university of groningen / campus fryslân

Master Thesis in Voice Technology

The relevance of using authentic laughter data in natural laughter synthesis: A case study on LaughNet

Master Candidate

Sjors Weggeman

Student ID 5007453

Supervisor

Dr. Matt Coler

University of Groningen

Second reader

Dr. Shekhar Nayak

University of Groningen

External supervisors

Dr. Aki Kunikoshi

ReadSpeaker

Dr. Jaebok Kim

ReadSpeaker

Academic Year
2022/2023

To my family
and friends

"Perhaps I know best why it is man alone who laughs;
he alone suffers so deeply that he had to invent laughter."

~

Der Wille zur Macht (1901)
by Friedrich Nietzsche

Abstract

Purpose: The purpose of this research was to enhance the naturalness of synthesised speech by incorporating authentic laughter data into the laughter synthesis process of the state-of-the-art model LaughNet (Luong & Yamagishi, 2021b).
Method: A Support Vector Machine (SVM) was trained to demonstrate the differences between acted and spontaneous human laughter at the acoustic level, by classifying them based on their acoustic features. Factor analysis was applied to identify the most relevant acoustic features in determining authenticity. Then the influence of the synthesis procedure of LaughNet on these features was researched by examining the waveform silhouette format and by generating synthetic laughter using LaughNet, classifying it with the SVM, and comparing the classification performance to that of human laughter. The ability of human listeners to recognise the difference between human and synthetic laughter was evaluated using a listening test.
Results: The results of this study show that acted and spontaneous laughter can be distinguished on the basis of their acoustic features. The most relevant acoustic features are: 1) the F0 mean, maximum, and variability, 2) the percentage of unvoiced segments and the intensity, and 3) the F0 minimum. Out of these factors, only the second one is captured in the waveform silhouette. The other factors have to be regenerated by the model for the synthetic laughter. This could not be confirmed through synthesis and classification, since I was unable to get sufficiently usable output from LaughNet. Consequently, synthetic laughter could not be evaluated. Human listeners were able to detect the authenticity of human laughter significantly above chance level, with female laughter being easier to classify than male laughter. However, the authenticity judgements were not generally agreed upon.
Conclusion: Laughter authenticity matters for the synthesis of natural laughter, but appears to impose little additional naturalness on the synthetic laughter of LaughNet, as the model generates the most important lower-level acoustic features. Due to insufficient authentic laughter data to fine-tune the generator, there is little control over the final authenticity of its synthetic laughter. Future research with sufficient lab-collected data may be able to overcome this limitation by carefully selecting the generative model, data format, and training- and fine-tuning data. Moreover, the perceived authenticity of isolated laughter appears to be contentious, suggesting the need for context to be taken into account in experimental designs as a way to disambiguate the authenticity judgments.

Keywords *— Natural, Laughter, Synthesis, Authentic, Acted, Spontaneous*

Contents

List of Figures **xi**

List of Tables **xiii**

List of Acronyms **xix**

1 Introduction **1**
- 1.1 Speech synthesis goals . . . 2
- 1.2 Status quo of the speech synthesis goals . . . 4
- 1.3 Emotion . . . 5
- 1.4 Synthetic laughter . . . 6

2 Background **7**
- 2.1 Basic terminology of laughter . . . 8
- 2.2 History of speech synthesis . . . 8
- 2.3 Literature review . . . 11
  - 2.3.1 Parametric laughter synthesis . . . 12
  - 2.3.2 Articulatory laughter synthesis . . . 14
  - 2.3.3 DNN laughter synthesis . . . 16
  - 2.3.4 Defining naturalness within laughter synthesis . . . 20
  - 2.3.5 Key findings . . . 22
- 2.4 Research questions and hypotheses . . . 23

3 Methodology **27**
- 3.1 Experiment 1 . . . 28
  - 3.1.1 Research design . . . 28
  - 3.1.2 Data . . . 28
  - 3.1.3 Preprocessing . . . 30
  - 3.1.4 Materials . . . 33
- 3.2 Experiment 2 . . . 34
  - 3.2.1 Research design . . . 34
  - 3.2.2 Data . . . 35
  - 3.2.3 Preprocessing . . . 35

3.2.4 Materials . . . 37
3.3 Experiment 3 . . . 38
3.3.1 Research design . . . 38
3.3.2 Data . . . 38
3.3.3 Preprocessing . . . 39
3.3.4 Materials . . . 39
3.3.5 Participants . . . 41

4 Results **43**
4.1 Experiment 1 . . . 44
4.1.1 Data distribution . . . 44
4.1.2 Hyperparameter settings . . . 45
4.1.3 Evaluation . . . 45
4.2 Experiment 2 . . . 51
4.2.1 Theoretical analysis . . . 51
4.2.2 Practical analysis . . . 52
4.3 Experiment 3 . . . 53
4.3.1 Participants . . . 53
4.3.2 Reliability . . . 53
4.3.3 Accuracy . . . 54
4.3.4 Bias checking . . . 54

5 Discussion **55**
5.1 Experiment 1 . . . 56
5.2 Experiment 2 . . . 57
5.3 Experiment 3 . . . 58
5.4 Implications . . . 59
5.5 Limitations . . . 60
5.6 Future research . . . 60

6 Conclusion **61**

References **62**

Disclaimers **69**

Acknowledgments **71**

A Plots **75**

B Questionnaire **79**
B.1 Index page 79
B.2 Introduction page 80
B.3 Quiz 80
B.4 Results page 81

List of Figures

1.1 The uncanny valley (M. Mori, 1970) 2

2.1 Laughter segmentation; cf. Trouvain (2003) (Juhitha et al., 2018) . 8

3.1 Waveform decomposition into temporal features; cf. Rosen (1992) (Lizarazu, 2017) 34

4.1 Histogram of projections with density plots per split 46
4.2 Histogram of projections of acted data with density plots per split 47
4.3 Histogram of projections of spontaneous data with density plots per split 47
4.4 Histogram of projections stacked per authenticity with density plots and without gender separation per split 48
4.5 Histogram of projections stacked per authenticity without density plots per split 48
4.6 Scree plot 49
4.7 General Loss Total 52
4.8 Mel-Spectrogram Error 52
4.9 Validation Mel- Spectrogram Error 52

A.1 Parameter distribution of misclassified training files relative to the class means 76
A.2 Parameter distribution of misclassified test files relative to the class means 77

B.1 Screenshot from the index page of the questionnaire 79
B.2 Screenshot from the introduction page of the questionnaire 80
B.3 Screenshot from one of the quiz pages of the questionnaire 80
B.4 Screenshot from the results page of the questionnaire 81

List of Tables

2.1 Speech synthesis methods and their strengths and weaknesses in chronological order of conception from top to bottom per guiding principle; extrapolated and synthesised from Story (2019) and their sources . . . 10
2.2 Related work and their respective approaches to synthesising laughter . . . 12
2.3 Best average naturalness MOS (5-point Likert scale) per synthesis method for every parametric synthesis publication (Urbain et al., 2013b) . . . 14
2.4 Best synthetic laughter naturalness MOS (5-point Likert scale) per synthesis method for every calibrated DNN synthesis publication 19
2.5 Laughter datasets from the literature review, their authenticity, and the publications that used them, in chronological order . . . 21

3.1 Spontaneous laughs extracted from the MULAI Corpus . . . 30
3.2 Acoustic features extracted . . . 32
3.3 Evaluation setup used by the authors of the guiding papers . . . 41

4.1 Starting data distribution across gender and authenticity . . . 44
4.2 Training data distribution across gender and authenticity . . . 44
4.3 Testing data distribution across gender and authenticity . . . 44
4.4 Confusion matrix training data classification . . . 45
4.5 Confusion matrix testing data classification . . . 45
4.6 Factor loadings after varimax rotation . . . 49
4.7 Variance in the data accounted for by factors . . . 49
4.8 Class means-based misclassified acoustic features of misclassified laughs . . . 50
4.9 Acoustic feature representation in temporal features; cf. Rosen (1992) . . . 51
4.10 Listening test participants by gender and age range . . . 53
4.11 Reliable participant accuracy per laughter type and gender . . . 54
4.12 Answer percentages . . . 54
4.13 Usage “I really don’t know” per authenticity . . . 54

List of Acronyms

AE Autoencoder

ANN Artificial Neural Network

AVLC AVLaughterCycle

CNN Convolutional Neural Network

DAE Deep Autoencoder

DNN Deep Neural Network

GAN Generative Adversarial Network

HNR Harmonics-to-noise ratio

HiFi High Fidelity

IVI Intervoicing interval

HMM Hidden Markov Models

MOS Mean-Opinion Score

OGVC Online Gaming Voice chat Corpus

RNN Recurrent Neural Network

SAE Shallow Autoencoder

Seq2seq Sequence-to-sequence

SOTA state-of-the-art

SVM Support Vector Machine

VAE Variational Autoencoder

VCTK Voice Cloning Toolkit

1

Introduction

Research in the field of voice technology can be roughly divided into the following research areas: speech recognition, text analysis, and speech synthesis. The focus of this master's thesis will be the latter, since the main topic of study is laughter synthesis. To solidly ground this study however, I will start out from the broader perspective of the main research goals in speech synthesis and their relevance (1.1), discuss the status quo of achieving them (1.2), and then explain through the concept of emotion (1.3) how synthetic laughter contributes to achieving those goals (1.4).

1.1 Speech synthesis goals

In broad strokes, most contemporary work in speech synthesis can be split into achieving two goals: making synthetic speech intelligible and making it sound natural – that is, as human-like as possible (Campbell, 2007a, p. 36; Taylor, 2009, p. 1; Baird et al., 2018, p. 2863). Note however, that this definition is a point of contention that will be discussed later on and for now will only be used as a working definition. Intelligibility matters because the receiver needs to be able to understand which words are being produced by the system (Campbell, 2007a, pp. 30, 36; Taylor, 2009, p. 48). Naturalness matters because it is highly preferable that speech produced by the systems sounds like it is spoken by a human (Campbell, 2007a, pp. 30, 36 (cf. Mattingly 1974), Taylor, 2009, p. 47).

Why speech needs to be intelligible does not require any additional explanation. Why speech needs to natural is not so straightforward however. This can best be explained through the concept of the "uncanny valley", which is closely related to "naturalness". The uncanny valley was first hypothesised by M. Mori in 1970. It describes the relationship between how closely a stimulus resembles human features (x-axis) and our level of emotional connection to it (y-axis) (see figure 1.1). The theory is named after the sudden drop in emotional connection that occurs when a stimulus closely resembles human features, but not quite perfectly. In this hypothesised valley our emotional connection switches from a positive to a negative state, in which we experience feelings of eeriness. The experienced affinity is amplified when the stimulus is moving compared to when it is still, as has been depicted by the dotted line (see figure 1.1).

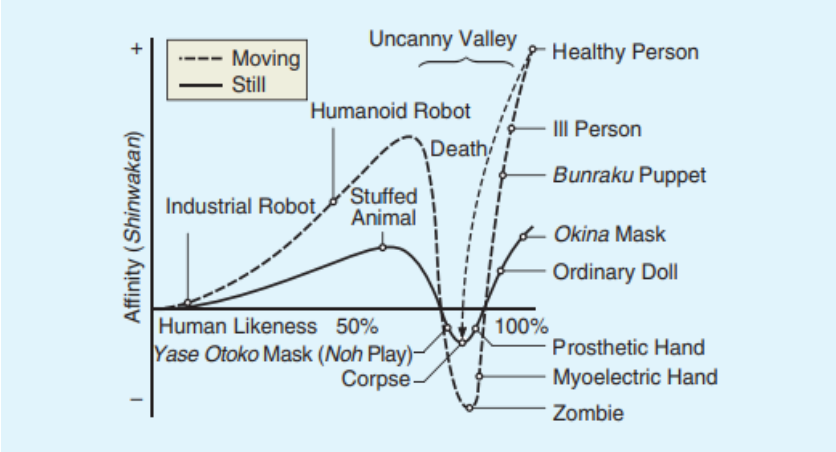


Figure 1.1: The uncanny valley (M. Mori, 1970)

A common mistake made by researchers researching the uncanny valley, is that they consider this exact graph to be the uncanny valley. (Tinwell & Grimshaw, 2009, p. 69). The graph however, was only used by M. Mori (1970) to illustrate the concept. Instead, researchers should determine *what* causes this phenomenon *when*, and *why*. Although the precise answer to this question remains yet to be found, it is known that expectations play a major role in it (Tinwell and Grimshaw, 2009, p. 67; Burleigh et al., 2013, p. 771).

Firstly, based on experience and knowledge we form expectations about the world. Expectations that belong together are then combined into frames of reference that we use to reason about the world. We have a frame for everything we have encountered in our lives, including humans. When human-like stimuli are perceived in unison with non-human-like stimuli, or when it is unclear which of these categories stimuli belong to, a conflict arises between the incompatible frames, causing feelings of uneasiness (Tinwell and Grimshaw, 2009, p. 67; Burleigh et al., 2013, pp. 761, 770–771). Secondly, being and living amongst humans, our frame of reference for humans is much more detailed than other frames. This means that we have more- and higher expectations concerning human-like traits (Burleigh et al., 2013, p. 771). Thirdly, the context (the collection of all frames that apply to your surroundings) in which the stimuli appear, also affects our expectations and thus our frames (Tinwell & Grimshaw, 2009, p. 71).

An example that can be used to illustrate each of these types of expectations for speech synthesis is the study by Mitchell et al. (2011), in which videos of a robot or a human were combined with either a robotic voice or a human voice: 1) When a video of a face was presented with a voice belonging to the opposing category, a feeling of eeriness was evoked (Mitchell et al., 2011, p. 11). 2) When a human speaks with a robotic voice, the total does not live up to the human frame (Mitchell et al., 2011, p. 11). This is likely due to the fact that speech is an almost uniquely human feature. 3) If the participants were told that the human was auditioning for the role of robot in a play, then the feelings of eeriness would not have been evoked as strongly, if evoked at all (Tinwell & Grimshaw, 2009, p. 71).

This means that in technology, including speech synthesis, expectations either need to be managed appropriately (see Romportl 2014), or need to be satisfied. In light of the human frame however, unimodal speech synthesis systems – that is, not in context with other modalities – are more likely to be expected to produce human-like speech, because their acceptability depends on it (Taylor, 2009, p. 1; Baird et al., 2018, p. 2863). Therefore, in order to surpass the uncanny valley, it is necessary to synthesise the laughter of a healthy person (see figure 1.1) – that is, both as intelligible and natural as possible.

1.2 Status quo of the speech synthesis goals

Having established the speech synthesis goals and why they matter, I will now discuss the status quo on achieving those goals: according to Taylor (2009, p. 48),[1] the intelligibility goal has been achieved as of the late 1970s. The naturalness goal however, has not yet been achieved (Schröder, 2001, p. 561; Taylor, 2009, p. 1; Baird et al., 2018, p. 1). To understand why this goal has still not been achieved five decades later and to understand what it takes to achieve it, it is necessary to have a better understanding of the terms and their coverage.

'***Intelligibility***' has been elaborately discussed by Miller (2013, p. 602), starting from the fact that communication is a multimodal process, in which all possible channels are being used synchronously to maximise the likelihood of successful transmission. Following from this is the notion that there are many different aspects, across different modalities, that affect intelligibility. Consequently, a distinction can be made between the aspects of intelligibility that are captured in the speech signal and the aspects that are not. He calls this ***signal-dependent intelligibility*** and ***signal-independent intelligibility*** respectively.

This distinction is especially useful for the field of voice technology: a voice technology system on its own does not have the capability to affect intelligibility through other means than the voice signal. Therefore, from hereon, intelligibility refers to signal-dependent intelligibility.

Accordingly, intelligibility is composed of things that affect the signal. The essence of this can be captured by the following aspects: content separation, speech rate, loudness (Miller, 2013, p. 602), intensity, noise (French & Steinberg, 1947, pp. 90–91), and distortion (Steinberg, 1929, p. 121). Each of these problems has already been dealt with in the field of telecommunication, allowing for easy transferring of the solutions, leaving only the synthesis of comprehensible speech sounds to solve in order to accomplish the intelligibility goal.

'***Naturalness***' on the other hand, has not yet been clearly defined and is consequently often interpreted differently (Dall et al., 2014, p. 1). It is unclear why it has not yet been clearly defined, but possibly it is because the term encompasses a wide variety of speech aspects. Examples of this are, but are not limited to: coherence, disfluencies, context awareness, references (Lustgarten & Juang, 2003), emotion (Sebe et al., 2005), and prosody (Dall et al., 2014), all of

[1]The claims made by Taylor hold specifically for the English language, but are extended to other languages given that the right data can be provided (Taylor, 2009, p. 7). At the time of writing (Feb. 2023) however, there are still many under-resourced languages for which this is not the case (Eberhard et al., 2022; UNESCO, 2023; University of Hawaii at Manoa, 2023).

which entail various other terms, illustrating the complexity of clearly defining naturalness. However, the field, amongst others, would greatly benefit from a precise definition.

Lastly, the ways in which intelligibility and naturalness function in language are also different. The former can be considered as a *multiplicative* property in the sense that when just one aspect is expressed poorly, then the speech becomes unintelligible. The latter can be considered as an *additive* property in the sense that when one aspect is expressed poorly, the speech just becomes less natural.[2] Therefore, in order to achieve as natural speech as possible, each aspect of naturalness needs to be dealt with individually.

1.3 Emotion

The most logical starting point is the aspect that is most evidently missing, since that is where the most can be gained. According to Schröder (2001, p. 561), this is emotion (Sebe et al., 2005). Henceforth, the focus of this thesis will be on the topic of emotion. The problem with emotions though, is that there is still a lack of consensuses and understanding on multiple levels (Nesse, 2020, p. 1). Most notoriously being the lack of consensus on *how many* emotions there are and *which* emotions they are, or as Nesse (2020, p. 2) worded it:

> *"[...] noting that consensus is lacking would be a vast understatement."*

Consequently, it is challenging to establish a sound scientific grounding for research concerning specific emotions. However, the manifestations of emotions: subjective experience, physical response, and expression (Lazarus, 1991), can still be researched. For the purposes of this thesis in the field of speech synthesis, the relevant parameter is "expression".

The idea of integrating emotion into speech synthesis is not new. The simplest manner was to add emotion words to the content. However, most meaning is not conveyed through what we say, but through how we say it (Mehrabian, 1971).[3]

What we say is the verbal content that we ascribed to a different part of the voice technology pipeline. *How* we say it is part of the subfield of non-verbal communication that is strictly concerned with the non-linguistic part of the speech signal: paralinguistics (Schuller et al., 2013, p. 5); cf. Crystal (1974).

[2] An illustration of this concept is Wernicke's aphasia: people with this disability produce incoherent sentences, in a natural sounding manner. Likewise, an incoherent synthetic sentence can still sound very natural, just less than if the sentence were coherent.

[3] An illustration of this concept is 'irony': we say one thing, but we mean the exact opposite.

1.4 Synthetic laughter

The first attempts to add emotion to synthesised speech through paralinguistics focused solely on prosodic features (Schröder, 2001, p. 2). Whilst prosody is certainly an important tool to communicate emotions, it is not the only way. Much less represented in speech synthesis systems and research are vocal affect bursts (Scherer, 1994; Schröder, 2003). One of the most flexibly used and therefore most often occurring vocal affect bursts is laughter (Urbain, 2014, p. 5).

Laughter comes in various forms: voiced "song-like", unvoiced "snort-like", and unvoiced "grunt-like" (Bachorowski & Owren, 1995). Within these categories there is a lot of variability that depends on the function of the laughter, the characteristics of the producer, their personal style (Urbain, 2014), the social context (Urbain, 2014; Wood, 2020), and their company (Campbell, 2007b; Farley et al., 2022), and possibly more. On top of that, these forms can occur isolated or intertwined with speech as so-called "speech-laughs". Taken together with the fact that the field of laughter synthesis is still very young and under-researched, with the first attempt being done by Sundaram and Narayanan in 2007, it should come as no surprise that natural laughter synthesis has not yet been achieved.

