1.0 Statement of Achievements

The following details achievements reached in this thesis.

- A thorough literature review was conducted to analyse voice-controlled digital assistants, investigating and justifying the importance of addressing related problems such as attribute inference attacks. A comparison of state-of-the-art methods was also conducted, revealing a direction to improve privacy-utility trade-offs utilising signal processing and disentangled representation learning.
- A sound-based variational autoencoder (VAE) model was created after heavily modifying an image-based VAE made following a tutorial in [24]. Using this VAE, I explored the latent space of pitched and unpitched samples and used this as a reference to disentangle pitch and control the reconstruction of the voice to conceal gender identity.
- Results in Section 7 revealed we are able to conceal gender identity by reducing classification accuracy to random guess (48.38%), however utility was compromised due to limitations in our disentangled representation learning methodology.
- We stated a direction to reconcile the benefits of signal processing and disentangled representation learning, and provided a model that can be further optimised to categorically separate more features and have higher performance overall.

2.0 Abstract

Our voice is a digital footprint that is embedded with a suite of sensitive information. Aside from linguistic content, these embeddings also contain vocal characteristics such as pitch, speech characteristics such as pronunciation and other acoustic elements that can be analysed through advanced methods to allow inferences about a user's biometric identity, gender, moods and emotions. As Voice-controlled Digital Assistants (VCDA's) such as Apple's Siri, Microsoft's Cortarna and Amazon's Alexa become increasingly popular and are integrated into a variety of Internet of Things devices including home assistants and smartphones - more of our personal data is stored and the threat of attribute inference attacks becomes a pressing issue due to the potential for the attacker to infer sensitive aspects from the user's voice. In this thesis, we compare the latent space of pitched and unpitched voice samples to independently encode pitch as its own latent vector and disentangle it from other features. We then modify pitch upon voice reconstruction to fool a gender classifier, thereby utilising signal processing and disentangled representation learning to improve user privacy through protecting gender identity, whilst attempting to preserve functional utility. Experiments on the Fluent Speech Commands dataset shows we are able to reduce gender classification accuracy to random guess, however utility is compromised due to our disentangled learning process. Our work provides an interesting direction for reconciling the benefits and drawbacks of signal processing and disentangled representation learning, as well as a model that can be further optimised for better performance and concealment of other attributes.

3.0 Table of Contents

4.0 Introduction

Voice-controlled digital assistants (VCDA's) such as Amazon Alexa, Microsoft Cortana, Apple's Siri and Google Assistant have become increasingly popular over the last decade - being embedded into many of our everyday devices such as home assistants, smartphones, and even TV's [1]. These devices enhance the user experience via a hands-off approach made possible through improvements to speech-recognition technology. For example, users can make phone calls, control media playback, set calendar entries or ask general information queries with simple commands [5]. However, frequent use of these devices often leads to the personalisation of the assistants to suit individual needs, requiring the storage of personal data.

Our voice input creates a digital footprint embedded with sensitive information that can be used against us. This footprint contains not only linguistic content, but voice characteristics such as loudness and pitch, speech characteristics such as pronunciation and expression, non-speech human sounds such as laughter, and other background or acoustic sounds [1]. Voice embeddings can be analysed through advanced methods and used to uncover inferences about a user. These methods are known as attribute inference attacks, and can uncover a user's gender, personality traits, emotions, and even inferences regarding socioeconomic status, body measures and physical health [1]. Hence, it is important to improve privacy mechanisms to avoid compromising sensitive user audio data, whilst also maintaining the user experience.

Current approaches can be broadly categorised into two methods - voice transformation through signal processing and disentangled representation learning [2]. Voice transformation involves modifying the sound of speech without actually changing the speaker, this may involve altering pitch, timber, tempo or other vocal characteristics. Disentangled representation learning involving categorically separating factors in raw speech data into separate encodings. Doing so allows manipulating features independently of one another, such as gender, speaker identity or background nose, enabling a controlled reconstruction of the voice to increase user privacy [2].

In this thesis, we provide an analysis and background to VCDA's and an investigation into the problems facing them today. We explore and compare the effectiveness of speech protection mechanisms in defending attribute inference attacks, namely signal processing via pitch standardisation and disentangled representation learning. We then use a variational

autoencoder (VAE) model to compare the latent space of pitched and unpitched voice samples using the Fluent Speech Commands dataset [3] to independently encode pitch as its own latent vector and disentangle it from other features. In depth understanding of the latent representations are then used to control the reconstruction of transformed speech through pitch modification to fool a gender classifier. We aim to improve privacy-utility trade offs through use of the memory efficient benefits of signal processing and the privacy control of disentangled representation learning, enabling the protection of gender identity whilst attempting to preserve functional utility. We then validate performance based on privacy and utility metrics such as gender classification accuracy and word error rates respectively.

We find that gender classification accuracy can be reduced to random guess, however utility is compromised due to limitations of our disentanglement learning process. However, our work provides an interesting direction for reconciling signal processing with disentangled representation learning, as well as a model that can be further optimised for better performance.

5.0 Literature Review

5.1 Background

From the invention of the first phonograph in 1877 that exclusively captured and played back sound, to the IBM Shoebox digital speech recognition tool that recognised 16 words and 9 digits [4], to modern day digital assistants such as Amazon Alexa that assist us with daily tasks - humans have sought to fuse technology and voices through microphone-equipped devices that aim to improve our lives. This has culminated in smart voice-activated assistants that interpret human speech and carry out instructions. [5]

Voice-controlled digital assistants (VCDA's) can be defined as software applications that interpret instructions and carry out tasks for individuals. Since the introduction of Apple's Siri, VCDA's have increased in popularity immensely - reaching 94.4 million users in the United States in 2022, up from 47.3 million users in 2018 [6]. This surge in usage can be attributed to the improvement in natural language processing and implementation of speech recognition technology into a variety of Internet of Things devices such as smartphones, smart speakers, personal computers and tablets [5]. This advancement enables users to interact with VCDA's in a conversational manner without obstructions such as large microphones, keyboards or even touchscreens.

The tasks of modern VCDA's can be broadly summarised into the following categories.

Table 1. Uses of VCDA's

Category	**Tasks** [5]
Communication	Making phone calls, sending and reading text messages and emails
Entertainment	Controlling media playback from connected services such as Spotify, Netflix and Pandora, tell jokes and stories
Utility	Information queries, control IoT devices, set reminders and calendar entries, asking directions

Modern VCDA's are often also customisable to user preferences. For example, Google Assistant allows individuals to assign tasks to certain phrases. Simply saying "Good morning" can activate a suite of instructions that control other IoT devices as well - such as turning on the lights, coffee maker, and opening the garage door. And saying "We are leaving!" can lock the doors, set the alarm and make adjustments to the thermostat [5]. The integration of VCDAs into households has also grown immensely, with 36.6% of US adults having smart speakers in their homes, up by nearly 17% from just 4 years prior [6]. This increase has enabled companies to take advantage of the large amounts of data collected and introduce technology that allows devices to make suggestions to users. For example, Amazon Alexa introduced Hunches [7], which enables Alexa to alert the user when their devices are not in their usual state - Alexa can even automatically make changes on the user's behalf based on information about device states, user interaction and signals such as your location [7].

Amazon Alexa, Apple's Siri and Microsoft's Cortana are all VCDA's that reflect the decades of improvements in speech recognition and natural language processing technology. These devices have provided benefits to hundreds of millions of people, fundamentally changing the way we interact with computers. However, with such a large amount of audio data stored and available to companies and third party organisations - the potential misuse of personal user data is a significant privacy concern.

5.2 The Problem

Our voice is a digital footprint. Through advanced data analysis methods of recorded audio data, sensitive information can be retrieved from human speech [8]. As device manufactures and third-party developers can be in possession of this data, user privacy can be compromised [9]. In this section, we investigate the privacy implications of VCDA's, such as voice and speech analysis. We also explore privacy perceptions from the public, the effectiveness of privacy policies, and the balancing of privacy-utility tradeoffs in protection mechanisms

5.2.1 Privacy implications of VCDA's

There are many privacy challenges associated with VCDA's. Most of these stem from the collection, processing and usage of audio data. For instance, people may have their conversations recorded due to false triggers that sound similar to device wake words - 'Alexa' vs 'election'. There are hundreds of accidental trigger words that can result in privacy implications [10]. In 2018, Amazon confirmed that the audio of a private conversation between an Alexa user and her husband regarding hardwood floors was recorded and

accidentally sent to a stored contact; the spokesperson's explanation was that it was due to a string of false word interpretations [11]. Malfunction or not, the possibilities of private conversations being wrongfully collected and processed are real.

The privacy implications that we focus on in this thesis are attribute inference attacks. These are aimed at inferring sensitive attributes (e.g gender, personality traits) from user's voice recordings.. Attackers may include any party that has an interest in the user's data; from the service provider, to advertiser, to surveillance agencies. For example, a service provider may use inferred information for targeting content, and surveillance agencies may use inferred information to recognise, track, and analyse user traits and behaviours [12].

To understand why attribute inference attacks are possible, we explore the different inferences possible from recorded speech. The following is an overview of some attributes that can be inferred from speech data [8].

