 18

[Du Bois *et al.* 2003] John W. Du Bois, Wallace L. Chafe, Charles Meyer, Sandra A. Thompson and Nii Martey. *Santa Barbara Corpus of Spoken American English, Part 2.* University of Pennsylvania: Linguistic Data Consortium, 2003. 18

[Dupont *et al.* 2009a] Stéphane Dupont, Thomas Dubuisson, John Anderson Mills III, Alexis Moinet, Xavier Siebert, Damien Tardieu and Jérôme Urbain. *LaughterCycle.* In Thierry Dutoit and Benoît Macq, editors, QPSR of the numediart research program, volume 2, pages 23–31. numediart, 6 2009. 104

[Dupont *et al.* 2009b] Stéphane Dupont, Thomas Dubuisson, Jérôme Urbain, Christian Frisson, Raphaël Sebbe and Nicolas D'Alessandro. *AudioCycle: Browsing Musical Loop Libraries.* In Proc. of IEEE Content Based Multimedia Indexing Conference (CBMI09), Chania, Greece, June 2009. 104, 105

[Edmonson 1987] Munro S. Edmonson. *Notes on laughter.* Anthropological linguistics, pages 23–34, 1987. 34, 43, 50, 51, 52, 58, 62, 63, 68, 69, 76, 115

[Ekman *et al.* 2002] Paul Ekman, Wallace V. Friesen and Joseph C. Hager. *Facial action coding system: A technique for the measurement of facial movement.* Consulting Psychologists Press, Palo Alto, CA, 2002. 27, 63

[Ekman 1992] Paul Ekman. *Are there basic emotions?* Psychological Review, vol. 99, pages 550–553, 1992. 29

[Ekman 2003] Paul Ekman. Emotions revealed: Recognizing faces and feelings to improve communication and emotional life. Times Books/Henry Holt and Co, 2003. 27

[Ellis & Poliner 2007] Daniel P.W. Ellis and Graham E. Poliner. *Identifying cover songs' with chroma features and dynamic programming beat tracking.* In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), volume 4, pages IV–1429. IEEE, 2007. 77, 109, 116

[Esling 2007] John H. Esling. *States of the larynx in laughter.* In Proceedings of the Interdisciplinary Workshop on the Phonetics of Laughter, pages 15–20, Saarbrücken, Germany, August 2007. 29, 53, 61

[Eyben *et al.* 2010] Florian Eyben, Martin Wöllmer and Björn Schuller. *Opensmile: the munich versatile and fast open-source audio feature extractor.* In Proceedings of the international conference on Multimedia, pages 1459–1462, Florence, Italy, 2010. ACM. 94

[Fant *et al.* 1985] Gunnar Fant, Johan Liljencrants and Qi-guang Lin. *A four-parameter model of glottal flow.* STL-QPSR, vol. 4, no. 1985, pages 1–13, 1985. 84

[Ferner & Aronson 2013] R. E. Ferner and J. K. Aronson. *Laughter and MIRTH (Methodical Investigation of Risibility, Therapeutic and Harmful): narrative synthesis.* BMJ: British Medical Journal, vol. 347, 2013. 8, 9, 11

[Filippelli *et al.* 2001] Mario Filippelli, Riccardo Pellegrino, Iacopo Iandelli, Gianni Misuri, Joseph R. Rodarte, Roberto Duranti, Vito Brusasco and Giorgio Scano. *Respiratory dynamics during laughter.* Journal of Applied Physiology, vol. 90, no. 4, page 1441, 2001. 47

[Fleiss *et al.* 1981] Joseph L. Fleiss, Bruce Levin and Myunghee Cho Paik. *The measurement of interrater agreement.* Statistical methods for rates and proportions, vol. 2, pages 212–236, 1981. 149

[Foot & Chapman 1976] Hugh C. Foot and Antony J. Chapman. *The social responsiveness of young children in humorous situations.* Humor and laughter: Theory, research, and applications, pages 187–214, 1976. 62

[Free Dictionary 2008] The Free Dictionary. *Online Dictionary, Encyclopedia and Thesaurus. Free access.* http://www.thefreedictionary.com/, Consulted on September 19, 2008. 2

[Fry 1994] William F. Fry. *The biology of humor.* Humor: International Journal of Humor Research, vol. 7, pages 111–126, 1994. 4, 9

[Fukushima *et al.* 2010] Shogo Fukushima, Yuki Hashimoto, Takashi Nozawa and Hiroyuki Kajimoto. *Laugh enhancer using laugh track synchronized with the user's laugh motion.* In Proceedings of the 28^{th} of the international conference on Human factors in computing systems (CHI'10), pages 3613–3618, 2010. 162

[Glenn 2003] Phillip J. Glenn. Laughter in interaction. Cambridge University Press, Cambridge, 2003. 5, 6, 7, 24, 46, 52, 62, 63, 66, 67

[Grammer & Eibl-Eibesfeldt 1990] Karl Grammer and Irenaus Eibl-Eibesfeldt. *The ritualisation of laughter.* Natürlichkeit der Sprache und der Kultur, pages 192–214, 1990. 59

[Greenfield 2002] Pete Greenfield. *The Role of Laughter in the Good Life: A Philosophical Examination.* In Sewanee Senior Essays on Philosophy. 2002. 61

[Gupta *et al.* 2013] Rahul Gupta, Kartik Audhkhasi, Sungbok Lee and Shrikanth Narayanan. *Speech Paralinguistic Event Detection Using Probabilistic Time-Series Smoothing and Masking.* In Proceedings of the 14^{th} Annual Conference of the International Speech Communication Association (INTERSPEECH), pages 173–177, Lyon, France, 2013. 97

[Hall & Alliń 1897] G. Stanley Hall and Arthur Alliń. *The psychology of tickling, laughing, and the comic.* The American Journal of Psychology, vol. 9, no. 1, pages 1–41, 1897. 59

[Hall *et al.* 2009] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann and Ian H. Witten. *The WEKA data mining software: an*

update. ACM SIGKDD Explorations Newsletter, vol. 11, no. 1, pages 10–18, 2009. 116

[Hall 1998] Mark A. Hall. *Correlation-based Feature Subset Selection for Machine Learning*. PhD thesis, University of Waikato, Hamilton, New Zealand, 1998. 116

[Haykin 1994] Simon Haykin. Neural networks: a comprehensive foundation. Prentice Hall PTR, 1994. 89

[Hillenbrand *et al.* 1995] James Hillenbrand, Laura A. Getty, Michael J. Clark and Kimberlee Wheeler. *Acoustic characteristics of American English vowels*. Journal of the Acoustical Society of America, vol. 97, no. 5, pages 3099–3111, May 1995. 50

[Hofmann *et al.* 2011] J. Hofmann, F. Stoffel, A. Weber and T. Platt. *The 16 Enjoyable emotions induction task (16 - EEIT)*. Unpublished Research instrument, Department of Psychology, University of Zurich, Switzerland, 2011. 27

[Hofmann *et al.* 2013] Jennifer Hofmann, Tracey Platt, Willibald Ruch, Gary McKeown, Will Curran, Lesley Storey, Nadia Berthouze and Harry Griffin. *ILHAIRE project, Deliverable 5.3: Quantitative and Qualitative Classification of and Responses to Laughter Stimuli*, 2013. 57, 62, 63, 64

[Hsu & Glass 2008] Bo-June Hsu and James Glass. *Iterative language model estimation: efficient data structure & algorithms*. In Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH), volume 8, pages 1–4, 2008. 150

[HUMAINE 2008] Association HUMAINE. *The HUMAINE Portal, emotional databases page*. http://emotion-research.net/wiki/Databases, Consulted on September 11, 2008. 17