Since natural laughter synthesis can contribute significantly to the naturalness of synthesised speech (Campbell, 2006) however, it is important to keep studying it. To understand what can be done to improve the naturalness of synthesised laughter, a detailed understanding of the field is required. Therefore, in chapter 2, I will provide the necessary background information and perform a literature review. After this my research questions and hypotheses will be posed. Then, in chapter 3, I will explain my plan of approach to answer the research questions, followed by the results in chapter 4. In chapter 5, I will discuss the results in light of the chosen procedure and point out directions of future research. Lastly, in chapter 6, a conclusion will be drawn from the complete picture.

2

Background

The aim of this thesis is to investigate how the naturalness of synthetic laughter can be improved. To this purpose, I will firstly discuss the basic terminology of laughter (2.1) and the different speech synthesis methods that have been developed throughout history (2.2). Both types of information are necessary to understand the literature review (2.3), in which I will show that LaughNet (Luong & Yamagishi, 2021b) is the state-of-the-art (SOTA). I will then build my case that its naturalness can be improved through the use of authentic laughter data, leading up to my research questions and hypotheses (2.4).

2.1 Basic terminology of laughter

Laughter is a multimodal phenomenon insofar that it is generally expressed through a combination of sound, facial expressions, and bodily movements (Urbain, 2014, pp. 3–5). Accordingly, it has been researched from different (sub-)disciplines which use varying terms and concepts, making it difficult to compare research across (sub-)disciplines. An important attempt to standardise the terms and concepts for the acoustic features of laughter was performed by Trouvain (2003), who asserted that: An instance of laughter is referred to as an ***episode***, an episode comprises one or more ***bouts***, and each bout consists of multiple ***calls***. A call usually consists of a fricative or silence and a vowel (Urbain, 2014). This has been visualised by Juhitha et al. (2018) (see figure 2.1).

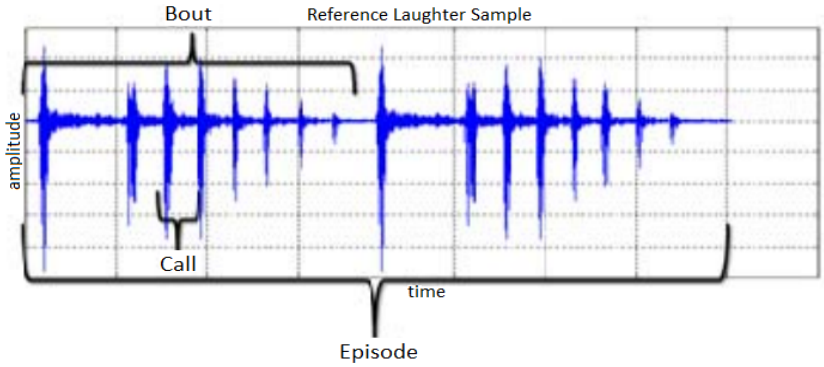


Figure 2.1: Laughter segmentation; cf. Trouvain (2003) (Juhitha et al., 2018)

2.2 History of speech synthesis

Since laughter synthesis has its background in speech synthesis, it helps to have a clear oversight of the different synthesis methods that have been developed throughout history, as well as their respective strengths and weaknesses. To this end, I have created an overview in table 2.1 below by extrapolating and synthesising information from Story (2019) and the sources they used.

The synthesis methods in this table have been divided into two categories: those that are guided by the physics of speech production and those that are guided by speech data itself. The methods belonging to the former category try to model speech sounds by focusing on the interplay between the sound signal and the vocal tract. Although this makes them ideal for studying the speech production process,[1] it is an indirect way of modeling speech sounds, leaving room for errors in the modelling process. This makes them suboptimal for synthesising speech that is both highly intelligible and highly natural. For

[1]Under the assumption that the model functions correctly.

the methods belonging to the latter category, which directly model the speech sounds instead, the opposite holds.[1] Because the speech sounds are modelled directly, there is only little room for errors. However, errors also contain information about the parts of the vocal tract and their respective contributions to the speech sound formation process. Furthermore, it is not feasible to reverse engineer the speech sound formation process due to the many factors that are involved in speech production.

Besides intelligibility and naturalness, the main speech synthesis goals (see 1.1), the table also contains information on: the computational cost, the modifiability, and, for the models guided by data, the database size for each method. This additional information provides insight into the feasibility of implementation. Although this is not a major concern for this research in particular, it will matter for future research and helps in determining the SOTA. Computational cost negatively affects the accessibility of the system (Campbell, 2007a, p. 35), modifiability positively affects the acceptability of the system (Klatt, 1987, p. 779; Campbell, 2007a, p. 35), and an increase in database size negatively affects the required storage capacity, as well as the computational cost and thus the accessibility of the system.

Lastly, following the flow of time from start to end, table 2.1 has been arranged chronologically from top to bottom. The only exception to this is the articulatory synthesis method, which was created between concatenative synthesis and Deep Neural Network (DNN)[2] synthesis, but belongs to the methods guided by the physics of speech production. Since advancements come with time, the overall quality of the synthesised speech improves when moving down the table. Contrasting with that is the worsening computational cost: as technology becomes more capable, we run more complex models with more data, which in turn improves the synthesised speech again.

From table 2.1 it becomes clear that, in concordance with Taylor (2009), the earliest conceived methods, guided by physics, generally reach high levels of intelligibility. In spite of that, there is always a trade-off between how natural the synthesised speech sounds and how well its voice characteristics can be changed. As established in the introduction however, very high levels of both intelligibility and naturalness are needed to surpass the uncanny valley (see section 1.1). Additionally, the voice characteristics should be modifiable to account for the large variety of preferences and situations in which synthesised speech is used. None of the methods guided by physics are capable of doing so.

[2]DNNs are Artificial Neural Network (ANN)s with multiple hidden layers.

	Synthesis method	Strengths	Weaknesses
Guided by physics	Vocal tract model	- High naturalness (Hill et al., 1995)	- Moderate intelligibility - Moderate computational cost (Hill et al., 1995) - Low modifiability
	Formant synthesis	- Very high intelligibility (Klatt & Klatt, 1990) - High modifiability - Low computational cost (Klatt, 1987)	- Low naturalness (Klatt, 1987; Klatt & Klatt, 1990)
	Articulatory synthesis	- High intelligibility (Birkholz, 2013) - High naturalness (Birkholz, 2013)	- Moderate computational cost (Birkholz, 2013) - Low modifiability (Birkholz, 2013)
Guided by data	Parametric synthesis	- Very high modifiability - Small database needed (Zen et al., 2009)	- Moderate intelligibility (Zen et al., 2009) - Moderate naturalness (Zen et al., 2009) - High computational cost
	Concatenative synthesis / Unit selection synthesis	- Very high intelligibility - Can achieve high naturalness	- Very low modifiability - Limited to content of database - Trade-off between naturalness and: • unit size • number of units • number of variations per unit • matching criteria between units • sound quality of the recordings Increasing any of these yields an increase in naturalness, but also in the size of the database and with it the computational cost, making very high naturalness unfeasible.
	DNN synthesis	- Very high intelligibility - Very high naturalness - High modifiability (Oord et al., 2016)	- Very high computational cost - Large database needed

Table 2.1: Speech synthesis methods and their strengths and weaknesses in chronological order of conception from top to bottom per guiding principle; extrapolated and synthesised from Story (2019) and their sources

The data-based methods, on the other hand, are capable of reaching higher levels of naturalness, modifiability, or both. Out of these three methods, concatenative synthesis is slightly different: whilst parametric synthesis and DNN synthesis learn features from the data, concatenative speech synthesis consists of speech data. Therefore, the former two generalise well, whilst concatenative speech synthesis is limited by the data captured in the database and can hardly be modified. Although this method is highly intelligible and capable of achieving near perfect naturalness, it is only feasible for domain-specific applications. The amount of storage and resources required to traverse the storage, would take on extreme values for more general, aspecific applications. Out of parametric synthesis and DNN synthesis, the latter is more advanced, as it learns features using a more complex and more refined analysis, based on more examples.

Various DNN architectures exist with their own respective strengths and weaknesses, but explaining all of them in detail is out of the scope of this thesis. Instead, I will provide a brief description and highlight the main improvements over previously encountered architectures. Additionally, I will provide references to papers that explain the architectures in detail.

2.3 Literature review

For the literature review, in June 2022, I searched three different databases for papers on 'laughter synthesis' published in the last ten years. The databases were: IEEEXplore (18 hits), Google Scholar (110 hits), and WorldCat (395 hits). For the latter I queried for libraries worldwide and sorted the results based on '*Best Match*'. Furthermore, I limited the results to the first 100 hits, because the relevance of the results dropped significantly after that. This left me with a starting database of 228 papers.

From this starting database I included all papers that had the words: 'laughter' and 'synthesis' in their title. Papers that also had the words: 'animation', 'motion', 'multimodal', 'virtual', or 'visual' in the title were excluded, because they refer to research that includes a wrong modality. To constrict the research area even further and reduce the number of confounding variables, I decided to focus on isolated laughter. Accordingly, I also excluded papers that had: 'amused speech', 'context', 'narrative' or 'speech-laughs' in their title, because they refer to research concerning the context in which laughter occurs.

This left a total of 12 papers to be included in the literature review. In addition to those, I manually added 2 papers from the starting database: Mansouri and Lachiri (2021) and Urbain (2014). Although the titles of neither of these papers contain the word 'synthesis', several other papers from these authors had already been included in the literature review, prompting me to look closer at the rest of the papers in the starting database. Both papers listed 'laughter synthesis' in their keywords and cited several papers that had already been included in the literature review, and were therefore included.

Table 2.2 below provides the results from the literature review. It shows that only 14 attempts have been done over the past decade to synthesise isolated natural laughter sounds or to ease the process of doing so. Each attempt can be divided into one of the categories discussed in the history of speech synthesis (see section 2.2). Coincidentally, this division corresponds roughly to the timeline of the history of speech synthesis, hence I will treat each category in a separate subsection (subsections 2.3.1-2.3.3). This means that the remainder of this section involves only what can be extrapolated from the content of table 2.2, discussing the content of each paper and comparing their evaluations. Additionally, I will refer back to the lacking definition of '*naturalness*' that I mentioned in section 1.2, and discuss how it is defined within laughter synthesis (2.3.4). After this I will distill the key findings from this literature review (2.3.5), based on which I will then pose my research questions and hypotheses (2.4).

Author	Year	Title	Model Type
Luong & Yamagishi	(2021)	LaughNet: Synthesizing Laughter Utterances from Waveform Silhouettes and a Single Laughter Example	DNN (GAN)
Mansouri & Lachiri		Human Laughter Generation using Hybrid Generative Models	DNN (VAE-CNN/VAE-RNN)
Mansouri & Lachiri	(2020)	Laughter synthesis: A Comparison Between Variational Autoencoder and Autoencoder	DNN (VAE/AE)
Tits et al.		Laughter Synthesis: Combining Seq2seq modelling with Transfer Learning	DNN (Seq2seq)
Mansouri & Lachiri	(2019)	DNN-Based Laughter Synthesis	DNN
Mori et al.		Conversational and Social Laughter Synthesis with WaveNet	DNN (CNN)
Juhitha et al.	(2018)	Laughter Synthesis using Mass-spring Model and Excitation Source Characteristics	Articulatory
Bollepalli et al.	(2014)	A Comparative Evaluation of Vocoding Techniques for HMM-based Laughter Synthesis	Parametric (HMM)
Oh et al.		Affective Analysis and Synthesis of Laughter	[Unavailable]
Urbain		Acoustic Laughter Processing	-
Urbain et al.		Arousal-Driven Synthesis of Laughter	Parametric (HMM)
Sathya et al.	(2013)	Synthesis of Laughter by Modifying Excitation Characteristics	Articulatory
Urbain et al.		Automatic Phonetic Transcription of Laughter and Its Application to Laughter Synthesis	Parametric (HMM)
Urbain et al.		Evaluation of HMM-Based Laughter Synthesis	Parametric (HMM)

Table 2.2: Related work and their respective approaches to synthesising laughter

2.3.1 Parametric laughter synthesis

The content of this section consists solely of research authored or co-authored by Urbain because, besides Oh et al. (2014), whose work was unavailable, he is the only person that performed research in this specific area. He performed or instigated this research in light of his dissertation on *Acoustic Laughter Processing* (Urbain, 2014), with which he tried to pave the way for natural laughter synthesis. In doing so he focused specifically on Hidden Markov Models (HMM), because they were the SOTA at that time. Since HMMs fall under the parametric synthesis category, we should expect highly natural laughter (see table 2.1).

Content of the publications

The first thing Urbain (2014) notes in his dissertation is the lack of databases containing clean recordings of natural laughter. The reason for this is that the acquisition of this data is very challenging, since natural laughter usually occurs in noisy environments. To solve this, the laughter should be recorded in a laboratory setting, but then the problem shifts towards not having the social context that caused the emotion giving rise to the laughter. There are two workarounds: 1) using methods to induce emotions in participants, and 2) using actors, who can induce emotions in themselves. Although neither of these solutions yields exactly the same emotions as in a natural situation, it is suspected that there are many significant differences between acted and spontaneous laughter. Accordingly, most researchers opt for induced emotions.

To ease the process of doing research in this field, Urbain et al. (2010) created the first database containing clean recordings of "as natural laughs as possible" (Urbain, 2014, p. 39): the AVLaughterCycle (AVLC) database. They do so by inducing emotions in the participants by having them watch funny videos in isolation in a laboratory setting. The most significant drawback of this method is that there is no way of knowing whether the participants laugh the same way in isolation as they would when having company. Urbain (2014, pp. 57–59) questions whether this is actually an issue, as it is believed that context is needed to accurately interpret laughter, but that more research needs to be done.

In the first paper Urbain et al. (2013a) designed a way to synthesise laughter from the original, hand-written transcriptions of the AVLC database (Urbain et al., 2010) using HMMs to model the laughter parameters. They then performed a subjective evaluation of the synthesised laughter using a 5-point Likert scale ranging from "very poor (1)" to "excellent (5)", achieving a Mean-Opinion Score (MOS) of 2.6.

In the second paper Urbain et al. (2013b) used HMMs to automatically generate phonetic laughter transcriptions to ease the process of creating new, larger laughter databases. They then tested the quality of the new transcriptions by synthesising it with the method of the first paper and comparing the naturalness scores. This time they achieved a significantly lower MOS of 2.2.

To improve the quality of the automatic transcriptions, in the third paper Urbain et al. (2014) extend their automatic transcription generation model with a module that estimates the intensity of the emotion giving rise to the laughter. The intensity of the emotion giving rise to the laughter is positively correlated with the intensity of the laugh, which in turn affects how far open the mouth is and thus which sound is produced. Therefore, taking this into account should produce laughter with more naturally accurate vowels. This is indeed what they found reaching naturalness levels similar to those of the original transcriptions. Additionally, Urbain et al. (2014) found that participants who used headphones during the evaluations generally gave slightly higher ratings than participants who did not use headphones.

Another important aspect for the naturalness of synthesised speech and laughter is the module that translates the parameters into sound: the vocoder. Several vocoders exist and each of them has their own strengths and weaknesses and unique artifacts. To optimise the naturalness, Bollepalli et al. (2014) evaluated the naturalness of four different HMM vocoders in the fourth paper. They found that vocoders with robust modelling techniques performed better, achieving a MOS of 2.6 for female laughter and a MOS of 2.3 for male laughter.

Evaluations

In every publication in this section, except Urbain et al. (2013b), the synthesised laughter was compared to human laughter samples and to copy-synthesised laughter samples during the evaluations. In copy-synthesis, parameters are extracted from a human laugh and then fed directly into the vocoder. Hence the result is a measure for the highest possible quality that can be achieved with a given vocoder. Table 2.3 below provides the best naturalness MOS achieved in each publication for each of these methods, as well as the standard deviation per score.

Publication	Laugh gender	Best avg. score synthesised	Avg. score copy-synthesis	Avg. score human
Urbain et al. (2013a)	-	2.7 (std: 1.1)	3.2 (1.2)	3.8 (1.2)
Urbain et al. (2013b)	-	2.2 (1.1)	-	4.3 (0.9)
Urbain et al. (2014)	-	2.5 (1.1)	3.3 (1.2)	4.0 (1.2)
Bollepalli et al. (2014)	F M	2.6 (-) 2.3 (-)	3.7 (-) 3.8 (-)	4.3 (-) 4.1 (-)

Table 2.3: Best average naturalness MOS (5-point Likert scale) per synthesis method for every parametric synthesis publication (Urbain et al., 2013b)

Although I draw conclusions based on all the results mentioned in table 2.3, the authors draw the same conclusions from the individual results:

- With MOS between 3.8 and 4.3, isolated human laughter is generally not perceived as perfectly natural.
- With all the reported standard deviations being very similar, the large variation in perceived naturalness has to stem from the human laughter.
- With 3 out of 4 papers that used copy-synthesis reporting copy-synthesis MOS values of at least 0.6 less than the MOS of human laughter, and 1 paper 0.3 less, HMMs are incapable of reaching naturalness levels comparable to human laughter.
- With the best average MOS for synthetic laughter being 2.7, there is still significant room for improvement within this category.

From the best case scenario (copy-synthesis) MOS of 3.8 on a scale from 1 to 5, I conclude that highly natural laughter can be achieved using HMM synthesis. This corresponds to the indication from table 2.1. With 2.7 being the best average naturalness MOS for synthetic laughter however, only moderate naturalness has been achieved.

2.3.2 Articulatory laughter synthesis

In the articulatory category are the papers by Sathya et al. (2013) and Juhitha et al. (2018). According to table 2.1, articulatory synthesis should be able to achieve high naturalness, but in reality it is quite hard because several intricate models need to be integrated into one coherent system (Birkholz, 2013, p. 1).

Content of the publications

Sathya et al. (2013) synthesised laughter by analysing and extracting the excitation characteristic from voiced laughter from the AVLC database, and used it to alter the excitation characteristics of vowels. They specifically opted for voiced laughter because it is more likely to induce a positive feeling in the listener (Bachorowski & Owren, 1995). This way they created laughter at call level (see figure 2.1). In their subjective evaluations, Sathya et al. (2013) evaluate the synthesis quality and acceptability of the synthesised laughter, achieving MOS of 3.57 and 3.38 respectively.

Juhitha et al. (2018) later extended the method of Sathya et al. (2013) with a mass-spring model, which made it possible to synthesise a complete bout at once (see figure 2.1). The result is a more natural energy loss over the course of the bout. In their subjective evaluations, Juhitha et al. (2018) only assessed the acceptability, achieving a MOS of 2.7. This is significantly worse than the 3.38 achieved by Sathya et al. (2013). For quality however, they used a direct comparison measure. Here Juhitha et al. (2018) found that their model performed significantly better than the 3.57 MOS of Sathya et al. (2013). The authors conclude from these results that their adaptation positively affected the naturalness, but that it is still far from perfect. They attribute this to the fact that articulatory synthesis is incapable of capturing the many, rapid variations that are present in laughter.

Evaluations

The evaluations in this section are significantly different from the one discussed in subsection 2.3.1, namely in *what* they evaluated and in the calibration of their evaluations. Firstly, Sathya et al. (2013) and Juhitha et al. (2018) assessed the synthesis quality and acceptability of the laughter, instead of naturalness. Whereas synthesis quality is literally an aspect or interpretation of naturalness, acceptability is a vague, subjective term that should perhaps be considered synonymous of *"affinity"* in light of the uncanny valley (see figure 1.1). Secondly, neither of the evaluations of Sathya et al. (2013) and Juhitha et al. (2018) involved human laughter as reference signal, meaning that their results are not calibrated to the ground truth. If such a reference signal had been implemented though, their results would likely have been attenuated. Because of these two reasons, their result cannot reliably be compared to the results of other research and have thus not been summarised in a table.

2.3.3 DNN laughter synthesis

Most attempts at laughter synthesis in the past decade have been done using DNNs, because of their good performance in speech synthesis (see table 2.1). DNNs can learn the relationships between the text and the corresponding speech sounds through training on large datasets of speech samples. Their relative success in synthesising speech is, in part, the outcome of the ability to handle large amounts of data – particularly important in speech synthesis as speech signals are typically high-dimensional and time-varying. However, the adequacy of DNNs for laughter synthesis is questionable insofar as there is a general paucity of suitable laughter data.