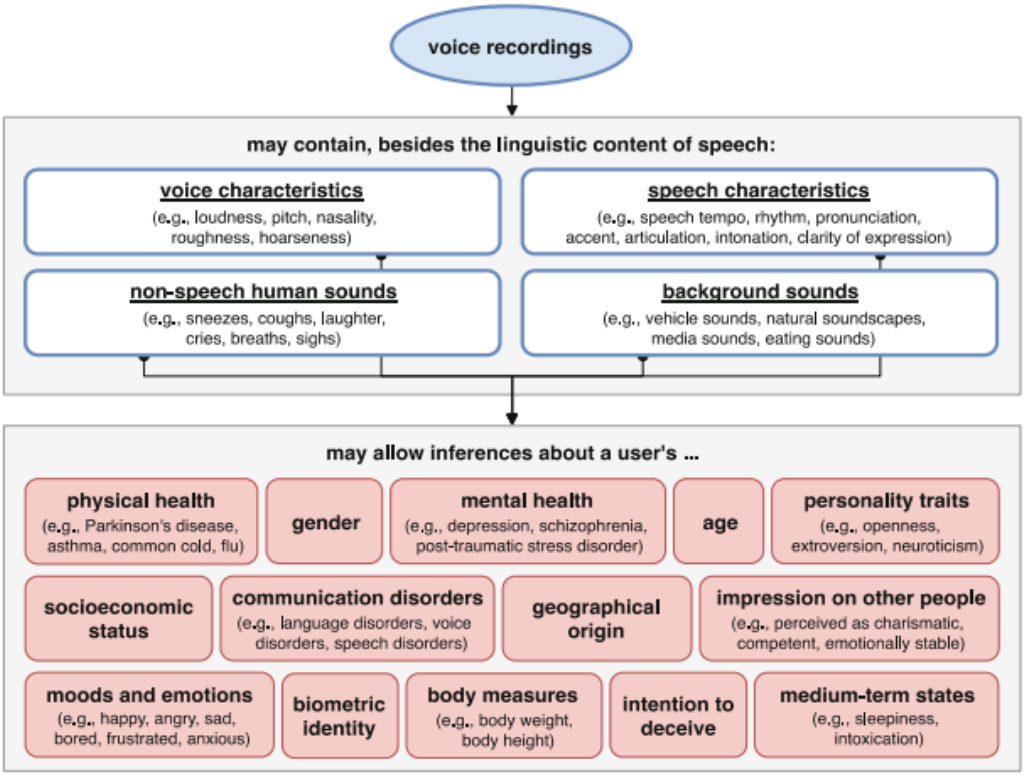


Figure 1. Inferences from speech data [8]

As shown in figure 1, without considering linguistic speech content many inferences of personal information can be disclosed based on voice/speech characteristics and the surrounding environment. For example, vocal parameters such as subharmonies, pitch, and frequency perturbation can lead to inferences regarding body measures (e.g height and

weight), sensitive information that can be used for automatic surveillance and profiling [8].

Voice variations such as speaking faster and non-speech sounds such as laughter, crying and sighing can be used to infer mood and emotion from the user's audio data. Automatic emotion analyser methods that recognise happiness, anger or sadness can even exceed human performance [8]. Mood and emotion analytics is information valued by even the largest of companies - such as Facebook, which manipulated over 600 000 user feeds to observe changes in mental states as part of a 2014 experiment. Reasons for this experiment were not made transparent, but was likely conducted to understand how to increase user screen time on the application [13]. Amazon recently patented a software that allows voice-based determination of physical and emotional characteristics - a possible use case is Amazon Alexa suggesting cough lozenges after hearing the individual cough [14]. The principle of suggesting products based on user characteristics is targeted advertising and as evidenced extends to voice analysis in VCDA's.

Personality traits can also be inferred from certain markers such as speaking rate, pitch, energy and linguistic expression. Using the "OCEAN model" (openness, conscientiousness, extroversion, agreeableness, and neuroticism) software can categorise traits into high/low categories or assign numerical scores [8]. Personality traits represent valuable information for customer profiling in many different industries. HireIQ [15] offers AI-driven virtual interview processes and according to a report on corporate surveillance, is used by many Fortune 500 companies in more than a million interviews to improve organisational outcomes [15]. HireIQ analyses and rates applicants after the interview process primarily through vocal characteristics, a method that can be discriminatory to applicants with accents, speech impediments or mental disorders [15]. Hence, inferring attributes based on audio data can be not only a violation of privacy and consent, but harmful to the user as well.

While attribute inferences through VCDA's can have some advantages such as convenience and increased productivity, the potential misuse of vast amounts of personal data warrants strong protection mechanisms in order to make the devices socially acceptable.

5.2.2 Public perception

There are multiple factors a user considers before purchasing a VCDA. These may include language performance, price and privacy. A study on acceptance relevant factors revealed privacy was the most important factor as shown in the figure below

Figure 2. Relative importance of acceptance relevant factors [16]

It was also found that a significant number of participants would protest the installation of a VCDA without their consent [16], highlighting the privacy concerns of individuals.

The privacy concerns typically associated with VCDA's are microphones that continuously listen and distrust with speaker companies. The right to privacy in one's home was a major theme brought up by participants in a study concerning privacy with smart speakers [17]. One user noted *"I think there's a capability there to listen in when you're not expecting it"* reflecting on the possibility of hackers or misuse. Other users were less concerned, stating *"Amazon, Apple and Google don't really seem to have a track record of being hacked"*. An important finding was the shared lack of understanding surrounding data collection, with one user stating *"I think it could definitely record what you're saying, but I don't think it's intelligent enough to remember."* [17]. Overall, privacy was a primary consideration that deterred non-users adopting smart speakers, and amongst users who had little privacy concerns, a lack of understanding of risks was often a justification used.

Another study found that there was generally low awareness amongst different backgrounds and people with varying education levels regarding inferencing power from audio recordings and attribute inference attacks [18]. A mere 18.7% of participants had knowledge of the physical and mental attributes that could be inferred from audio data [18]. Participants in IT sectors also had little awareness. An important point raised is the topic of informed consent, and the impact that awareness of attribute inferences could have on the adoption of VCDA's.

We can see that privacy is an important factor in the decision making process of adopting VCDA's. As use of these devices increase, it is important that user concerns are addressed and privacy protection mechanisms are in place to avoid sensitive data being compromised without user consent.

5.2.3 Privacy policies

Privacy policies play an important role in the regulation of user data, including how the data is collected, processed and used by service providers and other third party developers. However, privacy policies can be misleading, ambiguous or not exist at all for certain practices. Hence, it is important to measure the effectiveness of current privacy policies for VCDA's.

A large-scale study [19] was conducted analysing 64 720 Amazon Alexa skills and 2201 Google Assistant actions revealing a number of problematic and inconsistent policies, as well as software that violated their own privacy policies. The table below summarises findings on the presence of policies for these two devices.

Table 2. Privacy policies comparison [19]

	Amazon skills	Google actions
Without privacy policy	Total: 46 768 Percentage: 72%	Total: 234 Percentage: 11%
Valid privacy policy URL	Total: 16 197 Percentage: 25%	Total: 1887 Percentage: 85%
Broken privacy policy URL	Total: 1755 Percentage: 2%	Total:80 Percentage: 5%

As evident, many skills and actions by the two devices are missing privacy policies. Alexa has a staggeringly high 72% absence due to its lenient skill certification allowing developers to choose not to declare skills that collect personal information and bypass privacy policy requirements in certain circumstances [19]. Of actions and skills that have policies, many of the URLs are broken or redirect to the general policies page of the company. Many are also duplicates and redirect to the same policy.

There was also evidence of possible violations against legal regulations such as the HIPAA (Health Insurance Portability and Accountability) and CalOPPA (California Online Privacy Protection Act), which requires transparency regarding what data is collected by developers [19]. For example, Amazon Alexa skills under the "Kids" category lack a privacy policy, a violation of CalOPPA regulations regarding the collection of personal information from

children [19]. These are important to follow so parents have an understanding of what information they are allowing their children to share with service providers and third parties.

Throughout the study there were numerous examples of inconsistencies between privacy policies and their descriptions - up to 50 skills in Alexa. For example, a ‘Running Outfit Advisor” states that the gender of the user is collected, however this is not mentioned in the privacy policy for that skill [19]. Inconsistencies like this mean users are not able to be properly informed on the data collection and processing methods of the devices they may share personal information with.

Another issue noted is poor usability issues. For instance, 58% of the privacy policies observed had over 1500 words - a length too long for most users, and very difficult to be verbally expressed by the VCDA [19]. This, along with the difficulty to access privacy policies (such as broken URLs) can contribute to mistrust between the public and VCDAs [17][19]. Hence, by improving clarity of policies owners and non-owners of VCDAs can have assurance that their data is being handled responsibly.

5.2.4 Privacy-utility tradeoff

A privacy-utility tradeoff can be described as the act of users disclosing private information with the expectation of receiving utility. The extent to which users disclose information varies based on perceptions of risks and benefits, but ultimately the decision making process can be based on many factors, such as transparency of privacy policies, educational background, and the device itself.

The most attractive utility for most users is the convenience of VCDAs, such as controlling IoT devices like entertainment systems with their voice. A participant of a study with children stated *“it’s helpful being hands-free when you’re holding a baby.”* [17]. Another participant stated that if service providers like Google or Amazon were more honest about what is done with user data, *“then maybe I would feel a little more obliged (to purchase one).”*[17] Hence, privacy-utility trade offs are an important concept service providers should aim to balance to attract as many users as possible.

From the perspective of service providers, optimising privacy-utility in terms of developing software involves investigating methods that provide the best trade-off between performance, privacy and complexity [20]. In terms of voice and speech protection, developers may aim to

create protection mechanisms that improve user privacy by reducing the accuracy of classification accuracies on inferences such as gender or accents. An equal error rate of 50% should also be desired, as this would represent randomness in verifying speaker identity [21]. These methods would reduce the success rate of attribute inference attacks. When improving utility, developers often focus on automatic speech recognition performance, which can be achieved by reducing word error and character error rates. The challenge in VCDAs is to have strong voice and speech protection mechanisms whilst also having functional utility so that performance is not negatively impacted.