[Hunt & Black 1996] Andrew J. Hunt and Alan W. Black. *Unit selection in a concatenative speech synthesis system using a large speech database*. In Proceedings of the 1996 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 373–376, Atlanta, Georgia, 1996. 150

[Imperial College London 2014] Imperial College London. *MAHNOB Laughter Database*. http://mahnob-db.eu/laughter/, consulted on February 17, 2014. 27

[International Clinical Phonetics and Linguistics Association (ICPLA) 2008] International Clinical Phonetics and Linguistics Association (ICPLA). *Extended International Phonetic Alphabet for Disordered Speech*. http://www.langsci.ucl.ac.uk/ipa/extIPAChart2008.pdf, (Consulted on April 24, 2014), 2008. 68

[International Phonetic Association 1999] International Phonetic Association. Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet. Cambridge University Press, 1999. 52, 67, 109, 110, 134

[Janicki 2013] Artur Janicki. *Non-linguistic Vocalisation Recognition Based on Hybrid GMM-SVM Approach*. In Proceedings of the 14^{th} Annual Conference of the International Speech Communication Association (INTERSPEECH), pages 153–157, Lyon, France, 2013. 97

[Janin *et al.* 2003] Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke*et al.* *The ICSI meeting corpus*. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), volume 1, pages I–364, Hong-Kong, April 2003. IEEE. 20, 37

[Janin *et al.* 2004] Adam Janin, Jeremy Ang, Sonali Bhagat, Rajdip Dhillon, Jane Edwards, Javier Marcías-Guarasa, Nelson Morgan, Barbara Peskin, Elizabeth Shriberg, Andreas Stolcke*et al.* *The ICSI Meeting Project: Resources and Research*. In NIST ICASSP 2004 Meeting Recognition Workshop, Montreal, May 2004. 20, 22

[Jefferson *et al.* 1987] Gail Jefferson, Harvey Sacks and Emanuel A. Schegloff. *Notes on laughter in the pursuit of intimacy.* In G. Button and J.R.E. Lee, editors, Talk and Social Organization, pages 152–205. Multilingual Matters, 1987. 46

[Jefferson 1985] Gail Jefferson. *An exercise in the transcription and analysis of laughter.* In T. Van Dijk, editor, Handbook of discourse analysis, volume 3 of *Discourse and Dialogue*, pages 25–34. London, UK: Academic Press, 1985. 52, 67, 68

[Katz 1987] Slava Katz. *Estimation of probabilities from sparse data for the language model component of a speech recognizer.* IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 35, no. 3, pages 400–401, 1987. 150

[Kawahara *et al.* 1999] Hideki Kawahara, Ikuyo Masuda-Katsuse and Alain de Cheveigné. *Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds.* Speech Communication, vol. 27, no. 3–4, pages 187–207, 1999. 140, 141

[Kawahara *et al.* 2001] Hideki Kawahara, Jo Estill and Osamu Fujimura. *Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT.* In 2^{nd} International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA), 2001. 140, 141

[Kawahara 2006] Hideki Kawahara. *STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds.* Acoustical science and technology, vol. 27, no. 6, pages 349–353, 2006. 132

[Kaya *et al.* 2013] Heysem Kaya, Ali Mehdi Erçetin, Albert Ali Salah and Sadik Fikret Gürgen. *Random Forests for Laughter Detection.* In Proceedings of Workshop on Affective Social Speech Signals, satellite of INTERSPEECH, Grenoble, France, 2013. 98

[Kennedy & Ellis 2004] Lyndon S. Kennedy and Danial P.W. Ellis. *Laughter detection in meetings.* In NIST ICASSP 2004 Meeting Recognition Workshop, pages 118–121, Montreal, May 2004. 89

[Kennedy & Hauptmann 1999] Paul Kennedy and Alexander G. Hauptmann. *Laughter Extracted from Television Closed Captions as Speech Recognizer Training Data.* In Proceedings of the 6^{th} European Conference on Speech Communication and Technology (EUROSPEECH), Budapest, Hungary, 1999. 86

[Kipper & Todt 2001] Silke Kipper and Dietmar Todt. *Variation of sound parameters affects the evaluation of human laughter.* Behaviour, pages 1161–1178, 2001. 55, 122, 168, 170

[Kipper & Todt 2003] Silke Kipper and Dietmar Todt. *The role of rhythm and pitch in the evaluation of human laughter.* Journal of Nonverbal Behavior, vol. 27, no. 4, pages 255–272, 2003. 56, 122, 168, 170

[Kipper & Todt 2007] Silke Kipper and Dietmar Todt. *Series of similar vocal elements as a crucial acoustic structure in human laughter.* In Proceedings of the Interdisciplinary Workshop on the Phonetics of Laughter, pages 3–7, Saarbrücken, Germany, August 2007. 25, 45

[Knox & Mirghafori 2007] Mary T. Knox and Nikki Mirghafori. *Automatic Laughter Detection Using Neural Networks.* In Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH), pages 2973–2976, Antwerp, Belgium, August 2007. 66, 91

[Knox *et al.* 2008] Mary T. Knox, Nelson Morgan and Nikki Mirghafori. *Getting the last laugh: Automatic laughter segmentation in meetings.* In Proceedings of the 9^{th} Annual Conference of the International Speech Communication Association (Interspeech), pages 797–800, Brisbane, Australia, 2008. 92, 108

[Kori 1987] Shiro Kori. *Perceptual dimensions of laughter and their acoustic correlates.* In Proceedings of the 11^{th} International Conference on Phonetic Sciences, pages 255–258, Tallin, Estonian SSR, 1987. 57, 58, 168

[Krikke & Truong 2013] Teun F. Krikke and Khiet P. Truong. *Detection of nonverbal vocalizations using Gaussian Mixture Models: looking for fillers and laughter*

in conversational speech. In Proceedings of the 14^{th} Annual Conference of the International Speech Communication Association (INTERSPEECH), pages 163–167, Lyon, France, 2013. 95, 119

[Krippendorff 2007] Klaus Krippendorff. *Computing Krippendorff's alpha reliability.* Departmental Papers (ASC), page 43, 2007. 75

[Lafontaine & Todoroff 2007] Marie-Jo Lafontaine and Todor Todoroff. *The world starts every second.* Artistic Installation held at the "Musee des Beaux-Arts", Angers, France, December 2007. 29

[Lasarcyk & Trouvain 2007] Eva Lasarcyk and Jürgen Trouvain. *Imitating conversational laughter with an articulatory speech synthesis.* In Proceedings of the Interdisciplinary Workshop on the Phonetics of Laughter, pages 43–48, Saarbrücken, Germany, August 2007. 45, 122, 123

[Laskowski & Burger 2007a] Kornel Laskowski and Susanne Burger. *Analysis of the occurrence of laughter in meetings.* In Proceedings of INTERSPEECH, pages 1258–1261, Antwerp, Belgium, 2007. 54

[Laskowski & Burger 2007b] Kornel Laskowski and Susanne Burger. *On the Correlation between Perceptual and Contextual Aspects of Laughter in Meetings.* In Proceedings of the Interdisciplinary Workshop on the Phonetics of Laughter, pages 55–60, Saarbrücken, Germany, August 2007. 20, 21, 54, 59, 61

[Linguistic Data Consortium 2008] University of Pennsylvania Linguistic Data Consortium. *Linguistic Data Consortium webpage.* `http://www.ldc.upenn.edu/`, Consulted on September 10, 2008. 21

