

RADBOD UNIVERSITY NIJMEGEN



FACULTY OF SOCIAL SCIENCES

---

# The Vocal Adversary: Privacy Protection by Voice

---

MASTER'S THESIS  
IN ARTIFICIAL INTELLIGENCE

*Author*

L. VAN BEMMEL

s4574249

*Supervisor*

M.A. LARSON

*Second reader*

L.F.M. TEN BOSCH

December 2022

## Abstract

Many privacy sensitive attributes such as gender, age, emotion and health of a speaker can be obtained just by their voice. Users of voice controlled devices, i.e. Smart Voice Assistants (SVAs) are often unaware of the privacy risks of voice input. So-called inference attacks specifically target the privacy sensitive attributes from a voice and are very successful with Deep Neural Networks. Proposed protective measures against inference attacks often also rely on neural networks to obfuscate privacy sensitive attributes from speech. Neural-on-neural methods are successful in the white-box case where the attacker neural network is known. Here, the protective computational perturbations can be sufficiently small to not disrupt the utility of the Automatic Speech Recognition (ASR) system that is needed to use the SVA. However, we find that additionally to being unpractical and not realistic for the use case of SVAs, neural adversaries are not successful when trying to protect against inference attacks that are based on speech features.

Instead we propose the vocal adversary: a person using their voice to obfuscate privacy sensitive paralinguistic attributes.

The experiments in this thesis specifically focus on gender obfuscation and before-the-mic protection. By examining successful neural adversaries with the use of speech features historically developed by speech scientists that link back to the speech mechanisms, it is revealed what speech features are useful for gender obfuscation. The vocal adversary leverages these features to protect against both neural and feature-based gender inference attacks without losing utility of the voice control. The vocal adversary is intended to provide a realistic everyday protection against inference attacks without requiring extensive effort on behalf of the SVA user. While more research is necessary, this thesis provides a step away from solely neural methods and towards more interpretable non-computational methods that are realistic to use in a daily manner.

## Additional achievements

Several presentations over the course of the thesis project were given as part of the ELLIS Excellence Fellowship at the Nijmegen ELLIS unit.

The final presentation was given at the combined PI research meeting of Centre for Language Studies at Radboud University Nijmegen, 14 December 2022.

This thesis has lead to a pending for review conference paper for International Conference on Acoustics, Speech and Signal Processing (ICASSP) (van Bemmelen, L., Liu, Z., Vaessen, N., Larson, M. (2023) Beyond neural-on-neural approaches to speaker gender protection.).

## Acknowledgements

Finalizing my second MA thesis at Radboud University was a great experience mostly because of the support of people around me. First of all, I would like to thank my supervisors Martha Larson and Zhuoran Liu for their guidance, advice, motivation and enthusiasm for the project. Together with the Acoustic Attacks group I have learned a lot about privacy and security research. Despite the busy academic life, everyone made the effort to be there when it was needed and that is hugely appreciated. Additional thanks for Louis ten Bosch for agreeing to be the second reader quite last-minute. I would also like to thank the co-authors of the paper written for this thesis, Nik Vaessen, Zhuoran Liu and Martha Larson again. Additional thanks to Nik for training the models. A special thanks go to the ELLIS Unit Nijmegen and my fellow ELLIS Excellence students. Our meetings with the amazing view at the top of the Erasmus building surely made completing a thesis project less daunting.

Finally, I would like to specifically thank my brother Rik for being a great Male vocal adversary and taking time from his busy schedule to record the same 251 words in 20 weird voice adaptations. His patience and helpfulness were vital for the proof-of-concept of the vocal adversary.

# Contents

<b>1</b>	<b>Introduction</b>	<b>4</b>
1.1	Gender in voice . . . . .	6
1.2	Threat model . . . . .	7
1.2.1	Neural adversary . . . . .	8
1.2.2	Vocal adversary . . . . .	9
1.3	Research Questions . . . . .	9
<b>2</b>	<b>Background</b>	<b>10</b>
2.1	Gender inference from speech . . . . .	10
2.2	Privacy protection in speech . . . . .	11
2.2.1	Neural adversarial examples . . . . .	11
2.2.2	Privacy sensitive representations . . . . .	12
2.2.3	Voice disguise . . . . .	13
2.3	Shortcomings of literature . . . . .	14
2.4	Vocal adversary . . . . .	15
2.5	Hypotheses . . . . .	15
<b>3</b>	<b>Methodology</b>	<b>16</b>
3.1	Data . . . . .	16
3.1.1	Data format . . . . .	17
3.2	Neural models . . . . .	18
3.3	Neural adversary . . . . .	18
3.4	From Computational to Vocal adversary . . . . .	19
3.4.1	Speech features . . . . .	19
3.4.2	Feature selection . . . . .	19
3.5	Vocal adversary . . . . .	20
3.6	Evaluation metrics . . . . .	21
<b>4</b>	<b>Results</b>	<b>22</b>
4.1	Neural adversaries . . . . .	23
4.1.1	Perturbation rates . . . . .	23
4.1.2	Gender classification results . . . . .	24
4.1.3	Speech features . . . . .	25
4.2	Vocal adversaries . . . . .	27
4.2.1	Speech adaptations . . . . .	27
4.2.2	Gender classification results . . . . .	30
<b>5</b>	<b>Discussion</b>	<b>33</b>
5.1	Limitations . . . . .	34
5.2	Future work . . . . .	35
<b>6</b>	<b>Conclusion</b>	<b>37</b>
<b>A</b>	<b>Full results for neural adversaries with PGD set to 10 epochs</b>	<b>43</b>
<b>B</b>	<b>Full results for neural adversaries with PGD set to 100 epochs</b>	<b>45</b>
<b>C</b>	<b>Mel spectrograms for all neural adversaries</b>	<b>46</b>

# 1 Introduction

Speech contains a lot more information than just what is spoken. So-called paralinguistic attributes that can be obtained from just a voice can lead back to personal characteristics of the speaker. There are links between voice features and human characteristics such as gender and emotion, but also weight, height, heart rate, health, certain diseases, use of birth control pills, psychiatric illnesses, and the environment of the speaker can be inferred from voice alone [1,2]. Some of the claims that these characteristics can be linked back to voice are controversial, but others are widely accepted. Health, for example, can already be monitored by applications in the patient’s home using just voice recordings [3]. While less invasive and less disruptive than other measurements of health that do not use voice, the privacy risks are also undeniable. Generally, people are unaware that characteristics such as health can be derived from a voice [4].

Perhaps this ignorance is one of the reasons people give up their voice as easily as they do [1]. Voice control is named as one of the most important reasons to purchase an Intelligent Personal Home Assistant [5], as voice control is convenient and hands-free [6]. This is by design, of course, as physical closeness to a device is not needed in most Internet of Things (IoT) devices. These devices are defined by their connection with each other and to the internet, where communication does not require a direct link with anything other than WiFi. Having a (voice) input that is not limited to a certain place in the house makes perfect sense for these devices. In this thesis, we focus on these Smart Voice Assistants (SVAs), where the device uses the speech of the user as input.

Examples of SVAs likely exist in our daily lives, the two most common models in Europe being Amazon Alexa and Google Assistant in combination with an Amazon Echo and a Google Home device [5]. A typical use of an SVA is as follows: the user says the wake word to activate the device and make it start recording, the user then gives the device a voice command, the recording of the voice command is sent to the cloud for transcription with an Automatic Speech Recognition (ASR) system and for further processing and analysing, the cloud returns an explicit command for the SVA, and finally the SVA executes the command [2,5,7].

Unfortunately, there are security and privacy risks that come with the working of these SVAs. Security risks include false activations where the wake word is wrongfully detected when it is not spoken, or harmful actors called ‘adversaries’ that are attempting to intentionally change the output of the SVA [7,8]. To compromise privacy, an attacker could try to gain access to the data at one point in the pipeline and infer privacy sensitive attributes from the speaker. This particular attack is called an inference attack [9] and is what we will focus on in this project.

Luckily, some defenses against these risks have been developed over the years. Regulations are focusing more and more on privacy of individuals. In Europe, the General Data Protection Regulation (GDPR) is a law attempting to protect people from having their personal data exploited. While speech is not explicitly mentioned as personal or biometric data, since unique identification of a speaker can be obtained from speech, it does fall under this category. Thus the GDPR applies to speech processing, including the processing by SVAs [5,10].

Other than legal protection, SVA users themselves can also take measures to protect their privacy. Whereas the adversary mentioned above is a harmful party compromising the users safety, adversarial protection is also possible. In this case, the adversary is the user trying to attack the harmful inference attack model. Since neural networks are most successful in paralinguistic inference [11], the protection against these networks are often also produced by neural methods. By using the gradients of a trained neural model and adding specific computational perturbations to the speech input of the SVA, these adversarial methods attempt to ‘trick’ the inference model into giving an incorrect label to the speech, thus protecting paralinguistic qualities of the speaker. The speech with the computational perturbations added in an attempt to hide paralinguistic qualities are called ‘adversarial examples’. These so-called ‘neural-on-neural’ (created with a neural network against inference with a neural network) methods work well in white-box scenarios where every detail about the inference model is known. However, in grey- or black-box scenarios, where some details about the inference model are missing, the neural-on-neural methods are less successful [12].

Unfortunately, both the legal protection and the neural-on-neural protection have some shortcomings

and do not protect the privacy of an SVA user sufficiently. Flaws of both legal and neural-on-neural protection with regard to voice will be discussed briefly.

While the GDPR demands the consent in order for (additional) data processing to occur to be freely given, specific, informed and unambiguous (Art. 4.11)<sup>1</sup>, this is often not the case. Because users are ignorant of the possible privacy risks of using their voice they cannot give fully informed consent to their speech being controlled and processed. Zuiderveen Borgesius [13] elaborates that informed consent cannot be given when people do not know the consequences of future data usage, and that companies are intentionally vague and refer back to their privacy policy, which is generally not read because of its length.

Still, even when the privacy policy is fully read, the data usage intentions and practices of a company can be unclear. We take Amazon’s SVA Alexa as our example. In Amazon’s privacy policy, for which a user is referred to the Help and Customer Service webpage ‘Alexa, Echo Devices, and Your Privacy’<sup>2</sup>, some common privacy questions about Alexa and Echo devices are answered. Here, Amazon states that speech data is used to control the device and for improving their services. They state “we use your requests to Alexa to train our speech recognition and natural language understanding systems using machine learning” and that “Our supervised learning process includes multiple safeguards to protect customer privacy.”. Note that the same webpage is available in Dutch (“Amazon.nl Privacyverklaring”)<sup>3</sup>, implying that the same privacy policy is employed in America as in Europe and thus the GDPR applies. The exact (future) applications and uses of the collected speech data is unknown even when reading the privacy policy. Users are potentially giving consent to speech processing and “improving services” for Amazon that they are not fully informed about. The GDPR does not sufficiently protect users against these risks.

The neural adversary is not without flaws either. Regarding the scenario of privacy protection in voice when using an SVA, the most straightforward protection would be before-the-mic. Here, the user has full control over their speech input and no sophisticated models have to be added somewhere in the internal SVA pipeline. The neural adversary would be an intermediate step between the input (which is speech) and the SVA. This would introduce some delay and most likely need an additional system next to the SVA, making the use of the SVA less convenient for the user.

Furthermore, for the neural adversaries the protection against neural inference models is good in the white-box case where everything about the neural model is known. If only some properties like network architecture or training data are unknown, the adversarial privacy protection is often less effective. Additionally, if the computational perturbations generated with the neural adversary introduce too much noise to the speech input, the SVAs utility will drop. For the SVA to work correctly, the ASR system in the cloud should be able to correctly transcribe the audio input. If the perturbations are too large, the utility will likely drop. However, if the utility is kept perfect, perhaps the privacy protection will not protect sufficiently. This is known as the privacy-utility trade-off [14,15] and is present in most adversarial protection methods. Finally, adding computational perturbations to input speech has disadvantages beyond utility decrease. Some SVA systems have liveness-detection in place to avoid non-human or non-speech input for other security reasons [16]. Since the computational perturbations could introduce noise that is detected as non-human, because it is, the liveness-detection could unintentionally defeat the privacy protection of the neural adversary.

With all these disadvantages, we can question whether neural-on-neural adversarial examples are the only privacy protective measures that can be taken. We find inspiration for an alternative by returning to the underlying characteristics of speech as they have been historically developed by speech scientists. These speech features are interpretable and link back to speech production mechanisms. Upon first consideration, it might be natural to suspect that speech features are not as relevant anymore for inferring paralinguistic aspects from speech, since neural methods outperform the less sophisticated speech feature-based models. Often, the use of speech features is discarded completely.

However, our exploratory experiments revealed that the total discard of speech features is a huge oversight. Neural adversaries otherwise successful against a neural model fail to protect privacy sensitive

---

<sup>1</sup><https://gdpr-info.eu/art-4-gdpr/>

<sup>2</sup><https://www.amazon.com/gp/help/customer/display.html?nodeId=GVP69FUJ48X9DK8V>, visited 28-11-2022

<sup>3</sup><https://www.amazon.nl/gp/help/customer/display.html?nodeId=201909010>, visited 28-11-2022

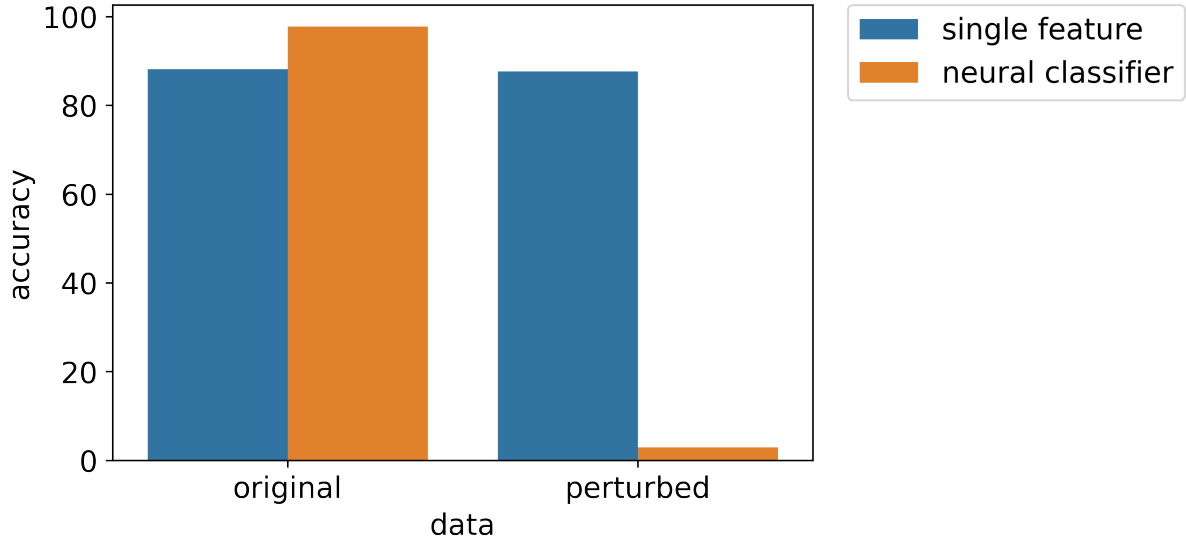


Figure 1: Gender classification accuracy on the VoxCeleb2 test set for a single feature classifier (a linear ridge classifier using the speech feature ‘mean pitch’) and a neural classifier using the raw waveform (WavLM) both trained on LibriSpeech. The neural classifier outperforms the single feature classifier on the original data, as can be seen by the left bars. The neural adversarial examples seen at the right bars (perturbed using WavLM as reference model) are successful in obfuscating gender in the neural classifier, but are not successful for the single feature classifier.

attributes in speech against a speech feature-based inference model. This effect can be seen in Figure 1, where the neural classifiers performance drops with the adversarial protection, while the single feature classifier does not have a change in performance.

Since both legal and neural protections do not seem to work, and the speech related aspects of voice are often not taken into consideration by a protective method, we propose another type of adversarial protection: the vocal adversary. The vocal adversary is a person using their voice to obfuscate privacy sensitive paralinguistic attributes.

The main goal of this thesis is to define the vocal adversary and give a proof of concept. There is a current gap in the privacy protection literature where the neural protections do not consider domain-specific speech feature inference which turns out to defeat neural adversaries, and the protections just by voice are not considered either. We intend to make the privacy protection accessible and realistic for every day use. The goal of the vocal adversary proposed in this thesis is to provide every day means of obfuscation before-the-mic to protect privacy sensitive attributes of a speaker against inference attacks. This can be done by means of voice alone or readily available help aids such as props.

While the vocal adversary is defined broadly, in this thesis we specifically give examples of a vocal adversary using non-electronic means of adapting their voice to obfuscate their gender.

## 1.1 Gender in voice

We focus on the privacy sensitive attribute ‘gender’ rather than other paralinguistic qualities that can be obtained from speech. We have chosen gender for a few reasons:

1. Gender is often reduced down to a binary class, grossly simplifying the socio-cultural concept of gender but also simplifying the inference task.
2. (Binary) Gender marketing is very prevalent in our society and automatic gender inference could lead to stereotypical profiling of users.
3. Gender labels in speech datasets are often available and are assumed to be less ambiguous than other labels such as emotion.

4. When creating the dataset for the vocal adversary, the ground truth is known and will not require any acting skills from the speakers.
5. Trans, non-binary and other ambiguous voices are often neglectfully excluded from the narrative, and we advocate that gender classification as a whole is less binary than assumed and should thus be obfuscated wherever possible.

The notion of binary gender has been historically rejected by many cultures, feminist and gender theorists. One of the most well-known authors, Judith Butler, has famously described both gender and sex as a constructed performance [17]. Gender is described as categories complicated by class, ethnicity and sexuality.

While gender is often attributed to physiological differences between people, the reality is more complicated than that. In speech, pitch is understood to be generally indicative of gender, where Female voices have a higher pitch and Male voices have a lower pitch [18, 19]. Biemans [20] finds that vocal qualities such as pitch are not solely related to typical physiological characteristics between genders. Cultural variability, age, and societal expectations will all influence pitch of voices. Even before the onset of puberty, differences in voices exist between genders. Interestingly, the qualities of the voice also differ between (perceived) genders of the conversation partner [20]. Also in voice, gender is more ambiguous than some might believe.

However, many marketing strategies use a binary view on gender in which research topics revolve around different responses of men and women to stimuli [21]. This reduction of gender undoubtedly results in stereotypical and sexist findings from marketing analysts, but a binary view on gender does seem to be the standard in marketing. Similarly, gender is often reduced down into a binary for machine learning research. This is because most data sets only contain a binary label for gender and strictly defining a class is convenient. According to Larson [22], Natural Language Processing research often lacks the clear definition of the gender categories. Either the definition is vague and problematic or there are no definitions given at all, and it is anyone’s guess how the gender label came to be.

If someone does not want their (possibly incorrectly perceived) gender to be exploited from their voice, they could try to obfuscate their gender. Brunton and Nissenbaum define obfuscation in one sentence in their book ‘Obfuscation: a user’s guide for privacy and protest’: “Obfuscation is the deliberate addition of ambiguous, confusing, or misleading information to interfere with surveillance and data collection” ([23], page 1).

In this thesis, a similar but more specific definition of obfuscation will be used. Obfuscation in our case means deliberately changing the before-the-mic input of an SVA (which is speech) to obtain an incorrect gender label by gender inference models. Because the gender label used in these models is assumedly binary, obfuscation in our case means ‘flipping the label’ from Female to Male or from Male to Female. More technical information about the specific scenario will be described in the next section.

## 1.2 Threat model

In security research, the details of the situation in which the protective adaptation is needed is described in the threat model. The threat model describes the specific situations that are in scope of this thesis and excludes the scenarios that are not considered.