Content of the publications

The first attempt at DNN laughter synthesis was that of Mansouri and Lachiri (2019), who tried to synthesise natural laughter from its acoustic and linguistic features, extracted from the AVLC database. In this attempt they compared standard DNN[3] architectures to Recurrent Neural Network (RNN)[3] architectures. Standard DNNs process information in a feed-forward manner, meaning that information flows in one direction from input to output. RNNs, on the other hand, process information in a recurrent manner, meaning that information from previous time steps is passed on and used in the computation of the current output. This makes them well suited for tasks involving sequential data, like speech synthesis, where the output at each time step (e.g. each syllable or sound) depends on what came before it.

In the subjective evaluation, Mansouri and Lachiri (2019) unexpectedly found that DNN models outperformed RNN models with a MOS of 2.8 over 2.32 for female laughter, and 3.64 over 2.8 for male laughter. These scores are similar to those of parametric synthesis (see table 2.3), whilst they should be able to achieve better. In the objective evaluation however, Mansouri and Lachiri (2019) found that the opposite holds for female laughter. The authors attribute these results to the shortage of laughter data. This contradiction will be discussed in more detail later on, together with their 2020, 2021[4] papers.

In the same year, H. Mori et al. (2019) tried to generate natural synthetic laughter by training a Convolutional Neural Network (CNN)[5], called WaveNet,

[3]The standard DNN architecture and the RNN architecture are explained by Mansouri and Lachiri (2019, p. 2).

[4]Mansouri and Lachiri (2020) is the preliminary version of Mansouri and Lachiri (2021).

[5]The CNN architecture is explained by Oord et al. (2016, pp. 2–3).

(Oord et al., 2016) on laughter transcriptions from the Online Gaming Voice chat Corpus (OGVC) dataset (Arimoto et al., 2012). They then compared the naturalness to that of an HMM (see subsection 2.3.1). CNNs scan a range of previous time steps for patterns. This allows them to learn the value of short-term, mid-term, and long-term dependencies as deemed fit, whilst RNNs are dependant on the information that is passed on through recurrence. This makes the dependency handling of CNNs more stable than that of RNNs.

H. Mori et al. (2019) found that the CNN outperformed the HMM and that male laughter was perceived as more natural than female laughter. For female laughter the CNN outperformed the HMM with a MOS of 2.16 over 1.97, and for male laughter the CNN outperformed the HMM with a MOS of 3.14 over 2.45. Compared to MOS of 4.50 and 4.74 for natural laughter however, there is still plenty of room for improvement.

The year after, Tits et al. (2020) attempted natural laughter synthesis from phonemes and transcriptions from the AmuS dataset (El Haddad et al., 2017), using a Sequence-to-sequence (Seq2seq) model with attention, called DCTTS (Tachibana et al., 2018).[6] Like H. Mori et al. (2019), they compared the naturalness of their model to that of a HMMs (see subsection 2.3.1). As the name already suggests, Seq2seq models map one sequence to another sequence. Its architecture consists of an encoder network, which reduces the input sequence to the essential information, and a a decoder network, which maps the essential information into another sequence. The attention module is located between the two networks and ensures that the output of the decoder is correctly aligned to the input of the encoder. Since both networks need to handle sequential data, both networks have to be a CNN or a RNN.

Tits et al. (2020) found that the Seq2seq model outperformed the HMM model with a MOS of 3.28 over 2.64, compared to a natural laughter MOS of 4.10. Additionally, it is important to note that Tits et al. (2020) attempted to bypass the scarcity problem (see subsection 2.3.1) by leveraging transfer learning from speech from the Acapela dataset. Laughter consists of the same sounds as speech, but plenty more clean speech data is available.

At the same time, Mansouri and Lachiri (2020, 2021)[4] explore the laughter synthesis capabilities of Shallow Autoencoder (SAE), Deep Autoencoder (DAE), and Variational Autoencoder (VAE) architectures on the basis of log magnitude spectrograms of laughter from the AVLC (Urbain et al., 2010) and AmuS (El

[6]The Seq2seq architecture is explained in a highly mathematical manner by Tachibana et al. (2018, pp. 2–3). A similar model, called Tacotron, is explained in a more comprehensible manner by Wang et al. (2017, pp. 3–5).

Haddad et al., 2017) databases.[7] All AE architectures are Seq2seq models, but rather than mapping a sequence to another sequence, they map to the same sequence. Consequently, they simply compresses and decompresses the data. Since there is always loss of data during compression however, the output is guaranteed to be slightly different. This behaviour of AEs can be stimulated by incorporating a chance element and some margins for change into the model, resulting in new data variants. Accordingly, AEs with this stimulated behaviour are called VAEs. These variations can be a solution to the scarcity problem (see subsection 2.3.1).

SAEs technically do not belong in this section because they are ANNs instead of DNNs, but this not a problem, because Mansouri and Lachiri (2020, 2021) found that DAE models outperformed SAE in both their subjective and objective evaluations. In the subjective evaluations they found that the DAEs outperformed VAEs with a MOS of 4.16 over 4. In their objective evaluations however, they found the opposite. Despite the higher MOS, the authors conclude that the VAE outperformed the DAE (Mansouri & Lachiri, 2021, p. 1606) based on their objective evaluation. Since their subjective and objective measurements contradict, there is no positive correlation between the used measures and the perceived naturalness. However, there is no negative correlation either. This means that the used measures are not adequate to evaluate the naturalness of synthetic laughter, rendering their conclusion incorrect. This also holds for the objective measures in their 2019 paper.

Luong and Yamagishi (2021b) then attempt to synthesise laughter from waveform silhouettes, which is synonymous for the acoustic envelope of a signal, from the Unity Laughs SFX package (Sound Ex Machina, 2018). For this attempt they used an architecture that, like the VAE, can create new data: the Generative Adversarial Network (GAN).[8] This architecture consists of a generator network and a discriminator network, that are trained against each other. The generator starts from random noise and tries to fool the discriminator, whilst the discriminator is trained on human data and distinguishes between fake and human data. Because these two networks compete they can achieve a better naturalness than the VAE, which is trained against a predefined loss function. This makes the GAN the SOTA model for laughter synthesis.

In their subjective evaluation, Luong and Yamagishi (2021b) evaluated quality and speaker similarity, rather than naturalness. For quality they achieved a MOS of 1.8 and for speaker similarity 2.6. Compared to a quality of 4.2 and a

[7]The Autoencoder (AE) and VAE architectures are explained by Mansouri and Lachiri (2021, pp. 1594–1598).

[8]The GAN architecture is explained by Kong et al. (2020, pp. 2–5).

speaker similarity of 3.5 for natural laughter. Additionally, it is important to note that, like Tits et al. (2020), Luong and Yamagishi (2021b) utilise transfer learning by pretraining the model with speech data. They used the Voice Cloning Toolkit (VCTK) dataset (Yamagishi et al., 2019) for this.

Evaluations

The evaluations in this section share their use of a 5-point Likert scale MOS on naturalness with the evaluations discussed in subsection 2.3.1. One major difference however, is that there is no copy-synthesis, because in DNN synthesis there is no explicit parameter extraction. Another major difference is that, with the exception of Tits et al. (2020), no standard deviations were reported.

Furthermore, the majority of papers from this section also have some similarities with the ones discussed in subsection 2.3.2, namely in *what* they evaluated and in their lack of calibration. Firstly, Luong and Yamagishi (2021b) evaluated quality and speaker similarity, instead of naturalness. Secondly, Mansouri and Lachiri (2019, 2020, 2021) did not include human laughter in their evaluations as reference signal, meaning that their results are likely higher than if a ground truth reference had been included. In contrast, they tried to find an objective measure for synthetic laughter, the absence of which is a limiting factor in the field of laughter synthesis (H. Mori et al., 2019, p. 520). However, by showing the absence of a correlation with naturalness, I have demonstrated the inadequacy of their objective measures and with that I have falsified their conclusions. Accordingly, their findings cannot reliably be compared to other research in the field and have thus not been summarised in a table.

Table 2.4 below provides the best naturalness evaluations achieved by publications in this section that did include human laughter.

Publication	Laugh gender	Best avg. score synthesised	Avg. score human	Model type
H. Mori et al. (2019)	F M	2.16 3.14	4.5 4.74	CNN CNN
Tits et al. (2020)	-	3.28 (std: 1.06)	4.10 (0.91)	Seq2seq

Table 2.4: Best synthetic laughter naturalness MOS (5-point Likert scale) per synthesis method for every calibrated DNN synthesis publication

From the best case scenario MOS of 3.28 on a scale from 1 to 5, I conclude that moderate naturalness has been achieved with DNN synthesis. Although this is an improvement over the naturalness achieved with parametric synthesis (see table 2.3), it is significantly lower than the very high naturalness of 4.74 that could hypothetically be achieved (see table 2.2).

2.3.4 Defining naturalness within laughter synthesis

The lack of a definition of '*naturalness*' I established in section 1.2 has also been deemed to be an issue in laughter synthesis by Tits et al. (2020, p. 3) and Urbain et al. (2013b, p. 157). Due to the lack of a definition, there are multiple ways in which '*naturalness*' can be interpreted. Examples of which are, but are not limited to (Urbain et al., 2013b, p. 157):

1. Whether or not a laugh sounds distorted.
2. Whether or not a laugh is perceived as being authentic.
3. Whether or not it matches to the listeners expectations (see section 1.1).

This interpretability issue renders the results unreliable, since it is unsure whether the participants evaluated the same aspects of naturalness (see section 1.2). Both papers deal with it in their own way, but neither solves the issue. Choosing the most frequent interpretation as definition however, reveals an issue with the methods used by Luong and Yamagishi (2021b).

Current definitions and why they do not work

The two ways the interpretability issue is dealt with in the current literature are as follows:

1. Urbain et al. (2013b, p. 157) decided, in line with previous research, not to provide a definition. This way the outcomes of their research can still be compared to the outcomes of the previous research.
2. Tits et al. (2020, p. 3) chose to define '*naturalness*' as '*human-likeness*' (cf. figure 1.1), which is defined by Merriam-Webster (n.d., 2023a, 2023b) as:

   Human
   1: of, relating to, or characteristic of humans
   Likeness
   3: the quality or state of being like: RESEMBLANCE

   The problem with this definition however, is that it suffers from the interpretability issue itself, due to the wide variety of characteristics we humans portray (Bertelsen et al., 2009, pp. 398–429). Additionally, it does not cover all interpretations of naturalness. Take for example the second aforementioned interpretation: an inauthentic laugh is still human-like.

Either way, the interpretability issue remains unresolved. Hence, these papers corroborate my earlier statement that the field would greatly benefit from a precise definition of '*naturalness*' (see section 1.2), since it will provide clarity and thus create reliable results across research.

USING THE MOST FREQUENT INTERPRETATION AS DEFINITION

The most frequent interpretation of '*naturalness*' in the literature review is '*authenticity*': the difference between acted and spontaneous laughter (Luong & Yamagishi, 2021b; H. Mori et al., 2019; Tits et al., 2020; Urbain, 2014; Urbain et al., 2013b).[9]

Table 2.5 below provides the datasets encountered during the literature review[10], their authenticity, and which publications used them. This shows that the SOTA LaughNet by Luong and Yamagishi (2021b) use acted laughter instead of spontaneous laughter.

Dataset	Authenticity of used data	Publications
AVLC (Urbain et al., 2010)	Induced	Urbain et al. (2013a) Urbain et al. (2013b) Urbain et al. (2014) Bollepalli et al. (2014) Urbain (2014) Sathya et al. (2013) Juhitha et al. (2018) Mansouri and Lachiri (2020, 2021) Mansouri and Lachiri (2021)
OGVC (Arimoto et al., 2012)	Spontaneous	H. Mori et al. (2019)
AmuS (El Haddad et al., 2017)	Induced	Tits et al. (2020) Mansouri and Lachiri (2020, 2021)
Unity Laughs SFX package (Sound Ex Machina, 2018)	Acted	Luong and Yamagishi (2021b)

Table 2.5: Laughter datasets from the literature review, their authenticity, and the publications that used them, in chronological order

It is common knowledge that acted and spontaneous emotional expressions, such as laughter, are not the same. This is quickly confirmed by querying Google Scholar for the combination of 'acted', 'spontaneous', and 'laughter'.[11] Taking the research with at least 100 citations[12] from the first 3 pages[12] results in (Bryant & Aktipis, 2014; Lavan et al., 2016). Both authors compared acted and spontaneous laughter by performing an objective acoustic analysis using *Praat* (Boersma, 2011), followed by a subjective perceptual analysis. Both authors found significant differences on either level.

After the acoustic analysis, Bryant and Aktipis (2014) performed three subjective experiments: firstly, they examined whether participants could distinguish between the two types of laughter in a forced-decision task. In the second and third experiment however, they altered the speed of the laughs in both directions. They found that participants could detect the authenticity well above

[9]These publications did not define '*naturalness*' as '*authenticity*', but merely mention the interpretation amongst others.

[10]Speech datasets used for transfer learning excluded.

[11]I deliberately used these terms instead of 'authenticity' to keep the query unbiased, since 'authenticity' contains 'authentic', which is synonymous for 'spontaneous'.

[12]These numbers are chosen semi-arbitrarily to ensure a workable amount of reliable results.

chance level, and that faster laughs were perceived as being more authentic than slower laughs. Furthermore, they found that participants could accurately detect acted laughter when it was slowed down, but not spontaneous laughter.

After the acoustic analysis, Lavan et al. (2016) performed two subjective experiments: firstly, they examined whether participants could distinguish between the two types of laughter by recording affective ratings of valence, arousal, and authenticity for all laughter samples. Secondly, they studied the relation between the Harmonics-to-noise ratio (HNR) and authenticity by having trained participants evaluate the breathiness, nasality, and mouth opening of all laughter samples. They too found that participants could detect the authenticity well above chance level. Furthermore, they found that spontaneous laughter is significantly more nasal than acted laughter, and that the objective acoustic features accurately predict the subjective affective ratings.

Despite these findings from Bryant and Aktipis (2014) and Lavan et al. (2016), Luong and Yamagishi (2021b) chose to use acted laughter data, generating synthetic laughter with suboptimal naturalness. To increase the naturalness of the synthetic laughter, their experiment should be redone using spontaneous laughter data.

2.3.5 Key findings

In this subsection, I summarise the key findings extrapolated from the content of the publications from the literature review, their evaluations, and the lacking definition of naturalness, leading up to my research questions.

Content of the publications

The most prominent problem in the field is the scarcity of clean recordings of isolated natural laughter and the difficulty of collecting more. This holds especially for DNN synthesis methods, which require a vast amount of data to learn from (see table 2.1). What makes this problem more complex, is that there is a large variability in natural laughter, meaning that, even with vast amounts of data, many types of laughter are likely not represented. Several attempts have been done to bypass this problem, with the most promising ideas being the use of generative models and transfer learning. Not only did Luong and Yamagishi (2021b) use the best generative model, but they also incorporated transfer learning into the synthesis procedure of LaughNet, making it the SOTA. Additionally, more research is required to determine which information can consistently be conveyed through isolated laughter (Urbain, 2014).

Evaluations

Even with an average naturalness MOS of 4.45 for human laughter, on a scale of 1 to 5, the highest MOS for synthetic laughter only reached 3.28 (see table 2.4). This means that natural laughter synthesis has not yet been achieved, despite the promising capabilities of DNNs (see table 2.2). Furthermore, the absence of an objective measure for the naturalness of synthetic laughter still poses a significant limitation for research into laughter synthesis (H. Mori et al., 2019) after the unsuccessful attempts from Mansouri and Lachiri (2019, 2020, 2021) to find such measure.

Definition of naturalness

Another obstruction to achieving natural laughter synthesis is the lack of a clear definition of the term: '*natural*' (see section 1.2 and subsection 2.3.4). The most common interpretation is that of '*authenticity*'. A brief inquiry of the most cited sources on Google Scholar shows that the acoustic and perceptual properties of isolated acted and spontaneous laughter are significantly different (Bryant & Aktipis, 2014; Lavan et al., 2016). Despite this, Luong and Yamagishi (2021b) used acted laughter instead of spontaneous laughter. This has likely negatively affected the naturalness of the synthetic laughter produced by LaughNet, but this cannot be checked since Luong and Yamagishi (2021b) evaluated quality and speaker similarity instead of naturalness.

Consequently, the naturalness of synthetic laughter from LaughNet, under the interpretation of '*authenticity*', remains to be discovered. Before this can be determined however, it should be researched whether or not the acoustic and perceptual differences between acted and spontaneous laughter actually affect the output of the model. It is very well possible that the input data is too high-level to capture the characteristics that are used to distinguish between the two types of laughter. To explore this possibility, I will research the difference between acted and spontaneous laughter and its relevance for laughter synthesis using LaughNet.

2.4 Research questions and hypotheses

The research goal of determining whether the difference between acted and spontaneous laughter is relevant for laughter synthesis can effectively be split up into three separate questions:

1. What is the difference between acted and spontaneous laughter on the acoustic level?
2. Are the acoustic differences affected by the synthesis process of LaughNet, and, if so, how?
3. Can the differences be detected by human listeners?

As mentioned in section 2.3, I will limit myself to isolated laughter to limit the number of possible confounding variables. I will also limit myself to voiced laughter in particular, in line with Sathya et al. (2013) (cf. Bachorowski and Owren (1995)), because a positive state of mind makes the listener more lenient in their judgments (Clore & Huntsinger, 2007) effectively lowering the subjective standard for acceptable naturalness. This should positively affect the evaluations, resulting in higher MOS.

Therefore, my first research question will be:

1. How do the acoustic features of isolated acted voiced laughter compare to the acoustic features of isolated spontaneous voiced laughter?

Based on the findings of Bryant and Aktipis (2014) and Lavan et al. (2016), I hypothesise that the acoustic features of the isolated acted voiced laughter used by Luong and Yamagishi (2021b), are significantly different from the acoustic features of isolated spontaneous voiced laughter.

If the hypothesis is confirmed, the findings of Bryant and Aktipis (2014) and Lavan et al. (2016) are strengthened and it can be researched if and how the two types of laughter are affected by the generative process. If the hypothesis is nullified however, their claims are weakened and the research goal will have been achieved, because there is no difference between acted and spontaneous laughter on the generation level. In the event of any significant differences found between the laughter data, further research is required to explain the reason behind it. Consequently, the difference that exists between the two types of laughter must be on a higher level and thus is irrelevant for laughter synthesis. In the case of confirmation however, the research goal will not yet have been achieved, giving rise to the second research question:

2. How are the acoustic features of isolated spontaneous voiced laughter and isolated acted voiced laughter affected by the generative process of LaughNet (Luong & Yamagishi, 2021b)?

Based on the fact that Luong and Yamagishi (2021b) feed the model waveform silhouettes of laughter, my hypothesis is that information from the lower-level relevant acoustic features (Lavan et al., 2016), is lost during the generative process of LaughNet (Luong & Yamagishi, 2021b). Accordingly, it will be more difficult

to distinguish between isolated acted voiced laughter and isolated spontaneous voiced laughter post-synthesis using LaughNet (Luong & Yamagishi, 2021b). If the hypothesis is nullified however, there should be no difference in the ability to distinguish between isolated acted and spontaneous laughter pre- and post-synthesis.

Lastly, regardless of the result, it matters whether or not people can accurately detect this result, leading to the last research question:

3. Can people accurately distinguish between isolated acted voiced laughter and isolated spontaneous laughter pre- and post-synthesis using LaughNet (Luong & Yamagishi, 2021b)?

Based on the findings of Lavan et al. (2016) I predict that people can distinguish isolated acted voiced laughter from isolated spontaneous voiced laughter pre-synthesis. Based on my hypothesis for research question 2 however, I hypothesise that people will not be able to distinguish between isolated acted voiced synthetic laughter and isolated spontaneous voiced synthetic laughter as well, because the distinction will be harder post-synthesis. If my hypothesis for this third question is rejected, that would imply that either type of laughter can be used to produce synthetic laughter, regardless of whether or not the lower-level acoustic features are affected by the synthesis procedure of LaughNet.