[Lockerd & Mueller 2002] Andrea Lockerd and Florian M. Mueller. *LAFCam: Leveraging affective feedback camcorder.* In CHI'02 extended abstracts on Human factors in computing systems, pages 574–575. ACM, 2002. 93

[Maekawa *et al.* 2003] Kikuo Maekawa, Hanae Koiso, Hideaki Kikuchi and Kiyoko Yoneyama. *Use of a large-scale spontaneous speech corpus in the study of linguistic variation.* In Proceedings of the 15^{th} International Congress of Phonetic Sciences (ICPhS 2003), pages 643–646, 2003. 18

[Mahony 2000] Diana L. Mahony. *Is Laughter the Best Medicine or Any Medicine at All?* Eye on Psi Chi, vol. 4, no. 3, pages 18–21, Spring 2000. 10

[Mancini *et al.* 2014] Maurizio Mancini, Laurent Ach, Eemline Bantegnie, Tobias Baur, Nadia Berthouze, Debajyoti Datta, Yu Ding, Stephane Dupont, Harry Griffin, Florian Lingenfelser, Radoslaw Niewiadomski, Catherine Pelachaud, Olivier Pietquin, Bilal Piot, Jérôme Urbain, Gualtiero Volpe and Johannes Wagner. *Laugh when you're winning (accepted for publication).* IFIP Advances in Information and Communication Technology, 2014. 165

[Martin & Lefcourt 2004] Rod A. Martin and Herbert M. Lefcourt. *Sense of humor and physical health: Theoretical issues, recent findings, and future directions.* Humor, vol. 17, no. 1/2, pages 1–20, 2004. 7, 10, 11, 62

[Martin 2001] Rod A. Martin. *Humor, laughter, and physical health: Methodological issues and research findings.* Psychological Bulletin, vol. 127, no. 4, page 504, 2001. 7, 9, 10

[Max Planck Institute for Psycholinguistics 2014] Nijmegen The Netherlands Max Planck Institute for Psycholinguistics The Language Archive. *ELAN, The Language Archive.* http://tla.mpi.nl/tools/tla-tools/elan/, consulted on February 17, 2014. 27

[McKeown *et al.* 2012a] Gary McKeown, Roddy Cowie, Will Curran, Willibald Ruch and Ellen Douglas-Cowie. *ILHAIRE Laughter Database.* In Proceedings of the 4^{th} International Workshop on Emotion Sentiment & Social Signals (ES^3 2012)-Corpora for Research on Emotion, Sentiment & Social Signals, held in conjunction with LREC 2012, ELRA, Istanbul, Turkey, pages 32–35, 2012. 17, 22, 24

[McKeown *et al.* 2012b] Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic and Marc Schröder. *The SEMAINE Database: Annotated Multimodal Records of Emotionally Colored Conversations between a Person and a Limited Agent.* IEEE Transactions on Affective Computing, vol. 3, no. 1, pages 5–17, 2012. 24

[McKeown *et al.* 2013] Gary McKeown, William Curran, Ciaran McLoughlin, Harry J Griffin and Nadia Bianchi-Berthouze. *Laughter Induction Techniques Suitable for Generating Motion Capture Data of Laughter Associated Body Movements.* IEEE Face and Gesture, 2013. 28, 62

[McKeown 2014] Gary McKeown. *The ILHAIRE Laughter Database.* http://qub.ac.uk/ilhairelaughter/, Consulted on February 5, 2014. 17, 25, 26, 28

[Melder *et al.* 2007] Willem A. Melder, Khiet P. Truong, Marten Den Uyl, David A. Van Leeuwen, Mark A. Neerincx, Lodewijk R. Loos and B. Plum. *Affective multimodal mirror: sensing and eliciting laughter.* In Proceedings of the international workshop on Human-centered multimedia, pages 31–40. ACM, 2007. 161

[Merriam-Webster 2009] Inc. Merriam-Webster. *Merriam-Webster Dictionary Online.* http://www.merriam-webster.com/, Consulted on January 14, 2009. 104

[MMI team 2014] MMI team. *MMI Facial Expression Database.* http://www.mmifacedb.com/, consulted on February 17, 2014. 25

[Moon 1996] Todd K. Moon. *The expectation-maximization algorithm.* IEEE Signal processing magazine, vol. 13, no. 6, pages 47–60, 1996. 205

[Morrison *et al.* 2007] Donn Morrison, Ruili Wang and Liyanage C. De Silva. *Ensemble methods for spoken emotion recognition in call-centres.* Speech Communication, vol. 49, pages 98–112, 2007. 16

[Natural Point, Inc. 2009] Natural Point, Inc. *OptiTrack - Optical motion tracking solutions.* http://www.naturalpoint.com/optitrack/, Consulted on October 20, 2009. 29, 30, 32

[Neuhoff & Schaefer 2002] Charles C. Neuhoff and Charles Schaefer. *Effects of laughing, smiling, and howling on mood.* Psychological reports, vol. 91, no. 3f, pages 1079–1080, 2002. 9

[Niewiadomski *et al.* 2009] Radoslaw Niewiadomski, Elisabetta Bevacqua, Maurizio Mancini and Catherine Pelachaud. *Greta: an interactive expressive ECA system.* In Carles Sierra, Cristiano Castelfranchi, Keith S. Decker and Jaime Simão Sichman, editors, Proccedings of the 8^{th} International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS), volume 2, pages 1399–1400, Budapest, Hungary, May 10-15 2009. IFAAMAS. 30, 162

[Niewiadomski *et al.* 2012] Radoslaw Niewiadomski, Jérôme Urbain, Catehrine Pelachaud and Thierry Dutoit. *Finding out the audio and visual features that influence the perception of laughter intensity and differ in inhalation and exhalation phases.* In Proceedings of the ES^3 2012 4^{th} International Workshop on Corpora for Research on EMOTION SENTIMENT & SOCIAL SIGNALS, Satellite of LREC 2012, Istanbul, Turkey, May 2012. 75, 78, 81, 115

[Niewiadomski *et al.* 2013a] Radoslaw Niewiadomski, Jennifer Hofmann, Jérôme Urbain, Tracey Platt, Johannes Wagner, Bilal Piot, Hüseyin Çakmak, Sathish Pammi, Tobias Baur, Stephane Dupont, Matthieu Geist, Florian Lingenfelser, Gary McKeown, Olivier Pietquin and Willibald Ruch. *Laugh-aware virtual agent and its impact on user amusement.* In Proceedings of the 12^{th} international conference on Autonomous Agents and MultiAgent Systems (AAMAS), Saint Paul, Minnesota, USA, May 2013. 67, 162, 164

[Niewiadomski *et al.* 2013b] Radoslaw Niewiadomski, Maurizio Mancini, Tobias Baur, Giovanna Varni, Harry Griffin and Min SH Aung. *MMLI: Multimodal multiperson corpus of laughter in interaction.* In Human Behavior Understanding, pages 184–195. Springer, 2013. 28

[Nwokah *et al.* 1999] Eva E Nwokah, Hui-Chin Hsu, Patricia Davies and Alan Fogel. *The integration of laughter and speech in vocal communication: A dynamic systems perspective.* Journal of Speech, Language, and Hearing Research, vol. 42, no. 4, page 880, 1999. 47

[Oh & Wang 2013a] Jieun Oh and Ge Wang. *Laughter Modulation: from Speech to Speech-Laugh.* In Proceedings of the 14^{th} Annual Conference of the Interna-

tional Speech Communication Association (INTERSPEECH), pages 754–755, Lyon, France, 2013. 130

