

Comparison Of Voice Changing Methods

by

Mehmet Arif Taşlı

Engineering Project Report

Yeditepe University

Faculty of Engineering


Department of Computer Engineering

2022

Comparison Of Voice Changing Methods

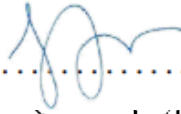
APPROVED BY:

Assist.Prof.Dr. Funda Yıldırım
(Supervisor)




.....

Prof. Dr. Sezer Gören Uğurdağ



.....

Assist.Prof.Dr. Mustafa Mutluoğlu



.....

DATE OF APPROVAL: 9/6/2022

ACKNOWLEDGEMENTS

First of all I would like to thank my advisor Funda Yıldırım for his guidance and support throughout my project.

Also I would like to thank my parents for their support and encouragement throughout my education up to the present and the students that participated in experimentation of this project.

ABSTRACT

Comparison Of Voice Changing Methods

Voice recognition systems provide human-like services but voice recording might create privacy problems. The recorded voice signals include more than speech information, such as emotional and health state of the user. This information can be used to identify the user. At a first glance this seems like a positive aspect but it is a huge problem in terms of privacy. There are multiple ways and layers of fixing this problem such as Acoustic and Textual Privacy But this project only focuses on Acoustic Privacy and compares 2 voice changing methods such as traditional pitch shifting method and machine learning based voice conversion method called Stargan-VC in terms of Voice Recognition(Lexical) accuracy and Speaker Identity Recognition Accuracy which is calculated by doing subjective testings among the students of Yeditepe University.

ÖZET

Ses Deęiřtirme Yöntemleri Arasında Karşılařtırma

Ses tanıma sistemleri, performans açısından insanlar ile karşılaştırılabilir hizmetler sağlarlar fakat birçok gizlilik sorunları yaratır. Bunun sebebi ise kaydedilen sesin içerisinde sözcük bilgisinin yanı sıra konuşan kişinin duyguları ve sağlık durumu gibi birçok bilginin çıkarılabilmesi ve bu bilgiler ile kişinin kimliğinin tanımlanabilmesidir. İlk başta bu güzel bir şey gibi gözükse de gizlilik anlamında büyük bir sorun oluşturmaktadır. Bu sorunu çözmenin birden fazla yönü ve odaklanılabileceęi yöntemleri vardır fakat bu proje sadece ses deęiřtirme yöntemlerine odaklanıp geleneksel ses perdesi deęiřtirme ve makine öğrenmesi destekli ses deęiřtirme yöntemleri üzerinde, ses deęiřtikten sonra sesin tanınma yüzdesi ve yeditepe öğrencilerinin yardımları ile birlikte öznel deneyler sonucunda ortaya çıkan ses kimliğinin tanınma yüzdesi hakkında bir karşılaştırma yapmaktır.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	iii
ABSTRACT	iv
ÖZET	v
LIST OF FIGURES	vii
1. INTRODUCTION	1
2. BACKGROUND	2
2.1. Previous works	2
3. ANALYSIS	4
3.1. Data Collection and Preprocessing	4
3.2. Pitch Shifting	5
3.3. Stargan-VC	6
3.4. Voice Recognition(Lexical) and Identity Recognition Statistics Collection	8
4. DESIGN AND IMPLEMENTATION	10
4.1. Data Collection And Preprocessing	10
4.2. Audacity Pitch Shifting	11
4.3. Stargan-VC Model	14
4.4. Voice Recognition Statistics Collection and Comparison	17
4.5. Experimentation	18
5. Test And Results	20
5.1. Voice Recognition(Lexical) Scale	20
5.2. Identity Recognisability Scale	25
5.3. Best of Both Worlds	27
6. Conclusion And Future Work	28
Bibliography	29

LIST OF FIGURES

Figure 1.1.	Overview Of The Project	I
Figure 2.1.	Pitch Shifting From The Paper[2]	2
Figure 2.2.	High-level overview of Preach, showing the knobs where a user can tune the associated trade-offs [1]	3
Figure 3.1.	Testing Data Collection State Diagram	4
Figure 3.2.	Effects of Pitch Shifting on Spectrograms	5
Figure 3.3.	How GANs do voice conversion	7
Figure 3.4.	Stargan-VC Model [3]	8
Figure 3.5.	Voice Recognition(Lexical) Statistics Collection State Diagram	8
Figure 3.6.	Identity Recognition Experiment State Diagram	9
Figure 4.1.	Testing Speech Dataset	10
Figure 4.2.	LP/HP Filter Demonstration [4]	10
Figure 4.3.	Audacity Pitch Shifting Interface [5]	11
Figure 4.4.	Audacity Editing Interface [5]	11
Figure 4.5.	Spectrogram Comparison from 0 to 40	12
Figure 4.6.	Spectrogram Comparison from 0 to 20	12
Figure 4.7.	Spectrogram Comparison from 0 to -20	13
Figure 4.8.	Spectrogram Comparison from 0 to -40	13

Figure 4.9.	Training And Validation Loss	14
Figure 4.10.	Stargan Spectrogram Comparisons for target speaker 1	15
Figure 4.11.	Stargan Spectrogram Comparisons for target speaker 2	16
Figure 4.12.	Stargan Spectrogram Comparisons for target speaker 3	16
Figure 4.13.	How the Voice Recognition API works [7]	17
Figure 4.14.	Google Voice Recognition API Output [7]	17
Figure 4.15.	Netbeans Interface [9]	18
Figure 4.16.	Correct And Incorrect Answers Output	19
Figure 4.17.	Experiment GUI	19
Figure 5.1.	Male and Female Natural Pitch Over Years [9]	21
Figure 5.2.	(%-40,%-20,%20%40 Pitch Shifts) Audacity Females / Voice Recognition(Lexical) Statistics	22
Figure 5.3.	(%-40,%-20,%20%40 Pitch Shifts) Audacity Males / Voice Recognition(Lexical) Statistics	22
Figure 5.4.	(Target Voices 1 to 10) Stargan-VC Females / Voice Recognition(Lexical) Statistics	23
Figure 5.5.	(Target Voices 1 to 10) Stargan-VC Males / Voice Recognition(Lexical) Statis- tics	24
Figure 5.6.	(%-40,%-20,%20%40 Pitch Shifts) Audacity Mixed / Voice Recognition(Lexical) Statistics	25
Figure 5.7.	(Target Voices 1 to 10) Stargan-VC Mixed / Voice Recognition(Lexical) Statis- tics	26

Figure 5.8.	(%-20,%20%40 Pitch Shifts) Audacity / Identity Recognition Statistics . . .	26
Figure 5.9.	(Target Voices 1 to 10) Stargan / Identity Recognition Statistics	27

I. INTRODUCTION

Automated Speech Recognition (ASR) technologies [6] provided human-like services and are available thanks to companies providing powerful APIs such as Google, Amazon. However, recording voice input causes privacy issues. When it comes to anonymity in terms of speech audio, there are multiple layers to speak of such as Textual Privacy and Acoustic Privacy. Acoustic elements in a speech recording can expose sensitive information about the user, such as age, gender, mood, accent, and health issues. The acoustic properties of the speakers are also biometric identifiers, allowing speaker identification and imitation. Furthermore, the linguistic substance of speech can be also private for example, business recordings can include proprietary information, or medical recordings can contain private health information about patients such as the medication that is taken or surgery details. The one which we will be focusing on is Acoustic Privacy which includes the acoustic features. To fix the issues in acoustic features, a simple pitch shifting software may be effective in some circumstances, but can lack the linguistic content in the end output, causing additional voice recognition issues. Our motivation is to compare the voice recognition accuracy and identity recognition accuracy of state-of-the-art Stargan Voice Conversion model [3] that can do any-to-many voice conversion against traditional pitch shifting [5] method while making sure that the resulted speech audio is still able to work on ASR Models [6] on a certain level. The evaluation data collection for the two methods will be done using subjective testing that is conducted among Yeditepe University students for Identity Recognition and by calculating the WER [12] for voice recognition(lexical) statistics.

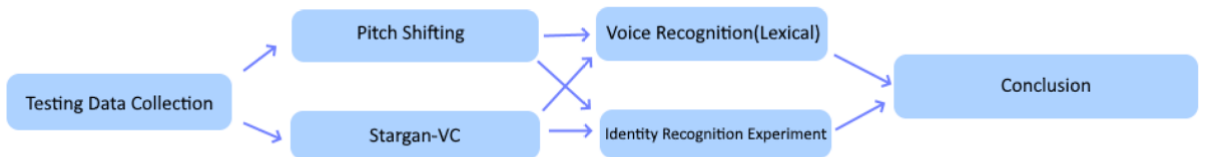


Figure 1.1. Overview Of The Project

2. BACKGROUND

Voice conversion is a field that has many ongoing and finished academic projects but I could not find any work in the literature that makes a comparison between pitch shifting and machine learning supported voice changing models but here are the two studies that work on pitch shifting and machine learning based Preech[1] voice changers.

2.1. Previous works

First one is **Voice Conversion Using Pitch Shifting Algorithm by Time Stretching with PSOLA and Re-Sampling [2]**

Voice changing has many applications in the industry and commercial field. This paper emphasizes voice conversion using a pitch shifting method which depends on detecting the pitch of the signal (fundamental frequency) using Simplified Inverse Filter Tracking (SIFT) and changing it according to the target pitch period using time stretching with Pitch Synchronous Overlap Add Algorithm (PSOLA), then re sampling the signal in order to have the same play rate. The same study was performed to see the effect of voice conversion when some Arabic speech signal is considered. Treatment of certain Arabic voiced vowels and the conversion between male and female speech has shown some expansion or compression in the resulting speech. Comparison in terms of pitch shifting is presented here. Analysis was performed for a single frame and a full segmentation of speech.