The findings of this research will contribute to the field by providing insight into whether or not actors must be used in the collection of clean recordings of data needed for the synthesis of laughter that is perceived as natural. Additionally, it will pave the way for the creation of an objective measure to evaluate the perceived naturalness of synthetic laughter, because I will have determined the acoustic features needed to distinguish between acted and spontaneous laughter, which is one part of naturalness.

3

Methodology

The aim of this study is to investigate whether the naturalness of synthetic laughter, with a focus on the interpretation of authenticity (i.e. acted versus spontaneous laughter), can be improved. This research is a case study that specifically examines the SOTA laughter synthesis model called LaughNet (Luong & Yamagishi, 2021b), which has been trained using acted laughter, despite earlier findings of differences between acted and spontaneous laughter (Bryant & Aktipis, 2014; Lavan et al., 2016).

To reduce the impact of confounding variables and align with the positive reception of "voiced laughter" as reported in prior research (Sathya et al. (2013); see 2.3.2), this study specifically focuses on "isolated voiced laughter". For the sake of readability, moving forward, this will be referred to simply as "laughter".

For clarity, I provide the research questions again:

1. How do the acoustic features of acted and spontaneous laughter differ and do these findings correspond to previous research (Bryant & Aktipis, 2014; Lavan et al., 2016)?
2. To what extent, if any, does the synthesis procedure of LaughNet affect the distinctive acoustic features in 1?
3. Can human listeners perceive the acoustic differences between acted and spontaneous laughter in human and synthetic laughter?

To answer these questions, the methodology involves analysing the acoustic features of both acted and spontaneous laughter, examining the influence of the synthesis procedure of LaughNet on these features, and conducting perceptual evaluations with human listeners to assess the naturalness of the synthesised laughter.

For clarity, each research question will be discussed in its own section (3.1, 3.2, and 3.3), with subsections dedicated to research design (3.x.1), data (3.x.2), preprocessing (3.x.3), and materials (3.x.4). For the third research question, which involved a survey, an additional subsection has been added, dedicated to participants (3.3.5).

3.1 Experiment 1

In experiment 1, I set out to establish a baseline for this study by confirming the findings from previous research Bryant and Aktipis (2014) and Lavan et al. (2016) regarding the differences between acted and spontaneous laughter, specifically in terms of their respective acoustic features.

3.1.1 Research design

To test the findings from Bryant and Aktipis (2014) and Lavan et al. (2016), I adopted their research methods. Lavan et al. (2016) found that subjective ratings of real laughter could be accurately predicted by objective features (see subsection 2.3.4), which suggests that this could also apply to synthetic laughter. Since there is no established objective measure for the naturalness of synthetic laughter, it was decided to use the objective acoustic features as a classifier for acted and spontaneous laughter, which would serve as an objective measure for naturalness in this study (see subsection 2.3.5).

If the classifier trained on the acoustic features would be able to accurately distinguish between the two types of laughter, the findings from Bryant and Aktipis (2014) and Lavan et al. (2016) would be confirmed. By analysing the classification boundary through factor analysis, and by studying the acoustic features of misclassified laughter samples, I would also gain insight about the proportional contribution of each acoustic feature to the laughter authenticity.

The classifier could be reused as a control experiment for the third research question. By comparing the classification accuracy pre- and post-synthesis, as well as the acoustic features and subjective authenticity ratios of the misclassified laughter samples, I would be able to confirm whether the synthesis procedure affected the naturalness of the laughter. The multifunctional applicability of this solution validates the chosen method.

3.1.2 Data

Determining the relation between the acoustic features of acted and spontaneous laughter, required data from both types of laughter. To limit the impact of confounding variables, the same acted data was used as Luong and Yamagishi (2021b): the Unity Laughs SFX package from Sound Ex Machina (2018). The spontaneous data from Bryant and Aktipis (2014) and Lavan et al. (2016) however, could not be reused, hence the MULAI Corpus (Jansen et al., 2018) was carefully selected as substitute.

Acted laughter

The acted laughter data used by Luong and Yamagishi (2021b) was the Unity Laughs SFX package from Sound Ex Machina (2018).[1] This package contains 300 recordings with a total duration of 72 minutes. Out of the 300 files, 109 contain laughs, cheers, and applause from groups of varying size and consistency, with sizes ranging up to over 200 people. The other 191 files contain laughs from various individuals that were deliberately varied in age, gender, and laughing style so as to capture a wide variety of laughter. Out of these 191 individual recordings, 80 are of a female and 111 are of a male. The sound files are stored in .wav format with a sampling rate of 48kHz and a bit-depth of 16 bits and have an average duration of 14.4 seconds per file.

Spontaneous laughter

For the spontaneous laughter database, there were three requirements:

1. Should contain spontaneous laughter.
2. The spontaneous laughter should occur in interaction, because a major limitation of the AVLC database (Urbain, 2014, p. 37) was that its laughter was induced in a non-interactive manner, whilst laughter is a social signal (Campbell, 2007a) (see subsection 2.3.1).
3. The database needs to be as recent as possible, because, as mentioned before: knowledge and improvements come with time.

The spontaneous laughter used by Lavan et al. (2016) did not adhere to requirement 2, and the spontaneous laughter used by Bryant and Aktipis (2014) came from another project and was therefore inaccessible. The database I found that satisfied all three requirements, was the MULAI Corpus (Jansen et al., 2018).

The MULAI corpus contains about 4.35 hours of annotated audio-, video-, and physiological data from 26 participants, recorded over 13 sessions in which the participants interacted with another participant on the basis of 3 specific tasks. From these participants 14 were male and 12 female. The audio files are stored in .wav format with a sampling rate of 16kHz and a bit-depth of 16 bits and have an average duration of 89 seconds per file. Additionally, the corpus contains personal evaluations from the participants about how funny they deemed themselves and their interlocutor during task 3: the joke-telling task. Lastly, the corpus also contains demographic information from each participant, as well as information regarding their personality.

[1] https://assetstore.unity.com/packages/audio/sound-fx/voices/laughs-sfx-111509 (last accessed: Nov. 1, 2022)

3.1.3 Preprocessing

To work with the acoustic features from the laughter, it first needed to be extracted from the recordings in the databases. This required the extraction of laughter in balanced amounts across gender and authenticity, which then needed to be matched in bit rate, sampling rate, and duration. Only then could the acoustic features needed to be selected and extracted from that laughter.

Laughter extraction

The acted laughter recordings from the individuals in the Laughs SFX package already contain only laughter, whereas the spontaneous laughter recordings from the MULAI corpus are captured in full dyadic interactions. Accordingly, the spontaneous laughter still needs to be extracted from the recordings.

To get the best recordings I first looked at the personal funniness evaluations to determine which laughs were most likely spontaneous and which ones were most least likely not spontaneous. The MULAI participants had to state how much they agreed with the following statements: "I think I was funny" and "I think the other was funny". They had to do so on a 5-point Likert scale ranging from 1 (completely disagree) to 5 (completely agree). If participants laughed in a recording whilst they indicated that they believed either person not to be funny, then the laugh was likely not spontaneous. Therefore, I only extracted laughs from recordings with a rating of 3 (neutral) or higher, focusing on the laughter around the joke of the person they indicated to be funny. This left me with a total of 75 files to extract laughter from. From these files, I only extracted isolated voiced laughs that were annotated, were not speech laughs, and did not overlap with sounds created by their interlocutor. Sometimes laughter was not annotated, but individual bouts (see figure 2.1) were. In these cases I extracted as much bouts as possible without capturing other sounds. This left me with the following array of spontaneous laughs, with a minimum duration of 0.5 seconds, an average duration of 2 seconds, and a maximum duration of 4 seconds:

Gender	Full laughs	Partial laughs	Total
F	9	8	**17**
M	11	6	**17**
Total	**20**	**14**	**34**

Table 3.1: Spontaneous laughs extracted from the MULAI Corpus

As can be seen in table 3.1, roughly 60% of the extracted spontaneous laughter consisted of full laughs and 40% of partial laughs. This ratio is slightly smaller for female laughs than for male laughs.

To discover patterns in the features from the data, neural networks need to compare data of equal proportions (see section 2.2). A little bit of variation in the length of the data is not necessarily problematic, as shorter samples can be padded with silence, but longer samples cannot be squeezed together. The amount of padding that can be added is not unrestricted however, as there is no laughter pattern in silence. Accordingly, I needed to extract parts of equal duration with the same amount of data per second. To achieve that the data firstly needed to be matched in sampling rate. This action was performed using SoX. To achieve the best possible quality audio, the global settings were set to guard against clipping and the resampling quality was set to 'very high'. A common sampling rate of 22.050kHz was used, so the acted laughter (48kHz) was downsampled and the spontaneous laughter (16kHz) was upsampled. Proceeding, the 4 second samples of acted laughter were extracted from the longer acted laughter recordings (avg. 14.4 s) in random fashion, conform Luong and Yamagishi (2021b, p. 3). Important to note here is that the 4 seconds are slightly shorter than the 6 second samples used by Luong and Yamagishi (2021b, p. 3).

To minimise the impact of confounding variables, I used the same recordings as Luong and Yamagishi (2021b). I determined which recordings those were by carefully listening to both the samples (Luong & Yamagishi, 2021a) and the acted laughter recordings. This resulted in 4 female samples and 4 male samples. Since this created an imbalance between acted and spontaneous laughter samples, I extracted 13 more samples from both female and male acted laughter recordings, such that I ended up with 17 female acted laughter recordings and 17 male acted laughter recordings. This left me with a total of 68 laughter samples.

Acoustic feature extraction

Bryant and Aktipis (2014) and Lavan et al. (2016) extracted objective acoustic features from the laughter through *Praat* (Boersma, 2011). Accordingly, I did so too, using the same settings as Lavan et al. (2016). To automate the extraction process I used a Python library for *Praat* called *Parselmouth*, version 0.4.3. This library runs on *Praat* version 6.1.38.

Table 3.2 below provides the acoustic features extracted by each author, divided into duration-, loudness-, and pitch related features. This shows that Bryant and Aktipis (2014) and Lavan et al. (2016) extract roughly the same acoustic features, with the main differences being the specific durations extracted by Bryant and Aktipis (2014), and the additional pitch features extracted by Lavan et al. (2016). The right most column shows the features used in this thesis for ease of comparison. The reasoning behind it follows after the table.

		Authors		
		Bryant and Aktipis (2014)	Lavan et al. (2016)	Weggeman (this thesis)
Acoustic features	**Duration**	- Call number - Bout duration (ms) - Mean call duration (ms) - Mean Intervoicing interval (IVI) (ms) - Mean rate of IVI per bout (%)	- Total duration (s) - Burst duration[2] (s)	- Total duration (s)
	Loudness	- Decibel standard deviation (dB)	- Intensity (dB)	- Intensity (dB)
	Pitch	- F0 mean (Hz) - F0 standard deviation (Hz) - F0 minimum (Hz) - F0 maximum (Hz) - F0 range (Hz)	- F0 mean (Hz) - F0 variability[3] (Hz) - F0 minimum (Hz) - F0 maximum (Hz) - **F0 range (Hz)**[4] - **F0 range (semitones)**[4] - Percentage unvoiced segments (%) - **Mean HNR (%)**[4] - **Spectral centre of gravity (Hz)**[4]	- F0 mean (Hz) - F0 variability[3] (Hz) - F0 minimum (Hz) - F0 maximum (Hz) - Percentage unvoiced segments (%)

Table 3.2: Acoustic features extracted

Post extraction, Lavan et al. (2016, p. 139) tested all their acoustic features for statistically significant distinctiveness for laughter authenticity by performing independent two-tailed t-tests between them. The features they reported to be statistically insignificant have been highlighted in table 3.2. This also means that features in their column that have not been highlighted are known to be statistically significantly distinctive for laughter authenticity.

For this thesis I extracted all the significant features from Lavan et al. (2016), with the exception of the burst duration. The reason why I did not extract the burst duration, nor any of the specific durations extracted by Bryant and Aktipis (2014), is that these durations require a detailed level of transcriptions. These transcriptions were not available for all data used in this thesis, since the acted laughter from the Laughs SFX package (Sound Ex Machina, 2018) has no annotations at all. Having no prior experience transcribing laughter, I did not feel comfortable dissecting the laughter into bouts and calls myself, as doing it incorrectly might adversely impact the results. Furthermore, I chose to extract intensity over the Decibel standard deviation extracted by Bryant and Aktipis (2014), because the intensity is known to be statistically significantly distinctive.

Before extracting the features, the leading and trailing silences had to be trimmed off the data, because the total duration and the percentage unvoiced

[2] 'Burst' is a synonym for 'call' (see figure 2.1).
[3] Variability is the standard deviation divided by the total duration.
[4] Statistically insignificantly distinctive acoustic features (see Lavan et al. 2016, p. 139).

segments are affected by it, since those silences do not occur within the laughter. Based on visual inspection I decided upon a 1% maximum amplitude cutoff.

After extraction, the data needed to be rid of outliers and needed to be normalised. For the outlier removal I used a z-score of 3, and for the normalisation I used min-max normalisation. To ensure that nothing was missed in this process, I followed up with a visual inspection.[5]

3.1.4 Materials

To create a classifier on the acoustic features of laughter, samples (3.1.4) are needed to extract the acoustic features from. Furthermore, a classification model (3.1.4) is required to classify the samples based on their acoustic features.

Laughter samples

The laughter samples used in this experiment were the 68 laughter samples extracted in subsection 3.1.3. These 68 laughter samples consisted of 17 acted female laughs, 17 acted male laughs, 17 spontaneous female laughs, and 17 spontaneous male laughs. This data was split into a train and test set, using a 75%-25% ratio.

Model

For the machine learning model I looked for inspiration in papers that classified different types of laughter using acoustic properties. This resulted in the papers by Ataollahi and Suarez (2019), Folorunso et al. (2020), Kantharaju et al. (2018), and Tanaka and Campbell (2014). Aside from Ataollahi and Suarez (2019), all papers used a SVM. Although the 3D-CNN used by Ataollahi and Suarez (2019) outperformed the SVMs, it comes with a highly increased computational cost. Since the SVMs performed well and this thesis is more exploratory in nature, the increased computational cost does not outweigh the higher classification accuracy. Accordingly, I opted for a SVM.

To optimise the hyperparameter settings of the SVM I used grid search over a range of kernels, regularisation parameter values, and gamma values for non-linear kernels.[6]

[5]For the Python implementation of the data preparation, see lines 46–249 on https://github.com/5weggeman/laughter_authenticity_classifier/blob/main/classifier.py (last accessed: Jun. 15, 2023).

[6]For the Python implementation of the training of the classifier, see lines 253–291 on https://github.com/5weggeman/laughter_authenticity_classifier/blob/main/classifier.py (last accessed: Jun. 15, 2023).

3.2 Experiment 2

In experiment 2, I set out to determine in what way the acoustic features relevant for the distinction between acted and spontaneous laughter, that were expected to be found in experiment 1 (see section 3.1), would be affected by the synthesis process of LaughNet.

3.2.1 Research design

Having established that LaughNet uses waveform silhouettes, also known as acoustic envelopes (see subsection 2.3.3), there were two ways to investigate the effect of its synthesis procedure on the distinctive acoustic features. Firstly, a theoretical acoustic analysis of the datatype could be performed to establish which acoustic features are captured in it. Secondly, a practical analysis could be performed by having synthetic laughter from LaughNet classified by the SVM classifier from experiment 1, and then comparing the performance to that of human laughter. To ensure robust findings both approaches were enlisted.

Theoretical approach

In the theoretical approach I analysed the acoustic envelope format to figure out which of the acoustic features from the factor analysis were captured in it. For this analysis I used the temporal framework of speech created by Rosen in 1992. This framework consists of three temporal features: the envelope, the periodicity, and the fine-structure. For each temporal feature, Rosen (1992) described *which* acoustic features are represented in it and *how* they are represented in it. A visual representation of this decomposition is provided in figure 3.1 below (Lizarazu, 2017). The envelope captures the slow modulations of the signal, whilst the fine-structure captures the fast modulations of the signal. Both of these temporal features can have signs of periodicity in them, shown by repetitive patterns.

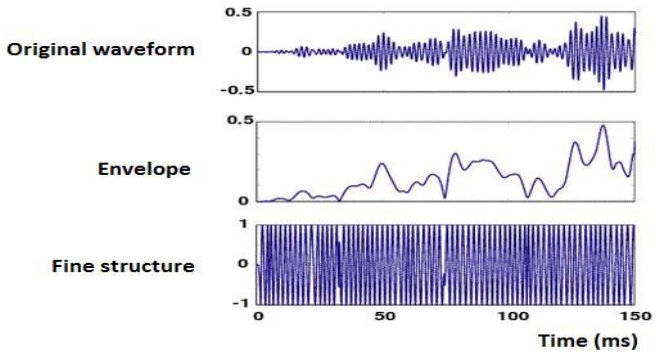


Figure 3.1: Waveform decomposition into temporal features; cf. Rosen (1992) (Lizarazu, 2017)

Practical approach

In the practical approach, I trained LaughNet, following the procedure of Luong and Yamagishi (2021b), and synthesised laughter with it. I then classified the laughter using the SVM classifier from experiment 1 (see section 3.1) and evaluated its performance using the exact same evaluation metrics as in experiment 1, such that their evaluations could be compared. This would provide insight into the alterations made by the synthesis procedure of LaughNet, or the absence thereof.

3.2.2 Data

The theoretical approach does not require any data, but the practical approach uses the same data as described in subsection 3.1.2. Additionally, this experiment requires speech data from the VCTK corpus (Yamagishi et al., 2019). Luong and Yamagishi (2021b) pretrained LaughNet with this data to apply transfer learning (see subsection 2.3.3).

Pretraining data

The VCTK corpus contains about 41.6 hours of stereo, annotated read-speech recordings from 110 participants, divided over 44455 files. The recordings are stored in .wav format with a sampling rate of 48kHz and a bit-depth of 16 bits, and have an average duration of 3.37 seconds per file. Furthermore, the corpus contains demographic information from each participant.

3.2.3 Preprocessing

Only the pretraining data for the practical approach has not yet undergone preprocessing. As mentioned in the research design (see section 3.2.1), the methods from Luong and Yamagishi (2021b) were leading. Accordingly, their preprocessing method was replicated. Furthermore, the pretraining data needed to be matched in bit rate, sampling rate, and duration to the laughter data (see subsection 3.1.3).

Speech extraction

I sought clarification from Luong and Yamagishi (2021b) to address specific ambiguities in the paper. Through this correspondence, I confirmed that the training setup for LaughNet was based on previous research, specifically:

- Only the recordings from microphone 1 were used.
- The leading and trailing silences from the VCTK dataset were trimmed.[7]
- The "common" utterances were used for validation, whilst "uncommon" utterances were used for training.

Accordingly, I implemented these actions in the preprocessing procedure as well. In the last point I interpreted "common" as shared: each participant read the rainbow passage[8] and elicitation paragraph[9]. Through manual comparison of the files from the first five speakers in the corpus I established that this concerned files 0 to 24. Therefore, I separated these files from the rest and stored them for the training validation.

Although the VCTK corpus contained about 41.6 hours of read-speech from 110 speakers, Luong and Yamagishi (2021b, p. 3) only used 24.4 hours from 100 speakers. This meant that I first had to drop the data from 10 of the speakers. Ideally the final data would generalise as well as possible. Therefore, to ease the working process, I firstly removed the one speaker with a diverging participant ID, and the two speakers with "incomplete" data. Secondly, I removed speakers with similar, frequently occurring demographics, such that the final data would capture the largest variety of accents in the dataset.

After reaching the 100 speaker requirement, I still needed to reduce the remainder of the data to 24.4 hours. Since 24.4 is roughly 60% from 41.6 hours of data, I needed to drop 40% of the data. The rainbow passage and elicitation paragraph however, cover about 5% of the data, so dropping 40% would result in less than 24.4 hours. Therefore I reduced the dropping percentage to 35%.

Then, I needed to split the data into training and testing data, for which I used an 85%-15% split. Since the rainbow passage and elicitation paragraph cover about 5% of the data, the validation set needed to be supplemented with 10% from the total number of files to achieve an 85-15 split. Accordingly, I supplemented the validation set with random files from the remaining data to the validation set by shuffling the remaining data using the Python *'random'* module, and then adding a partition equal to 10% from the total number of files to the validation set.