[Oh & Wang 2013b] Jieun Oh and Ge Wang. *LOLOL: Laugh Out Loud On Laptop.* In Proceedings of the 2013 International Conference on New Musical Instruments (NIME'13), Daejon, Korea, 2013. 129

[Oh *et al.* 2013] Jieun Oh, Eunjoon Cho and Malcolm Slaney. *Characteristic Contours of Syllabic-Level Units in Laughter.* In Proceedings of the 14^{th} Annual Conference of the International Speech Communication Association (INTERSPEECH), pages 158–162, Lyon, France, 2013. 95

[Oostdijk 2000] Nelleke Oostdijk. *The spoken Dutch Corpus: overview and first evaluation.* In Proceedings of LREC, pages 887–894, Athens, Greece, 2000. 18

[Oura 2011] Keiichiro Oura. *HMM-based Speech Synthesis System (HTS) [computer program webpage].* `http://hts.sp.nitech.ac.jp/`, consulted on June 22, 2011. 131

[Owren & Bachorowski 2003] Michael J. Owren and Jo-Anne Bachorowski. *Reconsidering the evolution of nonlinguistic communication: The case of laughter.* Journal of Nonverbal Behavior, vol. 27, no. 3, pages 183–200, 2003. 58

[Pammi *et al.* 2013] Sathish Pammi, Houssemeddine Khemiri, Dijana Petrovska-Delacrétaz and Gérard Chollet. *Detection of nonlinguistic vocalizations using ALISP sequencing.* In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 7557–7561. IEEE, 2013. 94, 108

[Peeters 2004] Geoffroy Peeters. *A large set of audio features for sound description (similarity and classification) in the CUIDADO project.* Technical report, Institut de Recherche et Coordination Acoustique/Musique (IRCAM), 2004. 76, 77, 109, 116

[Petridis & Pantic 2008a] Stavros Petridis and Maja Pantic. *Audiovisual discrimination between laughter and speech.* In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 5117–5120, Las Vegas, Nevada, 2008. 99

[Petridis & Pantic 2008b] Stavros Petridis and Maja Pantic. *Audiovisual laughter detection based on temporal features.* In Proceedings of the 10^{th} international conference on Multimodal interfaces, pages 37–44. ACM, 2008. 99

[Petridis & Pantic 2008c] Stavros Petridis and Maja Pantic. *Fusion of audio and visual cues for laughter detection.* In Proceedings of the 2008 international conference on Content-based image and video retrieval, pages 329–338. ACM, 2008. 99

[Petridis & Pantic 2009] Stavros Petridis and Maja Pantic. *Is This Joke Really funny? Judging the mirth by Audiovisual Laughter Analysis.* In Proceedings of the IEEE International Conference on Multimedia and Expo, pages 1444–1447, New York, USA, June 2009. 99, 101, 105

[Petridis & Pantic 2011] Stavros Petridis and Maja Pantic. *Audiovisual Discrimination Between Speech and Laughter: Why and When Visual Information Might Help.* IEEE Transactions on Multimedia, vol. 13, no. 2, pages 216–234, 2011. 60, 100

[Petridis *et al.* 2010] Stavros Petridis, Ali Asghar and Maja Pantic. *Classifying laughter and speech using audio-visual feature prediction.* In Proceedings of the 2010 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pages 5254–5257, Dallas, Texas, 2010. IEEE. 99

[Petridis *et al.* 2013a] Stavros Petridis, Maelle Leveque and Maja Pantic. *Audiovisual Detection of Laughter in Human-Machine Interaction.* In Proceedings of the 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII), pages 129–134, Geneva, Switzerland, 2013. IEEE. 100

[Petridis *et al.* 2013b] Stavros Petridis, Brais Martinez and Maja Pantic. *The MAHNOB laughter database.* Image and Vision Computing, vol. 31, pages 186–202, 2013. 23, 26, 60

[Pitt *et al.* 2007] Mark A. Pitt, Laura Dilley, Keith Johnson, Scott Kiesling, William Raymond, Elizabeth Hume and Eric Fosler-Lussier. *Buckeye Corpus of Conversational Speech (2^{nd} release).* `http://www.buckeyecorpus.osu.edu/`, 2007. 19

[Pompino-Marschall *et al.* 2007] Bernd Pompino-Marschall, Sabine Kowal and Daniel C. OâConnell. *Some phonetic notes on emotion: laughter, interjections and weeping.* In Proceedings of the Interdisciplinary Workshop on the Phonetics of Laughter, pages 41–42, Saarbrücken, Germany, August 2007. 53, 67

[Provine 1993] Robert R. Provine. *Laughter punctuates speech: Linguistic, social and gender contexts of laughter.* Ethology, vol. 95, no. 4, pages 291–298, 1993. 47, 62

[Pruthi & Espy-Wilson 2004] Tarun Pruthi and Carol Y. Espy-Wilson. *Acoustic parameters for automatic detection of nasal manner.* Speech Communication, vol. 43, no. 3, pages 225–239, 2004. 95

[Queen's University Belfast: School of Psychology 2014] Queen's University Belfast: School of Psychology. *Belfast Induced Natural Expression Database (BINED).* `http://www.psych.qub.ac.uk/BINED/`, consulted on February 17, 2014. 26

[Raitio *et al.* 2011] Tuomo Raitio, Antti Suni, Junichi Yamagishi, Hannu Pulakka, Jani Nurminen, Martti Vainio and Paavo Alku. *HMM-Based Speech Synthesis Utilizing Glottal Inverse Filtering.* IEEE Transactions on Audio, Speech and Language Processing, vol. 19, no. 1, pages 153–165, 2011. 140, 141, 142

[Raitio *et al.* 2013] Tuomo Raitio, John Kane, Thomas Drugman and Christer Gobl. *HMM-Based Synthesis of Creaky Voice.* In Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH), pages 2316–2320, 2013. 145

[Rajman *et al.* 2007] Martin Rajman, Romaric Besançon, Hervé Bourlard, Jean-Cédric Chappelier, Thierry Dutoit, Andrzej Drygajlo, Hynek Hermanski, Magharet King, Jacques Moeschler, Vincenzo Pallotta, Antoine Rozenknop, Jean-Philippe Thiran and Eric Wehrli. Speech and language engineering. Computer and communication sciences. EPFL Press, Distributed by CRC Press, 2007. 41

[Räsänen *et al.* 2009] Okko Johannes Räsänen, Unto Kalervo Laine and Toomas Altosaar. *An improved speech segmentation quality measure: the R-value.* In Proceedings of 10^{th} Annual Conference of the International Speech Communication Association (Interspeech), 2009. 113

[Reuderink *et al.* 2008] Boris Reuderink, Mannes Poel, Khiet P. Truong, Ronald Poppe and Maja Pantic. *Decision-level fusion for audio-visual laughter detection.* Machine Learning for Multimodal Interaction, pages 137–148, 2008. 99

[Ruch & Ekman 2001] Willibald Ruch and Paul Ekman. *The expressive pattern of laughter.* In A. Kaszniak, editor, Emotion, qualia and consciousness, pages 426–443. World Scientific Publishers, Tokyo, 2001. 1, 3, 4, 16, 44, 45, 46, 50, 54, 63, 69, 81, 115

[Ruch & Proyer 2008] Willibald Ruch and René T. Proyer. *The fear of being laughed at: Individual and group differences in gelotophobia.* Humor: International Journal of Humor Research, vol. 21, pages 47–67, 2008. 168

[Ruch & Proyer 2009] Willibal Ruch and René T. Proyer. *Who fears being laughed at? The location of gelotophobia in the Eysenckian PEN-model of personality.* Personality and Individual Differences, vol. 46, no. 5-6, pages 627–630, 2009. 5