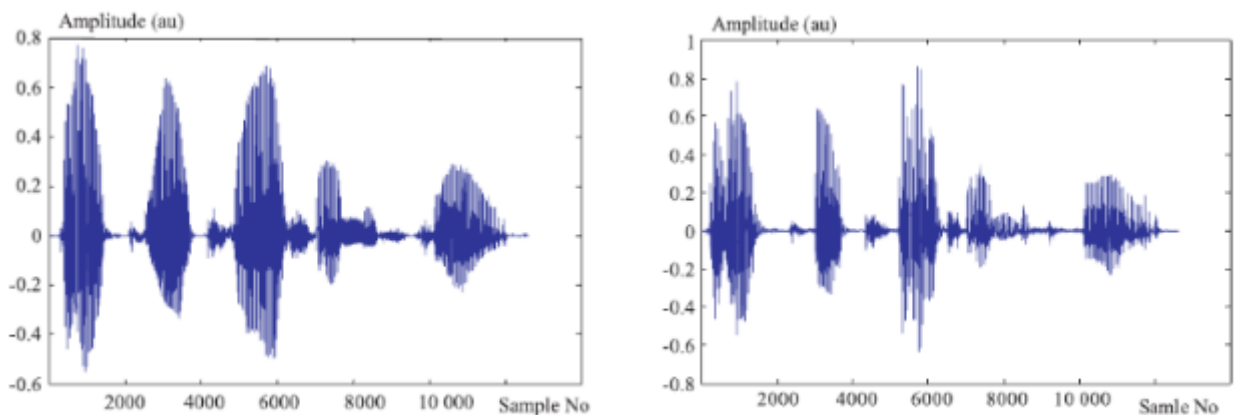


Figure 2.1. Pitch Shifting From The Paper[2]

Second one is **Preech: A System for Privacy-Preserving Speech Transcription**[1]

New advances in machine learning have made Automated Speech Recognition (ASR) systems practical and more scalable. These systems, however, pose serious privacy threats as speech is a rich source of sensitive acoustic and textual information. Although offline and open-source ASR eliminates the privacy risks, its transcription performance is inferior to that of cloud-based ASR systems, especially for real-world use cases. In this paper, we propose Preech, an end-to-end speech transcription system which lies at an intermediate point in the privacy-utility spectrum. It protects the acoustic features of the speakers' voices and protects the privacy of the textual content at an improved performance relative to offline ASR. Additionally, Preech provides several control knobs to allow customizable utility-usability-privacy tradeoff. It relies on cloud-based services to transcribe a speech file after applying a series of privacy-preserving operations on the user's side. We perform a comprehensive evaluation of Preech, using diverse real-world datasets, that demonstrates its effectiveness. Preech provides transcription at a 232.25rate over Deep Speech, while fully obfuscating the speakers' voice biometrics and allowing only a differentially private view of the textual content.

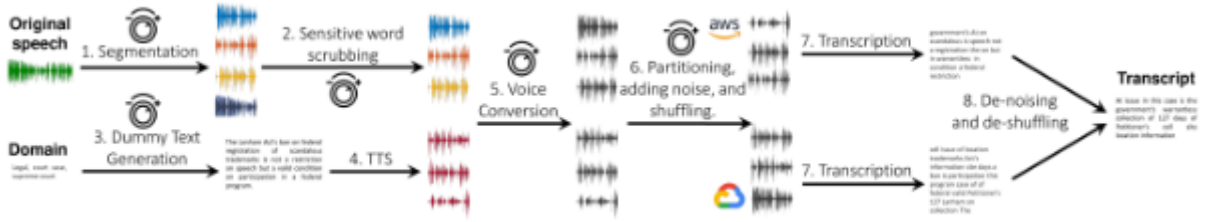


Figure 2.2. High-level overview of Preech, showing the knobs where a user can tune the associated trade-offs [1]

3. ANALYSIS

This project can be analyzed in four different sections, Data Collection and Preprocessing, Pitch Shifting, Stargan Voice Conversion, Voice Recognition and Identity Recognition Statistics Collection.

3.1. Data Collection and Preprocessing

In order to evaluate the proposed methodologies in terms of identity recognition, data must be acquired in a manner that meets the following conditions. Data must be acquired from people who can identify one another so that we may compare the Pitch Shifting method and the Stargan-VC method. The tools available are unimportant; however, employing the same recording and inferring procedures, as well as collecting data using the same words for each test subject is critical. After collecting the testing data, preprocessing to standardize the audio files may be needed.

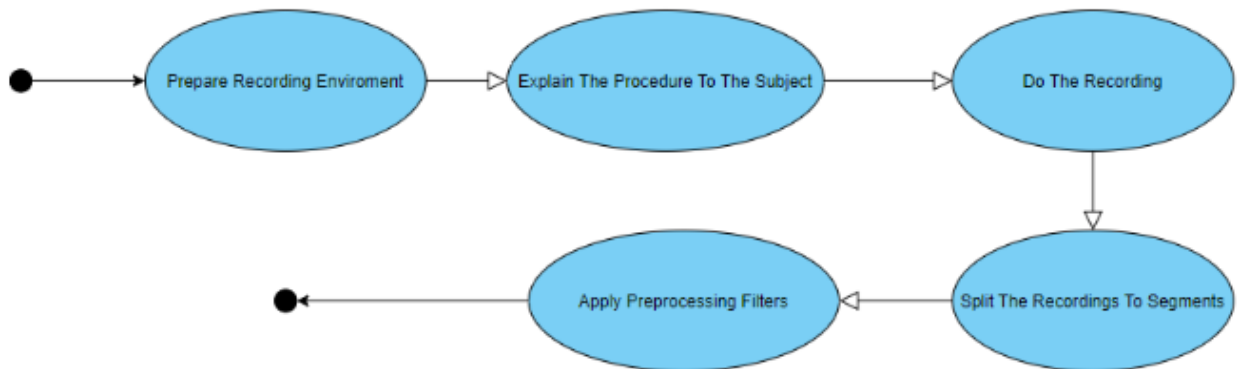


Figure 3.1. Testing Data Collection State Diagram

3.2. Pitch Shifting

Pitch Shifting is an important algorithm which is used in voice conversion and has many advantages such as being simple and rather fast compared to other voice conversion methods. To understand what pitch shifting does we first need to understand The pitch period. Pitch period determines whether some sounds are sharper than others. The pitch period is determined by the number of vibrations produced over a specific time period and the pace at which a sound vibrates is referred to as its frequency. The frequency determines the pitch; the higher the frequency, the higher the pitch. The goal of pitch shifting algorithms is to bring about a change in the pitch without affecting the replay rate and can be used as a voice changer in certain aspects.

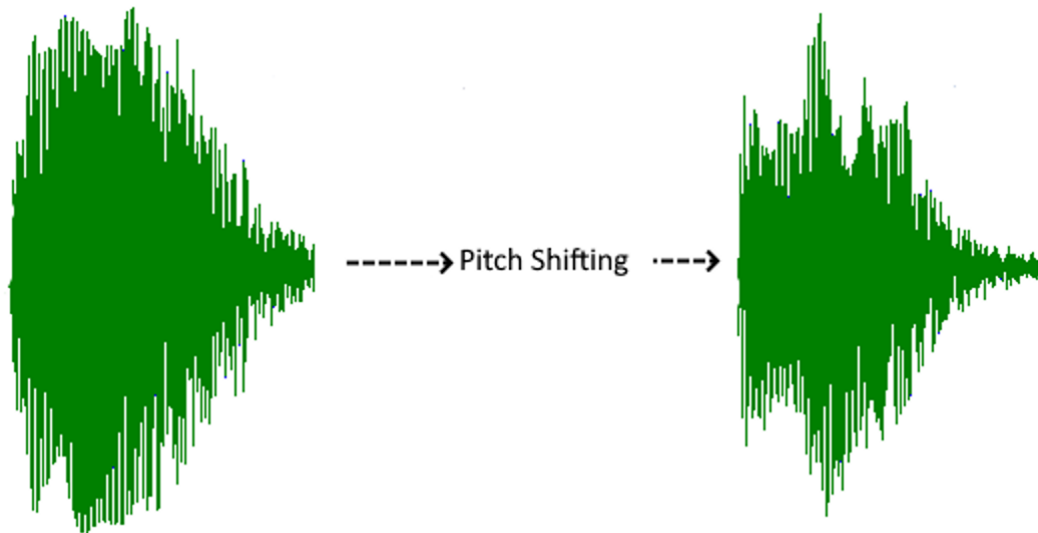


Figure 3.2. Effects of Pitch Shifting on Spectrograms

3.3. Stargan-VC

Typically, parallel data is required to train voice conversion models, which means that each speaker in the data-set must say the same thing in order to be used as a training data-set. However, when we apply the model in real life, the model may not be aware of the properties of the speaker, which is the speaker identity we are converting from, and the properties of the target speaker, which is the speaker identity we are converting to. This is referred to as the zero-shot requirement. The main reason I chose the Stargan-VC [3] model is that it can train on non-parallel data, work in a zero-shot situation, and work extremely fast relative to existing models.