Lastly, in line with the data manipulation performed in subsection 3.1.3, I resampled the files to 22.050kHz, such that the High Fidelity (HiFi)-GAN can learn patterns from the data. Before resampling, I first trimmed the leading and trailing silences and added 250ms of padding.[10]

[7]https://github.com/nii-yamagishilab/vctk-silence-labels (last accessed: Mar. 7, 2023)

[8]http://web.ku.edu/~idea/readings/rainbow.htm (last accessed: Mar. 7, 2023)

[9]http://accent.gmu.edu (last accessed: Mar. 7, 2023)

[10]For the implementation of the preprocessing see https://github.com/5weggeman/hifi-gan-laughnet/blob/master/preprocessing.py (last accessed: Mar. 7, 2023)

3.2.4 Materials

To test the influence of the synthesis procedure of LaughNet on the acoustic features of laughter in a practical manner, I synthesised laughter using LaughNet. This required laughter samples and the LaughNet model.

Laughter samples

The laughter samples used in the listening test were identical to the ones used in experiment 1 (see subsection 3.1.4).

Model

Luong and Yamagishi (2021b) created three different versions of LaughNet, with different quantisation methods and levels for the waveform-silhouettes extracted from laughter: a 256-bin linear quantisation model, an 8-bit (256-bin) mu-law quantisation[11] model, and a 4-bit (16-bin) mu-law quantisation model. Due to time and computing power limitations I could only train one model, therefore I opted for the model with the highest quality (see subsection 2.3.3) and the lowest computational cost: the 4-bit mu-law model.[12]

The waveform silhouette module had to be generalised, as it was only made to work for one example. Additionally, Luong and Yamagishi (2021b, p. 3) randomly scaled the laughter silhouettes, hence I added that as well. Using only the 4-bit quantisation, I commented the other two options out. Lastly, waveform silhouettes were embedded using a one-hot encoding.

From the contact with Luong and Yamagishi (2021b), I learned that the feature embeddings should have dimensions BxCxT, with B standing for batch size, C for channels, and T for the number of samples times the duration in seconds. Luong and Yamagishi (2021b) used a batch size of 16 and used 6 second segments. There are constantly 2 channels, because the waveform silhouette has an upper envelope and a lower envelope. The way in which the feature embeddings were reshaped by Luong and Yamagishi (2021b) to showcase the example, caused the data to have different dimensions than the ones mentioned above. Consequently, I had to adjust the notebook.py file and the silhouette.py file to arrive at the correct dimensions.[13]

[11] Mu-law quantisation is a variation on logarithmic quantisation that is more suitable for telecommunication purposes.

[12] For the implementation see https://github.com/5weggeman/hifi-gan-laughnet (last accessed: Mar. 7, 2023)

[13] For the correction see https://github.com/5weggeman/hifi-gan-laughnet/blob/master/waveform_silhouette.py (last accessed: Mar. 7, 2023) and https://github.com/5weggeman/hifi-gan-laughnet/blob/master/silhouette.py (last accessed: Mar. 7, 2023)

Due to aforementioned lack of computing power, I could only use a batch size of 8 and 3.75 second segments. Given that the leading and trailing silences of the initially extracted 4 second segments (see section 3.1.3) have been trimmed, most segments should already be slightly shorter than 4 seconds, hence this should not cause any problems. These 3.75 second segments were randomly extracted from the 4 second segments during the training of the model.

3.3 Experiment 3

In experiment 3, I set out to determine whether the acoustic differences between acted and spontaneous laughter, that were expected to be found in experiment 1 (see section 3.1), are detectable for human listeners in both human and synthetic laughter.

3.3.1 Research design

The most standard way to examine the capability of human listeners to detect the authenticity of laughter based on the acoustic features is through a subjective listening test. Lavan et al. (2016) also performed a subjective listening test (see section 2.3.4), hence I adopted their research methods. They recorded subjective affective ratings of valence, arousal, and authenticity. Valence and arousal are two dimensions used in the classification of emotions. Although most emotion classification models include at least these two dimensions, there is no consensus on the number of dimensions Kort et al. (2001), Plutchik (1991), and Russell (1980) (see section 1.3). Consequently, the ratings likely explain the findings insufficiently. To avoid this issue, I reduced the subjective affective evaluations to a forced-decision listening task between acted and spontaneous laughter.

In this experiment the independent variable was the authenticity of the laughter samples and the dependent variable was the perceived authenticity ratio per laughter sample. If the classification scores and authenticity ratios of the laughter samples would be correlated, the additional findings of Lavan et al. (2016) would also be confirmed.

3.3.2 Data

The data used in this experiment was the same as the data used in experiment 1 (see subsection 3.1.2).

3.3.3 Preprocessing

The preprocessing used in this experiment was the same as the preprocessing performed in experiment 1 (see subsection 3.1.3).

3.3.4 Materials

To find out if human listeners are capable of detecting the acoustic differences between acted and spontaneous laughter, that were expected to be found in experiment 1 (see section 3.1), participants needed to evaluate laughter samples. To collect these evaluations, as well as informed consent, and demographic data from the participants, a questionnaire was required (see Appendix B).

Laughter samples

The laughter samples used in the listening test were identical to the ones used in experiment 1 (see subsection 3.1.4). To minimise participant burden, the laughter samples were divided into 10 batches, with each participant evaluating a maximum of 28 samples. Out of these 10 batches, 9 batches included 7 samples each, while 1 batch had 5 samples. The batches were generated randomly, ensuring balanced representation of acted and spontaneous samples, as well as an equal distribution of female and male samples. This was achieved by initially dividing the samples into acted female, acted male, spontaneous female, and spontaneous male samples, and then constructing the batches.

Questionnaire

The questionnaire, which can be found in Appendix B, was designed such that each laughter sample was linked to a question in which participants had to rate the authenticity in a semi-forced-decision task, choosing between: "Yes, spontaneous", "No, acted", and "I really don't know".

The motivation for this design, specifically the value of "I really don't know", is reasoned from a format often encountered in these types of evaluations: the 5-point Likert scale (Chyung et al., 2017, p. 1). Firstly, having various functions, laughter is a very frequently occurring paralinguistic event. Therefore, it can be assumed that most people are inadvertently familiar with judging the authenticity of laughter. When participants are familiar with the topic, it is advised to provide an "I don't know" option instead of a midpoint (Chyung et al., 2017, p. 4; Alwin et al., 2018). This reduces the 5-point Likert scale to a 4-point Likert scale with a separate "I don't know" option, which reduces noise in the data.

Furthermore, a 4-point Likert scale provides different levels of intensity in the subjective experience. However, I only care about the subjective rating, regardless of its intensity, and using less anchors yields a higher reliability (Alwin et al., 2018). This reduces the 4-point Likert scale to a 2-point Likert scale, also known as a forced decision task, with a separate, neutral "I don't know" option. This also lowers the cognitive workload on the participants (Chyung et al., 2017), increasing the likelihood of completing the entire evaluation. Since there is also a risk that participants will misuse the "I don't know" option as an easy way out when the cognitive workload becomes too high, I decided to include the word "really", to dissuade people from misusing it.

To test for intra-participant reliability, each sample was evaluated twice by each participant resulting in batch sizes of 14 and 10 samples respectively. The presentation order of the samples was randomised as well.

The questionnaire was provided trough means of a digital survey. On the first page (see figure B.1), participants were informed of the purpose of this research and the task at hand, and were asked to provide their gender, age-range, and the country they spent most of their life in (unreported). This data was collected to determine whether there is a difference in the judgement of authenticity of laughter across different genders, age-ranges, and cultures. They were then informed of how the data would be stored, and informed about their rights and the means to exercise them. Lastly, they were informed that continuing with the test would be interpreted as providing informed consent for their data to be used.

If the participants continued to the second page (see figure B.2), they were firstly reminded of the research objective. Secondly, they were asked to only participate if they did not have any hearing impairments, and were asked to take the survey in a semi-controlled environment, namely: a quiet place and preferably with headphones. Lastly, the setup of the evaluations (see figure B.3) was explained to them. Participants could listen to each sample as many times as they wanted, but were instructed not to go back to a question once answered.

After completing all the evaluations, participants would move on to the last page (see figure B.4), where they were thanked for their effort and were shown their overall accuracy. Furthermore, they were provided with the contact details and their subject ID, which they needed to exercise their rights. Participants were given a full week to fill out the survey. After that the survey was closed.

To ensure reliable results, the data was checked for intra-participant reliability and entries from inconsistent participants were removed. To compute this

value, one needs to know what the odds are of picking the same option twice by chance. For consistency it does not matter which answer option is picked, so under the assumption that the "I really don't know" option is only used to reduce noise, the chance level is 50%. The chance level was then computed using a binomial distribution with a significance value of p=0.05, meaning that 1 out of 20 participants who pick the same option twice, likely did so by chance.

To account for task difficulty, laughter samples with an overall consistency rating below 50% were extracted. These laughter samples were deliberately not included in the determination of the intra-participant reliability. Entries from unreliable participants were removed from the data.

To make the results even more reliable, the entries from consistent participants were checked for inter-participant reliability. After the removal of erroneous entries and entries from inconsistent participants, the data was incomplete and samples were evaluated by different amounts of coders. The most suitable method to compute the inter-participant reliability with this type of data is Krippendorff's alpha. However, according to Zhao et al. (2022) it is a worse predictor than percent agreement when concerning evaluations based on subjective experience. Therefore, percent agreement was used with a standard threshold of 75% agreement.

3.3.5 Participants

Besides the materials, the listening test also required participants. To get accurate results from the listening test, an adequate sample size, sampling procedure, and inclusion criteria are required.

Sample size

For an indication about the sample size I looked to the total number of evaluations used by the guiding papers. This number was determined by multiplying the number of participants, the number of evaluations per participant per session, and the number of sessions. Additionally, It was checked whether participants were incentivised to participate using funding. Table 3.3 below provides a comprehensive overview of the evaluation setup per guiding paper.

Author	Participants	Evaluations	Sessions	Total eval. p.p.	Incentive provided
Bryant and Aktipis (2014)	63	36	1	36	Y
Lavan et al. (2016)	19	72	1	72	Y
Luong and Yamagishi (2021b)	16	32	8	256	Unknown

Table 3.3: Evaluation setup used by the authors of the guiding papers

As can be seen in table 3.3, Bryant and Aktipis (2014) had the largest number of participants, with the smallest number of total evaluations per participant. These participants were incentivised using funding. Since no funding was provided for this thesis, participants could not be compensated for participating. To ensure that participants completed evaluations nonetheless, I aimed for a low workload, with less total evaluations per participant than Bryant and Aktipis (2014).

With 68 samples needing to be evaluated twice for intra-participant reliability, by 10 participants for inter-participant reliability, I arrived at with 1360 evaluations total. To reliably determine whether the samples were affected by the synthesis procedure of LaughNet, the synthetic laughter samples needed to be evaluated by the same people, making this a within-participants design. Consequently, each participant will be asked to perform the test two times, making the final total of evaluations per participant: 2720. To arrive at a lower total number of evaluations per participant than Bryant and Aktipis (2014), I should aim for at least 100 participants. This puts the total number of evaluations per participant at 28. This level of participants was achieved, with a total of 104 responses.

Sampling procedure

To recruit at least 100 participants for the listening tests on a voluntary basis I used convenience sampling. This means that I approached them through different media, such as face-to-face conversation, a WhatsApp message, and LinkedIn. In the short conversation or message I explained the task and the expected duration and asked people to participate in it. Additionally, I asked people to share the message in their personal network so I could reach many people from all around the world in a short amount of time. This would increase the generalisability of the results.

Inclusion criteria

Any participant without hearing impairments that could legally give consent and did so was accepted for participation in this research. The participants were informed about the purposes of this research and about their rights prior to participating and prior to providing consent. Their data was anonymised and they could choose to withdraw from the research at any given moment without providing any reason. No extrinsic motivation was provided to the participants.

4

Results

In this case study, I researched whether the naturalness of synthetic laughter generated by the SOTA laughter synthesis model called LaughNet, could be improved by using spontaneous laughter instead of acted laughter to fine-tune the model. This was researched using three experiments: firstly, it was investigated which acoustic features were different between the acted and spontaneous laughter data used in this thesis (4.1). Secondly, it was examined which of these acoustic features were affected by the synthesis procedure of LaughNet (4.2). Thirdly, it was explored whether the acoustic differences could be noticed by human listeners in both human laughter and synthetic laughter (4.3).

The acted and spontaneous laughter data were separated by a classification boundary with a 90% accuracy on the training data and an 88% accuracy on the testing data. Out of the 7 extracted acoustic features, only duration was not included in any of the 3 factors. The other 6 acoustic features accounted for 77.7% of the variance in the data. Out of these distinctive acoustic features, only the intensity and the percentage unvoiced segments would theoretically be affected by the synthesis procedure of LaughNet. This could not be confirmed practically however, most likely due to a lack of processing power.

4.1 Experiment 1

In this experiment a SVM classifier was trained to compare the acoustic features from the acted and spontaneous laughter data, which were extracted and cleaned through preprocessing. The SVM was trained on the acoustic features from the starting data distribution (4.1.1). The results from the classifier with the optimal hyperparameter settings (4.1.2), are presented in the evaluation (4.1.3). For the Python implementation of the classifier, see https://github.com/5weggeman/laughter_authenticity_classifier/ (last accessed: Jun. 15, 2023).

4.1.1 Data distribution

Table 4.1 below provides the starting distribution of the laughter samples used in this experiment, post removal of outliers and incomplete data. This reveals that there is a slight imbalance between acted and spontaneous laughter, as well as a minimal imbalance between female and male laughter.

Gender	Acted	Spontaneous	Total
F	17	16	**33**
M	17	14	**31**
Total	**34**	**30**	**64**

Table 4.1: Starting data distribution across gender and authenticity

To ensure that the slight imbalance between acted and spontaneous samples did not induce a bias towards acted data in the classifier, the distributions of the train and test splits were retroactively checked for balancing. Table 4.2 below shows that the training split, comprising 75% of the data, contained both imbalances. Table 4.3 below shows that the testing split, comprising 25% of the data, was perfectly balanced. These results will be compared to those from the evaluation (4.1.3) in chapter 5.

Gender	Acted	Spontaneous	Total
F	13	12	**25**
M	13	10	**23**
Total	**26**	**22**	**48**

Table 4.2: Training data distribution across gender and authenticity

Gender	Acted	Spontaneous	Total
F	4	4	**8**
M	4	4	**8**
Total	**8**	**8**	**16**

Table 4.3: Testing data distribution across gender and authenticity

4.1.2 Hyperparameter settings

According to grid search, the optimal hyperparameter setting for the SVM was a linear kernel with a regularisation strength of 100.

4.1.3 Evaluation

The classifier performance was firstly evaluated using confusion matrices, to evaluate the accuracy and the error balance. Secondly, histograms of projections were created for detailed information regarding the data distributions with respect to the classification boundary. Lastly, factor analysis was performed to find the largest contributing factors to the authenticity of laughter.

Confusion matrices

Table 4.4 below provides the confusion matrix of the training data. It shows that the training data classification had an accuracy of 90% and that the misclassified 10% of consisted of 2 false positives (Type I error) and 3 false negatives (Type II error). With no preference for either type of error, this is well balanced.

		Predicted class		
		Acted	**Spontaneous**	
Actual class	**Acted**	23	3	Sensitivity: 0.88
	Spontaneous	2	20	Specificity: 0.91
		Precision: 0.92	Negative Predictive Value: 0.87	Accuracy: 0.90

Table 4.4: Confusion matrix training data classification

Table 4.5 below provides the confusion matrix of the testing data. It shows that the testing data classification had an accuracy of 88% and that the misclassified 12% consisted of 1 false positive (Type I error) and 1 false negative (Type II error). This closely resembles the training data classification.

		Predicted class		
		Acted	**Spontaneous**	
Actual class	**Acted**	7	1	Sensitivity: 0.88
	Spontaneous	1	7	Specificity: 0.88
		Precision: 0.88	Negative Predictive Value: 0.88	Accuracy: 0.88

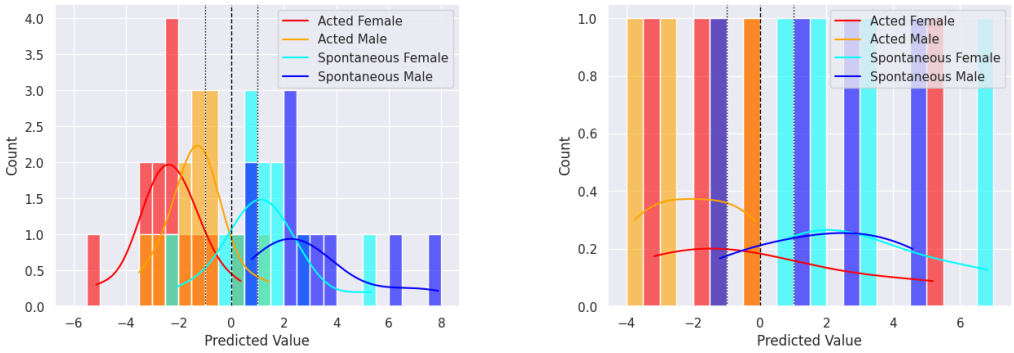
Table 4.5: Confusion matrix testing data classification

Histograms of projections

As a visual indication of the classification accuracy and as control of the confusion matrices (see tables 4.4 and 4.5), the predicted value of each sample was plotted in a histogram of projections relative to the decision boundary. The histograms of projections of the training data have been depicted in figure 4.1a below and those of the testing data have been depicted in figure 4.1b below.

The histograms of projections depict the predicted values (x-axis) assigned to each laughter sample by the classifier. These predicted values indicate the distance and position of said sample, relative to the decision boundary. The decision boundary is located at 0, indicated by the black dotted line. To either side of the decision boundary are the confidence intervals of 1, indicating the cleanness of fit. The confidence intervals are located at -1 and 1, indicated by the grey dotted lines. Grouped per 0.5, the laughter samples with a similar value are stacked on top of each other, indicated by the count (y-axis).

The first histogram (4.1a/4.1b) depicts all data samples from the respective data split, separated per authenticity and gender. Additionally, each of the depicted distributions has a density plot to ease comparison.

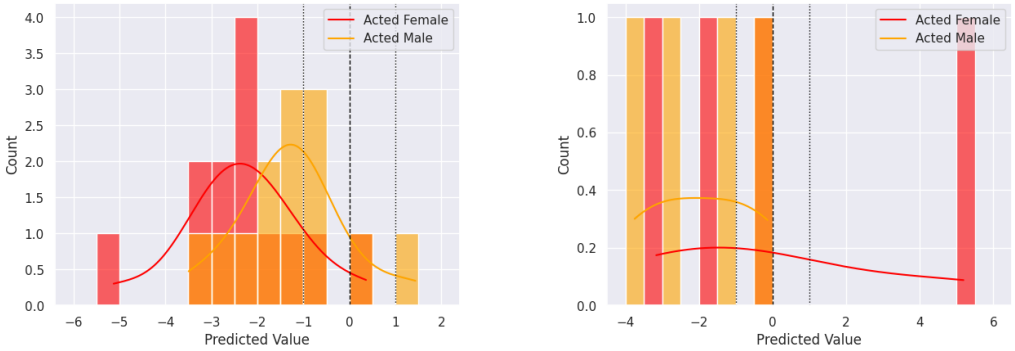


(a) Train split

(b) Test split

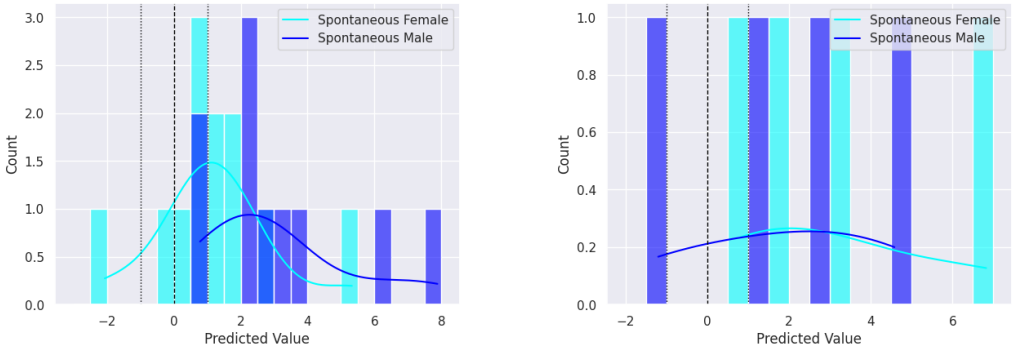
Figure 4.1: Histogram of projections with density plots per split

Due to the large amount of information in this single plot however, the readability has been compromised. For clarity, separate histograms have been created for acted data samples (4.2a/4.2b) and spontaneous data samples (4.3a/4.3b).