[Ruch *et al.* 2013] Willibald Ruch, Jennifer Hofmann and Tracey Platt. *Investigating facial features of four types of laughter in historic illustrations.* The European Journal of Humour Research, vol. 1, no. 1, pages 99–118, 2013. 81, 115

[Ruch 1993] Willibald Ruch. *Exhilaration and humor.* Handbook of emotions, vol. 1, pages 605–616, 1993. 4, 5, 46, 63, 65

[Ruch 1997] Willibald Ruch. *State and trait cheerfulness and the induction of exhilaration: A FACS study.* European psychologist, vol. 2, no. 4, page 328, 1997. 5

[Sathya *et al.* 2013] Adithya Thati Sathya, Kumar Sudheer and B Yegnanarayana. *Synthesis of laughter by modifying excitation characteristics.* The Journal of the Acoustical Society of America, vol. 133, pages 3072–3082, 2013. 43, 84, 107, 119, 127, 139

[Scherer *et al.* 2009] Stefan Scherer, Friedhelm Schwenker, Nick Campbell and Günther Palm. *Multimodal laughter detection in natural discourses.* In Human Centered Robot Systems, pages 111–120. Springer, 2009. 23, 92, 100

[Scherer *et al.* 2012] Stefan Scherer, Michael Glodek, Friedhelm Schwenker, Nick Campbell and Günther Palm. *Spotting laughter in natural multiparty conversations: A comparison of automatic online and offline approaches using audiovisual data.* ACM Transactions on Interactive Intelligent Systems (TiiS), vol. 2, no. 1, page 4, 2012. 101, 118

[Scherer 2003] Klaus R. Scherer. *Vocal communication of emotion: a review of research paradigms.* Speech Communication, vol. 40, pages 227–256, 2003. 16

[Schröder & Trouvain 2003] Marc Schröder and Jürgen Trouvain. *The German text-to-speech synthesis system MARY: A tool for research, development and teaching.* International Journal of Speech Technology, vol. 6, no. 4, pages 365–377, 2003. 123

[Schuller *et al.* 2007] Björn Schuller, Ronald Müeller, Benedikt Höernler, Anja Höethker, Hitoshi Konosu and Gerhard Rigoll. *Audiovisual recognition of spontaneous interest within conversations.* In Proceedings of the 9^{th} international conference on Multimodal interfaces, pages 30–37. ACM, 2007. 23

[Schuller *et al.* 2008] Björn Schuller, Florian Eyben and Gerhard Rigoll. *Static and dynamic modelling for the recognition of non-verbal vocalisations in conversational speech.* In Perception in multimodal dialogue systems, pages 99–110. Springer, 2008. 86, 90

[Schuller *et al.* 2011] Björn Schuller, Michel Valstar, Florian Eyben, Gary McKeown, Roddy Cowie and Maja Pantic. *AVEC 2011–the first international audio/visual emotion challenge.* In Affective Computing and Intelligent Interaction, pages 415–424. Springer, 2011. 24

[Schuller *et al.* 2013] Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi*et al.* *The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism.* In Proceedings of the 14^{th} Annual Conference of the International Speech Communication Association (INTERSPEECH), pages 148–152, Lyon, France, 2013. 19, 94

[Siebert *et al.* 2009] Xavier Siebert, Stéphane Dupont, Philippe Fortemps and Damien Tardieu. *MediaCycle: Browsing and Performing with Sound and Image libraries.* In Thierry Dutoit and Benoît Macq, editors, QPSR of the numediart research program, volume 2, pages 19–22. Numediart Research Program on Digital Art Technologies, 3 2009. 104

[Sjölander & Beskow 2011] Kåre Sjölander and Jonas Beskow. *Wavesurfer: open source tool for sound visualization and manipulation [computer program].* http://sourceforge.net/projects/wavesurfer/, June 21, 2011. 49, 68, 78

[Sjölander 2004] Kåre Sjölander. *The Snack Sound Toolkit [Computer program]. Retrieved February 10, 2011*, 2004. 78, 102, 116, 131

[Skype Communications 2009] Skype Communications. *The Skype Laughter Chain.* http://www.skypelaughterchain.com/, Consulted on January 22, 2009. 11

[Sloetjes & Wittenburg 2008] Han Sloetjes and Peter Wittenburg. *Annotation by Category: ELAN and ISO DCR.* In Proceedings of the 6^{th} International Conference on Language Resources and Evaluation (LREC), 2008. 27

[Sneddon *et al.* 2012] Ian Sneddon, Margaret McRorie, Gary McKeown and Jennifer Hanratty. *The belfast induced natural emotion database.* Affective Computing, IEEE Transactions on, vol. 3, no. 1, pages 32–41, 2012. 26

[Soong & Juang 1984] Frank K. Soong and Biing-Hwang Juang. *Line spectrum pair (LSP) and speech data compression.* In Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), volume 9, pages 37–40, San Diego, CA, USA, March 1984. IEEE. 142

[SPTK online 2013] SPTK online. *Speech Signal Processing Toolkit (SPTK) v. 3.6*, 2013. 131

[SSPNET 2014] SSPNET. *Green Persuasive Database.* http://sspnet.eu/2009/12/green-persuasive-database/, consulted on February 17, 2014. 24

[Stupakov *et al.* 2012] Alex Stupakov, Evan Hanusa, Deepak Vijaywargi, Dieter Fox and Jeff Bilmes. *The design and collection of COSINE, a multi-microphone in situ speech corpus recorded in noisy environments.* Computer Speech & Language, vol. 26, no. 1, pages 52–66, 2012. 19

[Suarez *et al.* 2012] Merlin Teodosia Suarez, Jocelynn Cu and Madelene Sta. *Building a Multimodal Laughter Database for Emotion Recognition.* In Proceedings of the ES^3 2012 4^{th} International Workshop on Corpora for Research on EMOTION SENTIMENT & SOCIAL SIGNALS, Satellite of LREC 2012, pages 2347–2350, Istanbul, Turkey, May 2012. 29, 58

[Sudheer *et al.* 2009] Kumar K. Sudheer, Sir Harish Reddy M., K. Sri Rama Murty and B. Yegnanarayana. *Analysis of laugh signals for detecting in continuous speech.* In Proceedings of the 10^{th} Annual Conference of the International Speech Communication Association (Interspeech), pages 1591–1594, Brighton, UK, September 2009. 34, 43, 93, 119, 127

[Sundaram & Narayanan 2007] Shiva Sundaram and Shrikanth Narayanan. *Automatic acoustic synthesis of human-like laughter.* Journal of the Acoustical Society of America, vol. 121, no. 1, pages 527–535, January 2007. 44, 122, 124, 126, 139, 148, 157

[Szameitat *et al.* 2007] Diana P. Szameitat, Chris J. Darwin, André J. Szameitat, Dirk Wildgruber, Annette Sterr, Susanne Dietrich and Kai Alter. *Formant characteristics of human laughter.* In Proceedings of the Interdisciplinary Workshop on the Phonetics of Laughter, pages 9–13, Saarbrücken, Germany, August 2007. 29, 49, 50, 51, 53

[Szameitat *et al.* 2009a] Diana P Szameitat, Kai Alter, André J Szameitat, Chris J Darwin, Dirk Wildgruber, Susanne Dietrich and Annette Sterr. *Differentiation of emotions in laughter at the behavioral level.* Emotion, vol. 9, no. 3, page 397, 2009. 29, 57, 58