In a standard GAN-VC model, we have the generator and the discriminator, and we use three loss terms to iteratively train the generator. We provide the generator the Source Spectrogram(S_{src}) and tell it to convert to Source Speaker(X_{src}), so we're just doing voice conversion from X to X , which should be equal to the source spectrogram(Σ_{idLid}). Next, we feed the Source Spectrogram (X_{src}) to the generator, which converts it to the target speaker ($X_{src} \rightarrow trg$). This converted spectrogram is then fed back to the generator, which converts it back to the source speaker ($X_{src} \rightarrow trg$), giving us a cyclic mapping where we've done voice conversion from X to Y and back to X again ($\Sigma_{cycLeye}$). Finally, we feed the transformed spectrogram to the discriminator and simply train it to make the discriminator believe the provided spectrogram is from the actual training data-set (L_g-adv).

For the discriminator loss, we feed the converted spectrogram ($X_{src} \rightarrow trg$) as well as the source spectrogram (X_{src}) from the dataset and train the discriminator to correctly distinguish if the converted spectrogram is fake or real (Ld-adv).

Following figure demonstrate how Stargan-VC [3] can be used for voice conversion

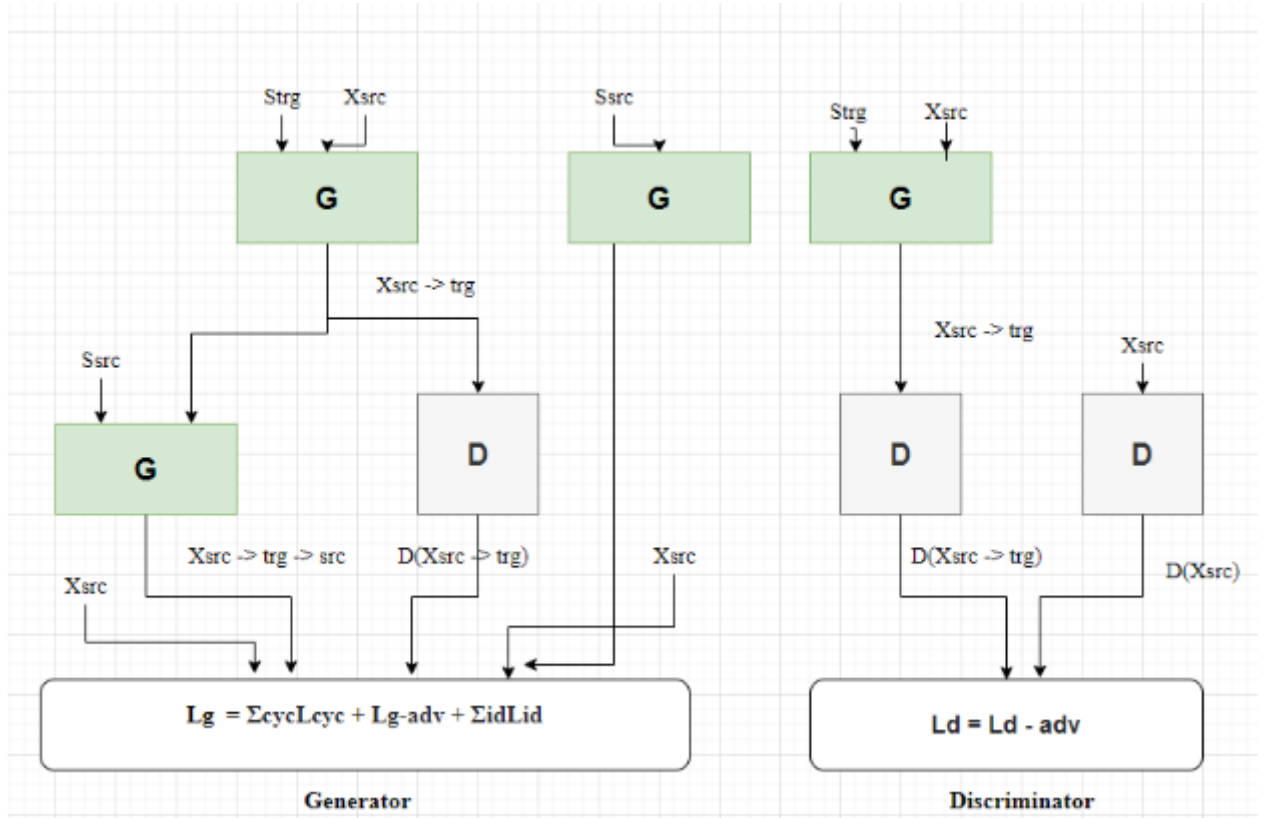


Figure 3.3. How GANs do voice conversion

Network architectures of generator G, real/fake discriminator D and domain classifier C. Here, the inputs and outputs of G, D and C are interpreted as images, where “h”, “w” and “c” denote the height, width and channel number, respectively.. IN, GLU, PS, and GSP indicate instance normalization, gated linear unit, pixel shuffler, and global sum pooling, respectively. The generator is fully convolutional. This allows an arbitrary length T which indicates the length of the audio to be input in inference

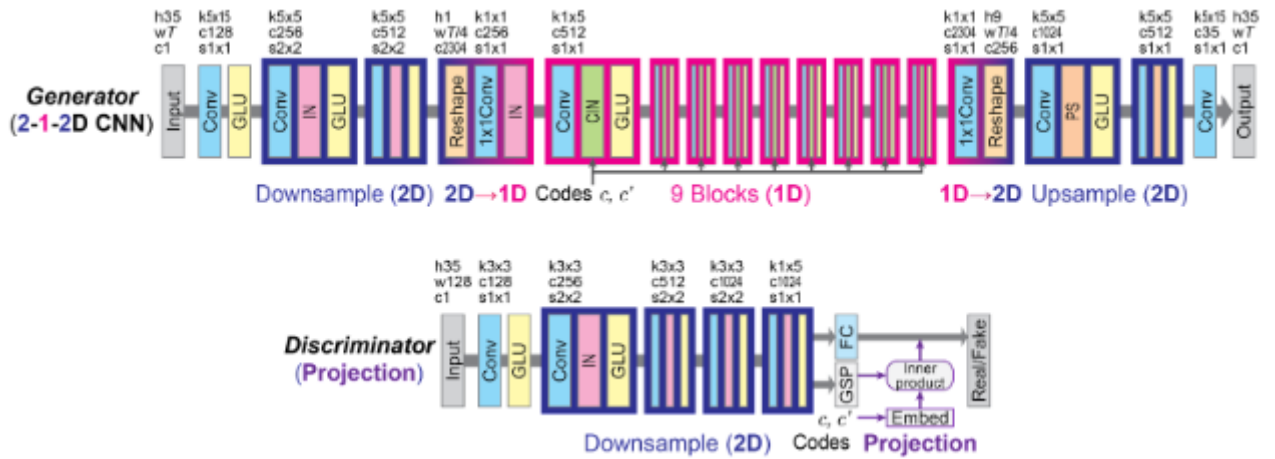


Figure 3.4. Stargan-VC Model [3]

3.4. Voice Recognition(Lexical) and Identity Recognition Statistics Collection

Evaluation in this project consists of 2 segments, one of which is voice recognition using any ASR system [6] provided that it can give us the WER(Word Error Rate) [12] on the type of testing data-set we collected. The second one is a bit tricky, we need to do an experiment with the subjects that participated in the data collection for the testing.

Following figures demonstrate the procedures taken when generating the statistics for evaluations.