(a) Train split (b) Test split

Figure 4.2: Histogram of projections of acted data with density plots per split

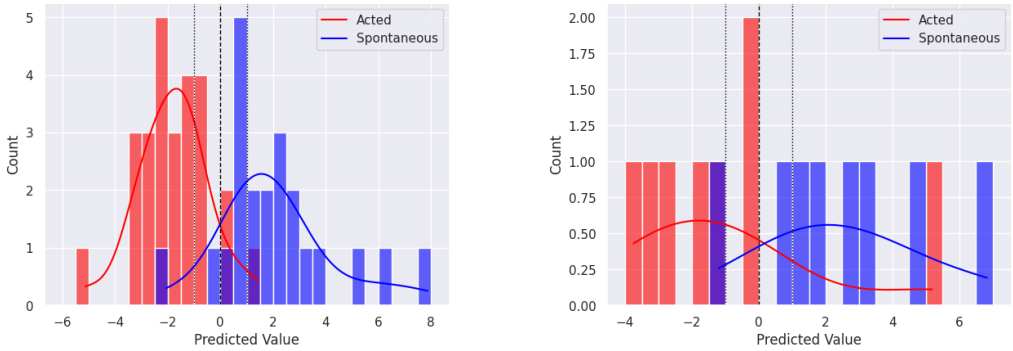


(a) Train split (b) Test split

Figure 4.3: Histogram of projections of spontaneous data with density plots per split

From the coloured bars to the right of the decision boundary in plot 4.2a, it becomes clear that 1 acted female and 2 acted male laughs have been misclassified in the training data. From the coloured bars to the left of the decision boundary in plot 4.3a, it becomes clear that from the spontaneous laughter, 2 female laughs have been misclassified in the training data. From the coloured bar to the right of the decision boundary in plot 4.2b, it becomes clear that 1 acted female laugh has been misclassified in the testing data. From the coloured bar to the left of the decision boundary in plot 4.3b, it becomes clear that 1 acted male file has been misclassified in the testing data.

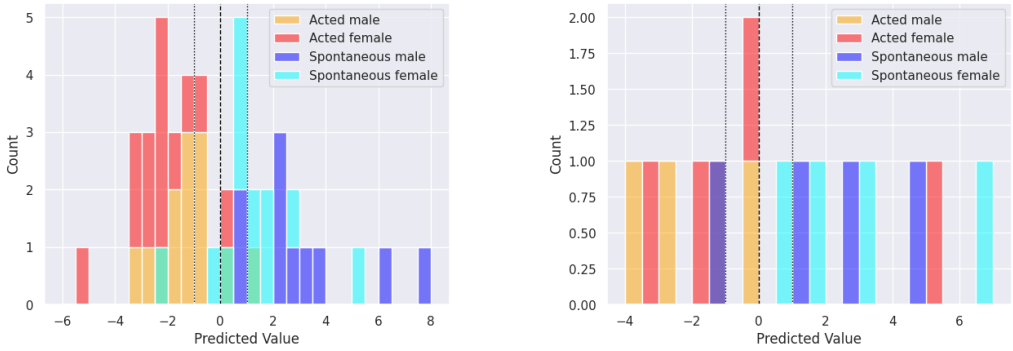
The manner in which the data has been represented in the previous three histograms – that is: separated per authenticity and gender, presented in parallel – provides little insight into the distribution per authenticity. Therefore, two more histograms were created in which the data was stacked per authenticity. The first one (4.4a/4.4b), meant to determine the authenticity distribution, had no gender separation and included density plots. The second one (4.5a/4.5b), meant to compare the contributions of each gender to the count per authenticity, had gender separation and no density plots.



(a) Train split (b) Test split

Figure 4.4: Histogram of projections stacked per authenticity with density plots and without gender separation per split

From the shapes of the stacked bars and the bell curves in plot 4.4a, it becomes clear that the acted and spontaneous laughter follow roughly the same distribution. The distribution of the spontaneous laughter is slightly more elongated however. From the places of the coloured bars and the bell curves in plot 4.4b, it becomes clear that the acted and spontaneous laughter follow roughly the same distribution as well.



(a) Train split (b) Test split

Figure 4.5: Histogram of projections stacked per authenticity without density plots per split

Lastly, from the gradually shifting colour distribution from left to right in plot 4.5a and from the bell shape curves in plot 4.1a, it becomes clear that the decision boundary is slightly favourable to acted female laughter and spontaneous male laughter. These two laughter types lie further from the decision boundaries compared to their respective counterparts. From the gradually shifting colour distribution from left to right in plot 4.5b and from the bell shape curves in plot 4.1b, it becomes clear that this is the other way around for the testing data, contrasting with the results from the training data.

Factor analysis

Bartlett's sphericity test validated the use of factor analysis with a statistically significant p-value ($p<0.05$) of 1.29e-55. The number of relevant factors corresponds to the number of Eigenvalues of the data with a value greater than 1. The scree plot in figure 4.6 below indicates that there are 3 relevant factors.

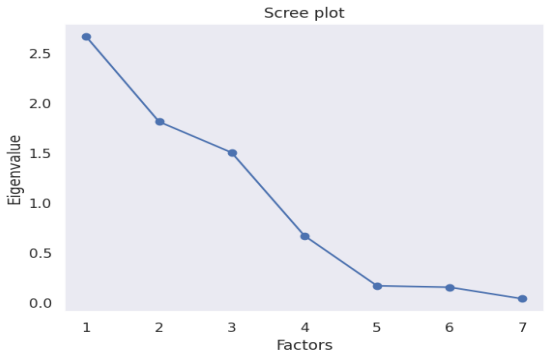


Figure 4.6: Scree plot

Table 4.6 below provides an oversight of the factor loadings. Since factor loadings indicate correlation between the acoustic features, large factor loadings have been highlighted. F0 mean, F0 maximum, and F0 variability have large loadings on factor 1, percentage unvoiced segments and intensity have large loadings on factor 2, and F0 minimum has a large loading on factor 3. The interpretation of these factor loadings will be discussed in chapter 5.

Acoustic features	Factor 1	Factor 2	Factor 3
Duration	0.274062	0.020453	-0.413713
Percentage unvoiced segments	-0.152140	**0.946545**	0.106704
F0 mean	**0.833731**	-0.101971	0.290323
F0 minimum	0.197921	0.052949	**0.986788**
F0 maximum	**0.982576**	-0.151460	-0.098001
F0 variability	**0.813476**	0.152264	-0.224312
Intensity	-0.074778	**-0.843017**	0.073254

Table 4.6: Factor loadings after varimax rotation

Table 4.7 below contains the variance in the data explained by each of these factors has been summarised in table below. Together, these 3 factors account for 77.7% of the variance in the data, almost half of which is explained solely by factor 1.

	Factor 1	Factor 2	Factor 3
SumSquare Loadings (Variance)	2.465326	1.666370	1.305868
Proportional Variance	0.352189	0.238053	0.186553
Cumulative Variance	0.352189	0.590242	0.776795

Table 4.7: Variance in the data accounted for by factors

These factors were then used to analyse the acoustic features of the misclassified laughter (see table 4.4 and plots 4.2a & 4.3a for the training data, and table 4.5 and plots 4.2b & 4.3b for the testing data). The acoustic features of the misclassified training samples have been plotted against the acted and spontaneous class means of the training samples in plots A.1a–A.1e, and those of the misclassified testing samples in plots A.2a–A.2b. Important to note here is that the values and their respective differences across acoustic features, cannot be accurately interpreted without the results from the factor analysis: a minor change in value in a relevant acoustic feature yields much more meaning than a major change in value in a less relevant acoustic feature.

Table 4.8 below provides the positions of the acoustic features of the misclassified laughs relative to the class-means of the training data, which act as an indication of the classification boundary. The misclassified acoustic features have been highlighted and counted per laughter sample and per feature. This shows that F0 variability is one of the most important acoustic features, since it has the highest count of all features and because it is the only misclassified feature in acted male laugh 2. It also shows that the misclassified male laughs are almost entirely accounted for by acoustic features from factor 1, with the percentage unvoiced segments of acted male laugh 1 being the only misclassified feature in a different factor for misclassified male laughs.

			Factor 1			**Factor 2**		**Factor 3**	
	Misclassified laugh	**Figure**	**F0 mean**	**F0 max.**	**F0 var.**	**% u. s.**	**Intensity**	**F0 min.**	**Total**
Train	Acted male laugh 1	A.1a	A	**S**	A	**S**	A	A	**2**
	Acted male laugh 2	A.1b	A	A	**S**	A	A	A	**1**
	Acted female laugh	A.1c	**S**	A	**S**	**S**	**S**	**S**	**5**
	Spontaneous female laugh 1	A.1d	S	**A**	S	S	S	**A**	**2**
	Spontaneous female laugh 2	A.1e	**A**	**A**	**A**	**A**	**A**	**A**	**6**
Test	Acted female laugh	A.2a	**S**	A	**S**	A	**S**	**S**	**4**
	Spontaneous male laugh	A.2b	**A**	S	**A**	S	S	S	**2**
		Total	**4**	**3**	**5**	**3**	**3**	**4**	

A = Acted, S = Spontaneous

Table 4.8: Class means-based misclassified acoustic features of misclassified laughs

Through these evaluation measures I have created a complete oversight of the most relevant factors for the authenticity of laughter, as well as the relative contributions of individual acoustic features. Furthermore, the results show how well the SVM classifier was capable of capturing the laughter authenticity on the basis of the acoustic features, and which features were the most relevant for this distinction. Since most of the results have been captured by at least two different measures, the conclusions should be scientifically sound.

4.2 Experiment 2

To examine how the relevant acoustic features from the factor analysis (4.1.3) were affected by the synthesis procedure of LaughNet, a theoretical analysis was performed using the temporal acoustic framework from Rosen (1992), and a practical analysis was performed using LaughNet[1] and the classifier from experiment 1 (see https://github.com/5weggeman/laughter_authenticity_classifier/blob/main/classifier.py (last accessed: Jun. 15, 2023).).

4.2.1 Theoretical analysis

In the synthesis process of LaughNet, a waveform silhouette from a source laughter sample is used as a mould and the periodicity and fine-structure are filled up by the generator, which is pre-trained using VCTK and fine-tuned using target samples from the laughter data(Luong & Yamagishi, 2021b, pp. 2–3). Table 4.9 below provides the extracted acoustic features and the temporal features from Rosen (1992) in which they are represented. This shows that only the percentage unvoiced segments and the intensity are captured in the waveform silhouette. In theory, these acoustic features should thus be preserved during the synthesis of laughter using Laughnet, whilst the others need to be regenerated.

Acoustic feature	**Temporal features**
Percentage unvoiced segments	Periodicity, fine-structure & **envelope**[2]
F0 mean	Periodicity
F0 minimum	Periodicity
F0 maximum	Periodicity
F0 variability	Periodicity
Intensity	**Envelope**

Table 4.9: Acoustic feature representation in temporal features; cf. Rosen (1992)

[1]For the Python implementation used in this section see https://github.com/5weggeman/hifi-gan-laughnet (last accessed: Mar. 7, 2023)

[2]Ordered from strongest to weakest representation (see table 1 Rosen, 1992, p. 76).

4.2.2 Practical analysis

Figure 4.7 below provides the general loss total. The rising trend in the data indicates that the model started overfitting to the data after 10k steps.

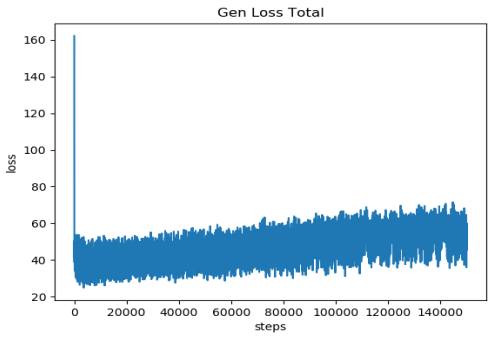


Figure 4.7: General Loss Total

Figures 4.8 and 4.9 below provide the Mel-spectrogram error and the validation Mel-spectrogram error. The rapidly decreasing error in the beginning, followed by the slowly declining decrease, and the stabilisation at roughly 10k steps, indicate that no abnormal learning behaviour occurred during the training and validation stages.

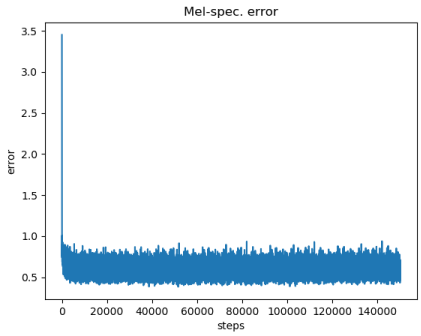


Figure 4.8: Mel-Spectrogram Error

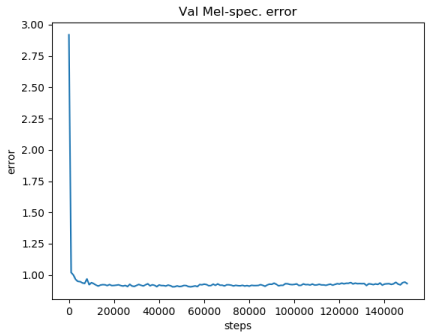


Figure 4.9: Validation Mel-Spectrogram Error

In practice, it could not be checked which acoustic features were preserved during the synthesis of laughter using LaughNet, since the output was not recognisable as laughter. This was likely a consequence of the overfitting. With no synthetic laughter, the practical analysis could not be performed.

4.3 Experiment 3

To check whether the acoustic differences between acted and spontaneous laughter could be detected by human listeners in both human and synthetic laughter, a listening test was performed. With no synthetic laughter however (see subsection 4.2.2), the listening test could only be performed with the human laughter. Although the participants would have been the same, their reliability, accuracy, and biases have only been reported for the human laughter evaluations.

4.3.1 Participants

The listening test was performed by 104 participants, 15 of which requested their data to be removed, resulting in 89 participants total. Some of these entries contained errors: some files were either rated less than two times or more than two times by the same participant. After removing the erroneous data from these entries, I also removed incomplete entries with less than 10 evaluations: 5 samples, evaluated twice. This left me with 83 entries total. Of these 83 participants, 43 indicated to identify as female, 36 as male, and 4 as other. The majority of the participants came from the age range between 18 and 30. For the specifics see table 4.10 below.

Age range	Female	Male	Other	Total
18-30	34	26	4	**64**
31-40	3	8	0	**11**
41-50	3	1	0	**4**
51-60	2	1	0	**3**
60+	1	0	0	**1**
Total	**43**	**36**	**4**	**83**

Table 4.10: Listening test participants by gender and age range

4.3.2 Reliability

Taking task difficulty into account, the intraparticipant reliability was computed: the number of statistically significant reliable participants was 48. Due to this reduction, 8 files were left with one or none reliable evaluations. Accordingly, these files were dropped, resulting in 60 laughter samples total, out of which 13 acted female, 16 acted male, 14 spontaneous female, and 17 spontaneous male laughter samples. Additionally, the number of reliable participants dropped to 47, out of which 25 female and 22 male.

The inter-participant reliability of these participants ranged from 37.1% to 61.5% agreement across batches, with an average of 44.6%, compared to the required 75% agreement.

4.3.3 Accuracy

The accuracy with which reliable participants evaluated the laughter samples have been tested for accuracy using two-sided 1-sample t-tests. Due to the double testing of each sample for intra-participant reliability, the probability of randomly guessing the correct answer on both trials was 25%. Table 4.11 below provides the accuracy per laughter type and gender, as well as their significance levels. The probability threshold was set at 0.05. Since each p-value in table 4.11 below is less than 0.05, all accuracies are statistically significant. This shows that it is unlikely that the reliable participants achieved these accuracies through random guessing.

		Gender	Accuracy	p-value
Laughter type	Acted	Female	46.4%	0.0188
		Male	43.2%	0.025
	Spontaneous	Female	59.9%	0.000
		Male	51.3%	0.000

Table 4.11: Reliable participant accuracy per laughter type and gender

4.3.4 Bias checking

Table 4.12 below provides the percentages of times participants picked each answer option. This shows that the "I really don't know" option was only constitutes 2.4% of all the answers given, post cleaning. This is a sharp contrast with the fact that 14.9% of all participants have used this option at least once.

Yes, spontaneous	No, acted	I really don't know
53.1%	44.5%	2.4%

Table 4.12: Answer percentages

Table 4.13 below provides the distribution of cases in which participants resorted to this answer option. This shows that this answer option was used quite inconsistently, and mostly when the laughter was actually spontaneous.

	Acted	Spontaneous	Total
Consistent	9.09%	27.27%	**36.4%**
Inconsistent	27.27%	36.36%	**63.6%**
Total	**36.4%**	**63.6%**	**100%**

Table 4.13: Usage "I really don't know" per authenticity

5

Discussion

The results suggest that the findings of Bryant and Aktipis (2014) and Lavan et al. (2016) hold between the acted laughter data used by Luong and Yamagishi (2021b) and the spontaneous laughter data from the MULAI Corpus (Jansen et al., 2018). The key factors in distinguishing acted from spontaneous laughter appear to be control, energy, and gender, with female laughter being easier to classify. Out of these factors, only the energy should remain unaffected by the synthesis procedure of LaughNet. This could not be confirmed however, since I was unable to get LaughNet to produce output that was recognisable as laughter due to overfitting. This was likely caused by either a shortage of human laughter data, or by mistakes made during the implementation of LaughNet. Furthermore, the interpretation of the authenticity of isolated laughter appears to be contentious.

5.1 Experiment 1

In section 2.4, I hypothesised that the acoustic features of isolated acted and spontaneous laughter would be significantly different, based on the findings from Bryant and Aktipis (2014) and Lavan et al. (2016). To test that I trained a SVM on the statistically significant acoustic features used by Lavan et al. (2016). This classifier demonstrated the separability of the acted and spontaneous laughter with a training accuracy of 90% and a testing accuracy of 88% (see tables 4.4 & 4.5), despite a slight imbalance in the data between acted and spontaneous laughter samples. Since these two accuracies are very close and the type I and type II errors are balanced for both the training and testing data, the model did not overfit. This is also reflected by the fact that all the sensitivity, specificity, precision, and negative predictive values have similar values. Accordingly, my hypothesis is confirmed.

Despite the fact that only statistically significant distinctive features were used (see table 3.2), factor analysis showed that the duration feature turned out to be irrelevant for distinguishing acted from spontaneous laughter (see table 4.6). The remaining six acoustic features, spread across three different factors, explain 77.7% of the variance in the data (see table 4.7), indicating that the authenticity of laughter can be described fairly accurately using just these three factors. The interpretation of these factors will be discussed below.

Starting with the factor that is easiest to explain: the third factor only consisted of the F0 minimum and explained 18.7% of the variance in the data. In 4.8 it can be seen that this feature is generally classified correctly for misclassified male laughter, but not for misclassified female laughter. This indicates that there is a difference in the authenticity interpretation of the other acoustic features between female and male laughter, which is corroborated by shifted distributions of female and male laughter in the histograms of projections in Appendix A. This is especially visible in plots 4.1a, 4.2a, and 4.3a. Accordingly, this factor likely describes gender.

The second factor consisted of the percentage unvoiced segments and the intensity and explained 23.8% of the variance in the data. In 4.8 it can be seen that, similarly to the F0 minimum, the intensity of the misclassified files is generally classified correctly for male laughter, but not for female laughter, with the exception of spontaneous female laugh 1. Additionally, it can be seen that both acoustic features in factor 2 are only misclassified 3 times, making this the least misclassified factor. Both features in this factor are related to the energy

contained in the laughter, with voiced laughter containing more energy than unvoiced laughter. However, in plots A.1a-A.1e and A.2a-A.2b it can be seen that for percentage unvoiced segments, the class mean of spontaneous laughter has a higher value than the class mean of acted laughter. Since spontaneous laughter is perceived as more natural, and natural is more positive, this contrasts with the finding from Bachorowski and Owren (1995) that voiced laughter is perceived more positively.