Commonly, the attacker is the party attempting to ‘trick’ a model into giving an incorrect label to an input sample by changing the input samples to adversarial examples. However, in this thesis we inspect the so-called ‘inference attack’, where an attacker is trying to obtain personal information from the input sample without the user’s knowledge. The victim in our case is the SVA user, from which the sample was taken. The user then attempts to obfuscate privacy sensitive information from their input using ‘adversarial protection’. In this case, the user is the one trying to ‘trick’ the attacker inference model into giving an incorrect label. The neural and vocal adversaries in this project are the user attempting to obfuscate gender in their voice and protect their privacy.

So, in our specific case: the attacker is using a gender inference model to obtain the user’s gender from their speech input. The user is using typical adversarial techniques to obfuscate gender from their speech so that the attacker cannot infer gender anymore.

We consider before-the-mic modifications of speech, meaning that we are not looking into disentangled representations or edge-privacy. Any gender obfuscation needs to be present in the audio input to the system. This was chosen to follow the use case of this thesis, the Smart Voice Assistant. We assume users of an SVA cannot control how their speech is represented in the system or what is sent to the cloud and what is kept locally. Rather, users can fully control what is given as audio input to the system before-the-mic.

Additionally, the privacy concern regarding gender inference is reduced down to voice only. We are not considering gender inference risks from language-specific inference or other downstream tasks such as i.e. user profiling based on search or purchase history. In this project we solely rely on speech and specifically voice itself to infer gender.

The VoicePrivacy Initiative [24] specifies a privacy-preservation scenario by five different factors:

1. Nature of data
2. Information seen as personal
3. Downstream goal(s)
4. Data accessed by the attacker
5. Attackers prior knowledge

The neural and vocal adversaries described and experimentally compared in this thesis follow two different threat models. Since they work with different resources (i.e. data and prior knowledge of the attacker), two different descriptions of the exact scenarios are needed. First we will discuss the threat model with the neural adversary, then the threat model with the vocal adversary.

### 1.2.1 Neural adversary

The neural adversary is described here as a person trying to protect their speech from inference attacks in a computational manner, i.e. using neural models to create computational perturbations that will obfuscate gender.

Initially, these computational perturbations will be computed in a white-box manner, meaning that the neural adversary has full access to the gender inference model, architecture, weights, and test data (which is our speech input).

While there is no direct access to the training data used to train the gender inference model, depending on the test data used the data might have the same distribution. For example, if the LibriSpeech dataset is used for both training and testing, the sets will likely have a very similar distribution even if there is no overlap between the train and test set.

We also assume that the neural adversary has unlimited tries and many resources available, meaning that they can create the computational perturbations over many iterations and tweak them without worry.

Using the five factors from the VoicePrivacy Initiative [24], the neural adversary in white-box setting is defined like this:

1. The nature of the data is the raw waveform.
2. The personal information is the gender of the speaker.
3. The downstream goal(s) are control of a SVA, or ASR performance.
4. The data accessed by the attackers is the raw waveform as well.
5. The neural adversaries' prior knowledge is the full trained model of gender inference, including training data, model weights and architecture.

Then, after the white-box scenario, the transferability of the computationally perturbed speech is also tested. In this case, not all of the aspects of the neural inference model are known, making it a grey-box scenario. The computationally perturbed audio is tested against different inference models than the reference model that the perturbation was created with. Different network architectures and different training sets are tested against.

For the five factors, the only change in the transferability scenario is in point 4 and 5. For point 4, the data accessed, other than the raw waveform the attackers have access to extracted speech features as well. For 5, the prior knowledge is not the full trained model of gender inference like it is in the white-box case. Depending on the exact scenario, training data, model architecture, and model weights are not known.

### 1.2.2 Vocal adversary

The threat model of the vocal adversary differs a bit from the threat model of the neural adversary described above. To create a clear picture of the scenario, the same aspects of the threat model will be repeated here with the differences that the vocal adversary brings. Most importantly, the vocal adversary has a different nature of the data and is black-box.

The vocal adversary is different from most adversarial examples in the literature. Rather than computationally perturbing the speech using a neural network, the vocal adversary leverages speech knowledge to adapt their voice in an attempt to obfuscate their gender.

Because no computational perturbations are directly used to create the vocal adversary, but rather different adaptations of voice similar to those of voice disguise techniques, the vocal adversary is by definition black-box.

Nothing about the gender inference model is known to the vocal adversary. The only thing the vocal adversary has access to is the output of the ASR system, likely in the form of whether the SVA does what is expected or not.

Following the five factors from the VoicePrivacy Initiative [24], the vocal adversary scenario is defined as such:

1. The nature of the data is the raw waveform.
2. The personal information is the gender of the speaker.
3. The downstream goal(s) are control of an SVA, or ASR performance.
4. The data accessed by the attackers is the raw waveform including extracted speech features.
5. The vocal adversary has no prior knowledge of model or training data, making this scenario black-box.

## 1.3 Research Questions

The main global research question in this thesis will be

(I) “How can a vocal adversary obfuscate gender from their speech in a non-computational manner against neural and speech feature-based inference models?”

Specifically;

[II] “What speech features does a neural adversary leverage to successfully obfuscate gender?”

II.I “In what settings (white-box or grey-box) is the neural adversary successful in obfuscating gender against a neural model?”

II.II “Are neural adversaries successful in a grey-box setting where the inference attack is done with extracted speech features and speech feature-based models rather than neural models?”

II.III “What is the overlap of speech features deemed useful for gender obfuscation and the ones leveraged by the neural adversary?”

(III) “How can a vocal adversary leverage useful features discovered from neural adversaries to obfuscate their gender in a non-computational way?”

III.I “What do speech adaptations change in the speech feature space? How to these features compare to the speech features leveraged by the neural adversary?”

III.II “How successful is a vocal adversary adapting their speech in obfuscating their gender in a black-box setting against neural and speech feature-based inference models?”

III.III “What is the utility of a vocal adversary measured with performance of an Automatic Speech Recognition model?”

## 2 Background

In the background section, relevant literature and previous research will be discussed. First, we start with briefly explaining the history and the state-of-the-art in gender inference from speech. Then we dive into privacy protection in speech and give several examples which will be used as inspiration for our neural and vocal adversaries. We will show the literature has some shortcomings and why the vocal adversary is a novel concept. Then we describe the vocal adversary in more detail and give the expectations of the thesis in the hypotheses.

### 2.1 Gender inference from speech

According to Harb et al. [18], inferring gender from speech is not a difficult task. Lebourdais et al. [25] even consider it a solved task, as many models have reached a classification accuracy above 90%. Although values of pitch in both Male and Female voices have some overlap and pitch is non-trivial to extract from speech, it is proven as a good discriminator for gender in voice. Wu and Childers [19] cite the physiological differences in the speech production mechanisms as the reason for the mean differences in pitch between Male and Female speakers. They state that the mean length of the vocal tract differs, as well as the length of the vocal fold, the thickness of the larynx, angle of the glottis, and more. These characteristics in the speech production mechanisms will all result in different values of the speech features. Harb et al. show a pitch histogram for Male and Female speech ([18], page 182), which is similar to the pitch histogram of Lebourdais et al. [25], and we have recreated it with LibriSpeech train in Figure 2 and gain very similar results as well. However, while the mean pitch might show these trends, it has also been found that the absolute pitch ranges between men and women do not show a significant difference [20]. Furthermore, as Wu and Childers state, different authors report different ranges of pitch values [19]. It is very likely that pitch is speaker-dependent and not indicative enough to infer gender with high accuracies.

Nowadays, gender inference is often used as a pre-processing step for other speech tasks and cascaded systems such as Emotion Recognition [26, 27], voice assistants [28], speaker recognition [29], and ASR systems [30, 31].

Among features that are often used for these other speech tasks, Mel-frequency cepstral coefficients (MFCCs) are the most common [25, 28, 32]. MFCC features are also reported to outperform classical pitch features in gender classification from speech [32]. For other paralinguistic tasks, specific feature sets have been made by speech scientists to capture interesting qualities in the speech. For example, the Geneva Minimalistic Acoustic Parameter Set (GeMAPS) [33] that contains 88 features has been specifically designed for clinical speech analysis including pathological speech. The Computational Paralinguistics Challenge (ComParE) [34] feature set contains 6373 features. Together with a Support Vector Machine (SVM) classifier, the ComParE feature set is used as the baseline for any ComParE challenges typically hosted at Interspeech conferences. Toolkits such as openSMILE [35] facilitate the automatic extraction of speech features from recordings. Praat is another popular program well-known among speech scientists to extract and analyse features from speech [36].

The main advantage of these features is that they are handcrafted to specifically capture important information from the speech data. Furthermore, they are interpretable and have a link back to the speech production mechanisms, meaning they actually tell us something about the speech itself [37]. The main

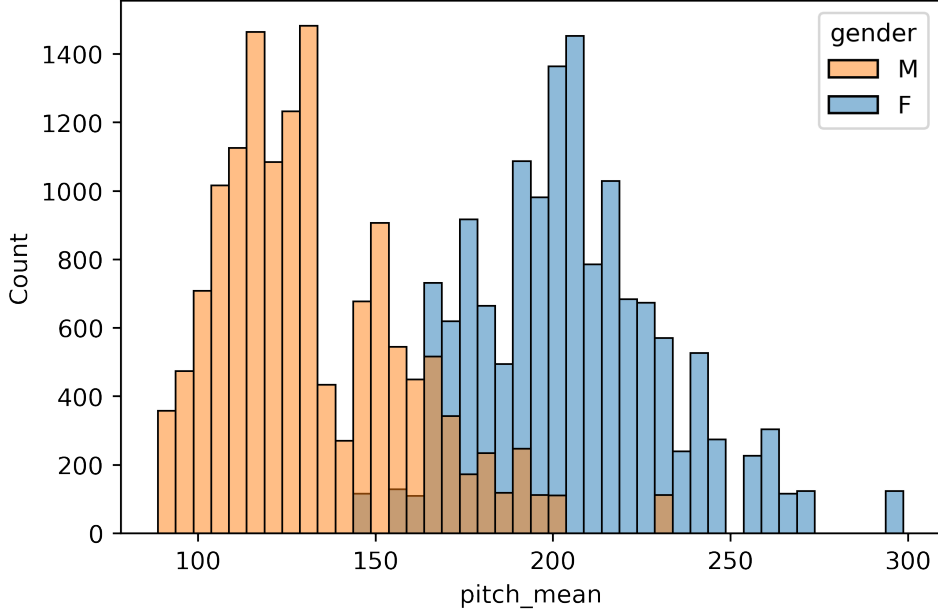


Figure 2: Histogram of speech feature ‘pitch\_mean’ split on Male (M) and Female (F) speakers from the LibriSpeech train set. Very similar to the findings by Harb et al. [18] and Lebourdais et al. [25].

downfall of these features is that conversion with them is lossy, meaning that the audio quality decreases with the conversion from MFCC or other features and back to audio. While the reconstructed audio is often still recognizable and intelligible, the audio quality itself has decreased in conversion [11, 15].

Because of the disadvantages of acoustic features in general, including their poor performance in noisy environments [38], state-of-the-art models in speech are so-called “End-to-End” models [11, 26, 32, 39, 40]. These models rely on layers in a Deep Neural Network (DNN) to automatically extract so-called ‘deep’ features from the raw audio [28, 39, 41]. Rather than using handcrafted speech features, these deep encoded features are supposed to capture variety in speech that in turn are useful for all kinds of analyses down the line. Lehmann and Stadelmann [42] have shown that deep features from different DNNs could also be used to distinguish gender. Kabil et al. [41] find that Convolutional Neural Networks (CNNs) with raw waveforms as input outperform models that use features for gender inference. Alnuaim et al. [28] find that pretrained networks fine-tuned on gender classification outperform traditional machine learning approaches. Lebourdais et al. [25] find the same for WavLM, where a pre-trained and fine-tuned WavLM outperforms the models using speech features. Lastow et al. [38] find that a fine-tuned WavLM on age prediction outperforms other models, and the embeddings obtained with the WavLM are well-suited for gender classification.

## 2.2 Privacy protection in speech

Projects like the VoicePrivacy initiative [24] have been very popular in the last few years. While most privacy initiatives focus on speaker identity protection, some methodology could be used for gender obfuscation as well. We focus on three main groups: neural adversarial examples, privacy sensitive representations, and voice disguises. Note that this is a non-exhaustive list of privacy protection methods, and there are many more that fall out of the scope of this thesis. The neural adversary used in this thesis will fall in the first category, while the vocal adversary examples in this thesis will fall under voice disguises and then specifically the non-electronic ones.

### 2.2.1 Neural adversarial examples

Adversarial examples created using neural networks have been used since 2014. They initially started in the domain of image classification and computer vision. Szegedy et al. [43] were the first to exploit

the inherent vulnerability of deep neural networks for image classification by adding carefully crafted perturbations to the images. By maximizing the networks prediction error, the perturbations achieve a high confidence score in an incorrect class. Furthermore, the perturbations are often small enough to be human imperceptible, or only add a little ‘noise’ to the data [12, 44–46]. Goodfellow et al. [8] went further and developed an efficient method to compute these perturbations for a given input and network. This method is called the Fast Gradient Sign Method (FGSM) and uses backpropagation to compute an effective yet small perturbation. Details on FGSM and iterative perturbation methods can be found in section 3.

Neural adversaries can be roughly split up into a few categories regardless of modality, defined by the specific threat model. Black-box or white-box signifies the amount of knowledge that an adversary has about the model and data. Black-box signifies that none of the model parameters, architecture, or training data is known and controlled. White-box signifies that everything about the model and training data is known. A more nuanced ‘grey-box’ scenario is also possible, which is a mix of white and black box where some aspects of the attacker model are known while others are not. Targeted or non-targeted signifies whether the adversary is attempting to misclassify the input sample with a specific label. In a targeted scenario, one specific output is aimed for and the input will be perturbed in such a way that this target class will now be the incorrect output. For a non-targeted scenario, the attempted output of the adversary is any label other than the true one. It is also possible for the adversary to target the confidence of the model rather than a specific output label.

Input-specific or universal signifies whether the perturbation only works for a specific input and model, or whether it works more universally. Ideally, an adversarial example would be universal and would work in both white, grey, and black-box scenarios for any input sample [47]. When a perturbation is computed with one network but tested with another, we also say that we test the transferability of the adversarial example on different networks [48]. Abdullah et al. [45] find that the adversary has a higher performance in transferability scenarios when a ‘harder’ model is used as reference model. They define the ‘harder’ model as using a higher threshold across samples. In practice, truly universal adversarial examples are difficult to create [12, 47].

Even white-box neural adversarial examples are not always effective. Adversarial speech privacy-preservation typically has a privacy-utility trade-off [14]. Often, privacy sensitive speech is not usable for further downstream tasks, and usable speech is not privacy sensitive.

Nevertheless, some efforts have been made for adversarial privacy protection in speech. Most adversarial examples in speech are attempting to obfuscate speaker identity [24, 44] or target ASR systems to result in an incorrect transcription of the audio [12, 46]. However, some efforts have been made to use adversarial examples to obfuscate paralinguistic qualities of speech. Gong and Poellabauer [11] were the first ones to attempt this, and they use an End-to-End approach to directly modify raw waveforms in order to obfuscate gender, emotion and speaker identity from speech. They report different results with different perturbation rates, but their best performing method still achieves a 70% gender classification accuracy, showing their obfuscation is not highly successful. Stoidis and Cavallaro [14] specifically target gender ambiguity for their adversarial speech, meaning they aim for a 50% classification accuracy rather than a 0% for different genders. They use a generative adversarial network to synthesize a voice that cannot be recognized as conclusively Male or Female. Their final classification accuracy using a CNN is 53.63% for the gender obfuscated speech.

Still, gender is not often considered as the main privacy sensitive attribute to obfuscate. In adversarial examples against speaker recognition, it is often found that inter-gender attacks where the targeted speaker is of another gender than the source speaker is more difficult [48]. Perhaps this, in combination with the perception that gender obfuscation is less interesting than speaker obfuscation, is why gender is an overlooked privacy sensitive quality of speech.

### 2.2.2 Privacy sensitive representations

Privacy sensitive representations are an on-device manner of ensuring privacy for an SVA user. The idea is that before any information is sent to the cloud for other speech tasks such as ASR, the speech is transformed into a representation from which paralinguistic attributes cannot be inferred anymore [7].

One of the privacy sensitive representations are so-called disentangled representations. These representations of speech are invariant to paralinguistic attributes of the speaker such as gender, accent, language, etc.. [7]. To make these representations fully privacy sensitive and yet still usable for downstream tasks in the cloud is challenging.

Stoidis and Cavallaro [49] have shown that disentangled representations of speech can protect a speaker from attribute inference attacks. Their disentangled representations of speech are generated with a variational autoencoder (VAE). They report a WER of  $> 65\%$  for their disentangled representations, while achieving a binary gender classification accuracy around  $50\%$  and reducing speaker identification. Interestingly for speaker identification, they show that encoding onto a random gender is more effective than encoding onto a random identity, meaning that gender obfuscation also obfuscates speaker identity in some way. Aloufi et al. [7] show that binary gender classification can be reduced to a random guess using EDGY’s disentangled representations.

All in all, disentangled representations work well for gender obfuscation.

### 2.2.3 Voice disguise

Voice disguise are techniques for a person to intentionally change their voice to hide their identity [50]. Voice disguise is categorized in four distinct types: deliberate, non-deliberate, electronic, and non-electronic [51, 52]. In this section, the focus will be on deliberate voice disguise techniques where the speaker is trying to obfuscate a paralinguistic attribute in their voice. Three examples are given: pitch shifting is an electronic yet simple technique, voice conversion is a more complicated technique and non-electronic voice disguises are most interesting for the development of our vocal adversary.

Similarly to other privacy-preserving techniques, there usually exists a trade-off between performance, computational costs and robustness for (automatic) voice disguise [53].

#### Pitch shifting

Since pitch has historically been an important speech feature in determining gender [18], and even as a single feature obtains a relatively high accuracy as seen in Figure 1, some privacy research in speech focuses on only changing pitch. Pitch shifting is a relatively simple process and does not require sophisticated neural methods.

Surprisingly, Wu et al. [15] find that pitch standardization (pitch shifting while preserving formants) is not very effective in gender obfuscation, as it creates artifacts that are somehow recognized by the system. Nair et al. [29] also find that pitch shifting does not impact gender classification performance much, regardless of extracted features or classification methods. When changing pitch in speech, other speech features are impacted as well, which could explain the lack of effectiveness [54]. This is found in both electronic and non-electronic change in pitch [55]. However, Zheng et al. [53] show that pitch scaling could be successful in obfuscating speaker identity rather than gender.

#### Voice conversion

Voice conversion (VC) is an electronic technique that converts the voice of a speaker to a chosen target speaker, ignoring ‘what is said’ in favour of converting ‘how’ it is said. In other words, the target voice is imitated while the linguistic content is unchanged [53]. While being more related to speaker identification, if the target and source speaker are not the same gender, gender obfuscation could be achieved with a VC method.

Usually, VC involves a decoding and encoding step. From the original speech, features such as F0 contour are extracted, then converted to the values of the target speaker, and then encoded back to a speech signal with a vocoder. These techniques require sophisticated models and are more complex than techniques such as pitch shifting [53]. Depending on the type of vocoder that is used, the conversion may be lossy and lose some quality in the audio signal with the VC.

Kaneko et al. [56] obtain similar results with inter and intra-gender VC, which is generally unusual in speaker identity obfuscation. However, they do use a vocoder for inter-gender conversion and no vocoder is needed for intra-gender conversion. The inter-gender conversion seems to require a more sophisticated method, even if the final performance with intra-gender conversion is similar.

Voice conversion can also be categorized as a type of voice disguise that is not necessarily electronic [50]. The electronic type of VC is the one described in this section, while the non-electronic type of VC will be briefly described in the next section.