Overall, privacy protection mechanisms are important to prevent the misuse of sensitive user data. It has been demonstrated that users hold concern regarding the way service providers and third parties can use their data, such as gaining private information through inferences. We also see that there are many gaps in privacy policies that can be exploited by service providers and third parties - there should be greater transparency and accessibility so users can make an informed choice on the VCDAs they use. Lastly, we provide a brief overview of privacy-utility tradeoffs and the motives to improve it for the benefit of both the user and service provider,

5.3 Related Work

In this section we explore different methods in voice and speech protection with a focus on defending attribute inference attacks via disentangled representations.

There are numerous types of solutions that have been proposed to improve upon user privacy, such as local data processing, user authentication, and even activation and recording control by implementing a camera to analyse the speaker’s gaze to determine if the VCDA should listen [9]. These methods however, are not defences against attribute inference attacks. Voice and speech protection is a method that involves altering characteristics of the speaker's voice so that personal attributes and identity cannot be inferred.

An example of such a method is VoiceMask [22], an app that acts as an intermediate between users and the cloud. VoiceMask uses keyword substitution for sensitive words identified by the user, and vocal tract length normalisation for voice conversion in order to prevent content (sensitive topics) and identify (age, sex) privacy breaches respectively. This method reduced speaker identification accuracy by 84%, with a 14.2% drop in speech recognition accuracy

[22]. However, this indicates a drop in usability due to reduced command recognition performance. Additionally, attribute inferences still remained relatively high [22].

Another method involves signal processing, such as through pitch standardisation to conceal gender [20]. This involves calculating the average pitch of each utterance and shifting it to a predefined value. Whilst this method is extremely efficient as it does not rely on a neural network, it is not foolproof as classifiers can learn to identify and separate male and female voices from the pitched samples.

5.3.1 Disentangled representations

In order to have a stronger defence against attribute inference attacks, finer control over the aspects of the voice is necessary. Hence, we leverage disentangled representations in order to achieve this. Disentangled representation learning is an unsupervised learning technique that breaks down or disentangles each feature into narrowly defined variables and encodes them as separate dimensions [12]. The goal is to improve quality of latent representations by explicitly separating the underlying factors of the observed data. In human recorded speech, this may be speaker identity, gender, accent, background noise, and more. Upon reconstruction of the voice, sensitive attributes are filtered to reduce the impact of attribute inference attack. An equally important goal is maintaining the functional utility of the application, so that automatic speech recognition is not impacted to an extent that individuals cannot make use of VCDA's.

The following is an analysis of current voice and speech protection mechanisms that leverage disentangled representations to reconstruct the voice and defend against attribute inference attacks. We compare performance through privacy-utility trade offs.

5.3.2 EDGY

Edgy is a configurable privacy-preserving framework that converts voice data by filtering sensitive attributes before it is sent to the cloud [23]. It leverages disentangled representations to learn independent factors in the raw data and provide configurable privacy to users based on their preferences. The framework [12] (visualised below) is based upon previous work in which the authors propose a dual phase disentanglement filter. The first phase (optimisation) allows users to choose a privacy preference (low, medium, high) that acts as a guide to reconstruct the output. Whilst the second phase (filtering) uses a vector quantised variational autoencoder structure.

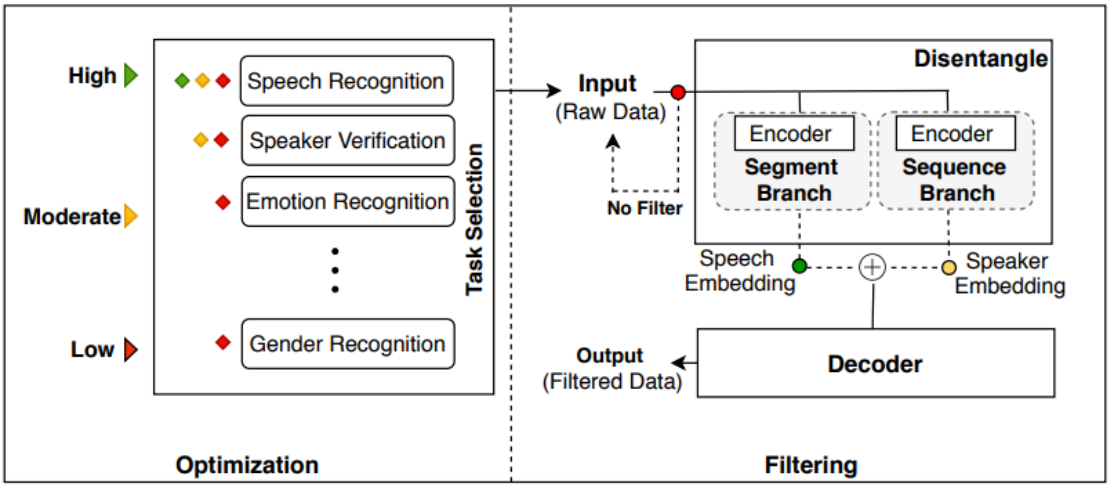


Figure 3. Edgy Dual Phased Disentanglement Module [12]

The disentangle module in figure 3 is split into two branches to learn independent factors for speech recognition and speaker verification. In doing so, diverse types of information can be learned to control the reconstruction during the decoder stage. The Edgy model expands upon this framework through optimisation to accelerate deep learning inferences at the edge prior to cloud offloading - however, the memory costs associated with deep learning are still high and make edge deployment difficult [23].

5.3.4 Client-VAE

This method uses a variational autoencoder (VAE) to learn latent representations to reconstruct the voice, similarly to Edgy. However, the VAE is convolutional and applies the Griffin-Lim algorithm for converting spectrograms into waveforms [20]. A VAE is used over a regular autoencoder to prevent issues of irregular latent distribution and enabling sound generative capabilities over the entire latent space. A normalisation layer is added to the decoder to improve reconstruction of speaker content. This paper also compares performance of disentanglement to that of signal processing and adversarial training.

It was found that as more of the encrypted speaker samples were given to a gender classifier for it to train on, disentanglement maintained best performance at preserving privacy. However, with minimal information exposed to the classifier, other methods such as signal processing and generative adversarial networks performed just as well. It was also found that signal processing outperformed the disentangled and adversarial training methods in terms of utility performance (word error and character error rates), most likely due to pitch standardisation not degrading sound quality as much as the other methods. Overall, this paper showed there is an opportunity to combine the privacy preserving performance of VAE's with

the utility and memory performance of signal processing.

5.3.5 Gender and identity protection framework

This method decodes linguistic content with gender information using a VAE to reconstruct speech. Disentangled latent representation is used to encode these different attributes independently. In doing so, gender information is used to protect identifying information as well. [21]

Table 3. Framework privacy settings [21]

	Identity	**Gender**
Same	SI	SG
Random	RI	RG
	RISG	SIRG

Table 8 illustrates how two biometric attributes (identify and gender) are combined into 5 privacy settings by keeping one attribute fixed and the other random. (S) means the biometric attribute is the same as the source speaker, while (R) means the biometric attribute is randomly assigned. The model is a Vector Quantised VAE, used to transform the speech content into a discrete latent space. The decoder is trained to disentangle identity and gender from the content information, then the voice is reconstructed according to the defined privacy settings.

The paper found that in order to protect speaker identity more effectively, it is preferable to map its contents to a randomly assigned gender, rather than a random speaker identity [21]. Doing this improved utility performance compared to the EDGY framework, with a word error rate of approximately 1.2% and demonstrated that gender is also an important feature in preserving speaker identity. With the same identity-random gender (SIRG) setting (equivalent to the moderate privacy setting on the EDGY framework) , gender classification accuracy was 59.89%. Whilst randomising identity brought this figure close to 50%, it shows that privacy improvements can be made if speaker identity is kept, which is important for speaker verification tasks.

This section has analysed different types of methods that defend against attribute inference attacks, from leverage disentangled representations to signal processing. What we have learned is that utilising disentanglement allows for finer control over the latent space, resulting in better privacy performance, namely concealing gender. However, other methods such as signal processing offer better memory performance and hence are easier for deployment in VCDAs. We will explore how signal processing, such as pitch modification can be used to learn disentangled representations and improve privacy-utility trade offs.

Through this literary review we have also provided insights into VCDA's and their privacy implications. In doing so, it is established that there is concern in the community and a problem that has merit in addressing.

6.0 Methodology

To improve upon the privacy-utility trade-offs of current voice protection methods, we propose a variational autoencoder (VAE) model and demonstrate a form of disentanglement to separate pitch from raw speech data. We then leverage disentangled latent representations to identify features responsible for pitch, and use these parameters to control the reconstruction of the voice in order to protect against attribute inference attacks, namely gender classification. Hence, we utilise pitch to protect the gender identity of speakers.

We use a VAE as it is most suitable for learning latent disentangled representations [12,17,20] and allows us to have fine control over the reconstruction of the voice. Since signal processing also has the benefits of memory efficiency, we use it for pitch modification to identify latent codes responsible for pitch.

6.1 High Level Overview

This section details a high level overview of the tasks, design and components involved in the implementation of the VAE and related tasks. An image based VAE was first constructed [24] which was then heavily modified to use audio.

6.1.1 Libraries

PyTorch [25] is an open source deep learning framework based on Python and the Torch library. Its high modularity, optimisation for python and use of dynamic computational graphs makes it preferable to other frameworks such as TensorFlow and Keras. PyTorch is used for the implementation of the VAE as it provides convenient abstraction to create the model architecture. This includes related algorithms such as the loss function, as the framework allows for configurable modifications. Pytorch is also used for the efficient computations of multi-dimensional matrices called tensors, as well as to automatically transition models and related data across GPUs and CPUs - optimising computer performance.

Librosa [26] is a powerful python library for music and audio analysis. Its tools to extract information from audio data make it extremely useful for sound generation, automatic speech recognition and other machine learning tasks.We use Librosa extensively in the processing and extraction of features from audio data, such as sample rates and short time fourier transforms. It is also used to visualise and display spectrograms, allowing for an analysis of

frequencies and amplitude over time period for a given signal.