[Szameitat *et al.* 2009b] Diana P. Szameitat, Kai Alter, André J. Szameitat, Dirk Wildgruber, Annette Sterr and Chris J. Darwin. *Acoustic profiles of distinct emotional expressions in laughter.* The Journal of the Acoustical Society of America, vol. 126, no. 1, pages 354–366, 2009. 29, 69, 103

[Talkin 1995] David Talkin. *A robust algorithm for pitch tracking (RAPT).* Speech coding and synthesis, vol. 495, page 518, 1995. 49, 77, 116, 132

[Tanaka & Campbell 2011] Hiroki Tanaka and Nick Campbell. *Acoustic Features of Four Types of Laughter in Natural Conversational Speech.* In Proceedings of the 17^{th} International Congress of Phonetic Sciences (ICPhS), Hong Kong, pages 1958–1961, 2011. 49, 50, 53, 64, 67, 91, 108

[Tokuda *et al.* 2000] Keiichi Tokuda, Takayoshi Yoshimura, Takashi Masuko, Takao Kobayashi and Tadashi Kitamura. *Speech parameter generation algorithms for HMM-based speech synthesis.* In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), volume 3, pages 1315–1318. IEEE, 2000. 131

[Toledano *et al.* 2003] Doroteo Torre Toledano, Luis A. Hernández Gómez and Luis Villarrubia Grande. *Automatic phonetic segmentation.* IEEE Transactions on Speech and Audio Processing, vol. 11, no. 6, pages 617–625, 2003. 108

[Trouvain & Schröder 2004] Jürgen Trouvain and Marc Schröder. *How (not) to add laughter to synthetic speech.* In Affective Dialogue Systems, pages 229–232. Springer, 2004. 62, 122

[Trouvain & Truong 2013] Jürgen Trouvain and Khiet P. Truong. *Exploring sequences of speech and laughter activity using visualisations of conversations.* In Proceedings of the Workshop on Affective Social Speech Signals, satellite of INTERSPEECH, Grenoble, France, 2013. 62

[Trouvain 2001] Jürgen Trouvain. *Phonetic Aspects of Speech-Laughs.* In Proceedings of the Conference on Orality & Gestuality (ORAGE), pages 634–639, Aix-en-Provence, France, 2001. Citeseer. 6, 47, 48

[Trouvain 2003] Jürgen Trouvain. *Segmenting Phonetic Units in Laughter.* In Proceedings of the 15^{th} International Congress of Phonetic Sciences, pages 2793–2796, Barcelona, Spain, August 2003. 34, 44

[Truong & Trouvain 2012a] Khiet P. Truong and Jürgen Trouvain. *Laughter annotations in conversational speech corpora-possibilities and limitations for phonetic analysis.* Proceedings of the 4^{th} International Worskhop on Corpora for Research on Emotion Sentiment and Social Signals, pages 20–24, 2012. 18, 19, 60

[Truong & Trouvain 2012b] Khiet P. Truong and Jürgen Trouvain. *On the acoustics of overlapping laughter in conversational speech.* In Proceedings of INTERSPEECH 2012, Portland, Oregon, USA, 2012. International Speech Communication Association. 66

[Truong & van Leeuwen 2007a] Khiet P. Truong and David A. van Leeuwen. *Automatic discrimination between laughter and speech.* Speech Communication, vol. 49, pages 144–158, 2007. 18, 21, 89, 90, 93, 161

[Truong & van Leeuwen 2007b] Khiet P. Truong and David A. van Leeuwen. *Evaluating automatic laughter segmentation in meetings using acoustic and acoustic-phonetic features.* In Proceedings of the Interdisciplinary Workshop on the Phonetics of Laughter, pages 49–53, Saarbrücken, Germany, August 2007. 91

[Universitat Augsburg 2013] Universitat Augsburg. *Social Signal Interpretation.* `http://hcm-lab.de/projects/ssi`, Consulted on December 11, 2013. 27, 28

[University of South California, Santa Barbara 2011] University of South California, Santa Barbara. *Santa Barbara Corpus of Spoken American English.* `http://www.linguistics.ucsb.edu/research/sbcorpus_contents.html`, Consulted on June 7, 2011. 18

[Urbain & Dutoit 2011] Jérôme Urbain and Thierry Dutoit. *A Phonetic Analysis of Natural Laughter, for Use in Automatic Laughter Processing Systems.* In Proceedings of the 4^{th} bi-annual International Conference of the HUMAINE Association on Affective Computing and Intelligent Interaction (ACII2011), pages 397–406, Memphis, Tennesse, October 2011. 54, 67, 132

[Urbain *et al.* 2009a] Jérôme Urbain, Elisabetta Bevacqua, Thierry Dutoit, Alexis Moinet, Radoslaw Niewiadomski, Catherine Pelachaud, Benjamin Picart, Joëlle Tilmanne and Johannes Wagner. *AVLaughterCycle: An audiovisual laughing machine.* In Thierry Dutoit and Benoît Macq, editors, QPSR of the numediart research program, volume 2, pages 97–104. Numediart Research Program on Digital Art Technologies, 9 2009. 104

[Urbain *et al.* 2009b] Jérôme Urbain, Thomas Dubuisson, Stéphane Dupont, Christian Frisson, Raphaël Sebbe and Nicolas D'Alessandro. *AudioCycle: A Similarity-Based Visualization of Musical Libraries.* In Proc. of IEEE International Conference on MultiMedia and Expo (ICME09), pages 1847–1848, New York City, USA, June 2009. 104

[Urbain *et al.* 2010a] 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 7^{th} conference on International Language Resources and Evaluation (LREC'10), Valletta, Malta, May 2010. 30, 54, 65, 67, 143

[Urbain *et al.* 2010b] Jérôme Urbain, Radoslaw Niewiadomski, Elisabetta Bevacqua, Thierry Dutoit, Alexis Moinet, Catherine Pelachaud, Benjamin Picart, Joëlle Tilmanne and Johannes Wagner. *AVLaughterCycle: Enabling a virtual agent to join in laughing with a conversational partner using a similarity-driven audiovisual laughter animation.* Journal on Multimodal User Interfaces, vol. 4, no. 1, pages 47–58, 2010. Special Issue: eNTERFACE'09. 30, 105, 107, 162

[Urbain *et al.* 2012] Jérôme Urbain, Radoslaw Niewiadomski, Emeline Bantegnie Jennifer Hofmann, Tobias Baur, Nadia Berthouze, Hüseyin Çakmak, Richard Thomas Cruz, Stéphane Dupont, Matthieu Geist, Harry Griffin, Florian Lingenfelser, Maurizio Mancini, Miguel Miranda, Gary McKeown, Sathish Pammi, Olivier Pietquin, Bilal Piot, Tracey Platt, Willibald Ruch, Abhisheck Sharma, Gualtiero Volpe and Johannes Wagner. *Laugh Machine.* In Proceedings of the 8^{th} International Summer Workshop on Multimodal Interfaces - eNTERFACE'12, pages 13–34, Metz, France, July 2012. 162

[Urbain *et al.* 2013a] Jérôme Urbain, Hüseyin Çakmak and Thierry Dutoit. *Automatic Phonetic Transcription of Laughter and its Application to Laughter Synthesis.* In Proceedings of the 5^{th} biannual Humaine Association Conference on Affective Computing and Intellignet Interaction (ACII), pages 153–158, Geneva, Switzerland, 2-5 September 2013. 108

[Urbain *et al.* 2013b] Jérôme Urbain, Hüseyin Çakmak and Thierry Dutoit. *Evaluation of HMM-Based laughter synthesis.* In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pages 7835–7839, Vancouver, Canada, May 2013. 109