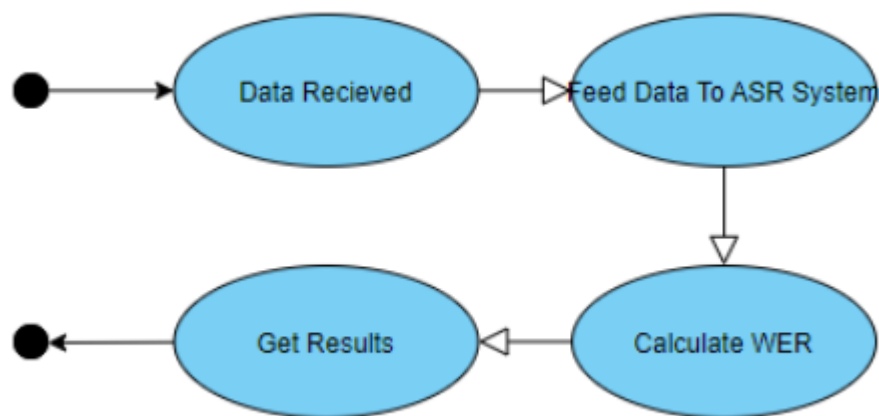


Figure 3.5. Voice Recognition(Lexical) Statistics Collection State Diagram

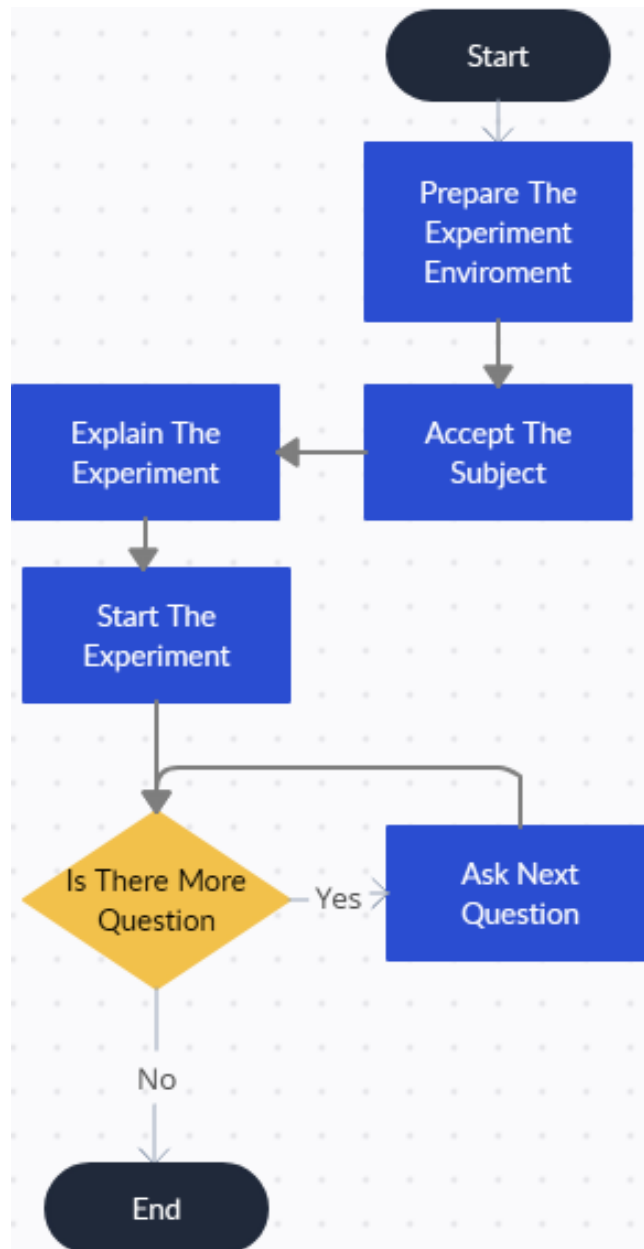


Figure 3.6. Identity Recognition Experiment State Diagram

4. DESIGN AND IMPLEMENTATION

The project can be viewed in five different parts in design and implementation chapter, sections 1,2,3,5 are the same as analysis chapter but section 4 includes additional information on the Voice Recognition API.

4.1. Data Collection And Preprocessing

The dataset collected from the subjects for testing has 73 words and 9 sentences in total.

```
Please call stella
Ask her to bring these things with her, from the store
Six spoons, a fresh apple and a watermelon
we also need a small plastic toy for the kids
Joe waited for the train.
The train was late.
Mary and Alice took the bus.
I looked for Mary and Alice at the bus station.
Mary and Alice arrived at the bus station early but waited until noon for the bus.
```

Figure 4.1. Testing Speech Dataset

I used high pass and low pass filters [4] to make the test subjects' voice audio significantly more crisp and less noisy. Low pass (LP) filters select low frequencies up to and including the cut-off frequency. HP filters choose frequencies that are higher than the cut-off frequency.

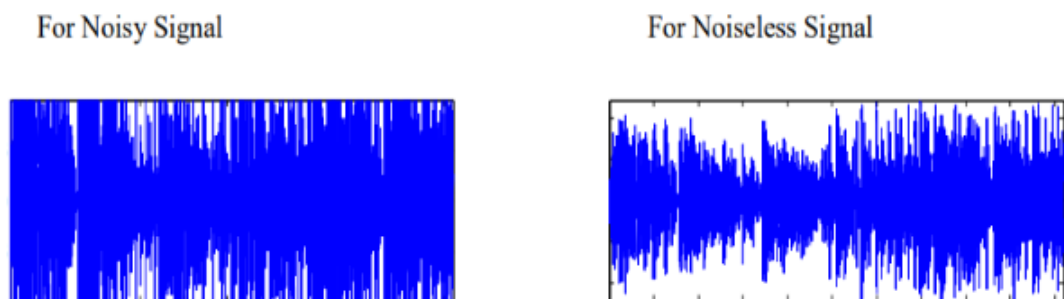


Figure 4.2. LP/HP Filter Demonstration [4]

4.2. Audacity Pitch Shifting

For the manual method I've used audacity's pitch shifter [5] because it allows me to change the pitch of the audio relative to its starting pitch which is crucial for making comparisons between different pitches. Also the starting pitch is estimated for each audio file by audacity. For accuracy reasons which will be discussed in further detail, I choose %20,%40, and %-20, %-40 relative pitch shifts.

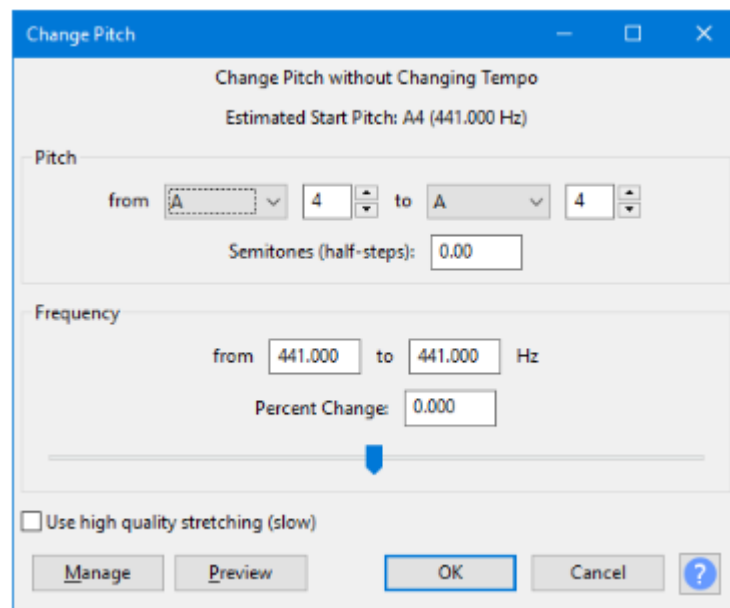


Figure 4.3. Audacity Pitch Shifting Interface [5]

Then processed each participant individually without any tempo changes and split the sentences to use in the experiment afterwards.

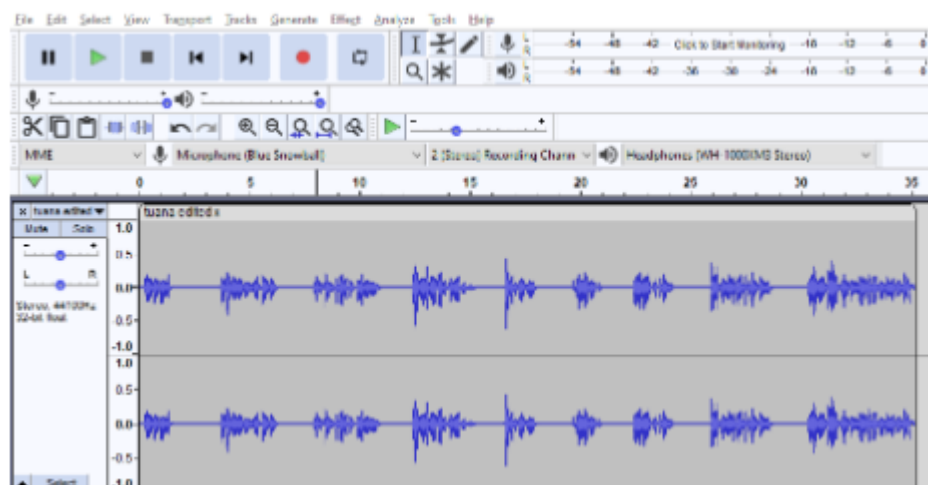


Figure 4.4. Audacity Editing Interface [5]

here are the original and pitch shifted spectrograms of the audio files to demons trade what pitch shifting does.