6.1.2 Audio Preprocessing Pipeline

Pre-processing is important in order to improve interpretability and usability of raw data. A pre-processing pipeline based on [27] was used for this purpose, taking raw audio files and converting them to normalised spectrograms through the following steps.

1. A loader class responsible for loading an audio file. Returns the signal
2. A padder class which performs left or right padding to the signal array if necessary. A zero padding technique is utilised to ensure greater frequency resolution in any resulting transform. Returns the padded array
3. A log spectrogram extractor class which performs a short time fourier transform on a time-series signal and converts it to a log spectrogram. Returns the log spectrogram
4. A min-max-normaliser class which normalises the log spectrogram by applying min-max normalisation (all elements within zero and one) to the array. Important to guarantee features have the same scale. Also performs denormalisation. Returns the normalised/denormalized array.
5. A saver class which is responsible for saving the normalised spectrogram features and associated min-max values

The pre-processing pipeline takes a dataset directory as input and performs the above steps, transforming all audio files into a suitable format for the model.

6.1.3 Training/Validation Split

The processed data is split into training and validation sets in order to improve model performance and fine-tuning after each epoch. Doing so can help prevent the model overfitting or underfitting. Twenty percent of the data is allocated for the validation set, whilst 80% is allocated for the training set.

6.1.4 Data Loaders

Data loaders are used to load the train and validation spectrogram sets in batches of 64, which is an efficient way to read the data due to reduced compute. The spectrograms are also transformed to tensors (a data structure similar to arrays) and resized to a 256 by 256 shape for optimised training. Upon loading, the spectrograms are shuffled to introduce randomness in training, which ensures sequential relations between the datasets for each epoch are not

learned

6.1.5 VAE Architecture

The goal in the architectural design of a variational autoencoder is to ensure the creation of a bottleneck which represents a lower dimensionality of the original data. This lower dimensional state can then be decompressed back to the original data. Hence, a VAE consists of three main components - an encoder, a lower dimensional representation, and a decoder.

The role of the encoder is to learn to compress data into a lower-dimensional representation. This lower-dimensional encoded representation is often called a latent space, which focuses on the most important features of the original data [12]. The role of the decoder is to decompress the representation from the latent space back to the original data. Hence, variational autoencoders appear to have a mirrored architecture, with the encoder compressing to the latent space, and the decoder decompressing back to the original domain.

VAE's can be fed any type of data, including images, audio or spectrograms [12]. My model inputs and outputs are spectrograms in the form of tensor arrays (a specialised data structure).

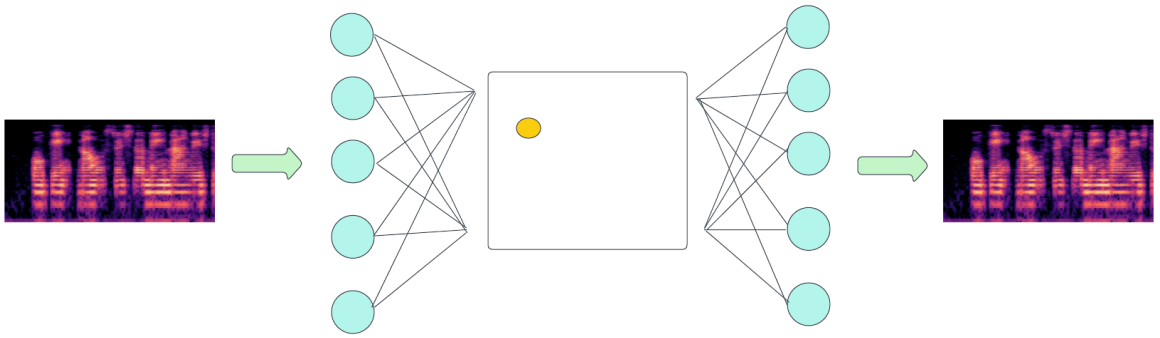


Figure 4. VAE architecture

Figure 4 visualises the architecture of my VAE. The encoder is fed spectrograms from the train set, which will compress data down to the latent space. Compressed spectrograms are represented as points on the latent space. These points can then be used and fed into the decoder which will output the reconstruction of the original spectrogram.

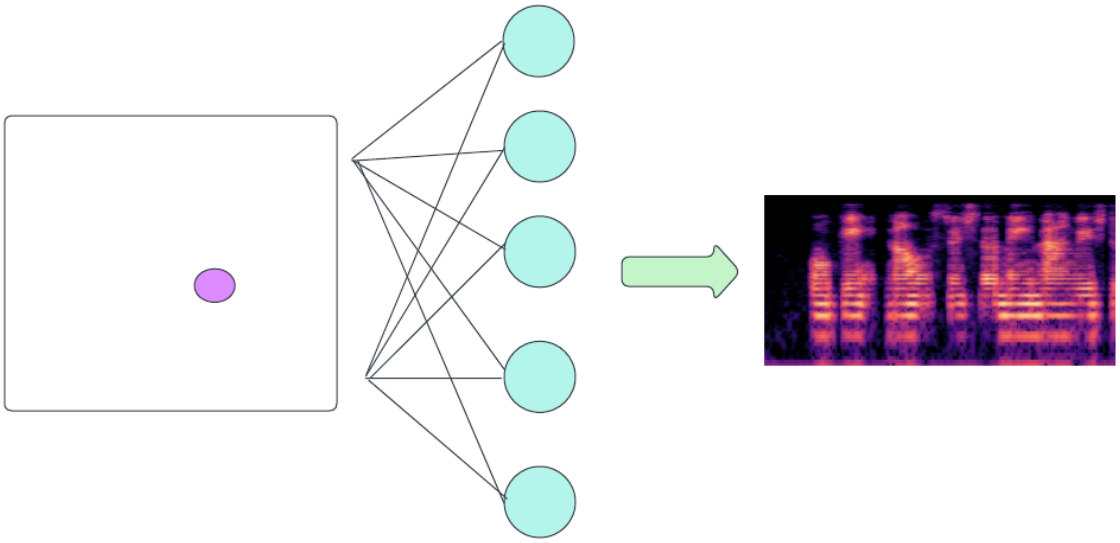


Figure 5. Sampling from latent space

The latent space can also be used for the generation of new data. Figure 5 illustrates sampling a point straight from the latent space, bypassing the encoder. This point can be fed to the decoder and can construct a spectrogram that has never been seen by the sound generator before. Hence, the VAE architecture allows for the reconstruction of compressed spectrograms or the generation of completely new or modified spectrograms.

6.1.6 Training Process

The training process involves encoding the input spectrogram data as distributions over the compressed latent space and then sampling a point from the latent space via the distribution and feeding it to the decoder. The reconstruction loss can then be determined from the decoded sample and then back propagated to the network. This loss is expanded upon in section 6.3.

The model is trained on over 1000 epochs. With each epoch and batch of spectrograms, gradients are set to 0 and the model is run. This involves encoding the data into lower dimensions through a nonlinearity function called relu, returning the mean and standard deviation vectors which is then reparameterized to deal with the randomness and distribution of the VAE latent space, and then finally returning the decoded data as part of the forward pass. After the training of each epoch, we generate a validation performance metric to see how the model performs on validation data alongside the training data. This is done by putting the model into evaluate mode to ensure gradients are frozen and aren't changed after the forward passes, avoiding training.

6.1.7 Sound Generation

Post processing is required to transform the output of the model back into a signal. The model output is a tensor array representing a log spectrogram, this is then denormalized using the saved min-max values, and then converted from a log spectrogram to a normal amplitude spectrogram. From here, an inverse short time fourier transform using the Griffin-Lim [20] algorithm is applied in order to reconstruct the phase and convert the spectrogram into a waveform. The figure below summarises this procedure.

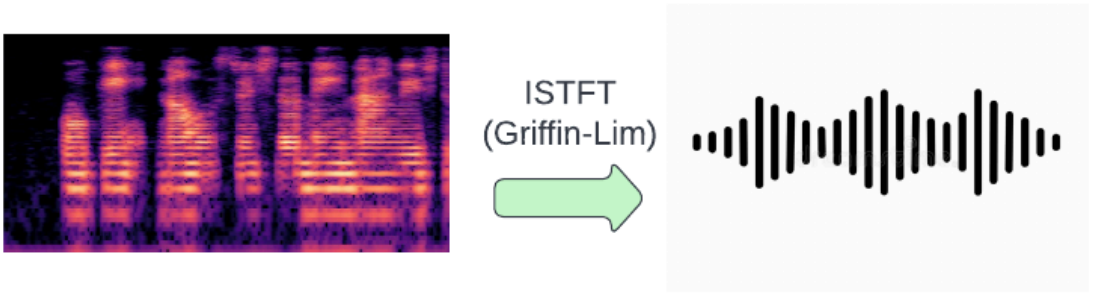


Figure 6. Spectrogram to wave-form

This waveform is then saved to a wav file with a sample rate of 11025 Hz where it can be played back as audio – which is done to identify how well the model can reconstruct and generate new voices.

6.1.8 Spectrogram Generation

We visualise spectrograms by extracting the short time fourier transform from a signal and taking the squared magnitude. We then convert this linear representation of amplitude to a logarithmic scale by applying a transformation to decibel units. This allows for a more realistic and useful representation of the audio data as humans use a base 10 logarithmic scale to perceive sound intensity [28]. After plotting these spectrograms, we are able to view the distribution and intensity of different frequencies. This is useful as we can analyse how modifying features of an audio sample alters its visual representation, as well as identify features of a voice from a visual representation alone [37].