The first factor consisted of the F0 mean, the F0 maximum, and the F0 variability and explained 35.2% of the variance in the data. In 4.8 it can be seen that for acted male laugh 2, only the F0 variability has been misclassified. The fact that that feature also happens to be the most misclassified feature shows the importance of this feature. Additionally, anywhere this feature is misclassified, the F0 mean is also misclassified, with the exception of acted male laugh 2. Since this factor contains the three remaining F0 statistics, this factor describes how much control is exerted over the laughter.

5.2 Experiment 2

In section 2.4, I hypothesised on the basis of Lavan et al. (2016) that the lower-level acoustic features from the distinctive acoustic features would get lost during the synthesis process of LaughNet. Using the three temporal features from Rosen (1992) as level indicators: the fine-structure is the lowest level, the periodicity the middle level, and the envelope the highest level. As hypothesised, the features from levels below the envelope are not passed on from the source laugh to the synthetic laughter. Instead, they are created from scratch by the generator. This implies that there is only little control over the authenticity of the synthetic laughter when using LaughNet, even if the fine-tuning laughter samples are selected very carefully. Consequently, there is no guarantee that a waveform silhouette from an acted source laugh also results in an acted synthetic laugh. Instead, there is a larger probability of getting synthetic laughter samples with parameter distributions not unlike those of the misclassified files. Since this likely yields more misclassifications, it is more difficult to determine the authenticity of synthetic laughter. This hypothesis could not be tested however, due to the fact that the output of my implementation of LaughNet was not recognisable as laughter. This was likely caused by a lack of sufficient human laughter data, or by mistakes made during the implementation of LaughNet.

5.3 Experiment 3

In section 2.4, I hypothesised on the basis of Lavan et al. (2016) that people would be able to accurately determine the authenticity of isolated voiced laughter. Next to the evaluations, I asked participants to provide their gender, age-range, and the country they spent most of their life in (see subsection 3.3.4 and figure B.1). This data was collected to explore any differences in the judgement of authenticity of laughter across different genders, age-ranges, and cultures. During the writing of this thesis however, Bryant and Bainbridge (2022) published a paper showing that the detection of laughter authenticity is equal across cultures. Since this rendered the privacy sensitive country data obsolete, I removed it.

To ensure reliable results, both intra-participant reliability and inter-participant reliability were computed, in that order. Out of 83 participants, 35 participants were removed for providing inconsistent evaluations on the intra-participant reliability retest samples. The remaining, reliable participants were indeed able to determine the authenticity of isolated voiced laughter with a significance level well above chance (see table 4.11), confirming the first part of the hypothesis. However, according to the inter-participant reliability, none of the batches reached statistical significance. This means that the perceived authenticity of isolated laughter is a point of contention. A possible explanation for this is that more information is needed to disambiguate the authenticity, such as context.

The second part of the hypothesis could again not be tested, due to the fact that the output of my implementation of LaughNet did not sound like laughter.

Two important things to note about the classification accuracies from the reliable participants in experiment 3 (see table 4.11) are that the accuracies for female laughter are both higher than those for male laughter, and that the accuracies for spontaneous laughter are both higher than those for acted laughter. This might explain an unknown phenomenon, but could also be caused by bias.

After removing the unreliable participants the only imbalance in the data was between female and male laughter samples, of which there were 27 and 33 respectively (see section 4.3.2). The difference of 8.6% in accuracy between spontaneous female and male laughter samples is too large to be explained by this slight imbalance. Therefore, it is apparently easier to determine the authenticity of isolated female laughter than of isolated male laughter.

The difference in accuracy between acted and spontaneous laughter however, is similar to the difference in how often each answer option was picked. The

answer percentages also show a bias towards "Yes, spontaneous" (see table 4.12). These two things combined indicate that the participants were likely biased.

Given that 14.9% of the participants used the "I really don't know" option, which make up 2.4% of the answers, indicates that there is no underlying pattern of misuse (see subsection 3.3.4). From table 4.13 however, it can be determined that participants used the option almost twice as often when the actual authenticity of the laughter sample was spontaneous. A logical explanation for this from an evolutionary perspective might be a sense of caution: in social situations false hope is generally more dangerous than being mistrusting.

5.4 Implications

The results from experiment 1 corroborate the findings from Bryant and Aktipis (2014) and Lavan et al. (2016), that the acoustic features from acted and spontaneous laughter are significantly different. This implies that the difference should be taken into account when researching or working with laughter.

The results from experiment 2 suggest that the acoustic features relevant for determining the authenticity of laughter, are mostly made up by the LaughNet model, with the only exceptions being the intensity and a part of the percentage unvoiced segments. This implies that, regardless of the results from experiment 3, the authenticity of the laughter data used to fine-tune LaughNet, does not significantly increase the capability of LaughNet to synthesise natural laughter.

The results from experiment 3 indicate that determining the authenticity of isolated laughter is an ambiguous task, which people perform well above chance level, but do not generally agree upon. This implies that more information is needed to disambiguate the task. Since I worked with isolated laughter, this additional information comes in the form of context.

Despite not being able to evaluate the ability of people to determine the authenticity of synthetic laughter in experiment 3, the findings of experiment 2 suggest that the authenticity of the synthetic laughter would mostly be decided arbitrarily by the generator in the model. Combined with the results from experiment 3, that the task of determining the authenticity of isolated laughter is already ambiguous, implies that people would likely not perform better in determining the authenticity of synthetic laughter.

5.5 Limitations

During the literature review (2.3) I encountered two papers that did not have the word 'synthesis' in their titles, but that did have a very similar word, namely: 'generation' (Mansouri & Lachiri, 2021) and 'processing' (Urbain et al., 2014). For scientific completeness the literature review search should have been performed again with those similar terms. This was not done due to time limitations.

Another limitation of this research is that the participants were likely biased, as was discovered in section 5.3. A possible source for this bias is the way in which the questions were framed. To improve the setup in the future, questions should be asked in a more neutral manner.

5.6 Future research

One of the first possible direction of future research is the delineation of the aspects of naturalness (see section 1.2 and subsection 2.3.4). Because of the additive nature of all these aspects, each individual aspect has to be researched in relation to speech synthesis. Furthermore, based on the result from experiment 3 (see section 5.3), laughter synthesis should be researched in the context of speech. Lastly, more research should be done to advanced source isolation techniques, such that laughter recordings can be made in noisy, social environments.

6

Conclusion

In this master thesis I have contributed to increasing the naturalness of synthetic laughter, by showing the relevance of using spontaneous laughter data instead of acted laughter data. I have demonstrated this by showing that an objective SVM is capable of accurately classifying human laughter authenticity based on its acoustic features, which likely extends to synthetic laughter.

Furthermore, the results suggest that LaughNet provides little control over the final authenticity of the synthetic laughter, making it suboptimal for synthesising natural laughter. This is due to the fact that the data format of the waveform silhouette only captures higher-level acoustic features, which only partially explain authenticity. Consequently, using authentic laughter data in the fine-tuning process imposes only little added naturalness. The more important lower-level acoustic features have to be generated by the generator, which depends purely on the training and fine-tuning data. There is, however, still a general paucity of authentic laughter data.

In the broader context, this means that certain limitations of lab-recorded data compared to real-world data, can be mitigated through careful selection of a generative model, data format, and training and fine-tuning data.

Lastly, I have found that the perceived authenticity of isolated laughter appears to be a point of contention. This suggests that contextual information is needed to further disambiguate the determination of laughter authenticity.

References

Alwin, D. F., Baumgartner, E. M., & Beattie, B. A. (2018). Number of response categories and reliability in attitude measurement. *Journal of Survey Statistics and Methodology*, *6*(2), 212–239.

Arimoto, Y., Kawatsu, H., Ohno, S., & Iida, H. (2012). Naturalistic emotional speech collection paradigm with online game and its psychological and acoustical assessment. *Acoustical science and technology*, *33*(6), 359–369.

Ataollahi, F., & Suarez, M. T. (2019). Laughter classification using 3d convolutional neural networks. *Proceedings of the 2019 3rd International Conference on Advances in Artificial Intelligence*, 47–51.

Bachorowski, J.-A., & Owren, M. J. (1995). "Not all laughs are alike: Voiced but not unvoiced laughter readily elicits positive affect in listeners". *Speech Transm. Lab. Q. Prog. Status Rep.*, *36*(2-3), 119–156.

Baird, A., Parada-Cabaleiro, E., Hantke, S., Burkhardt, F., Cummins, N., & Schuller, B. (2018). The Perception and Analysis of the Likeability and Human Likeness of Synthesized Speech. *Proc. Interspeech 2018*, 2863–2867. https://doi.org/10.21437/Interspeech.2018-1093

Bertelsen, P., Høgh-Olesen, H., & Tønnesvang, J. (2009). *Human characteristics: Evolutionary perspectives on human mind and kind*. Cambridge Scholars Publishing.

Birkholz, P. (2013). Modeling consonant-vowel coarticulation for articulatory speech synthesis. *PloS one*, *8*(4), e60603.

Boersma, P. (2011). Praat: Doing phonetics by computer (version 5.2. 19). Retrieved September 2021, from http://www.praat.org/

Bollepalli, B., Urbain, J., Raitio, T., Gustafson, J., & Cakmak, H. (2014). A comparative evaluation of vocoding techniques for hmm-based laughter synthesis. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 255–259. https://doi.org/10.1109/ICASSP.2014.6853597

Bryant, G. A., & Aktipis, C. A. (2014). The animal nature of spontaneous human laughter. *Evolution and Human Behavior*, *35*(4), 327–335.

Bryant, G. A., & Bainbridge, C. M. (2022). Laughter and culture. *Philosophical Transactions of the Royal Society B*, *377*(1863), 20210179.

Burleigh, T. J., Schoenherr, J. R., & Lacroix, G. L. (2013). Does the uncanny valley exist? an empirical test of the relationship between eeriness and the human likeness of digitally created faces. *Computers in human behavior*, *29*(3), 759–771.

Campbell, N. (2006). Conversational speech synthesis and the need for some laughter. *IEEE Transactions on Audio, Speech and Language Processing*, *14*, 1171–1178. https://doi.org/10.1109/TASL.2006.876131

Campbell, N. (2007a). Evaluation of speech synthesis. In *Evaluation of text and speech systems* (pp. 29–64). Springer.

Campbell, N. (2007b). Whom we laugh with affects how we laugh. *Proceedings of the Interdisciplinary Workshop on the Phonetics of Laughter*, 61–65.

Chyung, S. Y., Roberts, K., Swanson, I., & Hankinson, A. (2017). Evidence-based survey design: The use of a midpoint on the likert scale. *Performance Improvement*, *56*(10), 15–23.

Clore, G. L., & Huntsinger, J. R. (2007). How emotions inform judgment and regulate thought. *Trends in cognitive sciences*, *11*(9), 393–399.

Crystal, D. (1974). Paralanguage. *Current Trends in Linguistics*, *12*.

Dall, R., Yamagishi, J., & King, S. (2014). Rating naturalness in speech synthesis: The effect of style and expectation. *Proceedings of Speech Prosody*.

Eberhard, D. M., Simons, G. F., & Fennig, C. D. (Eds.). (2022). *Ethnologue: Languages of the world*. Retrieved February 6, 2023, from http://www.ethnologue.com/

El Haddad, K., Torre, I., Gilmartin, E., Çakmak, H., Dupont, S., Dutoit, T., & Campbell, N. (2017). Introducing amus: The amused speech database. *International Conference on Statistical Language and Speech Processing*, 229–240.

Farley, S. D., Carson, D., & Hughes, S. M. (2022). Just seconds of laughter reveals relationship status: Laughter with friends sounds more authentic and less vulnerable than laughter with romantic partners. *Journal of Nonverbal Behavior*, 1–28.

Folorunso, C. O., Asaolu, O. S., & Popoola, O. P. (2020). Laughter signature: A novel biometric trait for person identification. *International Journal of Biometrics*, *12*(3), 283–300.

French, N. R., & Steinberg, J. C. (1947). Factors governing the intelligibility of speech sounds. *The journal of the Acoustical society of America*, *19*(1), 90–119.

Hill, D., Manzara, L., & Schock, C. (1995). Real-time articulatory speech-synthesis-by-rules. *Proceedings of AVIOS*, *95*, 11–14.

Jansen, M.-P., Heylen, D., Truong, K. P., Englebienne, G., & Nazareth, D. S. (2018). The mulai corpus: Multimodal recordings of spontaneous laughter in dyadic interaction. *Proceedings of Laughter Workshop*, 58–63.

Juhitha, K., Vekkot, S., Yogesh, M. J., Tripathi, S., & Shashank, R. (2018). Laughter synthesis using mass-spring model and excitation source characteristics. *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 1102–1108. https://doi.org/10.1109/ICACCI.2018.8554573

Kantharaju, R. B., Ringeval, F., & Besacier, L. (2018). Automatic recognition of affective laughter in spontaneous dyadic interactions from audiovisual signals. *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, 220–228.

Klatt, D. H. (1987). Review of text-to-speech conversion for English. *The Journal of the Acoustical Society of America*, *82*(3), 737–793.

Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *the Journal of the Acoustical Society of America*, *87*(2), 820–857.

Kong, J., Kim, J., & Bae, J. (2020). Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems*, *33*, 17022–17033. https://doi.org/10.48550/arXiv.2010.05646

Kort, B., Reilly, R., & Picard, R. (2001). An affective model of interplay between emotions and learning: Reengineering educational pedagogy-building a learning companion. *Proceedings IEEE International Conference on Advanced Learning Technologies*, 43–46. https://doi.org/10.1109/ICALT.2001.943850

Lavan, N., Scott, S. K., & McGettigan, C. (2016). Laugh like you mean it: Authenticity modulates acoustic, physiological and perceptual properties of laughter. *Journal of Nonverbal Behaviour*, *40*, 133–149. https://doi.org/10.1007/s10919-015-0222-8

Lazarus, R. S. (1991). *Emotion and adaptation*. Oxford University Press.

Lizarazu, M. (2017). *Speech-brain synchronization: A possible cause for developmental dyslexia* (Doctoral dissertation).

Luong, H.-T., & Yamagishi, J. (2021a). *Samples for "laughnet: Synthesizing laughter utterances from waveform silhouettes and a single laughter example."* https://nii-yamagishilab.github.io/sample-laughnet-waveform-silhouette/

Luong, H.-T., & Yamagishi, J. (2021b). Laughnet: Synthesizing laughter utterances from waveform silhouettes and a single laughter example. https://arxiv.org/pdf/2110.04946.pdf

Lustgarten, P. C., & Juang, B.-H. (2003). Naturalness in speech communications. *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*.

Mansouri, N., & Lachiri, Z. (2019). *Dnn-based laughter synthesis*.

Mansouri, N., & Lachiri, Z. (2020). *Laughter synthesis: A comparison between variational autoencoder and autoencoder*. https://doi.org/10.1109/ATSIP49331.2020.9231607

Mansouri, N., & Lachiri, Z. (2021). Human laughter generation using hybrid generative models. *KSII Transactions on Internet & Information Systems*, *15*(5). https://doi.org/10.3837/tiis.2021.05.001

Mattingly, I. G. (1974). *Developing models of human speech*.

Mehrabian, A. (1971). *Silent messages*. Wadsworth Publishing Company.

Miller, N. (2013). Measuring up to speech intelligibility. *International Journal of Language & Communication Disorders*, *48*(6), 601–612.

Mitchell, W. J., Szerszen Sr, K. A., Lu, A. S., Schermerhorn, P. W., Scheutz, M., & MacDorman, K. F. (2011). A mismatch in the human realism of face and voice produces an uncanny valley. *i-Perception*, *2*(1), 10–12.

Mori, H., Nagata, T., & Arimoto, Y. (2019). Conversational and social laughter synthesis with wavenet. *Interspeech 2019*, 520–523. https://doi.org/10.21437/Interspeech.2019-2131

Mori, M. (1970). Bukimi no tani [The uncanny valley]. *Energy*, *7*, 33–35.

n.d. (2023a). Human [Retrieved February 28, 20223 from https://www.merriam-webster.com/dictionary/human]. In *Merriam-webster.com dictionary*. Merriam-Webster.

n.d. (2023b). Likeness [Retrieved February 28, 2023, from https://www.merriam-webster.com/dictionary/likeness]. In *Merriam-webster.com dictionary*. Merriam-Webster.

Nesse, R. M. (2020). Tacit creationism in emotions research.

Oh, J., Wang, G., Berger, J., & Chafe, C. (2014). *Affective analysis and synthesis of laughter*. Stanford University.

Oord, A. v. d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., & Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

Plutchik, R. (1991). *The emotions*. University Press of America.

Romportl, J. (2014). Speech synthesis and uncanny valley. *Text, Speech and Dialogue: 17th International Conference, TSD 2014, Brno, Czech Republic, September 8-12, 2014. Proceedings 17*, 595–602.

Rosen, S. (1992). Temporal information in speech: Acoustic, auditory and linguistic aspects. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *336*(1278), 367–373.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*, 1161–1178. https://doi.org/10.1037/h0077714

Sathya, A., Sudheer, K., & Yegnanarayana, B. (2013). Synthesis of laughter by modifying excitation characteristics. *The Journal of the Acoustical Society of America*, *133*, 3072–3082. https://doi.org/10.1121/1.4798664

Scherer, K. (1994). Affect bursts, in emotions: Essays on emotion theory.

Schröder, M. (2001). Emotional speech synthesis: a review. *Proc. 7th European Conference on Speech Communication and Technology (Eurospeech 2001)*, 561–564. https://doi.org/10.21437/Eurospeech.2001-150

Schröder, M. (2003). Experimental study of affect bursts. *Speech communication*, *40*(1-2), 99–116.

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., MüLler, C., & Narayanan, S. (2013). Paralinguistics in speech and language—state-of-the-art and the challenge. *Computer Speech & Language*, *27*(1), 4–39.

Sebe, N., Cohen, I., & Huang, T. S. (2005). Multimodal emotion recognition. In *Handbook of pattern recognition and computer vision* (pp. 387–409). World Scientific.

Steinberg, J. C. (1929). Effects of distortion upon the recognition of speech sounds. *The Journal of the Acoustical Society of America*, *1*(1), 121–137.

Story, B. H. (2019). History of speech synthesis. In *The routledge handbook of phonetics* (pp. 9–33). Routledge.

Sundaram, S., & Narayanan, S. (2007). Automatic acoustic synthesis of human-like laughter. *The Journal of the Acoustical Society of America*, *121*, 527–535. https://doi.org/10.1121/1.2390679

Tachibana, H., Uenoyama, K., & Aihara, S. (2018). Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention. *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 4784–4788.

Tanaka, H., & Campbell, N. (2014). Classification of social laughter in natural conversational speech. *Computer Speech & Language*, *28*(1), 314–325.

Taylor, P. (2009). *Text-to-speech synthesis*. Cambridge university press.

Tinwell, A., & Grimshaw, M. (2009). Bridging the uncanny: An impossible traverse? *Proceedings of the 13th international MindTrek conference: Everyday life in the ubiquitous era*, 66–73.

Tits, N., Haddad, K. E., & Dutoit, T. (2020). *Laughter synthesis: Combining seq2seq modeling with transfer learning*. https://github.com/CSTR-Edinburgh/ophelia

Trouvain, J. (2003). Segmenting phonetic units in laughter. *Proc. 15th International Conference of the Phonetic Sciences, Barcelona, Spain*, 2793–2796.

UNESCO. (2023). *World atlas of languages*. Retrieved February 6, 2023, from https://en.wal.unesco.org/

University of Hawaii at Manoa (Ed.). (2023). *Catalogue of endangered languages*. Retrieved February 6, 2023, from https://www.endangeredlanguages.com/

Urbain, J. (2014). *Acoustic laughter processing*. University of Mons.