### Non-electronic voice disguise

Most non-electronic types of voice disguise follow directly from the voice production mechanism of a speaker. They intentionally change their voice to hide their identity or hide some paralinguistic characteristics. Perrot et al. [50] categorize three different types of non-electronic voice disguises: prosody, deformation, and phonemic. Examples are speaking slow or fast for prosody, pinching nostrils for deformation and speaking hyper-nasally for phonemic.

Targeted voice disguise such as imitation (non-electronic VC as mentioned above) has some disadvantages. Most importantly, successfully mimicking someone’s voice is very difficult for an untrained speaker. Some difficulty in successfully imitating a different gender arises due to physiological differences in the speech production mechanisms between speakers [52]. According to Perrot et al. [50] and Hautamäki et al. [55], one of the important factors in a successful imitation is the fundamental frequency. However, as Farrus [52] mentions, untrained speakers trying to change the F0 range in their voice are often unsuccessful while professional voice actors succeed.

Whispering is another way to disguise a voice that requires no professional training or practice. Since whispering omits vibrating of the vocal cords, the fundamental frequency of a whisper is absent. While Tartter et al. [57] show that small-scale speaker identification from whispered speech by human evaluation is still excellent, it could be the case that an automatic system is less successful.

Other than imitating and whispering, non-electronic voice disguise can also be achieved with help aids that could influence the speech. Examples of these are manual, pinching noses or cheeks while speaking to change the speech, or with an additional prop. Ahmed et al. [16] use a specifically selected tube to obfuscate speaker identity and create targeted adversarial examples. The tube functions as an extension of the vocal tract and causes a pitch shift. Ahmed et al. show that ‘analog audio adversarial examples’ could possibly bypass liveliness detection in SVAs. Sebastian and Mary [58] found that voice masking using a thick cloth over the microphone affected the energy features, but prosodic features like pitch are more robust against this type of voice masking.

So, for non-electronic voice disguise, some tasks like imitation are quite difficult. Simply whispering, changing pitch, speed or loudness of speech is easier for an untrained speaker. Still, inter-speaker variation is found to be extensive [51, 55] and some speakers are simply more skilled in disguising their voice than others.

## 2.3 Shortcomings of literature

While end-to-end models and DNNs are performing quite well in gender inference, gender obfuscation is often overlooked as a commendable privacy-preservation.

In general for privacy protection methods, what is actually changed in the speech with the privacy protection is mostly overlooked. Some authors examine human perceptions of the obfuscation based on intelligibility, understandability and naturalness [45], but a formal analysis using speech or audio features is not done. Stoidis and Cavallaro [14] do show the spectrograms of original and computationally perturbed speech per gender, but do not further interpret the findings. In the speech domain there are interpretable speech features available, and researchers could use them to reveal important changes in their speech data.

Most privacy-preservation methods deal with the privacy-utility trade-off, and the main three methods described in this section mostly have this trade-off as well.

Neural adversaries are generally preferred over traditional speech-feature based methods because of their apparent high performances. However, as Figure 1 shows, neural adversaries are insufficiently obfuscating gender in speech when faced with a speech feature-based model. Additionally, some neural adversaries fail when they are used in realistic over-the-air scenarios [44], showing that before-the-mic protection with neural adversaries is likely to be less successful than anticipated.

While privacy sensitive representations get a lot of attention [24] and seem to work well [7], converting

the representation back into audio is non-trivial [15]. Because of the trend of End-to-End models and using raw waveforms rather than representations, and the fact that these methods cannot be performed in an before-the-mic manner, the privacy sensitive representations are not considered as a viable option for gender obfuscation in this thesis.

Voice disguise techniques are not as successful as neural-on-neural methods in a white-box scenario, and methods like voice conversion also require a conversion step back to audio. Non-electronic types of voice disguise such as imitation are often difficult for a non-professional to perform successfully.

Ahmed et al. [16] set a step in the direction of what they call ‘analog adversaries’, not using computational perturbations but instead using a tube to defeat speaker authentication models. They mention the benefit of circumventing liveness detection, as the human voice is still the input to the system. We expand onto this concept with the vocal adversary.

## 2.4 Vocal adversary

In this thesis, the vocal adversary is proposed. Rather than using a particular model to obtain computational perturbations with to protect gender information in speech like the neural adversary, the vocal adversary instead leverages knowledge of speech features and voice disguise to protect their speech.

The specific vocal adversary in this thesis is a person wanting to obfuscate gender information from their voice, and doing so by means of their own speech production mechanisms in a before-the-mic protection. Additional readily available help aids such as fans or tubes are also in the scope of our vocal adversary in this thesis. However, in general the vocal adversary is defined by using voice and voice alone. What is most distinctive of a vocal adversary in comparison with a neural adversary is that the former does not require any model or calculations to protect privacy sensitive information from their voice. The vocal adversary does not have access to any inference models. Instead, they only have access to the input of this model, which is their speech before-the-mic.

Our vocal adversary will have different voice adaptations where they change their voice in specific ways. These are inspired by non-electronic voice disguise techniques and will be explored more in the methodology and results sections.

There are a few things to consider when designing privacy protections for a vocal adversary. First, as mentioned, some voice disguise techniques are difficult to perform for an untrained speaker. Important speech features such as fundamental frequency are difficult to change as they are related to physiological mechanisms. Furthermore, some voice adaptations could introduce embarrassment. The goal of the vocal adversary is to use voice adaptations in a daily manner without trouble, and if someone is too embarrassed to yell to their SVA, maybe yelling is not a suitable voice adaptation for the vocal adversary. Easwara Moorthy and Vu [59] performed a survey into people using SVAs on a smartphone in private and public setting for private and non-private data. They found that the location and type of input information were both influential in determining how likely someone is to speak to their phone. They mention social acceptability as a factor for the use of voice input. When designing the voice adaptations, social acceptability should be a consideration. Partly because of this, we exclude imitations of genders, accents, or speech impairments in our adaptations.

## 2.5 Hypotheses

Since the research questions are generally worded and will not be tested with a statistical test, there is no null hypothesis to accept or reject. Instead, some general expectations will be given as hypotheses.

(I) “How can a vocal adversary obfuscate gender from their speech in a non-computational manner against neural and speech feature-based inference models?”

H: A vocal adversary can use techniques from voice disguise and features obtained with successful neural adversaries to obfuscate gender in both neural and speech feature-based models.

Specifically;

(II) “What speech features does a neural adversary leverage to successfully obfuscate gender?”

H: We suspect that pitch or F0 will be one of the top features leveraged by the neural adversary, as this is found to be a characterizing feature for gender in speech [18].

II.I “In what settings (white-box or grey-box) is the neural adversary successful in obfuscating gender against a neural model?”

H: We suspect that a neural adversary is most successful in white-box scenarios, but some transferability will occur between architectures and training data.

II.II “Are neural adversaries successful in a grey-box setting where the inference attack is done with extracted speech features and speech feature-based models rather than neural models?”

H: Following the observation in Figure 1, we suspect that neural adversaries will not be successful against speech feature-based models.

II.III “What is the overlap of speech features deemed useful for gender obfuscation and the ones leveraged by the neural adversary?”

H: We suspect that pitch or F0 will be a useful feature for gender obfuscation leveraged by the neural adversary.

(III) “How can a vocal adversary leverage useful features discovered from neural adversaries to obfuscate their gender in a non-computational way?”

H: We suspect that pitch or F0 will be an important feature in characterizing the voice adaptations that can be used by a vocal adversary. We also predict that voice adaptations on their own will obfuscate gender in both neural and speech feature-based models.

III.I “What do speech adaptations change in the speech feature space? How do these features compare to the speech features leveraged by the neural adversary?”

H: We suspect that different voice adaptations will change the speech features in different ways, mostly impacting pitch and intensity. We predict that the vocal adversary will partly leverage the important features discovered from the neural adversary.

III.II “How successful is a vocal adversary adapting their speech in obfuscating their gender in a black-box setting against neural and speech feature-based inference models?”

H: We predict that a vocal adversary will be more successful than a neural adversary in speech feature-based inference models, as the speech features are directly changed per voice adaptation. We also predict that the vocal adversary will be successful in obfuscating gender against the neural inference models.

III.III “What is the utility of a vocal adversary measured with performance of an Automatic Speech Recognition model?”

H: We predict that the utility of a vocal adversary will be lower for some voice adaptations but acceptable for other voice adaptations.

## 3 Methodology

In the methodology section, we first describe the neural adversary using data, neural models, and how the computational perturbations are generated. Then we describe how we step away from the neural-to-neural methods and towards the vocal adversary using speech features and feature selection. Then we briefly describe the vocal adversary and finally we describe how we will evaluate the two adversaries.

### 3.1 Data

An overview of the data used in this project can be found in Table 1. All data is spoken in English and contains both speaker tags and gender information. Note that training data (on top), development data (in the middle), and test data (on the bottom) are all separate independent data sets with no overlap of recordings or speakers.

Table 1: Overview of datasets: training, development and testing from top to bottom.

data	avg. duration	# speakers		# utterances	
		F	M	F	M
LS100h	12.69 s	125	126	14 342	14 197
LS960h	12.30 s	1 128	1 210	135 889	143 352
Vox2-dev	7.78 s	2 312	3 682	397 032	694 977
LSdev-clean	7.18 s	20	20	1 374	1 329
LSdev-other	6.44 s	16	17	1 450	1 414
LStest-clean	7.42 s	20	20	1 389	1 231
Voxtest	7.78 s	39	79	10 711	25 526
IEMOCAP	4.46 s	5	5	4 800	5 329

LibriSpeech [60] consists of people reading passages of books out loud and was initially recorded for the purpose of developing Automatic Speech Recognition systems. It is split up into different sections (LS100h, 100 hours of clean training data; LS960h, 960 hours of clean and less clean training data; dev signifying the development set and test signifying the test set. ‘clean’ and ‘other’, denotes clean or less-clean and more-noisy data respectively). Note that LS100h is a subset of LS960h, while the other sets are separate and do not have any speaker overlap. In general, LibriSpeech is known as a ‘clean’ dataset, meaning that there is little background noise and the recording quality is generally high. Furthermore, LibriSpeech is often used because of the gender-balanced subsets which makes for a fair gender classification. The dataset ‘LibriSpeech test clean’ is one of the test sets used in this thesis and will be called ‘LStest’ from now on.

Voxceleb [61] is an audio-visual dataset created from celebrity interviews from YouTube. Different from the read LibriSpeech data, VoxCeleb contains spontaneous speech. It is described as an ‘in the wild’ dataset [42, 62], meaning there is more background noise present and the recording quality varies. The development set of Voxceleb 2, ‘vox2-dev’ is used as training set in this thesis and the test set of Voxceleb 2 will be called ‘voxtest’ from now on.

IEMOCAP [63] (interactive emotional dyadic motion capture database) contains both read speech and semi-spontaneous speech. IEMOCAP is commonly used for speech emotion recognition and contains recordings of dialogue between two actors acting out a scenario with emotion. There is some cross-talk present in the data set and while the speakers are all actors, all the emotion in the speech is acted with a prompt rather than genuine spontaneous emotion.

### 3.1.1 Data format

Since we intent our privacy protection to be before-the-mic, a raw waveform as input is most suited for this project. Luckily, as mentioned before, End-to-End models that use raw waveforms as input are shown to outperform other types of speech representations [41]. By using raw waveforms, the neural adversary will not have any conversion loss either [11, 54].

Unless otherwise specified, we use the raw speech waveform in this thesis. LibriSpeech is in the ‘flac’ file format which is already scaled between in a  $[0, 1]$  range, while Voxceleb is in the ‘wav’ file format and first needs to be scaled. To achieve this, the raw waveform must be divided by 32767 ( $2^{15}$ ) since it is originally a 16 bit integer. All models are trained and tested with the scaled data, including the generation of the computational perturbations by the neural adversary, and then converted back to a full wav file for speech feature extraction. All audio in this thesis is mono-channel and has a sampling rate of 16 kHz.

Since the duration of the different audio recordings varies, see Table 1, the data is padded or cut into segments of 6 seconds similar to other paralinguistic research [11, 27]. Xie et al. [62] and Lehmann and Stadelmann [42] have shown that larger segments results in higher performance in speaker recognition, and it is likely that this is also the case for gender inference. For training, a randomly selected 3 second chunk from the recordings was taken.

## 3.2 Neural models

For this thesis, three different neural architectures are used for gender inference: M5, WavLM and XVector. The code for the training of the M5, WavLM and XVector networks can be found at the GitHub page<sup>4</sup>. Nik Vaessen wrote and executed the training for the neural gender inference models.

The M5 model is a convolutional neural network (CNN) designed to classify raw waveforms based on Dai et al. [39]. It consists of four convolutional layers with batch normalization, ReLU activation and a pooling layer, followed by a fully connected layer for gender classification. Note that a very similar network with a different number of convolutional layers is used by Gong and Poellabauer [11], which they call the WaveCNN. Another similar network is used by Stoidis and Cavallaro [14], which they call GenderNet. CNNs are often used in speech classification as they are a specialized architecture that were considered the state-of-the-art for speech tasks [28].

The WavLM model [40] is a transformer network, pre-trained on LS960 for speaker recognition and fine-tuned in this project as a gender classifier. To fine-tune WavLM for gender classification, we average the output of the transformer network and add three fully connected layers to classify gender. During fine-tuning, everything but the last three fully connected layers are frozen in the first cycle, after which we unfreeze the transformer network. The feature extraction CNN part of the WavLM is kept frozen during the entire fine-tuning process. WavLM uses self-supervised learning and is currently the state-of-the-art for many speech processing tasks, as it is as time of writing the top-performing network architecture on the SUPERB benchmark<sup>5</sup>. Note that HuggingFace provides multiple pre-trained WavLM networks, and the one used in this project is ‘WavLM-Base’ pre-trained on LS960, at time of writing number 3 on the benchmark.

The X-Vector network [64] is widely known for the speaker recognition task [53]. Instead of raw waveforms it uses MFCC features as input, and we extract 40 coefficients with the default values of torchaudio [65] for this purpose.

Each of these three architectures are trained (or fine-tuned in the case of WavLM) on the three training sets used in this project shown at the top of Table 1: LS100h, LS960h and Vox2-dev. In the remainder of this thesis, the architecture name with the training or fine-tuning data as suffix will be used to describe a trained model, i.e. M5-vox is a model with the M5 architecture trained with the vox training data.

## 3.3 Neural adversary

The neural adversary used in this thesis is most similar to the neural adversary examples listed in section 2.2.1. Computational perturbations are added to input audio in an attempt to change the output label. The weights and architecture of a trained model called the ‘reference model’ is used to generate computational perturbations that are added to the original input in an attempt to misclassify the input.

As mentioned, we reduce gender down to a binary class and define a successful computational perturbation when the label has been completely ‘flipped’ from Male to Female or from Female to Male.

The neural adversary is created with use of two of the neural architectures trained for gender inference: the M5 models and the WavLM models. Unfortunately, other inference models used in this thesis are less straightforward to use as a reference model for the neural adversary, since they do not use raw waveforms and the conversion back to speech will introduce noise. Because the M5 and WavLM models are trained on different sets of training data, there are six reference models in total.

To create the computational perturbations, Projected Gradient Descent (PGD) [66] is used, which is an iterative version of the well-known Fast Gradient Sign Method (FGSM) [8]. PGD is able to create a more powerful perturbation with multiple steps. The Equation for FGSM can be found at 1, and PGD does one FGSM cycle each iteration to compute the perturbation that will be added to the original audio.

$$\mathbf{x}_{i+1} = \mathbf{x}_i + \epsilon \cdot \text{sign}(\nabla J(\mathbf{x}_i, y)), \quad (1)$$

The perturbed waveform in iteration  $i$  is  $\mathbf{x}_i$ , and  $y$  is the true label.  $J$  is a loss function, which

---

<sup>4</sup><https://github.com/Loes5307/VocalAdversary2022>

<sup>5</sup><https://superbbenchmark.org/leaderboard>

for  $x_i$  is calculated with the reference model, which is the trained model for gender inference. The  $\epsilon$  is the perturbation rate, which we will expand upon later, and there is also a clipping rate to ensure the perturbation does not exceed a certain limit. Finally, *sign* is the sign function denoting whether the perturbation is positive or negative in value. This perturbation is then added to the initial input  $x$ , after which the network is likely to misclassify  $x$  with a different label than  $y$ .

As the Equation shows, FGSM and thus also PGD use a loss function to compute the gradients of the model. In our case, this function is the Cross-Entropy loss. A small experiment with the perturbation rate that will be shown in section 4.1.1 resulted in an optimal perturbation rate of 0.0005, and the results from section 4.1.2 shows that 100 epochs are preferred over 10. A clipping rate of 0.1 is used to limit the size of the perturbation.

As mentioned in the Threat model in section 1.2, the white-box scenario is when the reference and inference model are one and the same. Transferability is tested with grey-box scenarios where the reference model and the inference model are different. Either the architecture could be different or the data used to train the models could be different. For the non-neural inference model, even the format of the test data is different since the speech features are used for inference rather than the raw waveform. This will be elaborated on in the next section.

### 3.4 From Computational to Vocal adversary

As mentioned before, the neural adversary has many downsides that we want to explore and possibly mitigate using a vocal adversary. However, we also want the vocal adversary to be successful against neural inference models. This is why we take inspiration from successful neural adversaries to create the vocal adversary.

#### 3.4.1 Speech features

With Praat [36], a total of 35 speech features were extracted from each recording including the computationally perturbed speech from the neural adversaries. An overview of the features can be found in Table 2 in the next section. To list them here, the 35 features are the ones from the Praat voicereport: number of Pulses, Periods and Voicebreaks; the degree of Voicebreaks, the fraction of Unvoiced parts, jitter (local, local absolute, rap, ppq5), shimmer (local, local dB, apq3, apq5, apq11), mean of the autocorrelation, Noise-to-Harmonics-Ratio (NHR), Harmonics-to-Noise-Ratio (HNR), mean and standard deviation of period and the min, max, mean, median and standard deviation of pitch. Other than the voice report features, we also extract duration, intensity (min, max, mean, standard deviation), the fundamental frequency F0, first three formants and the centre of gravity from the speech.

Note that the pitch used in this project is linear and estimated with Praat. No semitones or other log transforms were performed in the calculation of pitch or any other features unless otherwise specified.

Fundamental frequency (F0) and pitch are linked to the tension in the vocal folds, estimating the larynx frequency from speech. It is also said to be scaled with the lengths of the vocal folds [19]. Intensity is linked to the subglottic pressure, and is measured with the energy of a signal [50]. Formant values are known to be correlated with vocal tract length [19].

Finally, the gender inference with speech features is performed with a Support Vector Machine (SVM) with a linear kernel using the default values from Scikit-learn [67]. SVM is often used in combination with speech features for classification tasks [29, 34, 58].

#### 3.4.2 Feature selection

Feature selection is a standard pre-processing step for feature-based system to increase the performance. In the case of our vocal adversary, the feature selection has another reason. While the 35 speech features extracted with Praat are interpretable and link back to speech production mechanisms, having 35 features to change simultaneously in your speech is difficult in practice. Since we assume our vocal adversary should be able to be an untrained speaker, having a list of features that are important in successfully obfuscating gender is more useful.