6.2 Model Details

Below is a listing detailing the layers of our variational autoencoder.

```
VAE(
  (fc1): Linear(in_features=65536, out_features=1000, bias=True)
  (fc21): Linear(in_features=1000, out_features=16, bias=True)
  (fc22): Linear(in_features=1000, out_features=16, bias=True)
  (fc3): Linear(in_features=16, out_features=1000, bias=True)
  (fc4): Linear(in_features=1000, out_features=65536, bias=True)
)
```

As can be seen the model is relatively simple, with just one layer between the input and the latent dimension. The initial layer consists of 65536 features, representing the 256x256 array of spectrogram data. This is then compressed to 1000 features in the next layer, and finally to 16 in the latent space. It is then decompressed following the same dimensional path. The middle and latent layer dimensions were chosen based on a reasonable tradeoff between reconstruction accuracy and acceptable sampling results. However these values are often arbitrary and are based on experimentation or datasets.

We use the PyTorch Adam optimiser for updating weights due to it combining the best properties of other algorithms and it being suitable for data with lots of parameters [29], as well as a learning rate of 0.0005, slightly smaller than the default of 0.001 which was changed based on experimentation.

The model loss function is that of a standard VAEv, returning the sum of the binary cross entropy (BCE) of the original and reconstructed image, with the score of the Kullback-Leibler divergence (KLD) between a normal distribution and the distribution of the model's latent space. Hence, Loss = BCE + KLD. Using this reconstruction and regularisation term ensures the reconstructed image is as close to the original as possible, and sampling from the latent space is more consistent, overcoming issues of regular autoencoders.

6.3 Datasets

The Fluent Speech Commands [3] dataset is chosen to represent realistic speech commands spoken into VCDA's. The dataset consists of 30 043 utterances from 97 speakers, with each utterance consisting of up to three slots, 'action', 'object' and 'location'. For instance, "Switch on the kitchen lights' ' has slots {action:activate, object:lights,: location:kitchen}. Each speaker is anonymised, however demographic information is provided and relevant attributes are provided below.

Table 4. Dataset attributes and descriptions

Attribute	**Description**
ID	A unique randomised set of characters to identify an anonymous voice.
Fluency	Self reported proficiency in speaking English (native, advanced, intermediate, basic)
First language	The speakers native language
Current language	The language the speaker uses in day to day life
Gender	Self reported gender identify (male/female)
Age Range	The age category the user falls in (22-40, 41-65, > 65

The data was crowdsourced with speakers uttering phrases in random order. To ensure reliability and validity of the data, unintelligible audio samples were removed.

The Fluent Speech Commands dataset was chosen as it most closely represents spoken commands into VCDAs. Other datasets were less suitable, the Google Speech Commands [30] has single word commands such as "stop" and "go". Airline Travel Information Systems [31] has phrases of people asking for travel inquiries, and LibriSpeech [12] is general English speech. Other datasets introduced in [32] are either costly or have too few variety in speakers and phrases. Hence, Flute Speech Commands was the most appropriate choice for our purpose as it has a large variety of multi-word commands and speakers for digital assistants, and is free.

For the purpose of faster training and to demonstrate a form of disentanglement as explained in the following section, we train the model on a small sample of size 922 utterances from one

male and one female speaker. These speakers are native english speakers and are between the ages of 22-40. Whilst it is easier for deep models to learn hierarchical representations of data with large training sets, especially with the complexity of audio data, [3] showed maintaining model accuracy was still possible with very small subsets of the original data.

6.4 Disentanglement

To have more explicit control over the reconstruction of audio samples in order to conceal privacy, we disentangle pitch to hide gender identity. Rather than conditioning on a label that describes this particular feature during the training of the variational autoencoder, we attempt to disentangle pitch via signal processing and observing the latent space directly - a method based on the ideas of [33,34] in which the authors explore finding important latent codes by comparing transformed and original image data and identifying feature changes by individually changing latent vectors. Without conditioning, comparing the latent space of male and female samples directly to investigate pitch is not feasible due to the high degree of variability from speaker to speaker. Hence, to isolate the effect of pitch on the latent space we add processed versions of audio samples where only pitch is modified. To do this, we sample three audio utterances from the male speaker and three from the female speaker.

Table 5. Chosen male and female spoken samples

Male speaker samples	**Female speaker samples**
Bedroom Lights On	Turn off the lights in the bedroom
Make the music softer	Switch on the bathroom lights
Switch on the kitchen lights	Increase the temperature in the kitchen

The above samples are chosen based on possessing a variety of the action, object and location slots. For each of these samples, a duplicate pitched version is added to the training set. For the male samples, the equivalent samples are pitched up by 6 semitones, and the female sample equivalents pitched down 9 semitones. These figures were finalised based on the experimentation of fooling gender classifiers to the point where they would classify the pitched samples as the opposite gender.

After training, we observe the latent variables of the male and female audio samples and compare them to the latent variables of their respective pitched versions - this data is visualised on excel graphs to assist in analysis. The underlying goal is to observe which variables change and identify which are most sensitive to changes in pitch. We then use this as

a reference when reconstructing the original samples, and control pitch through manipulating the identified variables to a degree where gender classification accuracy is reduced to random chance.

6.5 Utility Testing

We measure utility through testing speech recognition performance. This is done through calculating word and character error rates, done by adding the substitutions, insertions and deletions of words or characters between the reference and predicted transcript, divided by the total number of words or characters in the reference transcript.

To retrieve predicted transcripts, we use the 6 identified samples in the previous section and play the audio of the privacy-preserved versions into Apple's Siri. These predictions are then used to calculate the error rates using PyTorch. We use Apple's Siri rather than an online API tool to get accurate results of the utility of our method when using a real voice assistant.

6.6 Privacy Testing

We measure privacy through the randomness of gender classification. The models are binary and indicate either male or female, however return a threshold value which indicates the degree of classification. 0-49% is classed as male, and 50-100% female. We subtract the male value from 100 to calculate confidence for male classification. Hence, for our privacy preserving samples we aim for a confidence of around 50% which would indicate random guess.

The gender classification models are accessed via a web-based application [35] and are trained on 3168 samples consisting of both male and female speech phrases. The following are details of each model used, and each simulates a potential attack method by a third party.

6.6.1 XGBoost

XGBoost is an optimised and scalable decision tree based algorithm that utilises gradient boosting for classification [35]. The decision tree is fed weights assigned to independent variables which aid in the prediction of results. The model provides results separately for tuning with a small or large number of decision trees [35].

6.6.2 Tuned Random Forest

Random forest is a machine learning algorithm that uses a bagging method for making

predictions on decision trees created on samples of data and classifies based on an aggregate of results. This number of decision trees created is tuned for optimal results [35].

6.6.3 Stacked

The stacked model is fed the results of the two previous models (as well as another model not included due to lacking a confidence score). This final classifier decides which model to apply a stronger weighting towards with an aim to increase overall accuracy [35].

6.7 Limitations

This section details potential limitations in the methodology of this study. The approach used to disentangle pitch from the latent space may not be reliable as the VAE is not being conditioned to map pitch as its own feature. However, although it may not completely disentangle the feature to a single variable, which would prevent unwanted noise when tinkering with the variables, our method of comparing pitched and unpitched signals will help us identify the latent variables more sensitive to this feature. To help make this process more efficient, the latent dimension is reduced to just 16 variables.

Aside from pitch, there are many factors that differentiate male voices from female voices, such as intonation, timber, inflections and frequency range [35]. Hence, altering pitch may not be enough to fool advanced inference attacks. For this study, however, the pitched samples added to the train set were finetuned such that they could fool the web based gender classification into classifying the speaker as the opposite gender.

An argument which may arise is that pitch alteration alone using signal processing, without the need of a neural network and disentanglement, can achieve the same purpose as this study. Whilst it is true that signal processing can alter pitch and hide gender identity very well, it cannot extract features as finely as a disentangled representation can [20], which would be useful when modifying features other than pitch. Hence, our study is a demonstration of concealing gender identity using the latent space and its potential. Additionally, our method does utilise the benefits of signal processing to disentangle pitch.

As we are using a small sample size of one male and one female speaker, our model is prone to overfitting. Whilst this is not ideal as the model would not generalise new data well, such as new speech commands spoken by a user, we can still demonstrate the purposes of our study with this limited data. Due to the sheer amount of features present in audio data, this problem can be overcome using a much larger data sample size and more computer processing power.

The process of using an inverse short time fourier transform to convert a spectrogram back to a signal can be imperfect and cause additional noise to be present in the signal, even making the voice robotic. We experimented with multiple algorithms and decided on Griffin-Lim [20], which was able to reconstruct the signal phase with minimal artefacts.

Lastly, there is limited research on the Fluent Speech Commands dataset [3], and to the best of our knowledge no prior work on the impact of gender concealment on speech recognition performance using this dataset. Hence, we will have to compare our results to the LibriSpeech dataset [12]. Although this is not ideal, both datasets are spoken English speech and can be used for similar purposes.

7.0 Results

In this section, we state the results based on the methodology of the previous section. Namely privacy, utility and model performance, as well the generated spectrograms.

7.1 Privacy Performance

Table 6 measures privacy performance based on gender classification confidence. 50% indicates random guess, whilst 0-49% Male and 50-100% female.

Table 6. Confidence classification of gender attribute on different models

Attack Model	Confidence (Male samples)	Confidence (Female samples)	Overall confidence
XGBoost Small	56%	63%	53%
Tuned Random Forest	41%	52%	55%
XGBoost Large	61%	41%	40%
Stacked	45%	28%	42%
Model Averages	**50.75%**	**46%**	**48.38%**

The results are averaged on the samples and models explained in Section 6.4 and 6.6 respectively. We observe across all samples and models an average confidence of 48.38%.