Urbain, J., Cakmak, H., Charlier, A., Denti, M., Dutoit, T., & Dupont, S. (2014). Arousal-driven synthesis of laughter. *IEEE Journal of Selected Topics in Signal Processing*, *8*, 273–284. https://doi.org/10.1109/JSTSP.2014.2309435

Urbain, J., Cakmak, H., & Dutoit, T. (2013a). Evaluation of hmm-based laughter synthesis. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 7835–7839. https://doi.org/10.1109/ICASSP.2013.6639189

Urbain, J., Cakmak, H., & Dutoit, T. (2013b). Automatic phonetic transcription of laughter and its application to laughter synthesis. *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, 153–158. https://doi.org/10.1109/ACII.2013.32

Urbain, J., Niewiadomski, R., Bevacqua, E., Dutoit, T., Moinet, A., Pelachaud, C., Picart, B., Tilmanne, J., & Wagner, J. (2010). Avlaughtercycle. *Journal on Multimodal User Interfaces*, *4*(1), 47–58.

Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., et al. (2017). Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*.

Wood, A. (2020). Social context influences the acoustic properties of laughter. *Affective Science*, *1*(4), 247–256.

Yamagishi, J., Veaux, C., & MacDonald, K. (2019). *Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92), [sound]*. https://doi.org/https://doi.org/10.7488/ds/2645

Zen, H., Tokuda, K., & Black, A. W. (2009). Statistical parametric speech synthesis. *speech communication*, *51*(11), 1039–1064.

Zhao, X., Feng, G. C., Ao, S. H., & Liu, P. L. (2022). Interrater reliability estimators tested against true interrater reliabilities. *BMC Medical Research Methodology*, *22*(1), 232.

Disclaimers

The author of this thesis pledges that:

- *No funding was provided for this research.*
- *No living beings were harmed during this research.*
- *No text or information was acquired using ChatGPT.*

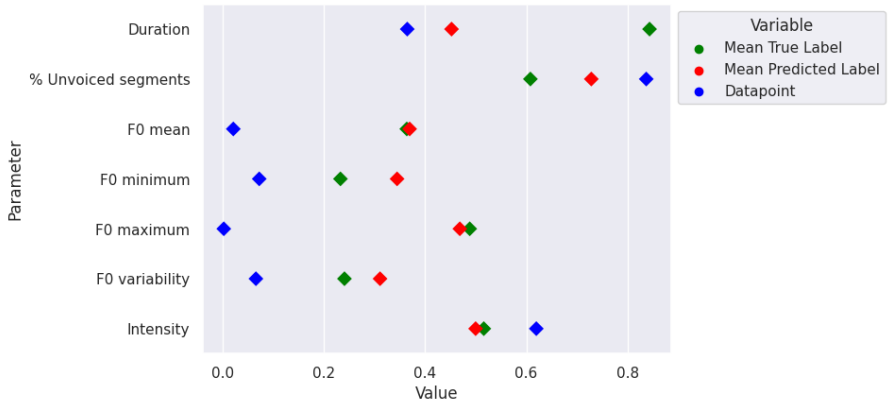
Acknowledgments

I would like to express my deepest gratitude to Matt Coler, Shekhar Nayak, Aki Kunikoshi, and Jaebok Kim for the direct supervision of this master thesis. I would also like to express my sincere gratitude towards the other staff members of the MSc Voice Technology programme for their endless enthusiasm to help us grow, both as students and as people. Furthermore, I am also thankful for work done by the Peregrine HPC cluster support staff to make this thesis possible. Special thanks go to my friends for their company and their invaluable support. Lastly, I'd like to mention my parents, brother, sister, and cat, for always being there for me.

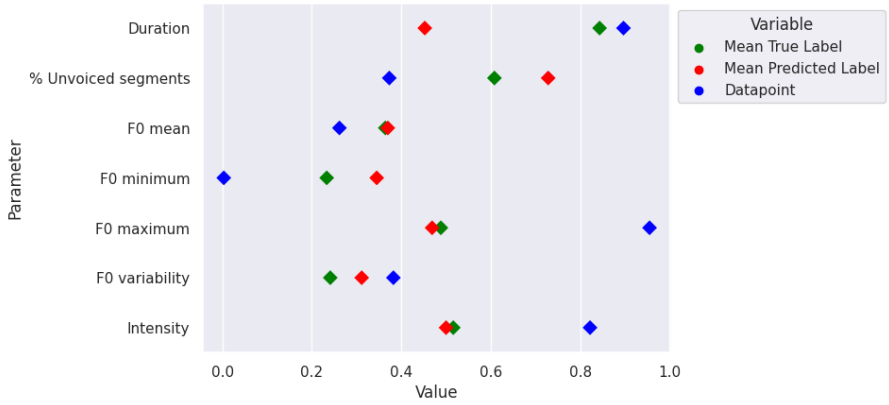
A

Plots

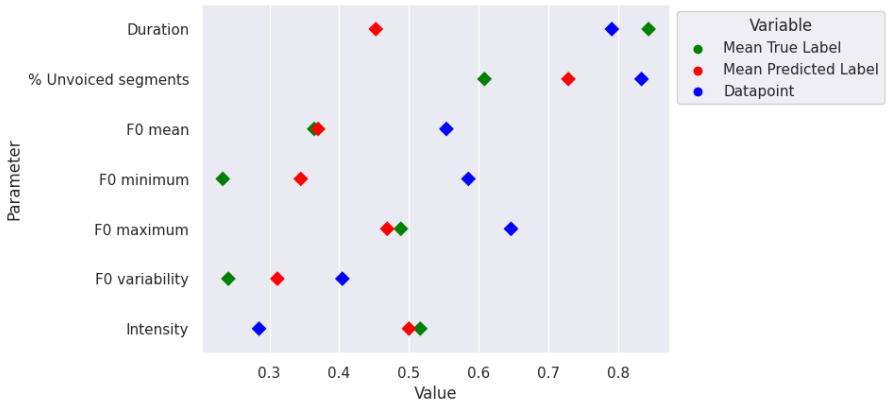
Parameter distributions misclassified laughter relative to class means – Training data



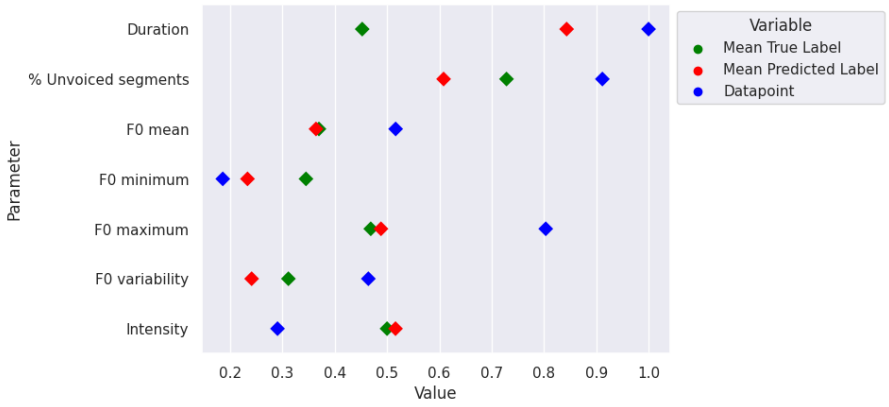
(a) Misclassified acted male laugh 1



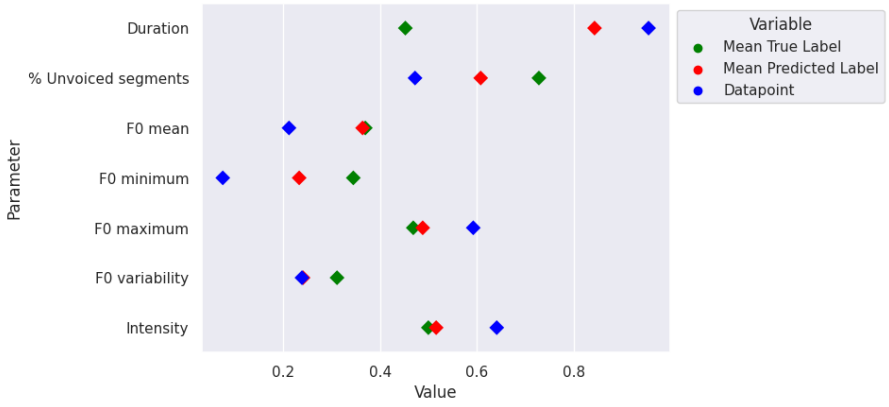
(b) Misclassified acted male laugh 2



(c) Misclassified acted female laugh

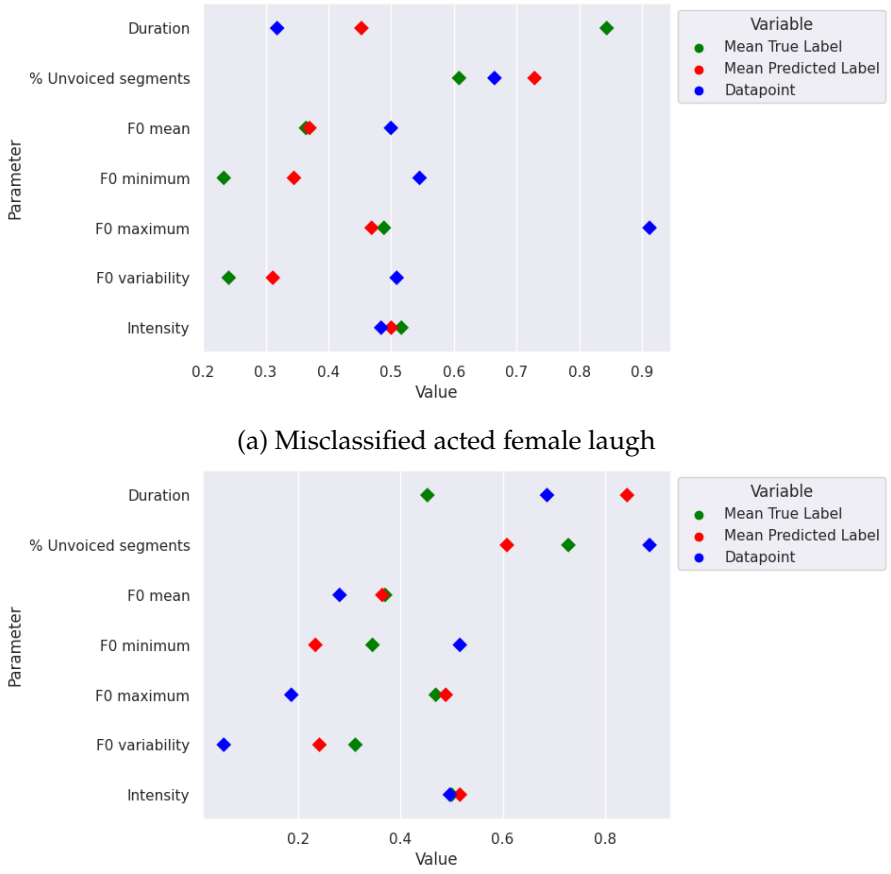


(d) Misclassified spontaneous female laugh 1



(e) Misclassified spontaneous female laugh 2

Figure A.1: Parameter distribution of misclassified training files relative to the class means



(a) Misclassified acted female laugh

(b) Misclassified spontaneous male laugh

Figure A.2: Parameter distribution of misclassified test files relative to the class means

B Questionnaire

B.1 Index page

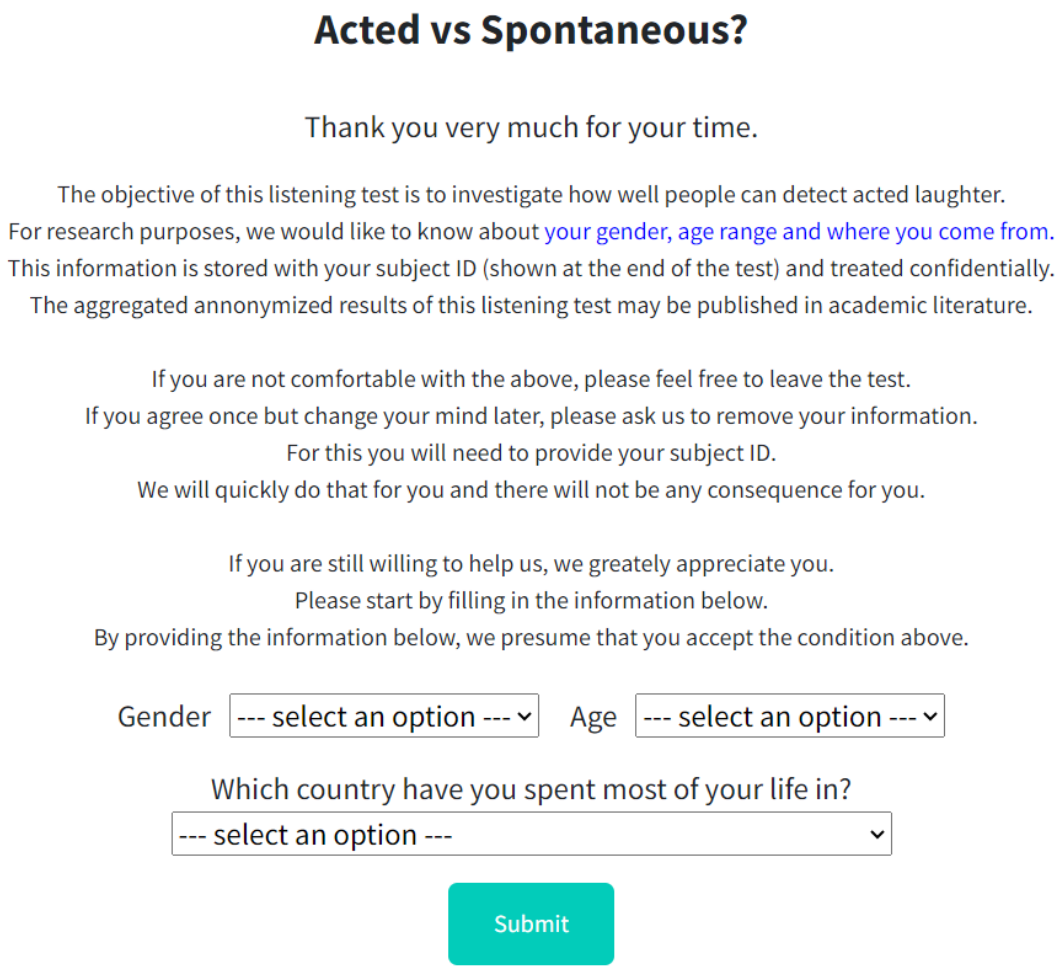


Figure B.1: Screenshot from the index page of the questionnaire

Answer options per question:

- What is your gender? **[Male / Female / Other / Prefer not to say]**
- What is your age range? **[18-30 / 31-40 / 41-50 / 51-60 / 60+ / Prefer not to say]**
- Which country have you spent most of your life in?
  [List of all countries[1]/ Prefer not to say]

[1]https://simple.wikipedia.org/wiki/List_of_countries

B.2 Introduction page

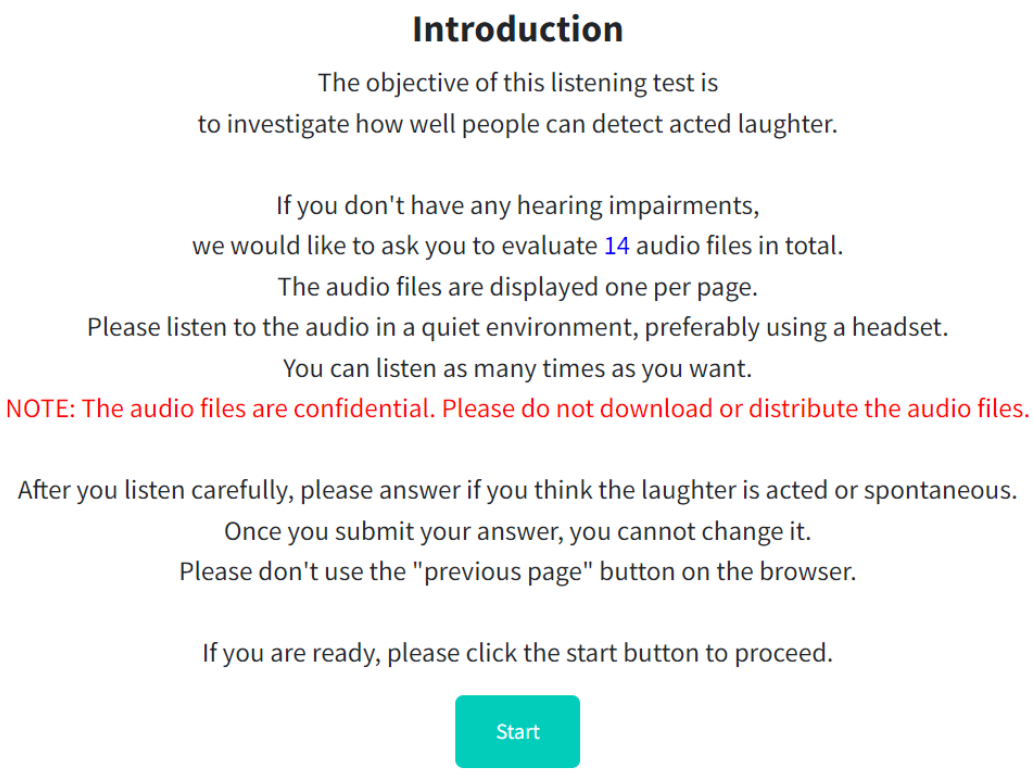


Figure B.2: Screenshot from the introduction page of the questionnaire

B.3 Quiz

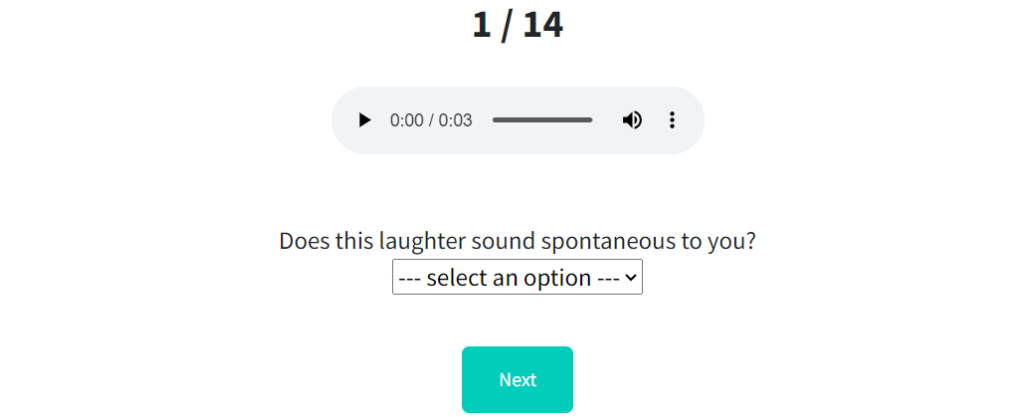


Figure B.3: Screenshot from one of the quiz pages of the questionnaire

Answer options:

- Does this laughter sound spontaneous to you?
  [Yes, spontaneous / No, acted / I really don't know]

B.4 Results page

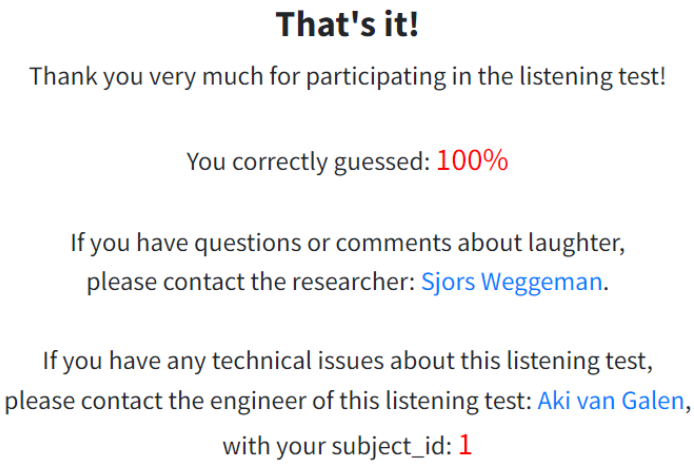


Figure B.4: Screenshot from the results page of the questionnaire