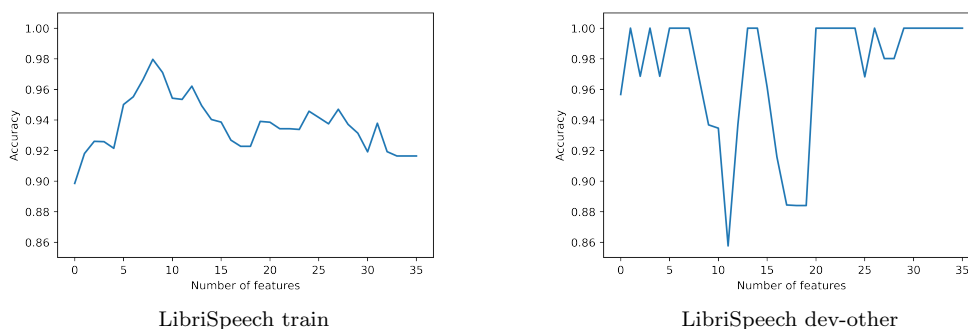


Figure 3: Accuracy scores for SVM gender classification on two LibriSpeech data sets. The SVM uses LOSO cross validation and the number of features on the bottom are ranked with the SVM-RFE ranking described in Table 2, where ‘1’ indicates only using the most important feature, ‘10’ indicates using the top 10 features, and the full 35 features are at the far right of the x-axis.

Additionally, having a feature selection step means that the speech feature set can be extended easily while the rest of the methodology of aiding a vocal adversary does not need to change. Instead of only Praat features, in the future this could be expanded to include openSMILE features such as eGeMAPS or ComPaRE features, which might give more insights into both the neural perturbations and the options for the vocal adversary.

The feature selection in this thesis is performed with Recursive Feature Elimination (RFE) in combination with an SVM, also called SVM-RFE and well-known as a feature selection method [68]. RFE is a wrapper feature selection algorithm that uses feature weights (support vectors in the case of SVM) to iteratively discard the least important feature until only one feature is left. The inverse of the ordered discard list is a ranking of the features based on importance to the classification task. This methodology has been used in previous research [69, 70] for similar experiments analysing speech. The SVM used to obtain the support vectors in the feature selection is a different SVM than the one used for gender inference. The SVM in the SVM-RFE has a linear kernel as well and otherwise uses the default settings from Sci-kit learn [67].

Table 2 shows the 35 features ranked on importance for the Male vs. Female classification, 1 being the most important and 35 being the least important. Note that duration is the least important as it has been padded or cut down to 6 seconds for all recordings.

Figure 3 shows the classification accuracy using this feature ranking. With these results, we can choose a suited number of features for the feature selection, as the smallest number of features together with the highest accuracy would be ideal. For LibriSpeech train, this seems to be at the top eight features, as the highest accuracy is reached with this feature set. The smaller set LibriSpeech dev-other seems to overfit. To obtain this Figure, every top number of features is taken as a set and tested separately. Because only one data set is used (no test-train split), the Leave-One-Speaker-Out (LOSO) cross validation scheme is used instead [3].

### 3.5 Vocal adversary

The vocal adversary recordings in this project consist of two speakers, one Male and one Female, both native in Dutch but with an academic proficiency in English. They read a part of “The patchwork girl of Oz”, a story by Lyman Frank Baum [71] in English. This story is also used as prompt in the LibriSpeech [60] test set and is publicly available online. The prompt is part of the first chapter of the story, “Ojo and Unc Nunkie”, and consists of 251 words split into 25 sentences.

The recordings were made with a Blue Snowball iCE microphone <sup>6</sup> set about 25cm from the speakers mouth using Audacity 3.1.3.0 [72] with a sample rate of 16 kHz. Similarly to the other speech data in this thesis, each sentence from the vocal adversary has been cut down or padded to 6 seconds to stay

<sup>6</sup><https://www.bluemic.com/en-us/products/snowball-ice/>

Table 2: The top 10 features ordered by importance for Male vs. Female classification obtained with SVM-RFE for LS100h, including the mean feature values for Female (F) and Male (M) speakers from LS100h.

Rank	Feature	Description	mean F	mean M	Higher for..
1	pitch_mean	mean of pitch	205.67	132.16	Female
2	autocor_mean	mean of autocorrelation	0.9	0.86	Female
3	nhr_mean	mean of Noise-to-Harmonics-Ratio	0.15	0.2	Male
4	pitch_std	standard deviation of pitch	53.43	51.25	Female
5	pitch_max	max of pitch	441.46	378.77	Female
6	intensity_mean	mean of intensity	68.28	69.18	Male
7	shimmer_apq11	shimmer computed with 11 neighbours	11.12	15.72	Male
8	shimmer_apq3	shimmer computed with 2 neighbours	3.92	5.26	Male
9	intensity_max	max of intensity	78.76	79.46	Male
10	jitter_local_absolute	mean absolute jitter	128.67	259.5	Male
11	jitter_rap	jitter computed with 2 neighbours	1.14	1.35	Male
12	fracUnvoiced	fraction of unvoiced pitch frames	43.25	44.13	Male
13	period_mean	mean of period	5	7.95	Male
14	pitch_median	median of pitch	195.57	117.55	Female
15	hnr_mean	mean of Harmonics-to-Noise-Ratio	13.13	10.67	Female
16	pitch_min	minimum of pitch	106.63	81.85	Female
17	shimmer_apq5	shimmer computed with 5 neighbours	5.81	8.02	Male
18	nrPeriods	number of periods	621.08	390.66	Female
19	pitch_var	variability of pitch	443.04	394.83	Female
20	degreeVoicebreaks	degree of voice breaks	34.73	36.2	Male
21	grav_center	centre of gravity	918.2	770.54	Female
22	shimmer_local	mean absolute shimmer	10.78	13.76	Male
23	nrPulses	number of pulses	638.1	407.44	Female
24	nrVoicebreaks	number of voice breaks	14.2	13.86	Female
25	jitter_ppq5	jitter computed with 5 neighbours	1.26	1.54	Male
26	jitter_local	absolute jitter	2.57	3.29	Male
27	period_std	standard deviation of periods	1.34	2.19	Male
28	f0	fundamental frequency	837.75	806.35	Female
29	f2	second formant	3044.48	2997.67	Female
30	intensity_std	standard deviation of intensity	40.14	38.41	Female
31	intensity_min	minimum of intensity	-57.48	-43.9	Male
32	f1	first formant	1964.6	1952.83	Female
33	shimmer_local_dB	log mean absolute shimmer	1.37	1.63	Male
34	f3	third formant	4098.89	4079.67	Female
35	dur	duration	6000	6000	-

consistent in fragment length.

The same exact prompt was read out with twenty voice adaptations, which will be elaborated upon in the Results section. The voice adaptations are chosen based on voice disguise literature and the computational perturbations from the neural adversary. By examining the change in speech features with a computational perturbation (that defeats a neural inference model), and the speech features important to classify gender in SVMs (seen in Table 2), a vocal adversary leverages the overlap of these features to defeat both neural and speech feature-based gender classification.

### 3.6 Evaluation metrics

Following our goal of designing a Vocal adversary that is able to protect their gender in speech from black-box inference attacks, some results are less interesting on their own and more interesting as a stepping stone to the next part of the project.

Classification accuracy of gender in the neural models and a Support Vector Machine (SVM) using the speech features is reported. Since our data sets are not all balanced on gender (see Table 1), both the full accuracy and the accuracy for Female and Male speakers separately are mentioned. Some papers state that true gender ambiguity in voices is reached when the final classification accuracy reaches 50% [14],

which would equate a random guess. This is only a valid goal if the data set is balanced and both Male and Female samples obtain around 50% accuracy. Since we follow the use case of a Smart Voice Assistant, the user would be most interested in fully obfuscating their gender by ‘flipping’ their gender label completely from Male to Female or from Female to Male 100% of the time. Note that the inverse of the classification accuracy shows the success of the obfuscation.

The transferability of the obfuscated speech is important. It is not exactly known which, if any, gender classification systems SVAs use and it is likely they vary system to system. Since we have seen that a simple linear model can defeat a successful neural adversary of a state-of-the-art gender classifier, we must take care that the obfuscation for the vocal adversary works for multiple models.

The utility of the system is the second most important metric. Utility indicates how well the Automatic Speech Recognition (ASR) system, the original intended system for the voice input in an SVA, can handle the privacy protected speech. This is often measured in Word Error Rate (WER) [7, 14, 44, 48], the number of words that are incorrectly transcribed with an ASR system. If the accuracy for the gender classification is low (indicating a successful obfuscation), but the WER is very high (indicating an unsuccessful ASR transcription), the input speech might be protected in terms of privacy but completely useless in terms of the utility of an SVA.

Equation 2 describes the calculation of the WER using the substitutions ( $S$ ), deletions ( $D$ ), insertions ( $I$ ) and the total number of words ( $N_w$ ) in the reference compared to the hypothesis. Note that a lower WER indicates a higher utility.

$$WER = \frac{S + D + I}{N_w} \times 100\% \quad (2)$$

Since not all data in the test datasets have a transcription available and 6 second segments are used for testing, the ASR transcription of the original data is taken as ground truth. This means the reported WER will be a relative WER for the neural adversary. For the vocal adversary, the comparison with the original prompt as ground truth is made.

There are multiple open-source ASR systems for transcribing English speech available online. For this thesis, both DeepSpeech 2 [73] and ESPNet [74] were considered. DeepSpeech 2 performance has been used before as a utility metric for privacy sensitive speech [7, 49], and ESPNet has been used for this purpose as well [15], implying that both models could be suitable.

To determine which ASR would be most suitable, the IEMOCAP dataset was used. IEMOCAP is a noisy dataset containing cross-talk, meaning that creating an automatic transcription is challenging. Since it is assumed that creating an automatic transcription for neurally perturbed speech is challenging as well, the ASR with the lowest WER for IEMOCAP was chosen for this project. For DeepSpeech 2, the average WER for IEMOCAP was 40.62%. For ESPNet, the average WER for IEMOCAP was 65.71%. Both of these WER results are quite high, but it is expected for noisy speech. Stoidis and Cavallaro [49] report a 6.75% WER on LS100h with DeepSpeech 2, which indicates a much better performance on a less noisy data set.

Following these results, the ASR used in the remainder of this project is DeepSpeech 2 [73].

We need to keep in mind that utility of the ASR system is just as important as protecting the gender of the speaker. The user is already using an SVA and they will likely not switch to a more privacy sensitive input modality such as pressing buttons. If the vocal adversary is to be employed, it should be employed in an easy non-annoying daily manner of the users of SVAs.

## 4 Results

The result section is split up into two subsections: the neural adversary and the vocal adversary. Extended results can be found in the Appendices of this thesis or at the GitHub page<sup>7</sup>.

<sup>7</sup><https://github.com/Loes5307/VocalAdversary2022>

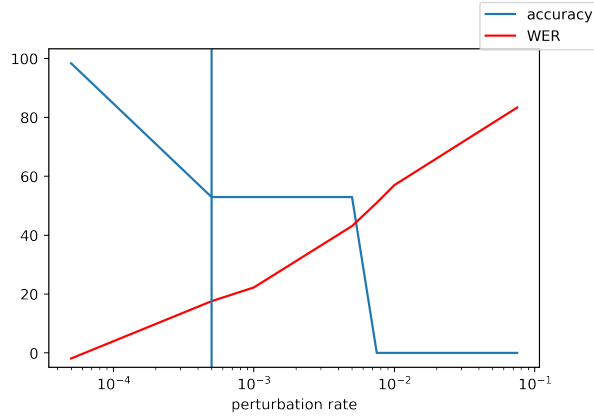


Figure 4: Gender classification accuracy and WER (relative WER, DeepSpeech 2 transcription of perturbed audio compared to DeepSpeech 2 transcription of original audio) of computationally perturbed LibriSpeech dev-other with reference model m5-LS100h with different perturbation rates. Note that the x-axis is in log scale since the perturbation values range from 0.075 to 0.00005. The other settings for the PGD were: 0.1 cutoff and 100 epochs. The chosen perturbation rate (0.0005) is highlighted with a vertical line on the axis.

## 4.1 Neural adversaries

The results for the neural adversaries will be described in this section. First, the perturbation rates and the parameters for the PGD will be explored. Then, the gender inference results will be given for the neural models. Finally, the speech features will be explored and the gender inference results using these speech features will be given.

### 4.1.1 Perturbation rates

To explore a suitable perturbation rate, the other parameters of the PGD have to be consistent. Initially, the number of epochs for PGD was set to 10. Especially for the larger Voxceleb test set, and the larger WavLM model, computing the neural perturbations with 100 epochs was not feasible in time for the publication we were trying to wrap up. However, in this thesis project, there were sufficient resources and time to recompute every neural adversary with 100 epochs. Since our Threat model (see section 1.2) states that there are no to little restrictions for our neural adversary, 100 epochs is more suitable than 10. The results with 10 epochs are reported in Appendix A.

For the perturbation rate, a small experiment with the development set LibriSpeech dev-other was performed. The PGD was tested with different perturbation rates to find a suitable rate that allowed for gender obfuscation (low inference accuracy) yet still an acceptable utility (low WER). The results of this experiment can be seen in Figure 4. As expected, a low perturbation rate results in less obfuscation and a higher utility while a high perturbation rate results in full obfuscation (accuracy at 0%) but also a low utility (WER above 80%). Following the Figure, it was decided that a perturbation rate of 0.0005 (see vertical line) was optimal for the experiments in this thesis. Note that ‘optimal perturbation rate’ will differ between data sets and other factors, so the value of 0.0005 was an informed decision rather than a perfect conclusion.

To briefly see the impact of the computational perturbations on the data, the mel-spectrograms for the original, perturbed with M5-LS960h and perturbed with WavLM-LS960h can be found in Figure 5 for one Female and one Male speaker from LStest. Spectrograms of the same speakers perturbed with all other neural adversaries can be found in Appendix C.

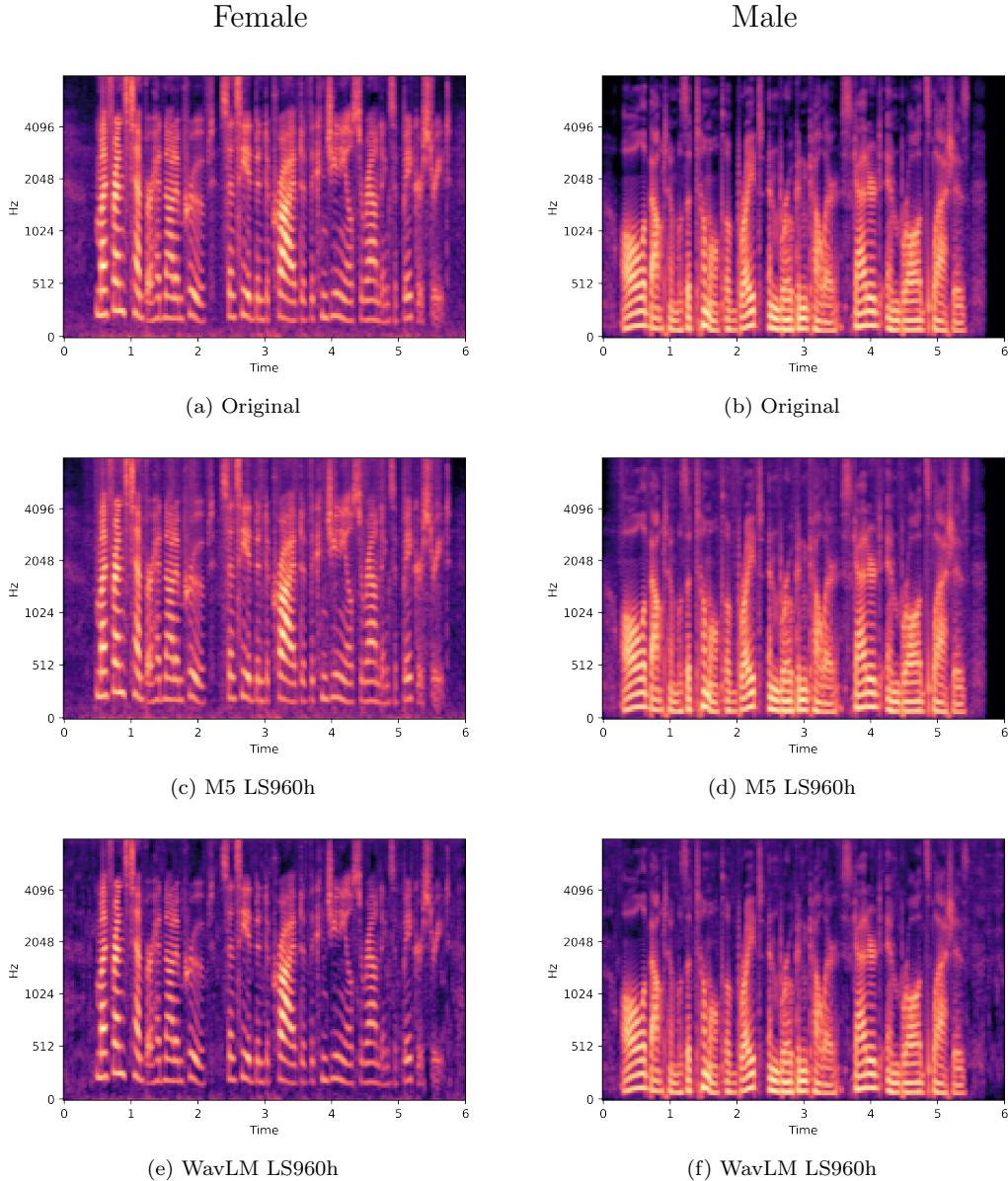


Figure 5: The Mel spectrograms for two speakers from LS960h computationally perturbed with different neural adversaries (reference model in subcaption). Inspired by Stoidis and Cavallaro [14].

#### 4.1.2 Gender classification results

The white-box and some grey-box gender classification results for the neural adversaries can be found in Table 3.

The scenarios where the Reference and Inference model are the same result in a low accuracy, meaning the gender obfuscation of the neural adversary was successful in a white-box setting. The grey-box scenarios where the model architecture is the same but the training data differs results in a higher accuracy, though it could still be stated that the gender obfuscation is successful as the majority of the accuracies are below 50%. Grey-box scenarios with different model architectures were less effective, as can be seen in Appendix B.

To inspect the utility of the neural adversaries we look towards the relative WER listed in the rightmost columns of Table 3. It can be seen that LibriSpeech, the cleaner dataset, results in relatively low WER regardless of perturbation. Note that Stoidis and Cavallaro [49] report a 6.75% WER on the original LibriSpeech dataset, and that our WER is relative to the 6 second fragment we use for the original dataset. VoxCeleb, a noisier dataset, obtains a higher WER for the perturbed speech. However, compared to the

Table 3: Gender classification accuracies on the LibriSpeech and Voxceleb test sets for the original and the neurally perturbed data with different reference models. The white-box case where the Reference and Inference model are the same are underlined. The WER with DeepSpeech 2 with regard to the DeepSpeech 2 transcription of the original data is given as a measure of utility. The PGD for the neural adversary was set to 100 epochs.