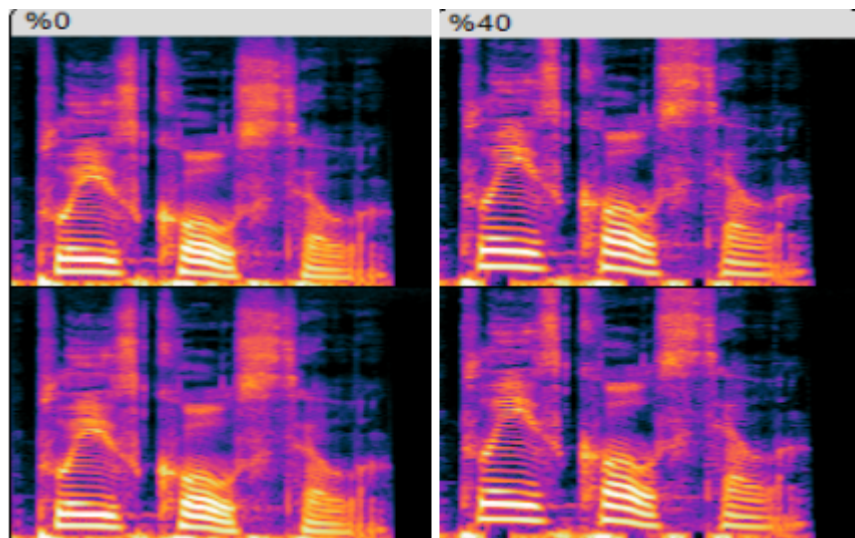


Figure 4.5. Spectrogram Comparison from 0 to 40

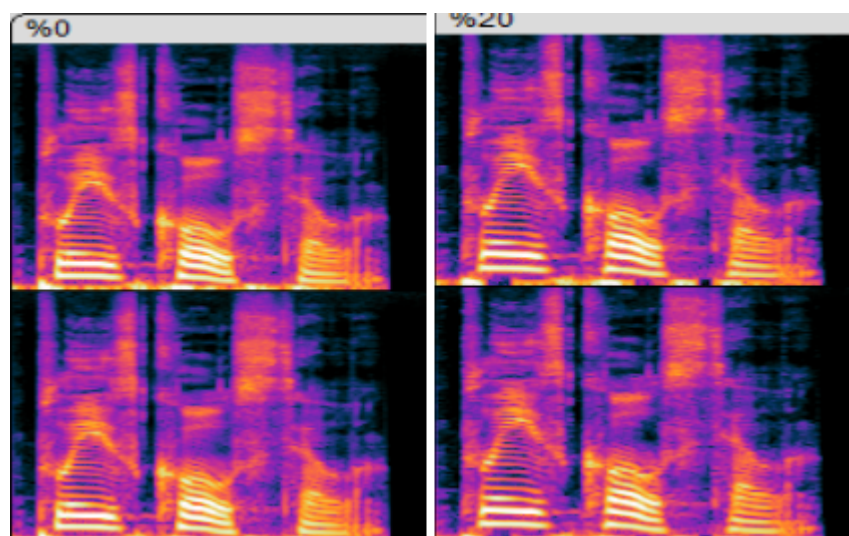


Figure 4.6. Spectrogram Comparison from 0 to 20

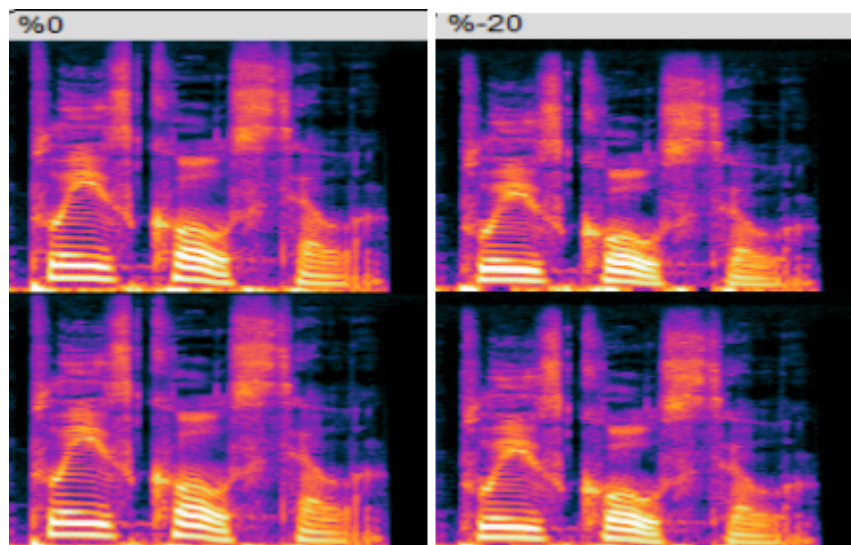


Figure 4.7. Spectrogram Comparison from 0 to -20

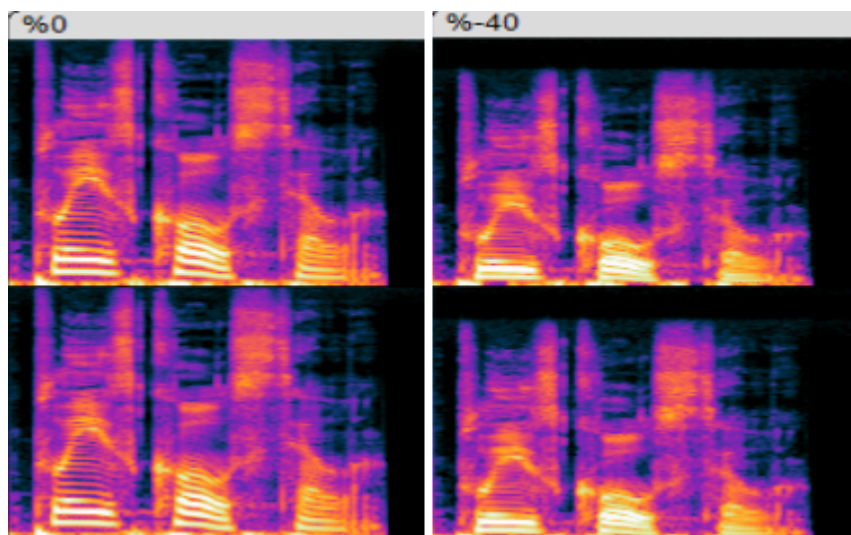


Figure 4.8. Spectrogram Comparison from 0 to -40

4.3. Stargan-VC Model

The model is trained on VCTK-Corpus data-set [7] which contains the selected 20 speakers, 10 male and 10 females. Each speaker reads out about 400 sentences, most of which were selected from a newspaper plus the Rainbow Passage and an elicitation paragraph intended to identify the speaker's accent. Each audio file is re-sampled to 24 kHz and split the data to %80, %10, %10 for Train/Test/Validation respectively. Before training the model we first need to have the features of the audio. To do this, the audio files are converted into melspectograms.

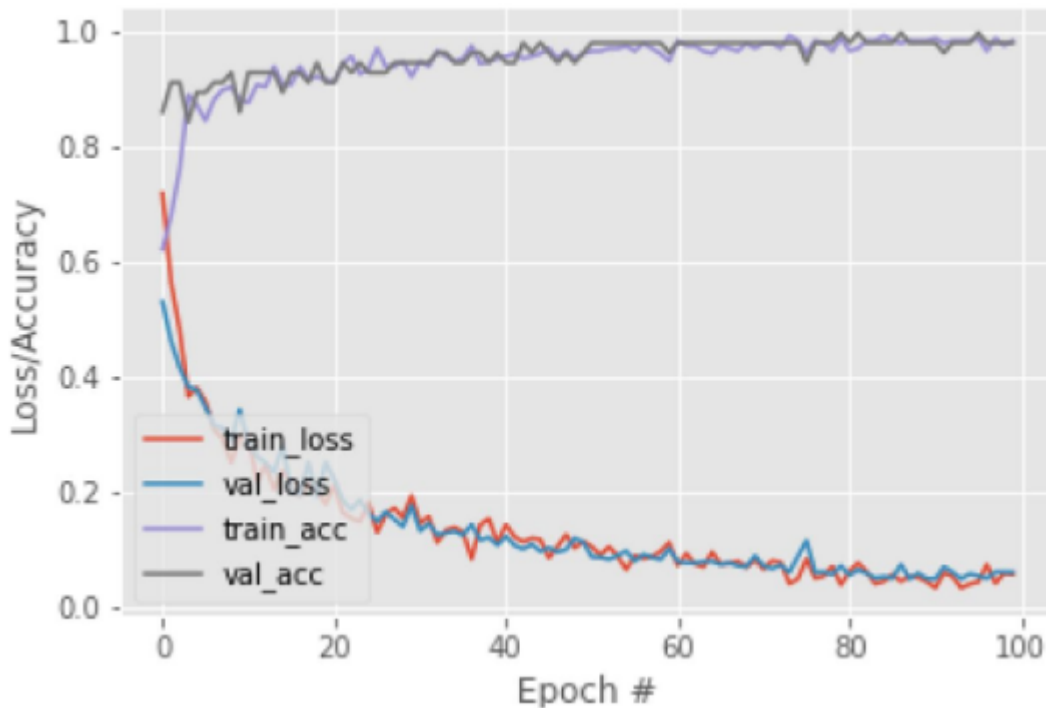


Figure 4.9. Training And Validation Loss

The model is trained using Adam optimizer with learning rate of 0.0001, and a batch size of 8. Discriminator is trained for 5 iterations and 1 iteration for encoder/generator. In the end the training loss converged at 100 epochs so i stopped the training and was left with a trained model with 0.083 training loss.

Each audio file from the test subjects is converted to 10 different target voices which are 5 male and 5 female voices. This adds a total of 486 audio files to the experiment.

Following figures are the original and Stargan-VC Converted spectrograms of the audio files for the target speakers 1 to 3.