[Urbain *et al.* 2014] Jérôme Urbain, Hüseyin Çakmak, Aurélie Charlier, Maxime Denti, Thierry Dutoit and Stéphane Dupont. *Arousal-driven Synthesis of Laughter.* IEEE Journal of Selected Topics in Signal Processing, vol. 8, 2014. 149

[Valstar & Pantic 2010] Michel Valstar and Maja Pantic. *Induced disgust, happiness and surprise: an addition to the MMI facial expression database.* In Proceedings of the 7^{th} conference on International Language Resources and Evaluation (LREC'10), Workshop on EMOTION, pages 65–70, Valletta, Malta, May 2010. 25

[Valstar *et al.* 2007] Michel F. Valstar, Hatice Gunes and Maja Pantic. *How to distinguish posed from spontaneous smiles using geometric features.* In Proceedings of the 9^{th} international conference on Multimodal interfaces, pages 38–45. ACM, 2007. 16

[Vettin & Todt 2004] Julia Vettin and Dietmar Todt. *Laughter in conversation: Features of occurrence and acoustic structure.* Journal of Nonverbal Behavior, vol. 28, no. 2, pages 93–115, 2004. 18, 49, 54, 55, 62, 66

[Wagner *et al.* 2009] Johannes Wagner, Elisabeth André and Frank Jung. *Smart sensor integration: A framework for multimodal emotion recognition in real-time.* In Proceedings of Affective Computing and Intelligent Interaction (ACII), pages 1–8, Amsterdam, The Netherlands, 2009. 31, 34

[Wagner *et al.* 2011] Johannes Wagner, Florian Lingenfelser and Elisabeth André. *The Social Signal Interpretation Framework (SSI) for Real Time Signal Processing and Recognition.* In Proceedings of INTERSPEECH, pages 3245–3248, 2011. 27

[Wagner *et al.* 2013] Johannes Wagner, Florian Lingenfelser and Elisabeth André. *Using Phonetic Patterns for Detecting Social Cues in Natural Conversations.* In Proceedings of the 14^{th} Annual Conference of the International Speech Communication Association (INTERSPEECH), pages 168–172, Lyon, France, 2013. 71, 98, 108

[Weninger & Schuller 2012] Felix Weninger and Björn Schuller. *Discrimination of Linguistic and Non-Linguistic Vocalizations in Spontaneous Speech: Intra-and Inter-Corpus Perspectives.* In Proceedings of INTERSPEECH, 2012. 19, 90

[Wilting *et al.* 2006] Janneke Wilting, Emiel Krahmer and Marc Swerts. *Real vs. acted emotional speech.* In Proceedings of the 9^{th} Annual Conference of the

International Speech Communication Association (INTERSPEECH), pages 805–808, Pittsburgh, USA, September 2006. 16

[Yamagishi *et al.* 2009] Junichi Yamagishi, Takashi Nose, Heiga Zen, Zhen-Hua Ling, Tomoki Toda, Keiichi Tokuda, Simon King and Steve Renals. *Robust speaker-adaptive HMM-based text-to-speech synthesis.* IEEE Transactions on Audio, Speech, and Language Processing, vol. 17, no. 6, pages 1208–1230, 2009. 139

[Yoshimura *et al.* 1999] Takayoshi Yoshimura, Keiichi Tokuda, Takashi Masuko, Takao Kobayashi and Tadashi Kitamura. *Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis.* In Proceedings of Eurospeech, Budapest, Hungary, 1999. 131

[Yoshimura *et al.* 2001] Takayoshi Yoshimura, Keiichi Tokuda, Takashi Masuko and Tadashi Kitamura. *Mixed-excitation for HMM-based speech synthesis.* In Proceedings of the European Conference on Speech Communication and Technology (Eurospeech), pages 2259–2262, 2001. 141

[Young & Young 1994] Steve J. Young and Sj. Young. *The HTK Hidden Markov Model toolkit: Design and philosophy.* In Entropic Cambridge Research Laboratory, Ltd. Citeseer, 1994. 108, 131

[Young *et al.* 2006] Steve J. Young, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Xunying Liu, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey*et al. The HTK book version 3.4.* Cambridge University Engineering Department, vol. 2, no. 2, pages 2–3, 2006. 111, 112

[Yovetich *et al.* 1990] Nancy A. Yovetich, T. Alexander DALE and Mary A. Hudak. *Benefits of humor in reduction of threat-induced anxiety.* Psychological Reports, vol. 66, pages 51–58, 1990. 9

[Zen *et al.* 2007a] Heida Zen, Tomoki Toda, Masaru Nakamura and Keiichi Tokuda. *Details of Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005.* IEICE Transactions on Information and Systems, vol. E90-D, no. 1, pages 325–333, 2007. 141

[Zen *et al.* 2007b] Heiga Zen, Takashi Nose, Junichi Yamagishi, Shinji Sako, Takashi Masuko, Alan Black and Keiichi Tokuda. *The HMM-based speech synthesis system (HTS) version 2.0.* In Proceedings of the 6^{th} ISCA Workshop on Speech Synthesis, pages 294–299, 2007. 133

[Zen 2006] Heiga Zen. *An example of context-dependent label format for HMM-based speech synthesis in English.* The HTS CMUARCTIC demo, 2006. 133

[Zign Creations 2009] Zign Creations. *Zign Track.* `http://www.zigncreations.com/zigntrack.html`, Consulted on October 20, 2009. 30, 31

Appendix A

Summary table of databases in which acoustic laughs have been spotted

---

Table A gives a summary of the features of the databases presented in Chapter 2. Meanings of the columns and abbreviations are as follows:

1. Database: name of the database (see Chapter 2 for the corresponding references).

2. Type of Data: **S** for Spontaneous laughs, **I** for Induced laughs, **Ac** for acted laughs. Precisions about the recording scenario are provided within parentheses. Some databases combine different types of data recording, in which case they are separated by a "+" sign.

3. Modalities: recorded signals: **A** for audio (all databases in this list, as we focused on audio), **V** for video, **K** for Kinect, **MC** for Motion Capture. ST denotes the use of a far-field microphone to record all participant on a Single Track.

4. Fs: sampling frequency of the audio recordings, in kHz.

5. #part.: total number of participants in the database.

6. #part. / rec.: number of participants simultaneously involved in recordings. This provides information about the social context. As the presence of experimenters also influences the social context, it is also denoted when useful.

7. #laughs: total number of laughter episodes in the database.

8. Anno.: laughter annotations, i.e. whether the presence of laughs was simply indicated (without accurate time boundaries), which is denoted **Ind.**, or segmented (i.e., with time boundaries, **Seg.**), or other relevant information: **Emo.** for labeling in Emotional categories, **V**/**U** for voiced/unvoiced annotations, **SL** for speech-laughs.

9. Relations (between participants): whether participants simultaneously involved in recordings knew each other (and if so, what is the rapport between them:

Acq. for Acquaintances, friends, colleagues, etc.) or were strangers. This is also important for social context.

Note that the information presented in Table A corresponds to the information we could retrieve from the papers cited in Chapter 2. We made not attempt to verify the given figures.

Table A.1: Features of the audio databases in which laughter has been spotted.