Reference model	Inference model											
	M5-LS100h			M5-LS960h			M5-vox			rel. WER		
	tot	F	M	tot	F	M	tot	F	M	tot	F	M
Original LStest	96.56	96.69	96.43	96.83	96.24	97.32	92.98	87.09	99.25			
M5-LS100h	<u>0</u>	<u>0</u>	<u>0</u>	8.27	0.37	17.22	48.28	25.07	74.6	5.96	6.75	5.06
M5-LS960h	13.52	1.11	27.59	<u>0</u>	<u>0</u>	<u>0</u>	29.82	6.56	56.19	5.73	6.55	4.79
M5-vox	57.37	42.7	74	38.68	15.56	64.88	<u>0</u>	<u>0</u>	<u>0</u>	5.44	5.69	5.17
Original Voxtest	92.52	95.49	91.31	95.47	93.81	96.14	97.11	94.58	98.14			
M5-LS100h	<u>0.2</u>	<u>0.25</u>	<u>0.18</u>	11.67	2.08	15.58	51.1	37.73	56.55	27.87	25.11	29.17
M5-LS960h	14.21	4.12	18.33	<u>0.84</u>	<u>0.06</u>	<u>1.16</u>	31.88	16.53	38.15	26.75	24.43	27.73
M5-vox	59.86	49.54	64.08	48.8	18.33	61.23	<u>0.26</u>	<u>0.22</u>	<u>0.27</u>	26.81	22.77	28.5
Reference model	WavLM-LS100h			WavLM-LS960h			WavLM-vox					
	tot	F	M	tot	F	M	tot	F	M	tot	F	M
	Original LStest	97.81	98.6	96.91	98.71	98.53	98.91	98.35	98.01	98.75		
WavLM-LS100h	<u>0</u>	<u>0</u>	<u>0</u>	13.24	2.29	25.67	14.54	16.67	12.12	7.85	8.63	6.96
WavLM-LS960h	37.26	57.45	14.38	<u>0</u>	<u>0</u>	<u>0</u>	31.15	50.29	9.45	7.11	7.86	6.27
WavLM-vox	67.4	59.81	76	55.05	26.99	86.87	<u>0</u>	<u>0</u>	<u>0</u>	6.3	7.28	5.18
Original Voxtest	97.73	97.47	97.85	97	97.62	96.75	99.02	98.2	99.36			
WavLM-LS100h	<u>0.14</u>	<u>0.39</u>	<u>0.04</u>	17.7	7.39	22	23.25	28.01	21.26	34.84	30.9	36.5
WavLM-LS960h	34.45	75.99	17.05	<u>0.07</u>	<u>0.08</u>	<u>0.06</u>	35.29	65.49	22.63	33.33	28.21	35.48
WavLM-vox	81.72	78.25	83.18	75.06	55.8	83.13	<u>0.15</u>	<u>0.19</u>	<u>0.13</u>	29.65	28.02	30.33

baseline WER of 40.62% that IEMOCAP obtained, the neurally perturbed Voxceleb obtains a lower WER.

Our models with the original data outperform other gender classification models based on different architectures or non-raw waveforms, indicating that raw waveforms indeed outperform other speech representations and that WavLM in particular is successful in gender classification [11, 14, 32].

Extended results including more grey-box scenarios for the neural adversary with PGD set to 100 epochs can be found in Appendix B. Similarly, the full results for the neural adversary with PGD set to 10 epochs can be found in Appendix A. As expected, the neural adversary using 100 epochs is more successful in obfuscating the gender in a white-box scenario than the neural adversary using 10 epochs.

However, the WER of the neural adversaries using 10 epochs, as seen in Appendix A Table 14, is lower than the WER for the neural adversaries using 100 epochs. Intuitively this makes sense, since a lower number of epochs is likely to mean a smaller perturbation, thus less ‘noise’ introduced. Since the WER is relative to the original data, the less noise introduced the better the ASR will perform and the lower the WER will be. Especially for the clean LibriSpeech test set, the WER seems to increase with more epochs. For the Voxceleb test set this effect seems to be less substantial.

From this point, only results of the neural adversary with PGD set to 100 epochs will be given.

#### 4.1.3 Speech features

From the 35 features described in section 3.4.1 and shown in Table 2, the most prominent features that changed with each neural perturbation (according to the SVM-RFE) are noted here.

Table 4 shows the increase or decrease of the top 10 speech features with the computational perturbations for different neural adversaries. Features that overlap between the top 10 important features for the Male vs. Female classification and the original vs. perturbed classification are underlined. These features show which features that are important in distinguishing gender in speech are also changed with the perturbation. Note that to determine an increase or decrease, the mean feature value is taken for

Table 4: The top ten features from Table 2 and whether the computational perturbation increases (+) or decreases (-) the value of that particular speech feature compared to the original speech. The computational perturbation is generated with the neural adversary using the Reference model and the Original data (LS = LStest 100h, vox = voxtest). The features have been split up in Female and Male speech, shown in the format (Female/Male). Underlined features are the overlapping features for the top 10 Female vs. Male and the top 10 perturbed vs. original.

Original data	Reference model	<u>pitch</u> _mean	<u>autocor</u> _mean	<u>nhr</u> _mean	<u>pitch</u> _std	<u>pitch</u> _max	<u>intensity</u> _mean	<u>shimmer</u> _apq11	<u>shimmer</u> _apq3	<u>intensity</u> _max	<u>jitter_local</u> _absolute
LS	M5-LS100h	-/+	-/-	<u>+/+</u>	+/+	-/+	+/-	+/+	+/-	+/+	+/-
LS	M5-LS960h	-/+	-/-	<u>+/+</u>	+/+	-/+	+/-	+/+	+/+	+/+	+/-
LS	M5-vox	-/+	-/-	<u>+/+</u>	+/+	-/+	+/-	+/+	+/-	+/+	+/-
vox	M5-LS100h	-/-	-/-	+/+	-/-	-/-	+/+	-/-	<u>+/+</u>	-/+	-/-
vox	M5-LS960h	-/-	-/-	<u>+/+</u>	-/-	-/-	+/+	-/-	<u>+/+</u>	-/+	-/-
vox	M5-vox	<u>+/+</u>	-/-	<u>+/+</u>	-/-	-/-	+/+	-/-	<u>+/+</u>	-/+	-/-
LS	WavLM-LS100h	-/+	-/-	+/+	<u>+/+</u>	-/+	+/-	+/+	<u>+/-</u>	+/+	<u>+/+</u>
LS	WavLM-LS960h	-/+	-/-	+/+	<u>+/+</u>	-/+	+/-	+/+	<u>+/-</u>	+/+	<u>+/+</u>
LS	WavLM-vox	-/+	-/-	+/+	<u>+/+</u>	-/+	+/-	+/+	<u>+/-</u>	+/+	<u>+/+</u>
vox	WavLM-LS100h	+/+	-/-	+/+	-/+	-/-	<u>+/+</u>	-/-	<u>+/+</u>	+/+	-/-
vox	WavLM-LS960h	+/+	-/-	+/+	-/-	-/-	<u>+/+</u>	-/-	<u>+/+</u>	+/+	-/-
vox	WavLM-vox	+/+	-/-	<u>+/+</u>	-/+	-/-	<u>+/+</u>	-/-	<u>+/-</u>	-/+	-/-

the original and the perturbed speech. The actual value change of the speech feature is not examined in more detail, but a brief manual check shows that most feature changes are relatively small.

Figure 6 shows another insight into the perturbation and the extracted features. It is similar to the spectrograms in Figure 5, but the pitch, intensity and formants are drawn on top of the spectrograms, showing how the neural adversary changes the way these speech features are computed with Praat. For example, at the very beginning of the fragment, for the original speech no features seem to show up, but for the perturbed speech some formants are detected. Mean pitch is one of the overlapping features, as seen in Table 6, but the pitch in the spectrogram does not seem to change much by eye. Mean intensity was not an overlapping feature, and it also does not seem to change when comparing the two spectrograms.

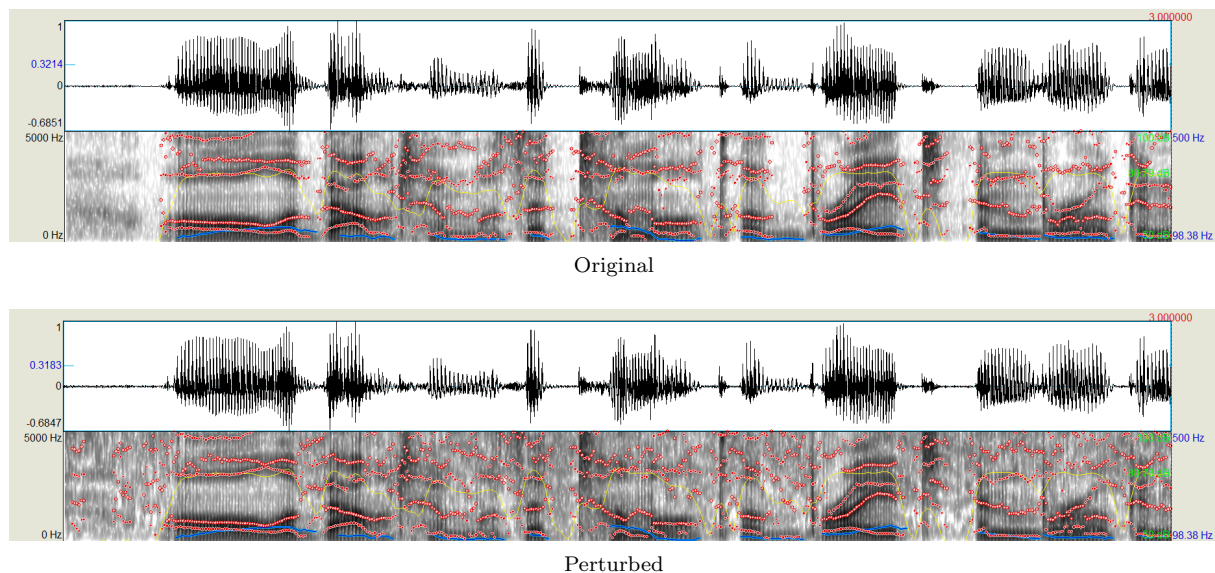


Figure 6: Screenshots from Praat [36] with the original speech (top) and perturbed speech (bottom) from the first 3 second of the same Male speaker from LStest as in Figure 5. The reference model for the perturbed speech is WavLM-LS960h, making these spectrograms simply a different style and from a different program than the spectrograms in Figure 5, where (b) is ‘original’ here and (f) is ‘perturbed’ here. Pitch (in blue), intensity (in yellow) and the formants (in red) are also drawn by Praat.

Table 5: Gender classification accuracy results with linear SVMs trained on and using the feature ranking obtained with LibriSpeech-train (seen in Table 2) for neural adversaries described with the reference model.

Reference model	Inference model								
	SVM top 8			SVM top 10			SVM full		
	tot	F	M	tot	F	M	tot	F	M
<b>Original LStest</b>	89.34	95.9	81.44	94.22	95.9	92.19	92.29	94.67	89.42
M5-LS100h	89.42	95.27	86.19	89.42	94.89	86.62	86.35	94.06	80.87
M5-LS960h	90.6	95.46	86.51	90.36	95.01	86.51	87.49	92.04	83.72
M5-vox	91.03	95.86	86.41	90.99	95.34	86.92	87.73	92.09	83.6
WavLM-LS100h	89.84	95.75	83.11	89.72	95.53	83.11	86.3	90.01	82.09
WavLM-LS960h	91.74	96.14	85.8	91.66	95.77	86.05	87.93	92	82.42
WavLM-vox	91.51	95.85	86.53	91.48	95.4	86.96	88.34	93.11	82.86
<b>Original Voxtest</b>	85.17	91.59	82.4	87.59	91.27	86.03	79.7	95.86	72.88
M5-LS100h	86.2	90.9	84.2	88.1	90.73	86.99	80.93	95.19	74.9
M5-LS960h	85.81	91	83.63	88.03	90.66	86.92	82.66	93.97	77.9
M5-vox	86.15	91.24	84	88.28	90.9	87.18	82.94	94.02	78.27
WavLM-LS100h	85.58	91.31	83.16	87.63	91.04	86.19	82.84	94.13	78.1
WavLM-LS960h	85.56	91.53	83.04	87.42	91.11	85.86	82.51	94.5	77.46
WavLM-vox	85.7	91.5	83.25	87.86	91.11	86.5	82.25	94.58	77.06

The gender inference accuracy of the neural adversaries using speech features and SVMs can be found in Table 5. As expected, the neural adversary does not decrease the accuracy in the SVM. In some cases, the neural adversary even increases the accuracy scores. Note that the feature selection seems to be successful, as the top 8 and top 10 features often obtain a higher accuracy than when using the full feature set.

## 4.2 Vocal adversaries

In this section, the results from the vocal adversary will be given. First we look into the speech adaptations used by the vocal adversary, and how these adaptations impact the speech features. Then, we show the gender inference results of the vocal adversary together with the utility.

### 4.2.1 Speech adaptations

The twenty voice adaptations used in this project can be found in Table 7. As mentioned before, the adaptations used by the vocal adversary are inspired by voice disguise literature [50–52, 55, 58] and the important features to distinguish Male and Female speech seen in Table 2. We see that the mean of pitch is the most important speech feature, but also the standard deviation in pitch (meaning how much pitch varies) and the maximum of pitch seem to be indicative of gender in speech. Furthermore, the mean and maximum of intensity are distinguishing, which can be interpreted as a measure of loudness. Shimmer and jitter are both less easy to directly influence in a voice, but they are influenced by vibration of the vocal cords and sometimes indicative of pathologies [37]. Sounding ‘more pathological’ such as increasing breathiness, hoarseness and roughness in a voice could potentially be obfuscating gender as well.

Voice adaptations made by the vocal adversary are in this thesis split up into three categories: only voice, manual and with an object. ‘Only voice’ describes when no additional methods other than voice can be used to adapt a voice. Lowrobot and Highrobot are both related to prosody and the phonemic category [55]. The whispers and overlyhappy voice adaptations could be categorized as prosody related. ‘Manual’ describes when the vocal adversary temporarily deforms their face using their own hands, for example pinching their nose, similar to the ‘deformation’ category in Perrot et al. [50]. ‘With an object’ describes the case in which readily available objects, such as a pen, are used to adapt the voice.

Table 7 also contains the WER scores per vocal adversary per voice adaptation. Note that the WER for the default voice is above 30% for both vocal adversaries. The WER is computed with the DeepSpeech 2 transcription compared to the original prompt.

Table 8 shows the top ten features most changed with different voice adaptations compared to default speech of the vocal adversary. These features describe what characteristics of the voice are changed with a voice adaptation in a technical way.

Table 6 shows whether the mean features increase (‘+’) or decrease (‘-’) with the adaptation compared to the default speech for a vocal adversary. Similarly to Table 4 for the neural adversary, Table 6 shows the top 10 most important features for distinguishing Male and Female speech and underlines the overlap with the top 10 features changed with the voice adaptation.

Table 6: The top 10 features for distinguishing Female and Male speech from Table 2 and whether the voice adaptation of the vocal adversary increases (‘+’) or decreases (‘-’) the value of that speech feature compared to the default speech. The value differences of Female and Male speech are given in the format (Female/Male). The overlap of features of the top 10 Male vs. Female and the adaptation vs. default of the vocal adversary are underlined.

Voice adaptation	<u>pitch</u> <u>_mean</u>	<u>autocor</u> <u>_mean</u>	<u>nhr</u> <u>_mean</u>	<u>pitch</u> <u>_std</u>	<u>pitch</u> <u>_max</u>	<u>intensity</u> <u>_mean</u>	<u>shimmer</u> <u>_apq11</u>	<u>shimmer</u> <u>_apq3</u>	<u>intensity</u> <u>_max</u>	<u>jitter</u> <u>_local</u> <u>_absolute</u>
mediumroomreverb	+/+	-/-	<u>+/+</u>	-/+	-/-	-/-	+/+	<u>+/+</u>	-/-	+/+
Whisper	+/+	-/-	+/+	+/-	+/+	-/-	+/+	<u>+/+</u>	-/-	+/-
whisperplus25db	<u>+/+</u>	-/-	<u>+/+</u>	+/-	+/+	<u>+/+</u>	+/+	<u>+/+</u>	+/+	+/-
stagewhisper	<u>+/+</u>	-/-	+/+	-/-	+/+	-/-	<u>+/+</u>	<u>+/+</u>	-/-	-/-
Overlyhappy	<u>+/+</u>	+/+	-/-	+/-	+/+	<u>+/+</u>	+/-	+/+	+/+	-/-
Lowrobot	-/-	+/+	-/-	+/-	+/-	<u>+/+</u>	-/-	-/-	-/-	-/+
Highrobot	+/-	+/+	-/-	+/-	+/-	<u>+/+</u>	-/-	-/-	+/-	-/-
handwhisper	<u>+/+</u>	-/-	+/+	+/+	+/+	-/-	+/+	+/+	-/-	+/-
hand	+/-	+/+	-/-	+/-	+/-	+/+	-/-	-/-	-/-	-/+
nose	<u>+/+</u>	-/-	<u>+/+</u>	-/-	-/-	<u>+/+</u>	+/+	+/+	+/+	-/+
cheek	+/+	<u>+/+</u>	-/-	+/+	+/+	<u>+/+</u>	+/-	<u>+/+</u>	+/+	-/+
penfront	+/-	<u>+/+</u>	-/+	-/-	-/-	<u>+/+</u>	-/-	-/-	-/-	-/+
penback	+/-	<u>+/+</u>	-/+	-/-	+/-	+/+	-/-	-/+	+/-	-/-
towel	+/+	+/+	-/+	-/+	-/+	-/-	<u>+/+</u>	+/+	-/-	-/+
cup	+/+	-/-	+/+	+/+	+/+	<u>+/+</u>	+/+	<u>+/+</u>	+/+	-/+
Fan	+/-	-/-	<u>+/+</u>	+/-	+/-	<u>+/+</u>	+/+	<u>+/+</u>	+/+	+/+
fanlowrobot	+/-	-/-	<u>+/+</u>	+/-	+/-	<u>+/+</u>	+/+	<u>+/+</u>	+/+	-/-
fanhighrobot	+/-	-/-	<u>+/+</u>	-/-	+/-	+/+	<u>+/+</u>	<u>+/+</u>	+/+	-/-
fanstagewhisper	+/-	-/-	+/+	-/-	+/-	+/+	+/+	<u>+/+</u>	+/-	-/+

Table 7: The twenty voice adaptations used by the vocal adversary in this project with the WER (DeepSpeech2 transcription compared to the prompt) and a short description of each adaptation. Split into the three groups of adaptations.

tot	WER		Voice adaptation	description
	F	M		
<b>Only voice</b>				
33.055	34.69	31.42	default	Talking normally
33.975	35.71	32.24	mediumroomreverb	Talking normally. artificially added room reverb (with Audacity. on the standard 'medium room' setting)
77.55	66.53	88.57	whisper	Whispering
77.955	72.24	83.67	whisperplus25db	Whispering and artificially added 25dB with Audacity. vaguely mimicking a megaphone
56.62	38.55	74.69	stagewhisper	Loudly whispering
44.9	38.78	51.02	overlyhappy	Talking with a high pitch while smiling broadly. resembling how you would speak to a kittycat
31.43	24.9	37.96	lowrobot	Talking like a robot in a low pitch
34.08	26.53	41.63	highrobot	Talking like a robot in a high pitch
<b>Manual</b>				
96.935	96.73	97.14	handwhisper	Whispering while covering your mouth with your hand
62.855	73.06	52.65	hand	Talking normally while covering your mouth with your hand
54.695	57.55	51.84	nose	Talking nasally while pinching your nose with your hand
52.005	42.97	61.04	cheek	Talking normally while pulling both cheeks with your hands
<b>With an object</b>				
69.595	71.43	67.76	penfront	Talking normally with a pen between your teeth at the front of your mouth
92.45	93.47	91.43	penback	Talking normally with a pen between your teeth at the back of your mouth. between your molars
40	33.47	46.53	towel	Talking normally while the microphone is covered with a towel
73.265	64.49	82.04	cup	Talking normally while covering your mouth with a porcelain cup
47.55	46.12	48.98	fan	Talking normally into a fan that is running on the lowest setting between the mouth of the speaker and the microphone
42.855	29.8	55.91	fanlowrobot	Talking like a robot in a low pitch while talking into a fan that is running on the lowest setting between the mouth of the speaker and the microphone
40	33.06	46.94	fanhighrobot	Talking like a robot in a high pitch while talking into a fan that is running on the lowest setting between the mouth of the speaker and the microphone
70.61	46.12	95.1	fanstagewhisper	Loudly whispering into a fan that is running on the lowest setting between the mouth of the speaker and the microphone

Table 8: The top ten speech features for all vocal adaptations in distinguishing the adaptation and the default speech of the vocal adversary obtained with SVM-RFE.