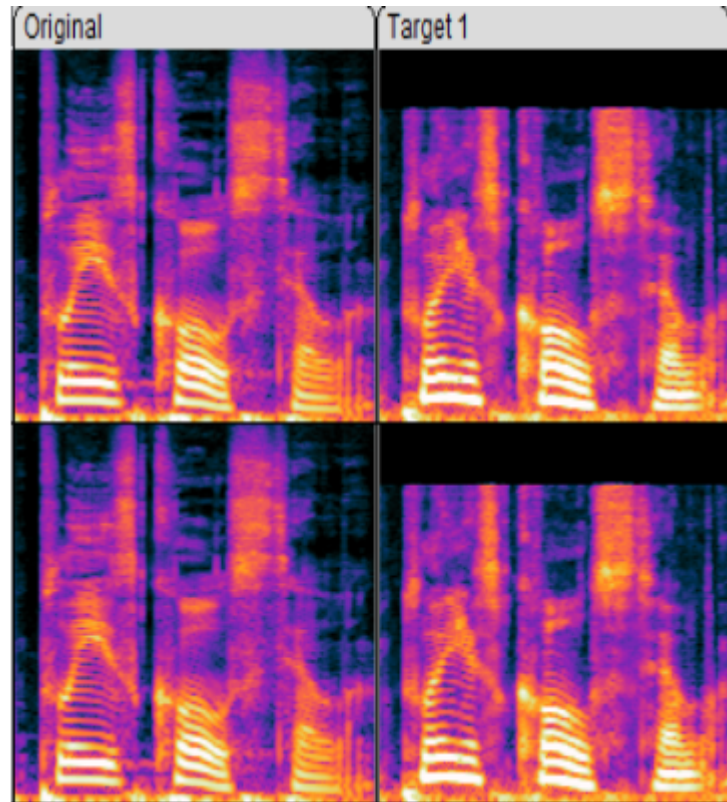


Figure 4.10. Stargan Spectrogram Comparisons for target speaker 1

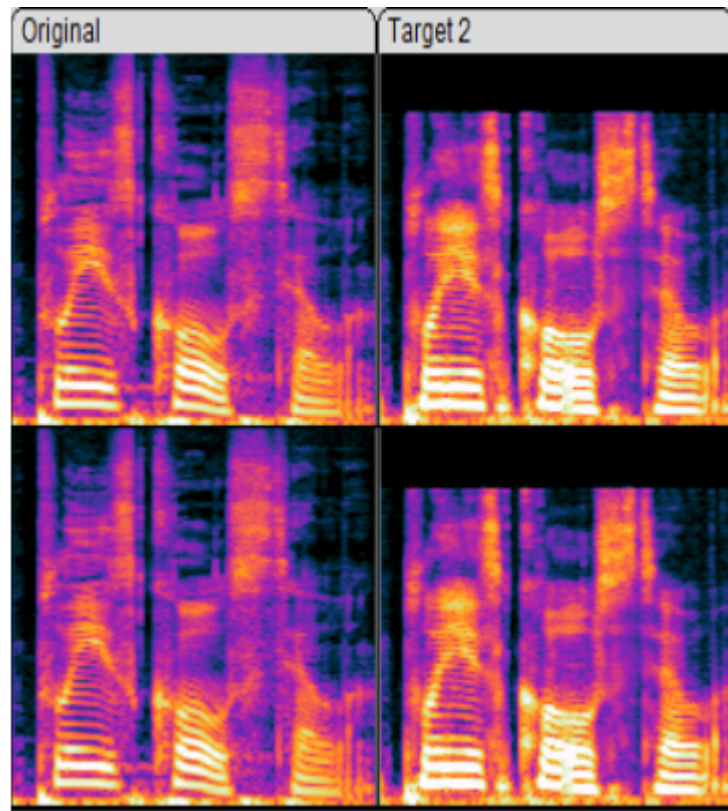


Figure 4.11. Stargan Spectrogram Comparisons for target speaker 2

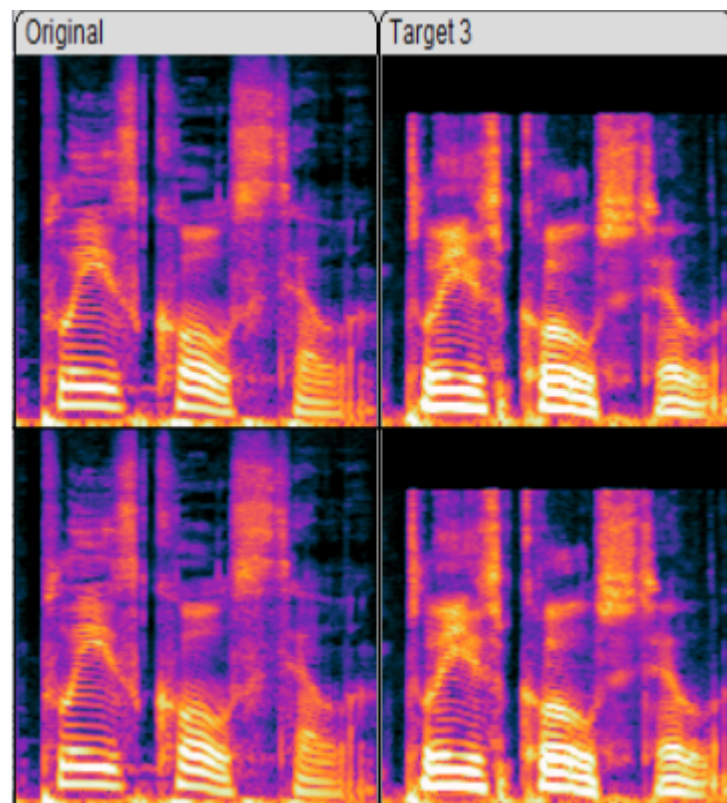


Figure 4.12. Stargan Spectrogram Comparisons for target speaker 3

4.4. Voice Recognition Statistics Collection and Comparison

We need a voice recognition model that works on our sort of data-set so that we can compare the two models in terms of accuracy. In our situation, the Google Voice Recognition API [8] works perfectly. The API automatically sets the sample rate and encoding type of the audio, and we are left with the output shown in the figure below.



Figure 4.13. How the Voice Recognition API works [7]

Given the original audio files and the correct settings, google voice recognition API was able to identify the sentences in the audio file with %99 accuracy. This is crucial for selecting the correct dataset because any audio file that cannot be recognized by voice recognition model is not usable for further experimentation.

Transcription ⬇				
Time	Channel	Language	Confidence	Transcript
00:00 - 00:01	0	en-us	0.96	please call Stella
00:04 - 00:06	0	en-us	0.77	escort to bring this thing with hard from the store
00:10 - 00:13	0	en-us	0.87	six pools and fresh apple and watermelon
00:16 - 00:19	0	en-us	0.96	we also need a small plastic toy for the kids
00:22 - 00:24	0	en-us	0.91	yo and waited for the train
00:27 - 00:28	0	en-us	0.83	the train was late
00:32 - 00:34	0	en-us	0.68	Mary and Alice Cooper divorce
00:37 - 00:40	0	en-us	0.86	I look for many and there was at the bus station
00:43 - 00:48	0	en-us	0.82	Mary and Alice and and the bus station already but waited long until noon for the bus

Figure 4.14. Google Voice Recognition API Output [7]

With that done, Voice recognition is applied to the Audacity Pitch shifted and Stargan-VC voice converted files. Then the WER(Word Error Rate) from the resulting voice recognition inferring is calculated using any method that is valid. After Looking at the results from the voice recognition API, i've decided to not include the %-40 pitch shifted method in the experiment to decrease the experiment duration.

4.5. Experimentation

The experiment is designed using java in Netbeans [9].

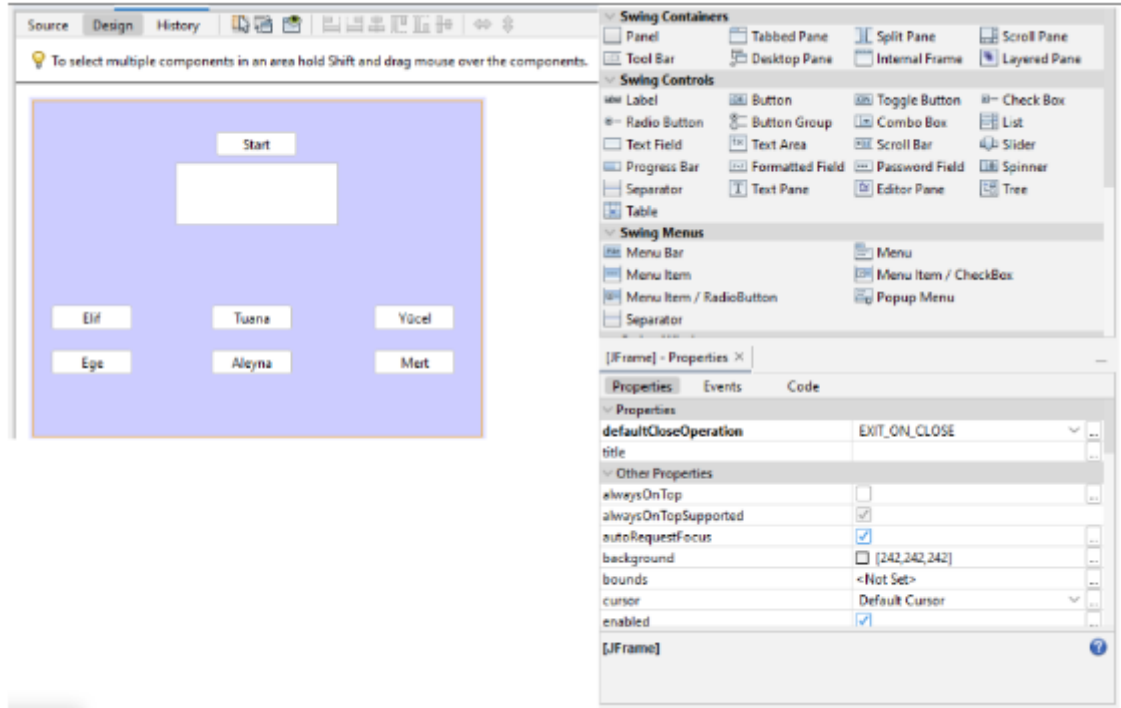
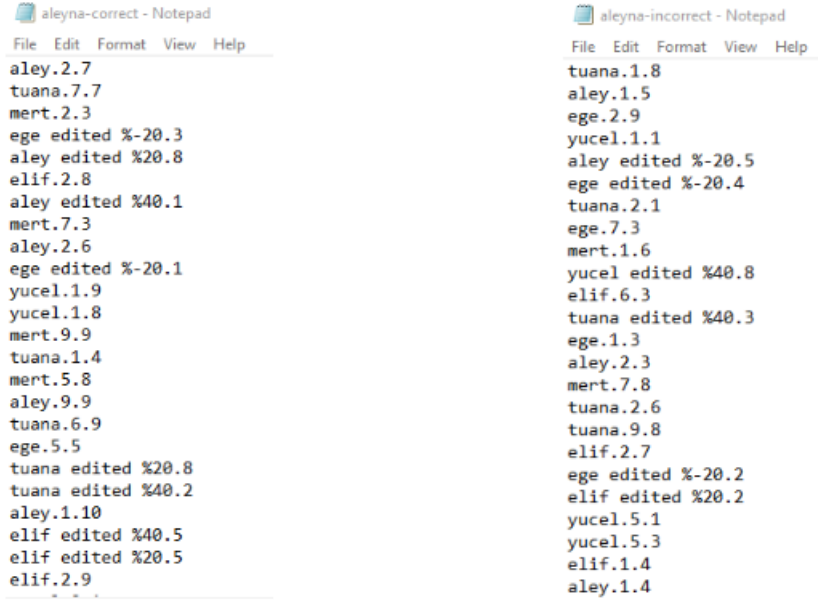