7.2 Utility Performance

Table 7 measure utility performance based on word error rates (WER) and character error rates (CER) on automatic speech recognition using Apple's Siri using the samples stated in Section 6.4

Table 7. Word error rates and character error rates for selected samples

Privacy preserving sample	Voice assistant transcription	WER	CER
"Bedroom lights on"	"put the lights on"	67%	41%
"Make the music softer"	"make some music softer"	25%	19%
"Switch on the kitchen lights"	"Switch kitchen lights"	40%	25%
"Turn off the lights in the bedroom	"Hello in the basement"	71%	71%
"Switch on the bathroom lights	"on the bathroom lights"	20%	28%
"Increase the temperature in the kitchen	"Increase the temperature ketchup	50%	26%
		45.5%	**35%**

Based on the 6 samples, we observe an overall word error rate of 45.5% and an overall character error rate of 35%.

7.3 Privacy-Utility Trade-Off

In table 8 we compare our privacy and utility performance on the Fluent Speech Commands dataset compared to state-of-the-art methods using the LibriSpeech dataset.

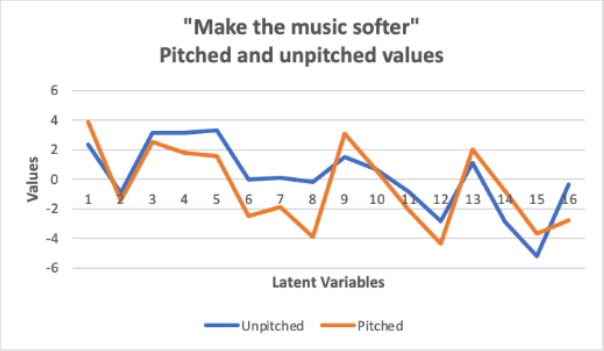
Table 8. Comparison of Utility (WER & CER) and Privacy (Gender classification confidence) performance obtained through different disentangled representation methods (EDGY, Gender-Identity framework) and signal processing using the LibriSpeech dataset, compared to our model using the Fluent Speech Commands dataset.

Model	WER	CER	Gender Classification Confidence
EDGY (high privacy setting) [12]	1.16%	N/A	51.36%
EDGY (medium privacy setting) [12]	0.32%	N/A	55.43%
Gender-Identity framework (SIRG) [21]	≈1.20%	N/A	59.89%
Signal Processing [20]	15.0%	24.7%	86.9%
OURS	45.5%	35%	48.38%

Comparisons are made against the EDGY framework on two privacy settings, to the Gender-Identity framework on the SIRG (same identity random gender) setting, and the signal processing via pitch standardisation method.

7.4 Disentanglement Performance

To measure disentanglement performance we graph the latent variable distribution for the pitched and unpitched samples, and then graph the significance of each variable. Figure 7 below is for the male spoken sample “Make the music softer”. We observe variable 8 to be most sensitive to pitch changes.



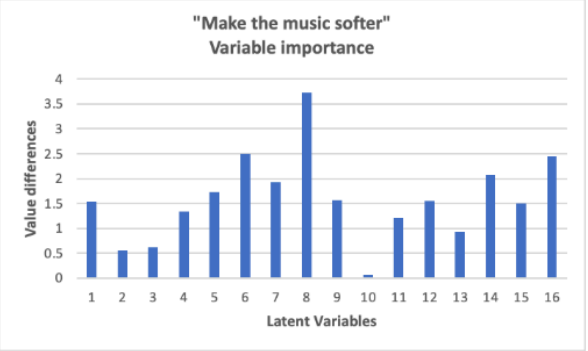


Figure 7. Latent variable distribution (left) and significance (right) for male sample

The same graphs but for the female spoken sample “Increase the temperature in the kitchen”. We observe variable 14 to be most sensitive to pitch modification.

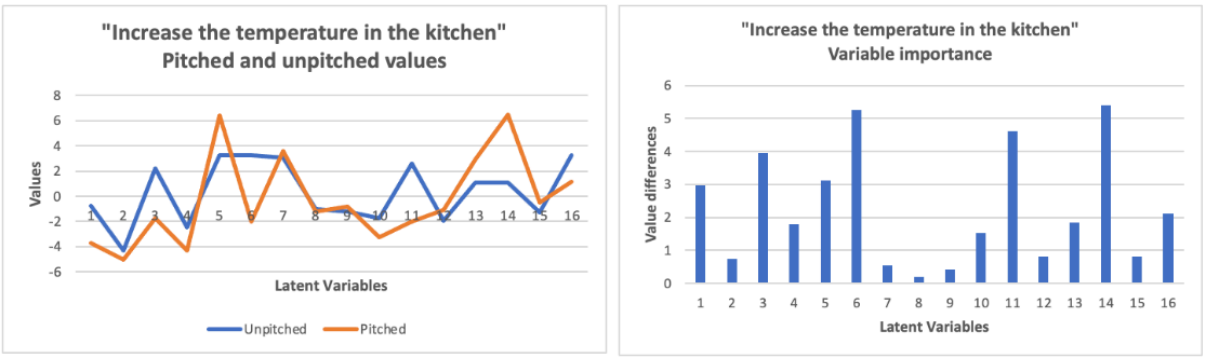


Figure 8. Latent variable distribution (left) and significance (right) female sample

7.4 Spectrogram Results

In this section we show the spectrograms for the raw speech, reconstruction speech without concealing gender, and then reconstructed speech protective gender identity via pitch. Figure Figure 9 shows these spectrograms for the male spoken sample “Make the music softer”.

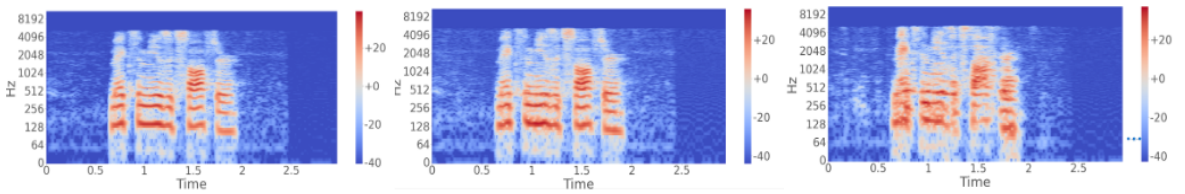


Figure 9. Spectrogram analysis for(left) raw speech, (middle) reconstructed speech without concealing gender, (right) reconstructed speech protecting gender identity using pitch, male sample.

Similarly, figure 10 below shows the spectrograms for the female spoken sample “Increase the temperature in the kitchen”

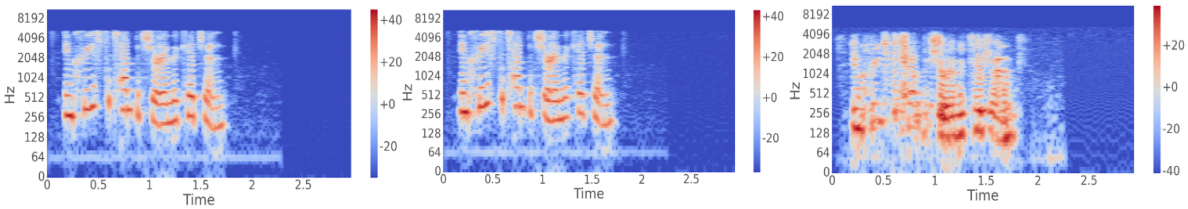


Figure 10. Spectrogram analysis for(left) raw speech, (middle) reconstructed speech without concealing gender, (right) reconstructed speech protecting gender identity using pitch, female sample.

7.5 Model Performance

In this section we show our model performance, observable in figure 11 below.

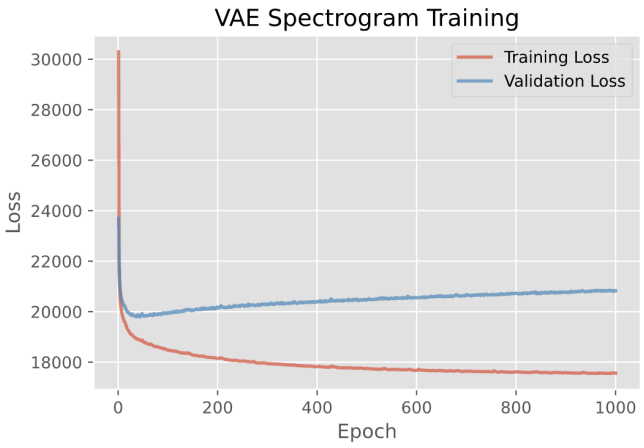


Figure 11. Model performance over 1000 epochs

Figure 11 shows the training and validation loss of our model over 1000 epochs. We observe a training loss of approximately 17500 and a validation loss of 20800.

8.0 Discussion

8.1 Privacy Performance

With an overall confidence of 48.36% based on an average of all the models and male and female samples, table 6 shows we are able to successfully conceal gender identity utilising pitch, reducing classification approximately to random guess. When looking at the results of individual models, however, we see high variability. For example, XGBoost Small on average guessed female samples with 63% confidence, whilst the stacked model 28%. This indicates that the models are interpreting pitch with different weights, and shows that pitch alone can potentially be an unreliable method of gender concealment if the attacker is aware it is the feature being used for protection. Additionally, we see classification of male samples (50.75%) to be more random than female samples (46%). This will be explored in Section 8.4, however reflects having finer control over controlling pitch in the male samples.

8.2 Utility Performance

We observe high word and character error rates as seen in table 7 for our privacy preserved samples overall, with 45.5% and 35% respectively, a markable drop in utility performance. This is due to extra noise added to the audio when modifying the variables in the latent space. However, WER and CER's don't necessarily tell us how 'good' a translation is, rather how 'different' it is to the original sample [36]. For instance, comparing "Bedroom lights on" to the transcribed "put the lights on" results in a WER of 67%, despite the overall meaning of the sentence being similar. Therefore, as long as a voice assistant, such as Apple's Siri, can carry out the intended task of the user, then strict emphasis on WER and CER's are not as useful in determining speech recognition performance and overall utility. However, there were instances in which the meaning of the command was completely lost, "Turn off the lights in the bedroom" vs the transcribed "Hello in the basement". In this case, all three command slots (action, object, location) were incorrectly transcribed. Hence, there is strong evidence that although privacy is preserved, utility has been significantly compromised.