Database	**Type of data**	**Modalities**	**Fs**	**#part.**	**#part. / rec.**	**#laughs**	**Anno.**	**Relations**
Corpus of Spontaneous Japanese	S (academic lectures and presentations)	A	?	1395	1 (+ audience)	?	Ind.	Class or Acq.
Campbell's telephone conversations	S (phone calls)	A	48	10	2	2001 (+ 1129 SL)	Seg. (+ SL)	Strangers
Santa Barbara Corpus of American English	S (various sources)	A (ST)	22.05	>>>	Variable (2 - ?)	?	Seg.	Variable
Corpus Gesproken Nederlands	S (various sources)	A	16[1]	>>>	Variable (1 -?)	?	Non-verbal segments annotated	Variable
Vettin & Todt	S (free conversations)	A	6	10	2	1921	Seg.	Acq.
Devillers & Vidrascu	S (call centers)	A	?	791 [2]	2	119	Ind.	Strangers

[1] 8kHz for telephone conversations
[2] 86 in dialogs containing laughter

COSINE	S (conversations, various noise)	A	48	91	2-7	326?[3]	Seg. (+ SL)	?
ICSI	S (meetings)	A	16	53	3-10	11515 (+ 1000 BL + 1000 SL)	Seg. (+ SL)	Colleagues
HCRC	S (map task scenario)	A	20	64	2	1000	Seg.	Friends or strangers
Buckeye	S (interviews)	A	?	40	1 (+1 interviewer)	1874	Seg.	Strangers
Diapix	S (game scenario)	A	44.1	40	2	582	Seg.	Friends
SVC	S (phone calls with game scenario)	A	?	120	2	1200	Seg.	Strangers
Belfast Naturalistic Database	S (TV excerpts + interviews)	A + V	44.1	125	Variable (2-?)	>53	Seg.	Strangers
HUMAINE	S + I (various sources)	A + V	geq22.05	>>	Variable (1 -?)	46	Seg.	Variable
AMI	S (meetings + role play)	A + V + Written notes	48	>>>	3-4	16477	Seg.	Mainly strangers
AVIC	S (role play)	A + V	44.1	21	2	294	Seg.	?
FreeTalk	S (conversations)	A (ST) + V	16	5	4-5	> 300	Seg.	Colleagues
Green Persuasive	S (role playing)	A + V	44.1	8 (+ professor)	2	280	Seg.	Professor and student

[3] in transcribed sessions

SEMAINE	S (video conference)	A + V	48	150	1 (+ 1 operator)	443	Seg.	Strangers
Belfast conversational dyads	S (conversations)	A + V + K	48	18	2	?	Seg. (ongoing)	Friends
Bachorowksi et al.	I (film watching)	A	?	97	1-2	1024	Seg.	Friends or strangers
MMI	I (film watching)	A	48	9	1	164	Seg. (+ V/U)	/
BINED	I (film watching and emotional tasks)	A + V	48	256	1 (+ 1-2 experimenters)	> 340	Seg.	/
MAHNOB	I + S + Ac	A + V + Th V	44.1	22	1 (+ 2 experimenters)	563 I/S + 51 Ac + 67 SL	Seg. (+ V/U) (+ SL)	/
Belfast Story-telling	I (16 emotions)	A + V + K	48	21	3-4	>934	Seg.	?
Belfast and UCL Motion Capture	I (film watching and amusing tasks) + Ac	A + V + Body MC	48	26	2 (+ experimenters)	>500	Seg.	Friends
MMLI	I (film watching and amusing tasks)	A + K + Body MC	16	16	2-3	520	Seg.	Friends
Pinoy Laughter 2	Mostly I (film watching and emotional tasks)	A + V	?	4	2-3	?	Seg. + syllables	?

AV-LASYN	I (film watching)	A + V + Facial MC	44.1	1	1	256	Seg. + phonetic	/
Szameitat et al.	Ac	A	48	8	1	429	Emo.	/
Pinoy Laughter	Ac + I	A	44.1	2 (Ac) + 3 (I)	1	497 Ac + ? I	Seg. + Emo.	/
Lafontaine & Todoroff	Ac	A	48	Singers + 7 children	1-3	447	Emo.	Colleagues
AVLC	I (film watching)	A + V + Facial MC	16	24	1	1021 I + 27 Ac	Seg. + phonetic + arousal	/

Appendix B

Introduction to Hidden Markov Models

Hidden Markov Models (HMMs) enable to model the temporal evolution of features. To do this, one HMM is composed of several states. A schematic HMM is displayed in Figure B.1. This HMM contains three states, numbered 1, 2 and 3. Each state is associated to a different probabilistic distribution of the features, called the *observations*. The probability of observing a feature vector x in state i is given by the function $O_i(x)$ and called *emission* probability. The emission probabilities are generally modeled with GMMs (as shown in the figure) or neural networks. Only one state of the HMM is occupied at each time frame. At the next time frame, the HMM can either stay in the same state, or move to another state. HMMs are Markovian processes, also called memoryless, which means that the future evolution only depends on the current state (past states do not have any impact on the future if we know the present state). The probability of going to state j when immediately coming from state i is given by the *transition probability* t_{ij}.

The HMM is fully described by its number of states, the transition probabilities t_{ij}, the emission probabilities $O_i(x)$ as well as the probabilities to occupy each state at the first frame.

Typically for speech recognition or synthesis, one HMM is built for each phoneme. The models are generally trained in a supervised way (i.e., transcriptions of the utterances are provided along with the data) with the Expectation-Maximization algorithm [Moon 1996]. Using an initial estimate of the HMMs (which can be randomly or uniformly initialized), the observations are assigned to each phoneme HMM according to the given phonetic transcription[1] and all the HMM parameters are estimated thanks to the available observations for each state. Then, the estimated parameters of the HMMs are used to compute the most likely sequence of states respecting the phonetic transcription and the HMM parameters are re-estimated according to the observations assigned to each state. Several steps of re-estimation are performed, until parameters do not change anymore between two iterations[2].

The different states of each HMM enable to model the evolution of the features across each phone. Transition probabilities between two different HMMs can also

[1] If the phonetic transcription does not specify the time boundaries between the phonemes, observations are uniformly divided for the initialization step.

[2] In practice, re-estimation is usually stopped after either a pre-defined number of steps, or when parameter changes between two steps fall below a pre-defined threshold.

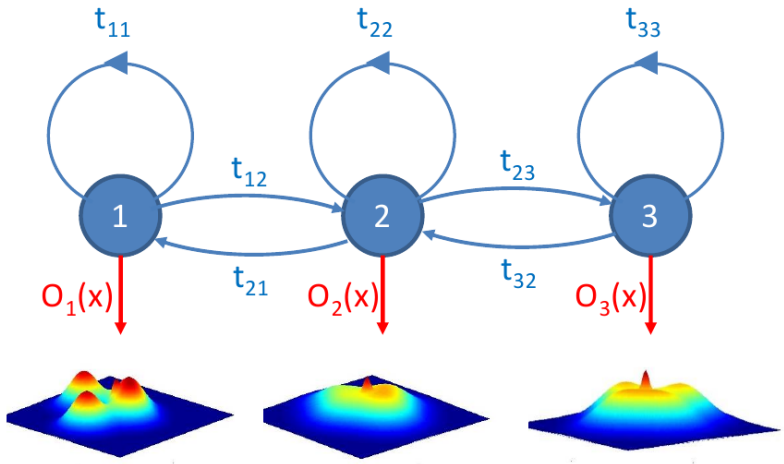

Figure B.1: Scheme of a three-state HMM

be computed on the training data: in the *n-gram* model [Brown *et al.* 1992], the probabilities of a phoneme Z_n to immediately follow a sequence of $n-1$ phonemes Y_1^{n-1} is estimated for all the phoneme combinations. More complex models can be used in speech recognition to encode lexical (only a finite number of words are allowed) and grammatical (specifying which sequences of words are permitted) rules, but this goes beyond the scope of this dissertation.

Once trained, the HMMs can be used to transcribe sequences of observations. The most likely sequence of states—given the models (HMMs, n-grams)—conducting to the observations is computed.