Figure 4.15. Netbeans Interface [9]

Key part of this experiment is to make people take the experiment under the same conditions every time. Six people were requested to take part in the experiment. The experiment dataset includes three Audacity Pitch Shifted audio variations and ten StarGAN-VC audio variations. Each test subject is offered 518 questions, each of which contains speech data from six different people, including themselves. Since asking 518 questions at once would cause our respondents to lose interest, the question data-set was randomly divided into two. There is also no time limit, and subjects can answer the questions before fully listening to the speech audio, with the key question being "Who owns this audio?" If the individual does not know the answer, they may make a wild guess. This is taken into account when drawing conclusions from the results.

With that said it is also important to log the answers to each question, for this an ArrayList is implemented and printed to a text file at the end of the experiment.



The figure shows two Notepad windows side-by-side. The left window, titled 'aleyna-correct - Notepad', contains a list of names and scores, some followed by 'edited' percentages. The right window, titled 'aleyna-incorrect - Notepad', contains a similar list of names and scores, also with some 'edited' percentages.

Correct Answers (aleyna-correct - Notepad):

```

aley.2.7
tuana.7.7
mert.2.3
ege edited %-20.3
aley edited %20.8
elif.2.8
aley edited %40.1
mert.7.3
aley.2.6
ege edited %-20.1
yucel.1.9
yucel.1.8
mert.9.9
tuana.1.4
mert.5.8
aley.9.9
tuana.6.9
ege.5.5
tuana edited %20.8
tuana edited %40.2
aley.1.10
elif edited %40.5
elif edited %20.5
elif.2.9

```

Incorrect Answers (aleyna-incorrect - Notepad):

```

tuana.1.8
aley.1.5
ege.2.9
yucel.1.1
aley edited %-20.5
ege edited %-20.4
tuana.2.1
ege.7.3
mert.1.6
yucel edited %40.8
elif.6.3
tuana edited %40.3
ege.1.3
aley.2.3
mert.7.8
tuana.2.6
tuana.9.8
elif.2.7
ege edited %-20.2
elif edited %20.2
yucel.5.1
yucel.5.3
elif.1.4
aley.1.4

```

Figure 4.16. Correct And Incorrect Answers Output

Following Figure is the Interface of the experiment.



Figure 4.17. Experiment GUI

5. Test And Results

Results we obtained from the implantation can be analysed in 2 different sections. One of which is Voice Recognition(Lexical) scale which tells us if the methods disrupt the linguistic information in the speech and the second one is the Identity Recognition scale which tells us if the subjects are successfully identify the speakers..

5.1. Voice Recognition(Lexical) Scale

This chapter focuses entirely on speech recognition(lexical) statistics we obtained from google's speech recognition API and ignores additional considerations such as identity recognition and processing needs.

To validate any of our work as usable, the output of the audacity method or the Stargan-VC approach must be able to preserve the input audio's linguistic qualities.

In the case of audacity pitch shifting results, boosting the pitch for male voices produces a much more natural voice that also performs better in terms of speech recognition. – i.e., males had an accuracy rate of 84 percent while females had an accuracy rate of 53 percent. And, for female voices, lowering the pitch creates a more natural sound that also works better in terms of speech recognition. This equates to a female accuracy rate of 53% and a male accuracy rate of 21. This could be because the male voice is already much lower than the female voice, and vice versa.

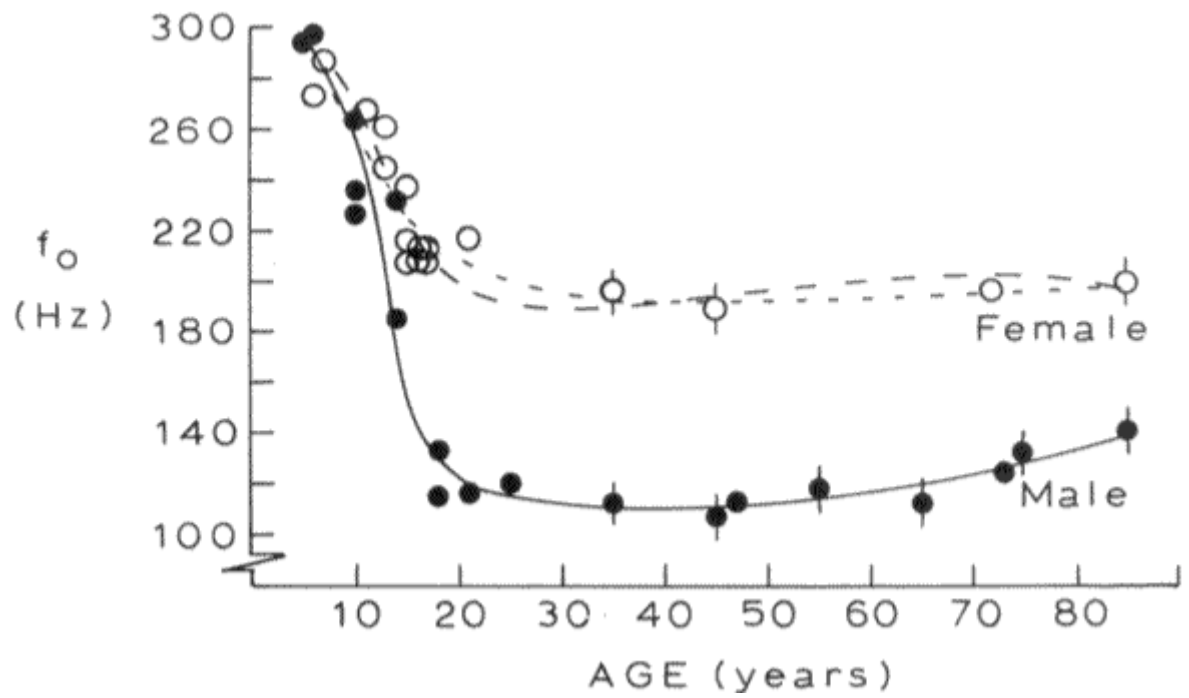


Figure 5.1. Male and Female Natural Pitch Over Years [9]

Following figures compare the 4 variations (%-40,%-20,%20%40 Pitch Shifts) of the speech input in terms of voice recognition accuracy.

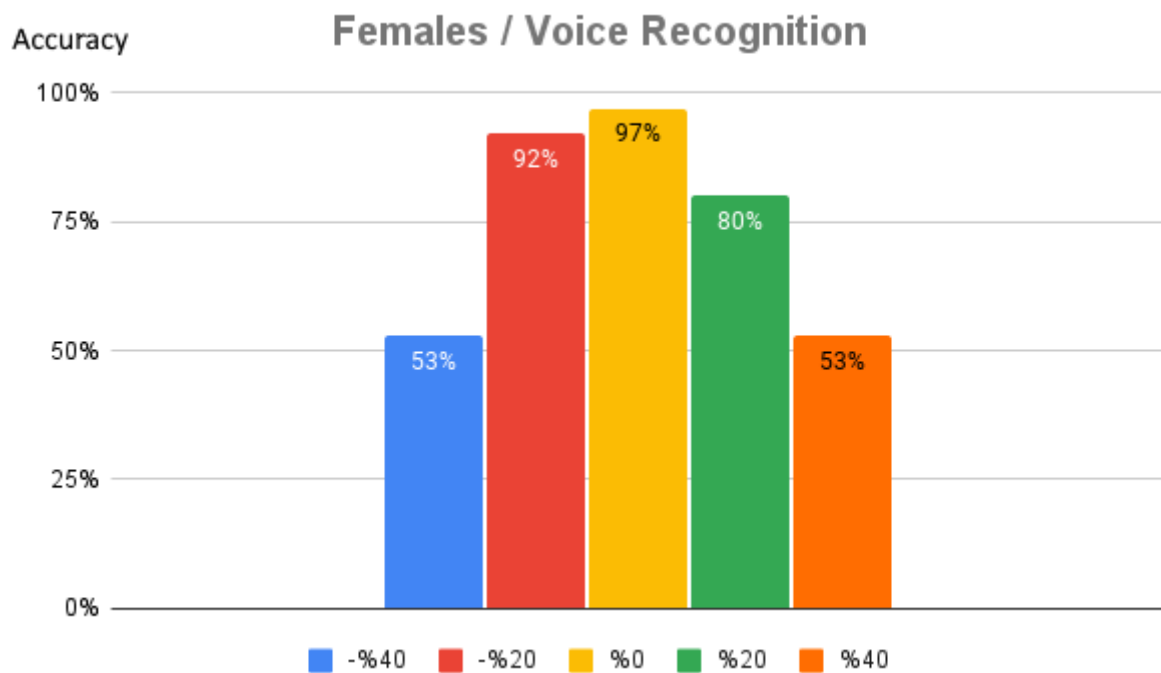


Figure 5.2. (-%40,-%20,%20%40 Pitch Shifts) Audacity Females / Voice Recognition(Lexical) Statistics