8.3 Privacy-Utility Trade-off

Table 14 compares the privacy-utility performance of our method compared to two other disentangled representations and a signal processing method. We observe that our gender classification confidence is roughly on par with the high privacy setting EDGY model [12] and slightly better than the setting chosen for the Gender-Identity framework [21], and better

than the signal processing method [20], whilst our utility performance is worse than all three. It is important to note these methods employ protection for gender in different ways. EDGY employs a configurable framework to learn speaker embeddings - we compare against the high privacy setting in which both speaker identity and gender identity is concealed, as well as the medium setting in which identity is preserved and gender is concealed. For the gender-identity framework, we compare against the same identity-random (SIRG) gender setting in which identity is preserved but gender is randomly mapped, equivalent to the EDGY medium setting and we believe to ours as well, as we aim to conceal only gender. Despite our method not concealing identity, we believe our gender classification performance is concealed as well as the EDGY high privacy setting due to us manually adjusting the latent space for each of the selected samples, rather than applying a general method like the two disentangled frameworks mentioned. If we apply a general method for manipulating the latent space, it is likely our gender confidence (48.38%) will be closer to the medium EDGY setting (55.43%) and SIRG framework setting (59.80%).

The signal processing method [20] uses a pitch standardisation approach to conceal gender, shifting the average pitch of each utterance to a predefined value. This approach has lower gender concealing performance (86.9%) compared to ours (48.38%), but better performing utility (WER) metrics (15% vs ours 45.5%) . As this method shifts pitch based on the average pitch of each utterance rather than a flat rate, it is better able to preserve automatic speech recognition. Our gender concealing performance is again, likely better due to manually configuring the latent variables until gender classification was reduced to random guess. If a gender classifier was trained on our data, it is likely it will learn patterns in the pitched sample to identify gender, similar to what occurred in this method [20]. However, a pitch standardisation approach is much more memory efficient due to the absence of a neural network. Hence, it is worth exploring how to further reconcile using disentanglement for pitch modification with the low memory costs of signal processing. Since we have demonstrated to an extent it is possible to disentangle pitch, we believe exploring this method further could result in improvements to privacy-utility trade offs, which is a fundamental problem in VCDA's [20].

8.4 Disentanglement Performance

We observe in figure 7 the values of the latent space variables and their sensitivity to pitch change for the male spoken sample "Make the music softer". Comparing the unpitched to the pitched up version, we can see the latent space has a similar distribution with the variables

following roughly the same trend. However, all values have slightly changed, demonstrating there is no single variable 100% responsible for pitch and the feature is not completely disentangled in our latent space - which was expected due to lack of conditioning during training. Variable 8 was most sensitive by far and we altered this variable only by a few units to achieve the desired privacy result. It did however add noise to the sample which is why the WER was 25% for the transcription.

Comparatively, figure 8 shows the values for the female spoken sample “Increase the temperature in the kitchen”. We observe that the value distribution for the pitched down and unpitched sample is not as closely followed compared to the male sample. In addition, there were multiple variables that were sensitive to the change in pitch, indicating the feature was not able to be disentangled as well using our method. Resulting in multiple variables needing to be changed to achieve the desired privacy performance. This was a pattern seen throughout our female spoken samples, possibly due to the model disentangling higher frequencies better than lower ones.

We observed that changing the variables sensitive to pitch sometimes added an overtone of the altered pitch, rather than a smooth transition. This is likely due to a limited sample size, as a limited pool of speakers would mean little variety in different styles of pitch in speech. Hence, a larger sample size would resolve this issue.

By comparing the latent space of pitched and unpitched samples and identifying feature changes with manual intervention based on the ideas seen in [33,34], we have demonstrated that pitch can be disentangled to some extent and speech can be reconstructed to preserve gender identity. To improve the quality of latent representations and have finer control over the reconstruction, a larger latent dimension can be experimented. However, this would make manual identification of pitch sensitive variables less efficient due to the amount of variables. A dimensional size of 16 was utilised for demonstration, however a size of 32 or 64 is suggested to reduce noise when changing features.

8.5 Is Disentanglement Necessary?

From this study, one may ask themself the purpose of disentangling pitch when it can be modified with signal processing through pitch standardisation or even directly through software such as Audacity. We have already shown that using pitch alone to conceal gender for voice assistants may not be a reliable choice due to significant degradation in utility and

the ability for classifiers to learn patterns in the pitched data, such as in [20]. However, there are limits to how finely signal processing can be used to alter the voice, such as masking accents, as there is no way to separate complicated entangled features [20]. Hence, we have used pitch as a demonstration for disentangling features in the latent space to control the reconstruction of the voice to improve privacy, and to attempt to control utility.

As human speech data is very complex, not separating underlying features or considering variability when training for downstream tasks can impact automatic speech recognition (ASR) performance. For example, [38] showed how this could result in racial or gender biased systems. In [39], it was revealed that state-of-the-art ASR systems by Amazon and Google, amongst others, varied in performance for different population subgroups such as white and black speakers due to flaws in the acoustic models used. Hence, utilising disentanglement can aid in factoring relevant features when training on highly variable data, such as speech [12]. It can also allow for tunable privacy configurations, providing the users the choice of what parts of their voice they permit being shared [12].

8.6 Spectrogram Analysis

Figure 9 shows the spectrograms of the raw, reconstructed and privacy preserving speech for the male spoken sample “make the music softer”. Due to the use of the Griffin Lim [20] algorithm, there is minimal noise added between the raw and reconstructed audio, which can be noted through near identical spectrograms. Between the original and privacy preserving samples the intensity of lower frequency bands such as at 128 Hz are lower (which can be identified through the lighter orange). This is due to the pitch being increased on the sample. The noise added between the original and privacy preserving sample, however, is noticeable and can be seen through increased intensity between certain bands, making the spectrogram shape less rigid.

The same spectrograms can be seen in figure 10 for the female spoken sample “Increase the temperature in the kitchen”. The original and reconstructed spectrograms, like the male sample, are nearly identical. We see from the privacy preserving spectrogram that many lower frequency bands in the 64-128 Hz range have a much stronger intensity. This is due to lowering the pitch. The difference between the raw and privacy preserving spectrograms is more stark compared to figure 9 as the gender classification was more sensitive to female frequencies, hence female samples needed to be pitched down further to fool the classifier - and like figure 9, the added noise can be seen through a more irregular shape.

Through an analysis of the distribution and intensity of frequency bands in spectrograms, we are able to identify pitch differences and noise in audio. This could be extended to identifying utterances and speaker gender as well [37]. This knowledge can be used to further improve privacy-utility trade offs as it opens an avenue for investigating the effects of latent values on features through spectrogram analysis.

8.7 Model Performance

We measure model performance through loss on the training and validation loss. As seen in figure 11, there is strong evidence of overtraining after around 50 epochs due to validation loss slowly increasing. Efforts were made to reduce this through lowering the model complexity and learning rate, however ultimately it is likely due to the small sample size used for training. This resulted in the samples from the validation set being quite noisy, which is not ideal as the model would need to generalise new commands and new speakers in a real world scenario. Hence, for demonstration purposes we performed our analysis of samples from the training set after extensive training so the samples had minimal noise after reconstruction.

9.0 Conclusion and Future Directions

In this thesis, we showed how the voice is a digital footprint that can be exploited through attribute inference attacks and that there exists public appetite to address privacy concerns in VCDA's [8,9]. Based on an investigation of privacy-utility trade offs in current voice protection methods, we proposed a variational autoencoder model that utilises signal processing and disentangled representations to separate pitch and used this feature to conceal gender identity upon reconstruction of the voice. We performed our analysis on the Fluent Speech Commands dataset [3] and demonstrated that gender classification could be reduced to random guess, however functional utility was compromised due to pitch not being completely disentangled from other features, as noise was introduced even when the most pitch sensitive latent codes were modified. Hence, manually comparing latent spaces of original and transformed data and experimenting with each variable may not be enough to disentangle features. We also demonstrated how pitch modification impacts features present on a spectrogram and how this could be taken advantage of to improve knowledge of the latent space.

Our work can be extended to further analyse how to best reconcile the low memory benefits of signal processing with the reconstruction capabilities of disentangled representation learning. As our method of disentanglement does not reliably encode pitch independently as its own feature, investigations can be conducted into alternative methods of learning latent representations. Lastly, our model can be optimised to have finer control over the latent space to reduce degradation of utility and to categorically separate more features other than pitch to guarantee privacy, such as gender directly, age or emotion.

10.0 References

[1] J. Kroger, O. Lutz and P. Raschke, "*Privacy Implications of Voice and Speech Analysis – Information Disclosure by Inference*", 2020. [Online]. Available:https://link.springer.com/chapter/10.1007/978-3-030-42504-3_16 [Accessed: 20- May- 2022].

[2] P. Wu, P. Liang, J. Shi, R. Salakhutdinov, S. Watanabe and L. Morency, "*Understanding the Tradeoffs in Client-side Privacy for Downstream Speech Tasks*", *arXiv.org*, 2021. [Online]. Available: https://arxiv.org/abs/2101.08919. [Accessed: 23- May- 2022].

[3] L. Lugosch, M. Ravanelli, P. Ignoto, V. S. Tomar, and Y. Bengio, *"Speech model pre-training for end-to-end spoken language understanding*," 2019. [Online]. Available: http://arxiv.org/abs/1904.03670. [Accessed: 01-Sep-2022].