Voice adaptation	Top ten features, from 1 to 10
mediumroomreverb	nhr_mean, jitter_local, shimmer_apq3, intensity_mean, autocor_mean, fracUnvoiced, degreeVoicebreaks, shimmer_local, hnr_mean, pitch_min
whisper	pitch_mean, period_mean, shimmer_apq3, shimmer_local, period_std, jitter_ppq5, intensity_max, shimmer_apq5, jitter_rap, intensity_mean
whisperplus25db	pitch_mean, grav_center, pitch_median, autocor_mean, period_std, hnr_mean, intensity_mean, period_mean, pitch_min, nhr_mean
stagewhisper	pitch_median, period_mean, pitch_mean, shimmer_apq3, intensity_mean, intensity_max, nrPeriods, fracUnvoiced, nrPulses, shimmer_apq11
overlyhappy	period_mean, pitch_mean, jitter_local_absolute, intensity_mean, intensity_max, period_std, pitch_var, pitch_min, intensity_std, nrPeriods
lowrobot	pitch_median, nrPeriods, intensity_mean, nrPulses, pitch_mean, pitch_min, shimmer_local_dB, jitter_rap, f3, intensity_max
highrobot	intensity_max, period_std, shimmer_apq11, shimmer_apq5, jitter_local_absolute, jitter_rap, jitter_ppq5, intensity_mean, shimmer_local_dB, pitch_mean
handwhisper	period_mean, intensity_mean, intensity_max, grav_center, pitch_mean, autocor_mean, shimmer_local, pitch_median, shimmer_local_dB, jitter_ppq5
hand	nhr_mean, shimmer_apq3, jitter_local_absolute, shimmer_local, shimmer_local_dB, degreeVoicebreaks, nrVoicebreaks, pitch_median, period_mean, period_std
nose	grav_center, hnr_mean, pitch_median, intensity_max, intensity_mean, f2, pitch_mean, shimmer_local_dB, period_std, pitch_max
cheek	intensity_max, period_mean, autocor_mean, jitter_rap, period_std, degreeVoicebreaks, intensity_mean, nrPulses, intensity_std, shimmer_apq3
penfront	shimmer_apq5, shimmer_apq3, shimmer_local_dB, period_std, intensity_max, pitch_max, pitch_var, pitch_mean, intensity_mean, fracUnvoiced
penback	jitter_ppq5, shimmer_local, shimmer_apq5, hnr_mean, nrPeriods, nrVoicebreaks, shimmer_apq3, fracUnvoiced, jitter_local_absolute, autocor_mean
towel	intensity_mean, pitch_median, fracUnvoiced, shimmer_apq11, intensity_min, intensity_max, f2, shimmer_apq3, jitter_rap, jitter_local_absolute
cup	shimmer_local_dB, hnr_mean, shimmer_local, jitter_rap, grav_center, pitch_median, intensity_mean, shimmer_apq3, pitch_std, pitch_var
fan	autocor_mean, hnr_mean, nhr_mean, intensity_mean, shimmer_apq3, shimmer_apq11, degreeVoicebreaks, nrPulses, grav_center, intensity_max
fanlowrobot	hnr_mean, jitter_local_absolute, autocor_mean, intensity_mean, period_std, nhr_mean, shimmer_apq3, jitter_local, pitch_median, shimmer_apq11
fanhighrobot	shimmer_apq3, period_std, autocor_mean, jitter_local, nhr_mean, pitch_median, intensity_max, hnr_mean, shimmer_apq11, shimmer_local_dB
fanstagewhisper	shimmer_apq5, period_std, shimmer_apq3, pitch_median, jitter_ppq5, intensity_mean, hnr_mean, pitch_mean, jitter_rap, pitch_min

#### 4.2.2 Gender classification results

The twenty voice adaptations used by the vocal adversary in an attempt to obfuscate gender from their voice are tested with neural and non-neural methods.

Interestingly, some of the voice adaptations are successful in obfuscating gender for a neural inference model. For M5 shown in Table 9, many adaptations reduce the accuracy of the inference model. Overall, the Male vocal adversary seems to obfuscate their gender less than the Female vocal adversary.

For WavLM shown in Table 10, the vocal adversaries seem less successful in obfuscating their genders. Here, especially the Female vocal adversary does not reduce their gender inference accuracy in many voice adaptations.

For X-Vector shown in Table 11 some of the adaptations are successful in obfuscating gender while others are not.

In general, the models trained with VoxCeleb seem to obtain the highest inference accuracies against vocal adversaries, indicating that a ‘noisy’ training data set could make the model more robust against vocal adversaries.

The vocal adversaries were also tested with an SVM using the 35 extracted speech features. Results for the SVM inference can be seen in Table 12. Differently from the neural adversaries, some voice adaptations from the vocal adversaries are successful in obfuscating gender even in the linear SVM case. The Female vocal adversary seems to be less successful against SVMs, while the Male vocal adversary is successful with some adaptations.

Table 9: Gender classification accuracy results for the vocal adversary with the M5 models.

Speech adaptation	Inference model								
	M5-LS100h			M5-LS960h			M5-vox		
	tot	F	M	tot	F	M	tot	F	M
default	100	100	100	95.83	91.67	100	62.5	25	100
whisper	50	95.83	4.17	50	0	100	50	0	100
lowrobot	100	100	100	62.5	25	100	50	0	100
highrobot	100	100	100	97.92	95.83	100	70.8	41.67	100
fan	100	100	100	58.33	16.67	100	58.33	16.67	100
overlyhappy	50	100	0	52.08	87.5	16.67	62.5	33.33	91.67
medium room reverb	100	100	100	95.74	91.3	100	59.57	17.39	100
whisper+25dB	72.92	62.5	83.33	66.67	91.67	41.67	79.17	83.33	75
stage whisper	73.47	100	45.83	100	100	100	75.51	52	100
handwhisper	64.58	95.83	33.33	50	0	100	50	0	100
hand	100	100	100	87.5	75	100	50	0	100
nose	100	100	100	95.83	91.67	100	75	50	100
cheek	100	100	100	94	88	100	54	8	100
pen front	100	100	100	85.41	70.83	100	64.17	8.33	100
pen back	100	100	100	85.41	70.83	100	54.17	8.33	100
towel	100	100	100	97.92	95.83	100	60.42	20.83	100
cup	97.92	100	95.83	95.83	100	91.67	62.5	25	100
fan lowrobot	100	100	100	56.25	12.5	100	54.17	8.33	100
fan highrobot	100	100	100	85.42	70.83	100	75	50	100
fan stagewhisper	100	100	100	75	50	100	64.58	29.17	100

Table 10: Gender classification accuracy results for the vocal adversary with the WavLM models.

Speech adaptation	Inference model								
	WavLM-LS100h			WavLM-LS960h			WavLM-vox		
	tot	F	M	tot	F	M	tot	F	M
default	100	100	100	100	100	100	100	100	100
whisper	93.75	100	87.5	93.75	100	87.5	97.92	95.83	100
lowrobot	100	100	100	100	100	100	100	100	100
highrobot	100	100	100	100	100	100	100	100	100
fan	97.92	100	95.83	100	100	100	97.92	100	95.83
overlyhappy	56.25	100	12.5	70.83	100	41.67	66.67	100	33.33
medium room reverb	100	100	100	100	100	100	100	100	100
whisper+25dB	89.58	100	79.17	75	100	50	100	100	100
stage whisper	100	100	100	89.58	100	79.17	100	100	100
handwhisper	89.58	100	79.17	87.5	91.67	83.33	52.08	4.17	100
hand	100	100	100	100	100	100	91.67	83.33	100
nose	100	100	100	100	100	100	100	100	100
cheek	100	100	100	100	100	100	100	100	100
pen front	100	100	100	100	100	100	95.83	91.67	100
pen back	100	100	100	100	100	100	89.58	79.17	100
towel	100	100	100	100	100	100	100	100	100
cup	97.92	100	95.83	100	100	100	100	100	100
fan lowrobot	100	100	100	100	100	100	100	100	100
fan highrobot	100	100	100	100	100	100	100	100	100
fan stagewhisper	50	100	0	62.5	100	25	58.33	100	16.67

Table 11: Gender classification accuracy results for the vocal adversary with the XVector models.

Speech adaptation	Inference model								
	XVector-LS100h			XVector-LS960h			XVector-vox		
	tot	F	M	tot	F	M	tot	F	M
default	100	100	100	87.5	75	100	100	100	100
whisper	54.17	100	8.33	60.42	20.83	100	100	100	100
lowrobot	100	100	100	89.58	79.17	100	95.83	91.67	100
highrobot	100	100	100	100	100	100	100	100	100
fan	52.08	4.17	100	54.17	8.33	100	100	100	100
overlyhappy	50	100	0	60.42	87.5	33.33	85.42	100	70.83
medium room reverb	100	100	100	82.98	65.22	100	100	100	100
whisper+25dB	93.75	87.5	100	95.83	91.67	100	100	100	100
stage whisper	69.39	100	37.5	91.84	88	95.83	100	100	100
handwhisper	72.92	100	45.83	50	4.17	95.83	52.08	4.17	100
hand	100	100	100	83.33	66.67	100	93.75	87.5	100
nose	100	100	100	97.92	95.83	100	100	100	100
cheek	100	100	100	92	84	100	100	100	100
pen front	100	100	100	85.42	70.83	100	91.67	83.33	100
pen back	97.92	95.83	100	87.5	75	100	81.25	62.5	100
towel	100	100	100	75	50	100	100	100	100
cup	91.67	100	83.33	79.17	58.33	100	100	100	100
fan lowrobot	58.33	16.67	100	77.08	54.17	100	100	100	100
fan highrobot	58.33	16.67	100	97.92	95.83	100	100	100	100
fan stagewhisper	58.33	25	91.67	95.83	91.67	100	87.5	100	75

Table 12: Gender classification accuracy results for the vocal adversary with the linear SVM models, trained on and using the feature ranking obtained with LibriSpeech-train (seen in Table 2).

Speech adaptation	Inference model					
	SVM-top10			SVM-full		
	tot	F	M	tot	F	M
default	100	100	100	100	100	100
whisper	50	100	0	50	100	0
lowrobot	100	100	100	100	100	100
highrobot	100	100	100	100	100	100
fan	98	100	96	98	100	96
overlyhappy	50	100	0	50	100	0
medium room reverb	100	100	100	97.5	100	95
whisper+25dB	50	100	0	50	100	0
stage whisper	50	100	0	50	100	0
handwhisper	50	100	0	50	100	0
hand	100	100	100	100	100	100
nose	100	100	100	98	100	96
cheek	96	96	96	98	100	96
pen front	100	100	100	100	100	100
pen back	98	100	96	100	100	100
towel	96	100	92	100	100	100
cup	92.5	100	85	95.5	96	95
fan lowrobot	100	100	100	100	100	100
fan highrobot	100	100	100	95.5	100	91
fan stagewhisper	60.5	100	21	73.5	100	47

## 5 Discussion

In this section, we will first discuss the results of the neural adversary, then the vocal adversary and then discuss some general choices made in this project. The limitations of this thesis will briefly be described and finally we will put some future perspectives in the future work section.

As expected, the neural adversaries are successful in obfuscating speech in white and grey-box scenarios where the model architecture is the same. However, the transferability of the neural adversaries show some interesting trends. While M5 was trained from scratch, the WavLM was pretrained using LS960h and only fine-tuned by us on different data sets. Additionally, LS100h is a subset of LS960h. It is expected that the neural adversary using LS100h or LS960h with the WavLM architecture as reference models would transfer well, following that both of those data sets are at least partially overlapping with the train set of the WavLM inference models. While the inference accuracy is quite low for data perturbed with WavLM-LS100h on all WavLM inference models, this does not hold for data perturbed with WavLM-LS960h.

The actual speech feature values and whether the change in them with a computational perturbation or voice adaptation is significant is out of the scope of this project. Furthermore, it is not that interesting whether a speech feature is significant, but rather what that means in terms of speech production and how that could be used for privacy protection. Little is known about the inner workings of a neural network, and statistical significance does not necessarily mean that a feature will be important for the classification.

The success of the vocal adversaries differed between models and voice adaptations. Additionally, the Female and Male vocal adversary obtain different results. While the conclusions that can be drawn are limited with only two vocal adversaries, it is interesting to see the differences that arise. Especially in the speech feature-based SVM, the Female vocal adversary is not very successful in obfuscating their gender. Only the ‘cheek’ or ‘cup’ voice adaptation seem to lower the inference accuracy, and then only by 4%.

Note that the training data for the SVM is balanced on gender, so it should not necessarily default to a Female classification. More research is necessary to draw any general conclusions, but it does seem like the Male vocal adversary is more successful in changing the speech feature values in their voice. It is known that untrained speakers might have trouble changing speech feature values such as fundamental frequency. While neither vocal adversary was trained, individual differences might have impacted the ability of the vocal adversary to change their voice.

The success of the vocal adversary is not diminished if only one of the speakers was successful with a certain voice adaptation. It is very likely that different voice adaptations will work for different people, and if one voice adaptation is found to successfully change a Male label to a Female label, that could already be used by a lot of SVA users to obfuscate gender in their voices. We do not expect one voice adaptation to work universally, which is why a broader range of voice adaptations need to be considered in future work.

Another interesting finding of the vocal adversary is that the voice adaptations do not necessarily reduce the utility of the ASR system. Both the low and high robot adaptation for the Female vocal adversary actually increase the utility compared to the default speech. This could be due to the monotone voice and slowly articulated words in the robot voice adaptations. It seems that the vocal adversary does not necessarily follow the same privacy-utility trade-off other methods have.

While DeepSpeech 2 generally obtains a low utility for most voice adaptations of the vocal adversary (seen in the high WER scores), it can be expected that other ASR methods perform better. Especially in SVAs, the input speech is often noisy and unclear as the user is not necessarily close to the microphone. The SVA likely has a more robust ASR method to deal with these types of problems. Furthermore, some voice adaptations like whispering could be built-in. Amazon Alexa, for example, has a whisper functionality where the user can whisper to the SVA and the ASR will have no problem analysing that speech. It is likely that most ASR systems actually used by SVAs will obtain a higher utility than the WER scores reported in this thesis.

## 5.1 Limitations

In this thesis, we kept the scope of the adversaries rather limited. This was mostly done because of the time constraints within a MA thesis. Furthermore, we based our methods on the use-case of an SVA and kept more elaborate scenarios out of scope.

Since there are only two speakers as vocal adversaries in this project, it is likely that the results are speaker-dependent rather than gender-dependent. The speakers are both in their early 20s and have the same first language, but age difference could also be a contributing factor to the differences found in their voices. Further exploration is needed to be able to draw any strong conclusions about the success of gender obfuscation in the vocal adversaries. Additionally, the Female vocal adversary does not always obtain a 100% accuracy using the default voice adaptation, suggesting that their voice might be more ambiguous than the Male vocal adversary who always obtains a 100% accuracy.

The prompt for the vocal adversary (“The patchwork girl of Oz”) had some out-of-vocabulary words which resulted in the relatively high WER even for the default recordings of the vocal adversary. A different prompt will likely result in different WER values that could be better compared with the utility of the neural adversaries, although between voice adaptations within the vocal adversaries the comparison is still valid.

Furthermore, read speech (which all our vocal adversary data consists of) is inherently different than the spontaneous speech or speech commands one would give an SVA. This difference in speech type could result in quite different results for both the neural and the vocal adversary. However, read speech has a lot of research benefits (prompt is known, articulation is relatively high, lots of data available, most models are trained on this data for these reasons as well). It is a trade-off between consistency and realistic everyday scenarios.

The 35 speech features used in this project likely do not encompass every part of variation in the speech. It could be that the computational perturbations from the neural adversary change a part of the speech that is not captured by any of the extracted speech features. If this was the case, it would explain why the SVM inference with the perturbed speech does not decrease the accuracy at all.

Furthermore, the feature extraction itself could be less reliable with the computational perturbations from the neural adversary and the voice adaptations from the vocal adversary. Both adversaries introduce some kind of noise or modification to the speech signal, which could worsen the feature extraction. For example, the pitch from Praat is only an estimation, as the actual pitch is notoriously difficult to measure from audio alone. Especially for the neural adversary, the computational perturbations introduce something that is typically outside the scope of human speech, so Praat could simply be unable to capture this in the audio.

Most of these limitations could be addressed by further research. Possibilities for future work will be explored in the next section.

## 5.2 Future work

Since this project is the first one formally defining a vocal adversary, we urge researchers to explore this concept in future work. The experiments with a vocal adversary performed as part of this thesis project are limited and much more experimentation is possible in the future.

Selecting the voice adaptations for the vocal adversary that are successful and do not decrease the ASR performance and collecting more data with these adaptations would be a logical next step. However, there might be some adaptation options that we missed. We have not explored the seemingly obvious ‘talk like another gender’ adaption, mostly because the interpretation of this instruction varies per speaker and imitating someone else (an individual or a group of individuals) could have cultural implications that speakers would rather avoid because of possible embarrassment. Both broader and more specific exploration of voice adaptations would be interesting for the future of vocal adversaries.

Furthermore, the data recorded for the vocal adversary in this thesis is very limited. Only two speakers (one Male and one Female) were recorded and findings are likely to be speaker-dependent as well as possibly gender-dependent. Also for the neural adversary, it seems that some findings are both data and model dependent. If more speakers are available, statistical testing becomes more feasible as well. There are many more data sets, models and combinations to explore.

Given the scope of this project, we only looked at gender as a privacy sensitive paralinguistic quality of speech. Furthermore, we considered the reductive binary gender (either Male or Female) rather than a more nuanced and realistic representation of gender. However, there are many other paralinguistic characteristics that would be interesting to protect. The same methods used in this thesis could be employed to explore other obfuscations in speech.

Additionally, it might be interesting to test the vocal adversary with gender obfuscations for other speech classification tasks. It has been found that gender obfuscation might also obfuscate speaker identity, and it would be interesting indeed if that were the case for our vocal adversary. The testing of the vocal adversary for other tasks than gender inference or ASR utility is out of the scope of this thesis, but inspecting i.e. MFCC features of the voice adaptations could give us even more insights into the potential uses of the vocal adversary.

The vocal adversary introduced in this thesis could potentially be mitigated by attackers. One possible mitigation against the vocal adversary could be training the gender inference on speech with voice adaptations to still obtain the correct gender label even with the adaptations. Since the inference model would be trained with adversarial examples (from the vocal adversary), this is called ‘adversarial training’. However, that is resource-intensive and likely requires a cascaded system to first classify the voice adaptation before inferring gender. Without the additional step, training with voice adaptations could decrease the inference performance on default speech as well, so either option is less than ideal for the harmful third party trying to infer gender. Adversarial training has been left out of scope for this thesis, but it would be very interesting to explore together with other possible mitigations against the vocal adversary in future work.

Most likely, the vocal adversary is also interesting for so-called ‘data poisoning’. Data poisoning is an obfuscation technique where the user makes sure their input is unusable for training. Since Amazons privacy policy states that speech input to Alexa could be used for “improving their services” with “machine learning”, and it is impossible to know which future techniques could compromise privacy even more, this might be an interesting application for the vocal adversary. Users cannot give informed consent to

future speech processing tasks, but they are currently doing this when using an SVA.

Other than the vocal and neural adversaries, it would be interesting to test other privacy protection methods with the speech features. Since we found out that our neural adversaries are fully defeated by speech-feature based models, it could be the case that other privacy protection methods also fall short. Voice conversion, for example, could be evaluated using the 35 speech features directly. This could potentially improve privacy protection methods and possibly facilitate creative voice imitations for i.e. voice actors.

Plus, the 35 speech features used in this thesis are hardly the only interpretable speech features that exist. There are many more feature sets that are easily extracted from speech data. It could be possible that computational perturbations from the neural adversary simply change something in the speech not captured by the 35 speech features. Extending the feature set does not necessarily change the methodology, as the feature ranking can still be employed to choose a suitable number of features to inspect in more detail. The RFE feature selection or ranking is not an exhaustive method, and is generally designed for bigger data sets than 35 features.