[4] "IBM Archives: IBM Shoebox", *Ibm.com*, 2022. [Online]. Available:https://www.ibm.com/ibm/history/exhibits/specialprod1/specialprod1_7.html [Accessed: 20-May- 2022].

[5] M. Hoy, "Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants", *Taylor & Francis*, 2022. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/02763869.2018.1404391?journalCode=wmrs20. [Accessed: 20- May- 2022].

[6] B. Kinsella, "The Rise and Stall of the U.S. Smart Speaker Market - New Report - Voicebot.ai", *Voicebot.ai*, 2022. [Online]. Available: https://voicebot.ai/2022/03/02/the-rise-and-stall-of-the-u-s-smart-speaker-market-new-report/. [Accessed: 20- May- 2022].

[7] "Alexa and Alexa Device FAQs", *Amazon.com*, 2022. [Online]. Available: https://www.amazon.com/gp/help/customer/display.html?nodeId=201602230. [Accessed: 20- May- 2022].

[8] J. Kroger, O. Lutz and P. Raschke, "*Privacy Implications of Voice and Speech Analysis – Information Disclosure by Inference*", 2020. [Online]. Available:https://link.springer.com/chapter/10.1007/978-3-030-42504-3_16 [Accessed: 20- May- 2022].

[9] L. Acosta and D. Reinhardt, "*A survey on privacy issues and solutions for Voice-controlled Digital Assistants*", *sciencedirect.com*, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S1574119221001449?via%3Dihub [Accessed: 21- May- 2022].

[10] L. Schönherr, M. Golla, T. Eisenhofer, J. Wiele, D. Kolossa and T. Holz, "*Unacceptable, where is my privacy? Exploring Accidental Triggers of Smart Speakers*", *arXiv.org*, 2020. [Online]. Available: https://arxiv.org/abs/2008.00508 [Accessed: 22- May- 2022].

[11] S. Wolfson, "Amazon's Alexa recorded private conversation and sent it to random contact", *theGuardian*, 2022. [Online]. Available: https://www.theguardian.com/technology/2018/may/24/amazon-alexa-recorded-conve

rsation. [Accessed: 22- May- 2022].

[12] R. Aloufi, H. Haddadi and D. Boyle, "*Privacy-preserving Voice Analysis via Disentangled Representations*", *arxic.org*, 2020. [Online]. Available: https://arxiv.org/abs/2007.15064. [Accessed: 22- May- 2022].

[13] R. Booth, "Facebook reveals news feed experiment to control emotions", *theGuardian.com*, 2014. [Online]. Available: https://www.theguardian.com/technology/2014/jun/29/facebook-users-emotions-news-feeds. [Accessed: 22- May- 2022].

[14] "Voice-based determination of physical and emotional characteristics of users-Google Patents", *Patents.google.com*, 2018. [Online]. Available: https://patents.google.com/patent/US10096319B1/en?oq=10096319. [Accessed: 23- May- 2022].

[15] S. Spiekermann and W. Christl, "Networks of Control – A Report on Corporate Surveillance, Digital Tracking", *Researchgate.net*, 2016. [Online]. Available: https://www.researchgate.net/publication/341293266_Networks_of_Control_-_A_Report_on_Corporate_Surveillance_Digital_Tracking. [Accessed: 23- May- 2022].

[16] L. Burbach, P. Halbach, N. Plettenberg, J. Nakayama, M. Ziefle and A. Valdez, "*"Hey, Siri", "Ok, Google", "Alexa". Acceptance-Relevant Factors of Virtual Voice-Assistants*", *Ieeexplore.ieee.org*, 2019. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8804568. [Accessed: 23- May- 2022].

[17] J. Lau, B. Zimmerman and F. Schaub, "*Alexa, Are You Listening?: Privacy Perceptions, Concerns and Privacy-seeking Behaviors with Smart Speakers: Proceedings of the ACM on Human-Computer Interaction: Vol 2, No CSCW*", *Proceedings of the ACM on Human-Computer Interaction*, 2018. [Online]. Available: https://dl.acm.org/doi/10.1145/3274371. [Accessed: 23- May- 2022].

[18] J. Kroger, L. Gellrich, S. Brause and S. Pape, "*Personal information inference from voice recordings: User awareness and privacy concerns*", *ResearchGate*, 2022. [Online]. Available: https://www.researchgate.net/publication/356267590_Personal_information_inference_from_voice_recordings_User_awareness_and_privacy_concerns. [Accessed: 23- May- 2022].

[19] S. Liao, C. Wilson, L. Cheng, H. Hu and H. Deng, "*Measuring the Effectiveness of Privacy Policies for Voice Assistant Applications | Annual Computer Security Applications Conference*", *ACM*, 2020. [Online]. Available: https://dl.acm.org/doi/10.1145/3427228.3427250. [Accessed: 23- May- 2022].

[20] P. Wu, P. Liang, J. Shi, R. Salakhutdinov, S. Watanabe and L. Morency, "*Understanding the Tradeoffs in Client-side Privacy for Downstream Speech Tasks*", *arXiv.org*, 2021. [Online]. Available: https://arxiv.org/abs/2101.08919. [Accessed: 23- May- 2022].

[21] D. Stoidis and A. Cavallaro, "*Protecting gender and identity with disentangled speech representations*", *arXiv.org*, 2021. [Online]. Available: https://arxiv.org/abs/2104.11051. [Accessed: 23- May- 2022].

[22] J. Qian et al., "*VoiceMask: Anonymize and Sanitize Voice Input on Mobile Devices*", *arXiv.org*, 2022. [Online]. Available: https://arxiv.org/abs/1711.11460#:~:text=VoiceMask%20adopts%20a%20carefully%20designed,sanitize%20the%20voice%20input%20content. [Accessed: 23- May-2022].

[23] R. Aloufi, H. Haddadi and D. Boyle, "*Paralinguistic Privacy Protection at the Edge*", *arXiv.org*, 2021. [Online]. Available: https://arxiv.org/abs/2011.02930#:~:text=Voice%20user%20interfaces%20and%20digital,further%20processing%20and%20subsequent%20actions. [Accessed: 25- May-2022].

[24] A. Anastassiou, "Image compression and generation using variational autoencoders in python," Coursera. [Online]. Available: https://www.coursera.org/projects/image-compression-generation-vae.[Accessed: 24-Mar-2022].

[25] "PyTorch," Pytorch.org. [Online]. Available: https://pytorch.org/.[Accessed: 24-Mar-2022].

[26] "Librosa — librosa 0.9.2 documentation," Librosa.org. [Online]. Available: https://librosa.org/doc/latest/index.html.[Accessed: 24-Mar-2022].

[27] V. Velardo, "Preprocessing Pipeline". [Online]. Available: https://librosa.org/doc/latest/index.html. [Accessed: 01-Apr-2022].

[28] "Understanding the decibel scale," Hearing Like Me, 27-Feb-2014. [Online]. Available: https://www.hearinglikeme.com/thats-intense-understanding-the-decibel-scale/. [Accessed: 10-Oct-2022].

[29] J. Brownlee, "Gentle Introduction to the Adam Optimization Algorithm for Deep Learning," Machinelearningmastery.com, 13-Jan-2021. [Online]. Available: https://machinelearningmastery.com/adam-optimization-algorithm-for-deep-learning/#:~:text=Adam%20combines%20the%20best%20properties,do%20well%20on%20most%20problems.[Accessed: 10-Oct-2022].

[30] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," Cornell University, 2018. [Online]. Available: http://arxiv.org/abs/1804.03209.[Accessed: 10-Oct-2022].

[31] C. Hemphill, "Papers with code - ATIS dataset," Paperswithcode.com, 1990. [Online]. Available: https://paperswithcode.com/dataset/atis.[Accessed: 10-Oct-2022].

[32] J. Poncelet and H. Van Hamme, "*Multitask learning with capsule networks for speech-to-intent applications*," in ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020. Available: https://arxiv.org/pdf/2002.07450.pdf [Accessed: 10-Oct-2022].

[33] S. Schwettmann, E. Hernandez, D. Bau, S. Klein, J. Andreas, and A. Torralba, "Toward a visual concept vocabulary for GAN latent space," Cornell University, 2021.

[Online]. Available: http://arxiv.org/abs/2110.04292.[Accessed: 11-Oct-2022]
.

[34] A. Jahanian, L. Chai, and P. Isola, "On the 'steerability' of generative adversarial networks," Cornell University, 2019. [Online]. Available: http://arxiv.org/abs/1907.07171.[Accessed: 11-Oct-2022].

[35] K. Becker, "What is Your Voice Gender?," Herokuapp.com, 29-Mar-2022. [Online]. Available: https://voicegender.herokuapp.com/.[Accessed: 12-Oct-2022].

[36] R. Taylor, "What is WER? What does word error rate mean?," Rev, 06-Nov-2019. [Online]. Available: https://www.rev.com/blog/resources/what-is-wer-what-does-word-error-rate-mean. [Accessed: 13-Oct-2022].

[37] A. Tursunov, Mustaqeem, J. Y. Choeh, and S. Kwon, "*Age and gender recognition using a convolutional neural network with a specially designed multi-attention module through speech spectrograms*," Sensors (Basel), vol. 21, no. 17, p. 5892, 2021. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8434188/. [Accessed: 13-Oct-2022]

[38] X. Zhang, M. M. Khalili, C. Tekin, and M. Liu, "*Group retention when using machine Learning in sequential decision making: The interplay between user dynamics and fairness*," 2019. [Online]. Available: http://arxiv.org/abs/1905.00569.[Accessed: 31-Oct-2022].

[39] A. Koenecke et al., "*Racial disparities in automated speech recognition*," 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/32205437.[Accessed: 14-Oct-2022].

11.0 Appendix

- Github repository containing code: https://github.com/Robrules/Thesis