For now, we used the performance of the ASR DeepSpeech 2 as a proxy of the utility for an SVA. However, it would be interesting to directly test the vocal adversaries with an SVA. Instead of using read speech as the prompt, voice commands to an SVA could be used and evaluated with whether the SVA performs the command afterwards. While human evaluation is less interesting than automatic evaluation in our project, human intelligibility evaluation could be used as an additional measure.

Since it has been found that qualities of voice depend on the gender of the conversation partner, and SVAs are often equipped with a 'Female' voice by default, it might be interesting to look more closely into how the auditory feedback of an SVA could influence the users' own voice. If anything, it could facilitate a vocal adversary into changing their voice.

Finally, actual testing with a real-life SVA system could provide a realistic testing scenario. As briefly mentioned before, SVAs do not require physical closeness to the device. Over-the-air testing could possibly introduce noise and reverb depending on the environment, but since this is the intended use of a vocal adversary it is important to test these realistic scenarios.

## 6 Conclusion

In this thesis, we have addressed all research questions (RQs) stated in earlier sections. In this conclusion, we will refer back to the RQs using their given numbers. We have shown with multiple experiments that the current trend of solely employing neural-on-neural privacy-preserving methods are insufficient in protecting against gender inference attacks in speech. We find that neural adversaries are successful in obfuscating gender in a white-box scenario in otherwise highly successful neural gender inference models such as WavLM, and there is some transferability in grey-box scenarios with the same neural architecture (II.I). However, the privacy protection plummets when extracted speech features and an SVM are used for inference instead (II.II). The most important features changed with the computational perturbations are pitch, NHR and shimmer, answering RQ II and II.III, but there seems to be no transferability for neural adversaries to speech feature-based models.

The novel vocal adversary we propose leverages interpretable speech features useful for gender obfuscation in both speech feature-based and neural models. The voice adaptations from the vocal adversary change features like pitch, intensity and shimmer in the voice, but these speech features differ per adaptation (III.I). Experiments with the twenty voice adaptations have shown the difference in success of the vocal adversary against gender inference models, answering RQ III. Some adaptations prove to be useful in obfuscating gender for both speech feature-based and neural inference models, answering our main RQ I and III.II.

Interestingly, the utility of the vocal adversary does not consistently worsen with the adaptations (III.III), suggesting that the vocal adversary might not follow the same strict privacy-utility trade-off that neural adversaries do have. More data and experimentation with the voice adaptations are necessary to draw any strong conclusions.

Since the concept of a vocal adversary has not been seriously considered in the literature before, we urge privacy researchers to not only look at neural-to-neural methodology and include more feasible methods for every day use. Additionally, we suggest that historical speech features are not forgotten among deep neural networks, as they provide extra insights.

Current privacy protection for speech is insufficient with both legal and adversarial methods. Because of this lacking protection, there is some sense of an inescapability of privacy risks among SVA users. This privacy resignation is unneeded however, as the experiments in this thesis have shown. Easy, free, and convenient protection of gender in voice is possible with the vocal adversary and could be used in every day scenarios.

## References

- [1] J. Turow, “The voice catchers,” in *The Voice Catchers*. Yale University Press, 2021.
- [2] L. Hernández Acosta and D. Reinhardt, “A survey on privacy issues and solutions for voice-controlled digital assistants,” *Pervasive and Mobile Computing*, vol. 80, p. 101523, 2022.
- [3] B. E. Sakar, M. E. Isenkul, C. O. Sakar, A. Sertbas, F. Gurgen, S. Delil, H. Apaydin, and O. Kursun, “Collection and analysis of a parkinson speech dataset with multiple types of sound recordings,” *Journal of Biomedical and Health Informatics*, vol. 17, no. 4, pp. 828–834, 2013.
- [4] J. L. Kröger, L. Gellrich, S. Pape, S. R. Brause, and S. Ullrich, “Personal information inference from voice recordings: User awareness and privacy concerns,” *Proceedings on Privacy Enhancing Technologies*, vol. 2022, no. 1, pp. 6–27, 2022.
- [5] S. de Conca, “The enchanted house: An analysis of the interaction of intelligent personal home assistants (IPHAs) with the private sphere and its legal protection,” Ph.D. dissertation, 2021.
- [6] J. Lau, B. Zimmerman, and F. Schaub, “Alexa, are you listening? privacy perceptions, concerns and privacy-seeking behaviors with smart speakers,” *Proceedings on Human-Computer Interaction*, vol. 2, no. CSCW, pp. 1–31, 2018.
- [7] R. Aloufi, H. Haddadi, and D. Boyle, “Paralinguistic privacy protection at the edge,” *Transactions on Privacy and Security*, 2022.
- [8] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” in *International Conference of Learning Representations*, 2015.
- [9] J. Krumm, “Inference attacks on location tracks,” vol. 6, 2007, pp. 127–143.
- [10] A. Nautsch, A. Jiménez, A. Treiber, J. Kolberg, C. Jasserand, E. Kindt, H. Delgado, M. Todisco, M. A. Hmani, A. Mtibaa *et al.*, “Preserving privacy in speaker and speech characterisation,” *Computer Speech & Language*, vol. 58, pp. 441–480, 2019.
- [11] Y. Gong and C. Poellabauer, “Crafting adversarial examples for speech paralinguistics applications,” *DYNAMICS*, 2018.
- [12] N. Carlini and D. Wagner, “Audio adversarial examples: Targeted attacks on speech-to-text,” in *security and privacy workshops*. IEEE, 2018, pp. 1–7.
- [13] F. Z. Borgesius, “Informed consent: We can do better to defend privacy,” *Security & Privacy*, vol. 13, no. 2, pp. 103–107, 2015.
- [14] D. Stoidis and A. Cavallaro, “Generating gender-ambiguous voices for privacy-preserving speech recognition,” *Proceedings of Interspeech*, pp. 4237–4241, 2022.
- [15] P. Wu, P. P. Liang, J. Shi, R. Salakhutdinov, S. Watanabe, and L.-P. Morency, “Understanding the tradeoffs in client-side privacy for downstream speech tasks,” in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*. IEEE, 2021, pp. 841–848.
- [16] S. Ahmed, Y. Wani, A. S. Shamsabadi, M. Yaghini, I. Shumailov, N. Papernot, and K. Fawaz, “Pipe overflow: Smashing voice authentication for fun and profit,” *arXiv preprint arXiv:2202.02751*, 2022.
- [17] J. Butler and G. Trouble, “Feminism and the subversion of identity,” *Gender trouble*, vol. 3, no. 1, 1990.
- [18] H. Harb and L. Chen, “Voice-based gender identification in multimedia applications,” *Journal of intelligent information systems*, vol. 24, no. 2, pp. 179–198, 2005.
- [19] K. Wu and D. G. Childers, “Gender recognition from speech. part i: Coarse analysis,” *The journal of the Acoustical society of America*, vol. 90, no. 4, pp. 1828–1840, 1991.

- [20] M. Biemans, “The effect of biological gender (sex) and social gender (gender identity) on three pitch measures,” *Linguistics in the Netherlands*, vol. 15, no. 1, pp. 41–52, 1998.
- [21] S. Bettany, S. Dobscha, L. O’Malley, and A. Prothero, “Moving beyond binary opposition: Exploring the tapestry of gender in consumer research and marketing,” *Marketing Theory*, vol. 10, no. 1, pp. 3–28, 2010.
- [22] B. N. Larson, “Gender as a variable in natural-language processing: Ethical considerations,” in *Proceedings of the First Workshop on Ethics in Natural Language Processing*, ACL. Association for Computational Linguistics, 2017, pp. 1–11.
- [23] F. Brunton and H. Nissenbaum, *Obfuscation: A user’s guide for privacy and protest*. Mit Press, 2015.
- [24] N. Tomashenko, B. M. L. Srivastava, X. Wang, E. Vincent, A. Nautsch, J. Yamagishi, N. Evans, J. Patino, J.-F. Bonastre, P.-G. Noé *et al.*, “Introducing the voiceprivacy initiative,” in *Proceedings of Interspeech*, 2020, pp. 1693–1697.
- [25] M. Lebourdais, M. Tahon, A. Laurent, and S. Meignier, “Overlapped speech and gender detection with wavlm pre-trained features,” *Proceedings of Interspeech*, 2022.
- [26] T.-W. Sun, “End-to-end speech emotion recognition with gender information,” *IEEE Access*, vol. 8, pp. 152 423–152 438, 2020.
- [27] A. Nediyanath, P. Paramasivam, and P. Yenigalla, “Multi-head attention for speech emotion recognition with auxiliary learning of gender recognition,” in *International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2020, pp. 7179–7183.
- [28] A. A. Alnuaim, M. Zakariah, C. Shashidhar, W. A. Hatamleh, H. Tarazi, P. K. Shukla, and R. Ratna, “Speaker gender recognition based on deep neural networks and resnet50,” *Wireless Communications and Mobile Computing*, vol. 2022, 2022.
- [29] A. M. Nair and S. P. Savithri, “Classification of pitch and gender of speakers for forensic speaker recognition from disguised voices using novel features learned by deep convolutional neural networks.” *Traitement du Signal*, vol. 38, no. 1, 2021.
- [30] D. Doukhan, J. Carrive, F. Vallet, A. Larcher, and S. Meignier, “An open-source speaker gender detection framework for monitoring gender equality,” in *International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2018, pp. 5214–5218.
- [31] T. Bocklet, A. Maier, J. G. Bauer, F. Burkhardt, and E. Noth, “Age and gender recognition for telephone applications based on gmm supervectors and support vector machines,” in *International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008, pp. 1605–1608.
- [32] M. M. Nasef, A. M. Sauber, and M. M. Nabil, “Voice gender recognition under unconstrained environments using self-attention,” *Applied Acoustics*, vol. 175, p. 107823, 2021.
- [33] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, “The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing,” *transactions on affective computing*, vol. 7, no. 2, pp. 190–202, 2015.
- [34] B. Schuller, F. Weninger, Y. Zhang, F. Ringeval, A. Batliner, S. Steidl, F. Eyben, E. Marchi, A. Vinciarelli, K. Scherer *et al.*, “Affective and behavioural computing: Lessons learnt from the first computational paralinguistics challenge,” *Computer Speech & Language*, vol. 53, pp. 156–180, 2019.
- [35] F. Eyben, M. Wöllmer, and B. Schuller, “Opensmile: the munich versatile and fast open-source audio feature extractor,” in *Proceedings of the 18th international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

- [36] P. Boersma and D. Weenik, “Praat: doing phonetics by computer (version 6.2.14),” <http://www.praat.org>, 2020.
- [37] J. P. Teixeira, C. Oliveira, and C. Lopes, “Vocal acoustic analysis–jitter, shimmer and hnr parameters,” *Procedia Technology*, vol. 9, pp. 1112–1122, 2013.
- [38] F. Lastow, E. Ekberg, and P. Nugues, “Language-agnostic age and gender classification of voice using self-supervised pre-training,” in *Swedish Artificial Intelligence Society Workshop*. IEEE, 2022, pp. 1–9.
- [39] W. Dai, C. Dai, S. Qu, J. Li, and S. Das, “Very deep convolutional neural networks for raw waveforms,” in *International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2017, pp. 421–425.
- [40] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, “Wavlm: Large-scale self-supervised pre-training for full stack speech processing,” *Journal of Selected Topics in Signal Processing*, 2022.
- [41] S. H. Kabil, H. Muckenhirn, and M. Magimai-Doss, “On learning to identify genders from raw speech signal using cnns.” in *Proceedings of Interspeech*, 2018, pp. 287–291.
- [42] C. Lehmann, T. Stadelmann, and Z. Datalab, “Real-world speaker recognition on voxceleb2 using angular margin losses,” 2020.
- [43] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, “Intriguing properties of neural networks,” *International Conference of Learning Representations*, 2014.
- [44] Y. Qin, N. Carlini, G. Cottrell, I. Goodfellow, and C. Raffel, “Imperceptible, robust, and targeted adversarial examples for automatic speech recognition,” in *International conference on machine learning*. PMLR, 2019, pp. 5231–5240.
- [45] H. Abdullah, M. S. Rahman, W. Garcia, K. Warren, A. S. Yadav, T. Shrimpton, and P. Traynor, “Hear” no evil”, see” kenansville”\*: Efficient and transferable black-box attacks on speech recognition and voice identification systems,” in *Symposium on Security and Privacy*. IEEE, 2021, pp. 712–729.
- [46] M. Alzantot, B. Balaji, and M. Srivastava, “Did you hear that? adversarial examples against automatic speech recognition,” *arXiv preprint arXiv:1801.00554*, 2018.
- [47] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, “Universal adversarial perturbations,” in *Proceedings of the conference on computer vision and pattern recognition*. IEEE, 2017, pp. 1765–1773.
- [48] G. Chen, S. Chenb, L. Fan, X. Du, Z. Zhao, F. Song, and Y. Liu, “Who is real bob? adversarial attacks on speaker recognition systems,” in *Symposium on Security and Privacy*. IEEE, 2021, pp. 694–711.
- [49] D. Stoidis and A. Cavallaro, “Protecting gender and identity with disentangled speech representations,” *Proceedings of Interspeech*, 2021.
- [50] P. Perrot, G. Aversano, and G. Chollet, “Voice disguise and automatic detection: review and perspectives,” *Progress in nonlinear speech processing*, pp. 101–117, 2007.
- [51] C. Zhang and T. Tan, “Voice disguise and automatic speaker recognition,” *Forensic science international*, vol. 175, no. 2-3, pp. 118–122, 2008.
- [52] M. Farrús, “Voice disguise in automatic speaker recognition,” *Computing Surveys*, vol. 51, no. 4, pp. 1–22, 2018.
- [53] L. Zheng, J. Li, M. Sun, X. Zhang, and T. F. Zheng, “When automatic voice disguise meets automatic speaker verification,” *Transactions on Information Forensics and Security*, vol. 16, pp. 824–837, 2020.

- [54] F. Haider, P. Albert, and S. Luz, “User identity protection in automatic emotion recognition through disguised speech,” *Artificial Intelligence*, vol. 2, no. 4, pp. 636–649, 2021.
- [55] R. G. Hautamäki, M. Sahidullah, V. Hautamäki, and T. Kinnunen, “Acoustical and perceptual study of voice disguise by age modification in speaker verification,” *Speech Communication*, vol. 95, pp. 1–15, 2017.
- [56] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, “Cyclegan-vc2: Improved cyclegan-based non-parallel voice conversion,” in *International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2019, pp. 6820–6824.
- [57] V. C. Tartter, “What’s in a whisper?” *The Journal of the Acoustical Society of America*, vol. 86, no. 5, pp. 1678–1683, 1989.
- [58] K. Sebastian and L. Mary, “Fasr: Effect of voice disguise,” in *International Conference on Emerging Technological Trends*. IEEE, 2016, pp. 1–4.
- [59] A. Easwara Moorthy and K.-P. L. Vu, “Privacy concerns for use of voice activated personal assistant in the public space,” *International Journal of Human-Computer Interaction*, vol. 31, no. 4, pp. 307–335, 2015.
- [60] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: an asr corpus based on public domain audio books,” in *International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2015, pp. 5206–5210.
- [61] J. S. Chung, A. Nagrani, and A. Zisserman, “Voxceleb2: Deep speaker recognition,” *Proceedings of Interspeech*, pp. 1086–1090, 2018.
- [62] W. Xie, A. Nagrani, J. S. Chung, and A. Zisserman, “Utterance-level aggregation for speaker recognition in the wild,” in *International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2019, pp. 5791–5795.
- [63] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, “IEMOCAP: Interactive emotional dyadic motion capture database,” *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.
- [64] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, “X-vectors: Robust dnn embeddings for speaker recognition,” in *International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2018, pp. 5329–5333.
- [65] Y.-Y. Yang, M. Hira, Z. Ni, A. Astafurov, C. Chen, C. Puhersch, D. Pollack, D. Genzel, D. Greenberg, E. Z. Yang *et al.*, “Torchaudio: Building blocks for audio and speech processing,” in *International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2022, pp. 6982–6986.
- [66] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards deep learning models resistant to adversarial attacks,” in *International Conference of Learning Representations*, 2018.
- [67] O. Kramer, “Scikit-learn,” in *Machine learning for evolution strategies*, 2016, pp. 45–53.
- [68] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, “Gene selection for cancer classification using support vector machines,” *Machine learning*, vol. 46, no. 1, pp. 389–422, 2002.
- [69] L. van Bemmelen, W. Harmsen, C. Cucchiari, and H. Strik, “Automatic selection of the most characterizing features for detecting COPD in speech,” in *International Conference on Speech and Computer*. Springer, 2021, pp. 737–748.
- [70] L. van Bemmelen, C. Cucchiari, and H. Strik, “Using feature selection to evaluate pathological speech after training with a serious game,” *ExLing 2021*, p. 245, 2021.
- [71] L. F. Baum, *Oz, the Complete Collection, Volume 3: The Patchwork Girl of Oz; Tik-Tok of Oz; The Scarecrow of Oz*. Simon and Schuster, 1913, vol. 3.

- [72] T. Audacity, “Audacity (version 3.1.3.0),” *The Name Audacity (R) Is a Registered Trademark of Dominic Mazzoni Retrieved from <http://audacity.sourceforge.net>*, 2017.
- [73] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen *et al.*, “Deep speech 2: End-to-end speech recognition in english and mandarin,” in *International conference on machine learning*. PMLR, 2016, pp. 173–182.
- [74] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Enrique Yalta Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, “ESPnet: End-to-end speech processing toolkit,” in *Proceedings of Interspeech*, 2018, pp. 2207–2211.

## A Full results for neural adversaries with PGD set to 10 epochs

Table 13: Gender classification accuracy results with linear SVMs trained on and using the feature ranking obtained with LibriSpeech-train for neural adversaries with PGD set to 10 epochs.

Reference model	Inference model								
	SVM top 8			SVM top 10			SVM full		
	tot	F	M	tot	F	M	tot	F	M
<b>Original LStest</b>	89.34	95.9	81.44	94.22	95.9	92.19	92.29	94.67	89.42
M5-LS100h	91.71	96.26	86.52	91.51	95.96	86.43	88.48	96.19	79.69
M5-LS960h	91.81	96.2	86.75	91.57	95.98	86.49	88.38	93.67	82.27
M5-vox	91.69	95.83	86.92	91.57	95.6	86.92	88.7	93	83.73
WavLM-LS100h	89.68	95.75	82.84	89.83	95.53	83.42	86.42	90.26	82.1
WavLM-LS960h	91.7	95.91	87.2	91.35	95.62	87.2	87.93	92.99	82.46
WavLM-vox	91.27	96.13	86.34	91.23	95.69	86.76	87.93	93.49	82.18
<b>Original Voxtest</b>	85.17	91.59	82.4	87.59	91.27	86.03	79.7	95.86	72.88
M5-LS100h	85.14	91.51	82.46	87.54	91.26	85.98	79.82	95.69	73.15
M5-LS960h	85,16	91,5	82,49	87,58	91,22	86,05	80,8	94,73	74,93
M5-vox	85,2	91,6	82,5	87,62	91,29	86,07	81,82	94,55	76,46
WavLM-LS100h	85.48	91.57	82.91	87.58	91.09	86.1	82.7	94.43	77.75
WavLM-LS960h	85.15	91.51	82.46	87.22	91.23	85.53	81.76	94.79	76.27
WavLM-vox	85.47	91.48	82.93	87.78	91.3	86.3	81.66	94.84	76.1

Table 14: Relative WER scores using DeepSpeech 2 with regard to the DeepSpeech 2 transcription of the original data for different neural adversaries using PGD set with 10 epochs for both LibriSpeech test and Voxceleb test.

Reference model	rel. WER		
	tot	F	M
<b>Original LStest</b>			
M5-LS100h	1.99	2.32	1.6
M5-LS960h	1.82	2.11	1.47
M5-vox	1.68	1.89	1.45
WavLM-LS100h	7	7.9	5.98
WavLM-LS960h	5.89	6.39	5.31
WavLM-vox	4	4.6	3.3
<b>Original Voxtest</b>			
M5-LS100h	12.77	10.81	13.59
M5-LS960h	11.7	9.67	12.55
M5-vox	11.9	9.47	12.92
WavLM-LS100h	31.46	27.48	33.12
WavLM-LS960h	27.15	22.60	29.06
WavLM-vox	22.1	18.86	23.46

Table 15: Gender classification accuracy results for neural adversaries with PGD set to 10 epochs.

Reference model	Inference model								
	M5-LS100h			M5-LS960h			M5-vox		
	tot	F	M	tot	F	M	tot	F	M
<b>Original LStest</b>	96.56	96.69	96.43	96.83	96.24	97.32	92.98	87.09	99.25
M5-LS100h	<u>52.35</u>	<u>35.55</u>	<u>71.4</u>	90.13	86.14	94.65	88.56	80.31	97.91
M5-LS960h	91.81	90.27	93.56	<u>69.12</u>	<u>57.3</u>	<u>82.53</u>	86.83	77.36	97.58
M5-vox	95.14	94.54	95.82	94.12	92.18	96.32	<u>54.08</u>	<u>34.66</u>	<u>76.09</u>
WavLM-LS100h	96.54	96.01	97.13	96.39	95.43	97.46	94.58	90.72	98.94
WavLM-LS960h	96.62	96.67	96.56	96.62	95.87	97.46	94.08	89.78	98.94
WavLM-vox	96.62	96.74	96.48	96.77	96.16	97.46	93.5	88.49	99.18
<b>Original Voxtest</b>	92.52	95.49	91.31	95.47	93.81	96.14	97.11	94.58	98.14
M5-LS100h	<u>46.69</u>	<u>31.67</u>	<u>52.68</u>	88.26	82.91	90.44	95.53	91.9	97.01
M5-LS960h	84.04	87.71	82.54	<u>64.39</u>	<u>46.44</u>	<u>71.72</u>	94.58	90.2	96.37
M5-vox	89.81	93.68	88.24	92.89	90.19	93.99	<u>55.22</u>	<u>48.58</u>	<u>57.94</u>
WavLM-LS100h	92.76	95	91.85	95.37	93.37	96.19	96.98	94.67	97.93
WavLM-LS960h	92.79	95.45	91.7	95.47	93.58	96.25	95.47	93.57	96.25
WavLM-vox	92.58	95.44	91.42	95.46	93.71	96.18	97.1	94.56	98.14
	WavLM-LS100h			WavLM-LS960h			WavLM-vox		
<b>Original LStest</b>	97.81	98.6	96.91	98.71	98.53	98.91	98.35	98.01	98.75
M5-LS100h	97.65	98.53	96.66	98.75	98.75	98.75	98.39	98.38	98.41
M5-LS960h	97.69	98.3	96.99	98.71	98.82	98.58	98.55	98.6	98.49
M5-vox	98.35	98.3	98.41	98.75	98.75	98.75	98.39	98.45	98.33
WavLM-LS100h	<u>1.65</u>	<u>0.22</u>	<u>3.26</u>	16.07	5.38	28.18	20.61	19.84	21.49
WavLM-LS960h	44.47	58.92	28.09	<u>0.94</u>	<u>0.52</u>	<u>1.42</u>	41.78	52.29	29.85
WavLM-vox	77.39	75	80.1	62.77	45.35	82.53	<u>0.39</u>	<u>0.52</u>	<u>0.25</u>
<b>Original Voxtest</b>	97.73	97.47	97.85	97	97.62	96.75	99.02	98.2	99.36
M5-LS100h	97.31	97.96	97.04	96.57	97.93	96	98.98	98.16	99.33
M5-LS960h	97.31	97.93	97.05	96.6	97.94	96.04	98.97	98.17	99.31
M5-vox	97.44	97.68	97.34	96.84	97.76	96.45	98.97	98.12	99.33
WavLM-LS100h	<u>3</u>	<u>5.81</u>	<u>1.82</u>	26.68	26.56	26.73	34.96	42.16	31.94
WavLM-LS960h	47.09	81.38	32.73	<u>3.93</u>	<u>5.46</u>	<u>3.28</u>	57.64	76.63	49.69
WavLM-vox	85.07	86.88	84.31	81.08	73.77	84.14	<u>6.25</u>	<u>9.43</u>	<u>4.91</u>
	XVector-LS100h			XVector-LS960h			XVector-vox		
<b>Original LStest</b>	96.68	96.18	97.24	97.79	97.26	98.38	97.56	96.26	99.03
M5-LS100h	96.59	95.65	97.66	97.88	97.57	98.24	97.45	96.24	98.83
M5-LS960h	96.39	95.5	97.41	97.81	97.42	98.24	97.49	96.24	98.91
M5-vox	96.55	95.5	97.74	97.88	97.57	98.24	97.57	96.46	98.83
WavLM-LS100h	87.89	82.61	93.86	96	96.52	95.42	94.93	96.3	93.37
WavLM-LS960h	92.7	89.06	96.81	96.46	96.88	95.99	95.73	97.46	93.78
WavLM-vox	93.58	90.22	97.38	97.46	97.39	97.55	95.08	96.45	93.54
<b>Original Voxtest</b>	89.44	96.59	86.44	97	94.57	98.03	98.45	97.69	98.77
M5-LS100h	90.57	96.18	88.22	97.01	94.3	98.15	98.45	97.55	98.83
M5-LS960h	90.72	96.1	88.46	96.96	94.41	98.03	98.45	97.53	98.84
M5-vox	91.05	95.93	89	96.9	94.24	98.01	98.4	97.52	98.77
WavLM-LS100h	87.26	91.76	85.38	93.4	95.01	92.72	96.08	97.13	95.64
WavLM-LS960h	90.61	94.05	89.17	94.15	95.25	93.69	96.47	97.29	96.13
WavLM-vox	92.01	94.05	91.15	96.26	95.33	96.65	96.21	96.84	95.94

## B Full results for neural adversaries with PGD set to 100 epochs

Table 16: Gender classification accuracy results for neural adversaries with PGD set to 100 epochs.

Reference model	Inference model								
	M5-LS100h			M5-LS960h			M5-vox		
	tot	F	M	tot	F	M	tot	F	M
<b>Original LStest</b>	96.56	96.69	96.43	96.83	96.24	97.32	92.98	87.09	99.25
M5-LS100h	<u>0</u>	<u>0</u>	<u>0</u>	8.27	0.37	17.22	48.28	25.07	74.6
M5-LS960h	13.52	1.11	27.59	<u>0</u>	<u>0</u>	<u>0</u>	29.82	6.56	56.19
M5-vox	57.37	42.7	74	38.68	15.56	64.88	<u>0</u>	<u>0</u>	<u>0</u>
WavLM-LS100h	96.43	95.72	97.24	96.12	94.91	97.49	94.71	91	98.91
WavLM-LS960h	96.55	96.46	96.66	96.39	95.35	97.58	94.32	90.49	98.66
WavLM-vox	96.55	96.53	96.57	96.55	95.8	97.4	94	89.53	99.08
<b>Original Voxtest</b>	92.52	95.49	91.31	95.47	93.81	96.14	97.11	94.58	98.14
M5-LS100h	<u>0.2</u>	<u>0.25</u>	<u>0.18</u>	11.67	2.08	15.58	51.1	37.73	56.55
M5-LS960h	14.21	4.12	18.33	<u>0.84</u>	<u>0.06</u>	<u>1.16</u>	31.88	16.53	38.15
M5-vox	59.86	49.54	64.08	48.8	18.33	61.23	<u>0.26</u>	<u>0.22</u>	<u>0.27</u>
WavLM-LS100h	92.76	94.96	91.87	95.34	93.19	96.22	96.95	94.54	97.94
WavLM-LS960h	92.78	95.34	91.74	95.49	93.45	96.32	97.03	94.54	98.05
WavLM-vox	92.57	95.35	91.43	95.44	93.5	96.23	97.1	94.57	98.13
	<b>WavLM-LS100h</b>			<b>WavLM-LS960h</b>			<b>WavLM-vox</b>		
<b>Original LStest</b>	97.81	98.6	96.91	98.71	98.53	98.91	98.35	98.01	98.75
M5-LS100h	97.1	97.57	96.57	97.8	97	98.7	98.08	98.23	97.91
M5-LS960h	97.77	97.35	98.24	98.2	97.79	98.66	97.77	97.79	97.74
M5-vox	98.12	97.57	98.76	94.24	89.53	99.58	97.34	95.87	99
WavLM-LS100h	<u>0</u>	<u>0</u>	<u>0</u>	13.24	2.29	25.67	14.54	16.67	12.12
WavLM-LS960h	37.26	57.45	14.38	<u>0</u>	<u>0</u>	<u>0</u>	31.15	50.29	9.45
WavLM-vox	67.4	59.81	76	55.05	26.99	86.87	<u>0</u>	<u>0</u>	<u>0</u>
<b>Original Voxtest</b>	97.73	97.47	97.85	97	97.62	96.75	99.02	98.2	99.36
M5-LS100h	97.8	96.79	98.22	96.69	97.69	96.27	98.93	97.67	99.45
M5-LS960h	97.92	96.84	98.37	96.83	97.59	96.51	98.95	97.75	99.45
M5-vox	97.82	96.63	98.31	97.1	96.09	97.53	98.86	97.35	99.49
WavLM-LS100h	<u>0.14</u>	<u>0.39</u>	<u>0.04</u>	17.7	7.39	22	23.25	28.01	21.26
WavLM-LS960h	34.45	75.99	17.05	<u>0.07</u>	<u>0.08</u>	<u>0.06</u>	35.29	65.49	22.63
WavLM-vox	81.72	78.25	83.18	75.06	55.8	83.13	<u>0.15</u>	<u>0.19</u>	<u>0.13</u>
	<b>XVector-LS100h</b>			<b>XVector-LS960h</b>			<b>XVector-vox</b>		
<b>Original LStest</b>	96.68	96.18	97.24	97.79	97.26	98.38	97.56	96.26	99.03
M5-LS100h	90.99	91	90.97	96.98	97.12	96.82	96.9	95.28	98.75
M5-LS960h	92.75	91.67	93.98	97.33	97.05	97.66	97.1	95.73	98.66
M5-vox	90.99	87.91	94.48	96.79	96.24	97.41	96.28	94.69	98.08
WavLM-LS100h	87.15	83.78	90.97	96.32	96.61	95.32	94	96.83	90.8
WavLM-LS960h	92.63	90.27	95.32	96.08	96.83	95.23	95.1	97.64	92.22
WavLM-vox	92.71	89.53	96.32	97.18	97.35	96.99	95.14	96.61	93.48
<b>Original Voxtest</b>	89.44	96.59	86.44	97	94.57	98.03	98.45	97.69	98.77
M5-LS100h	83.8	94.4	79.37	95.3	89.61	97.68	98.41	96.58	99.17
M5-LS960h	87.78	94.3	85.05	95.56	91.63	97.21	98.29	96.68	98.97
M5-vox	89.24	89.96	88.93	95.3	90.77	97.2	97.52	95.45	98.39
WavLM-LS100h	83.3	92.39	79.5	93.15	94.75	92.48	95.69	97.31	95.01
WavLM-LS960h	88.64	94.79	86.07	94.09	94.82	93.79	95.86	97.32	95.24
WavLM-vox	90.55	93.11	89.47	96.15	94.86	96.69	96.14	96.68	95.91

## C Mel spectrograms for all neural adversaries

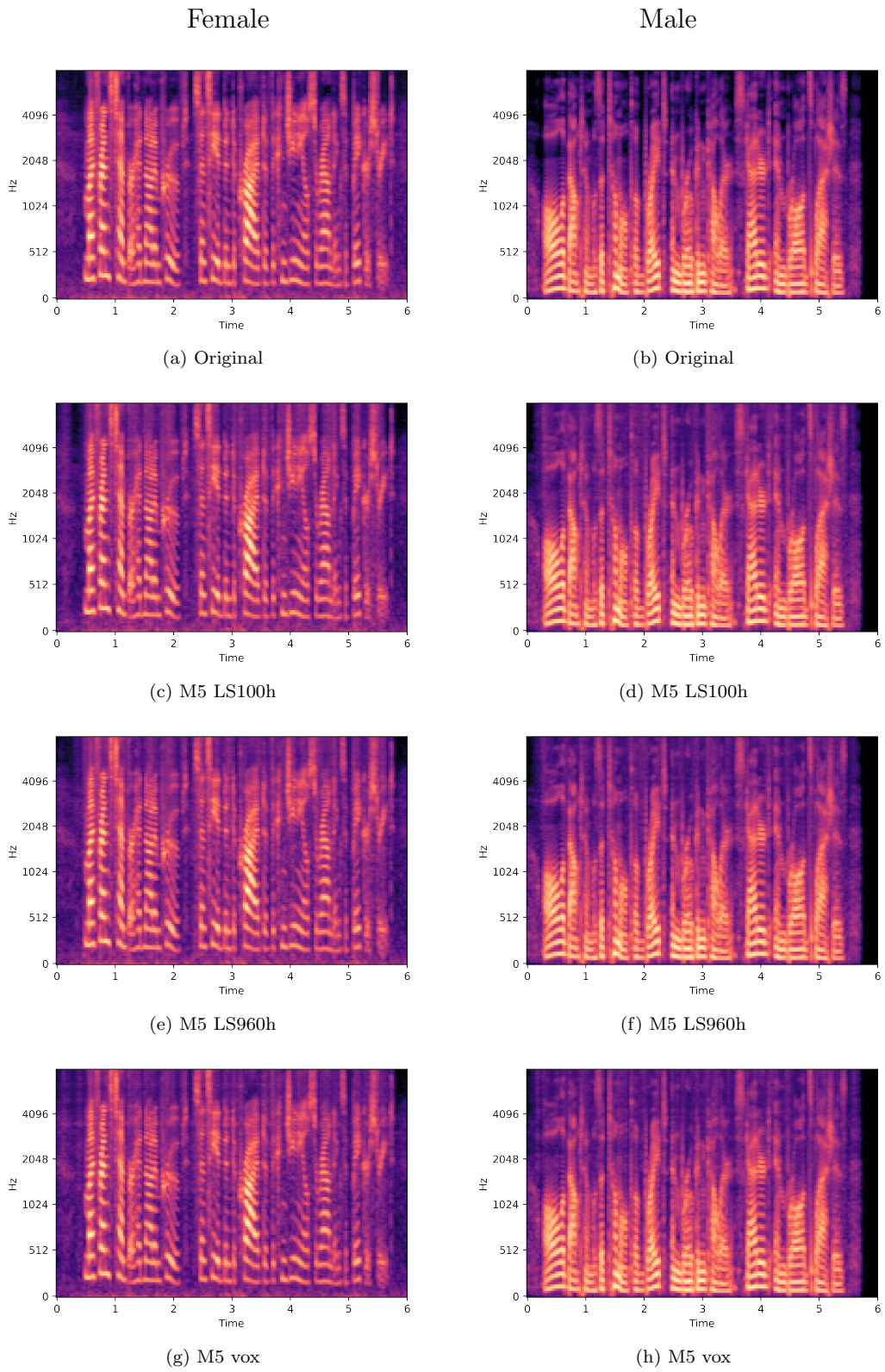


Figure 7: The Mel spectrograms for two speakers from LStest computationally perturbed with different neural adversaries using the M5 architecture (full reference model in subcaption). Inspired by Stoidis and Cavallaro [14].

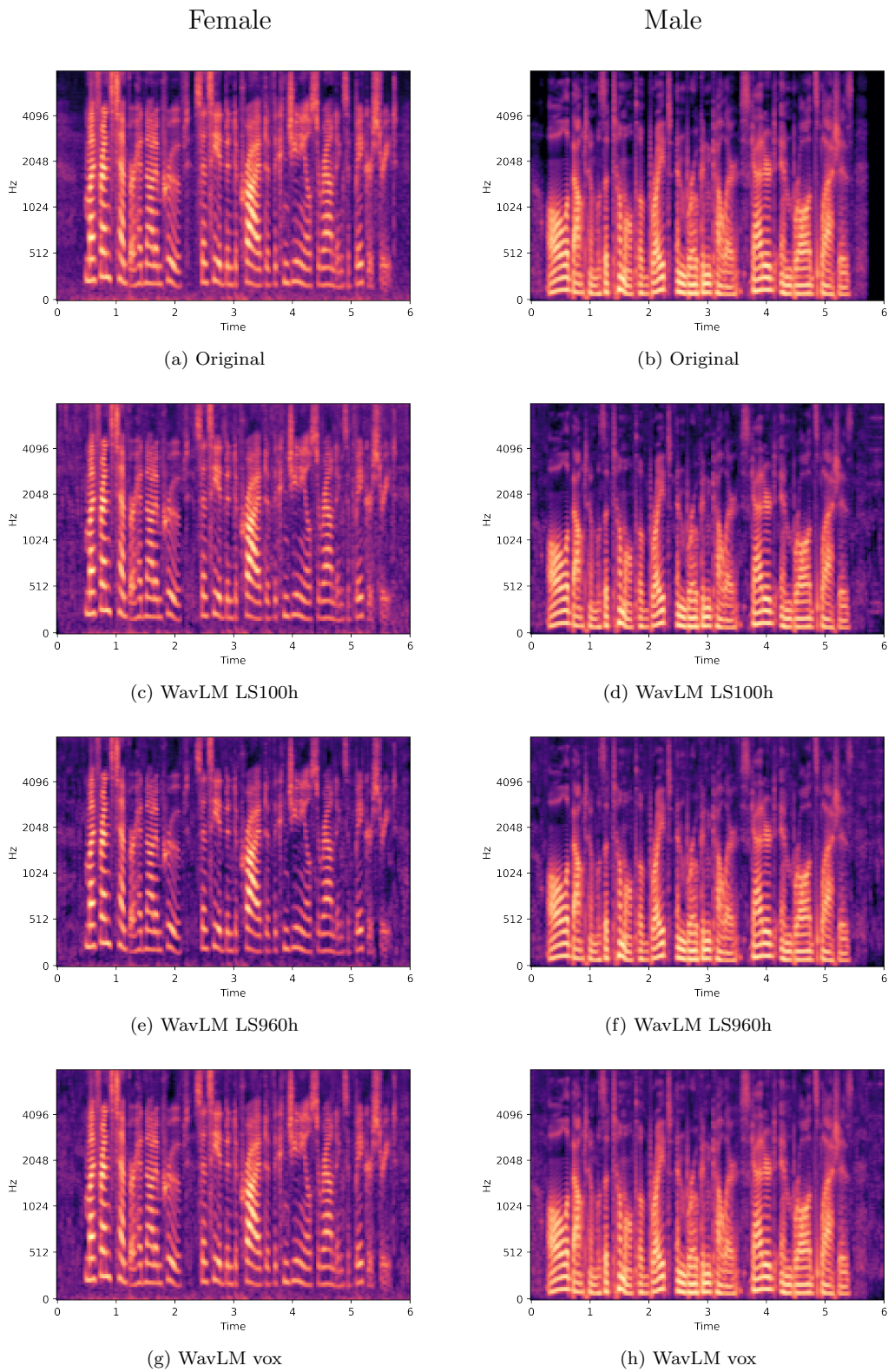


Figure 8: The Mel spectrograms for two speakers from LStest computationally perturbed with different neural adversaries using the WavLM architecture (full reference model in subcaption). Inspired by Stoidis and Cavallaro [14].