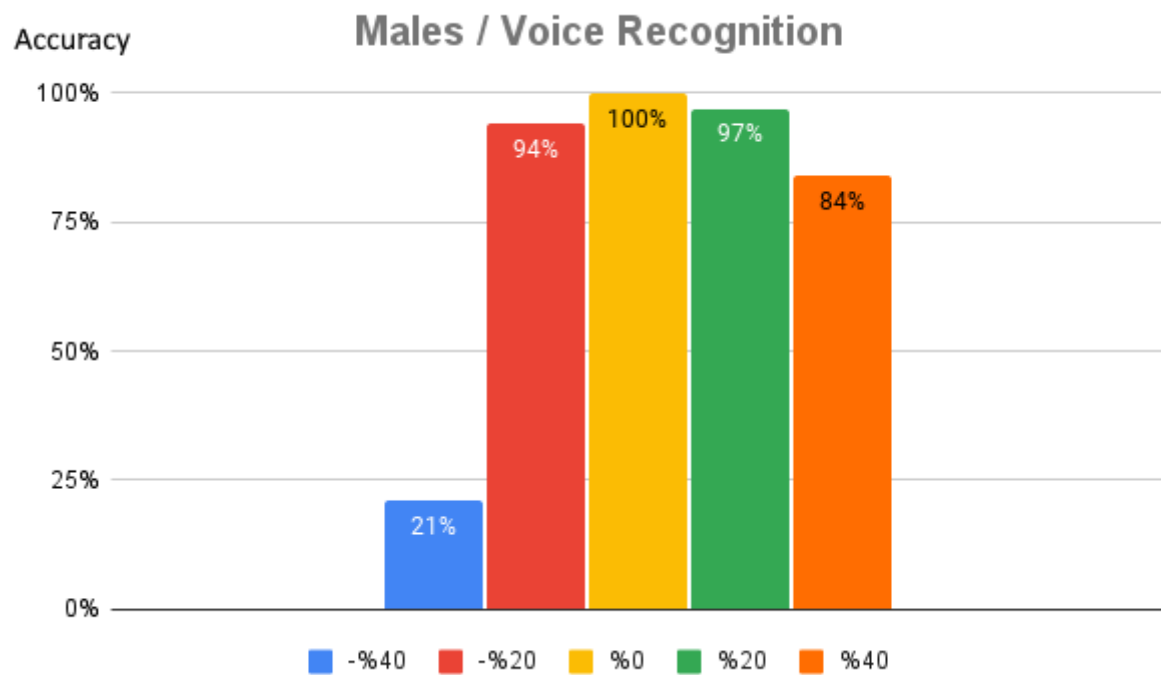


Figure 5.3. (-%40,-%20,%20%40 Pitch Shifts) Audacity Males / Voice Recognition(Lexical) Statistics

In the case of Stargan-VC, the model performs great in general but performs poorly with the female voices compared to male voices. I.e. maximum of 78% accuracy for female voices and maximum of 90% accuracy for male voices. This may be due to a training bias in stargan-VC model that could potentially make the output audio jittery thus performing bad in google voice recognition. But in general the Stargan-VC model produces voice recognition results that could potentially be used as an end product for a privacy voice changer.

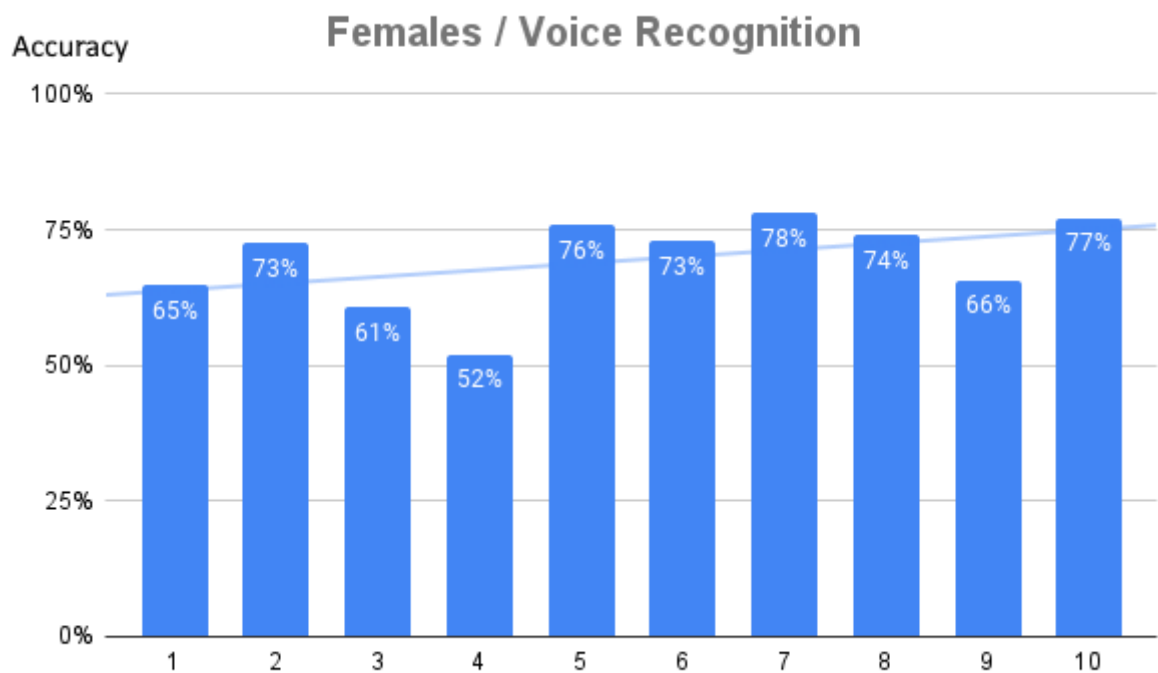


Figure 5.4. (Target Voices 1 to 10) Stargan-VC Females / Voice Recognition(Lexical) Statistics

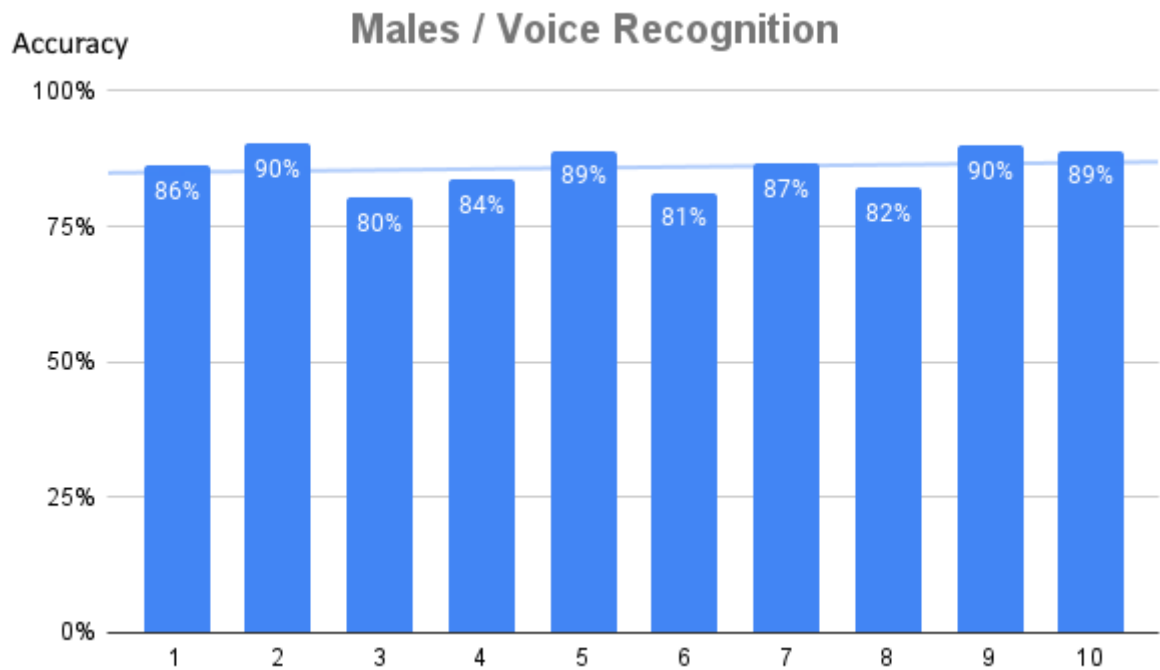


Figure 5.5. (Target Voices 1 to 10) Stargan-VC Males / Voice Recognition(Lexical) Statistics

When we look at the results of the two methods without considering any gender settings. the Audacity pitch shifting can compare to Stargan-VC at %20 and %-20 pitch shifts but performs very poorly at %40 and %-40 and in general Stargan-VC is the superior method with as high as 90% accuracy in terms of voice recognition. This shows that while Audacity pitch shifting can keep the linguistic contents at certain levels after the conversion but Stargan-VC is able to keep the linguistics concents at almost every target speaker(1 to 10).

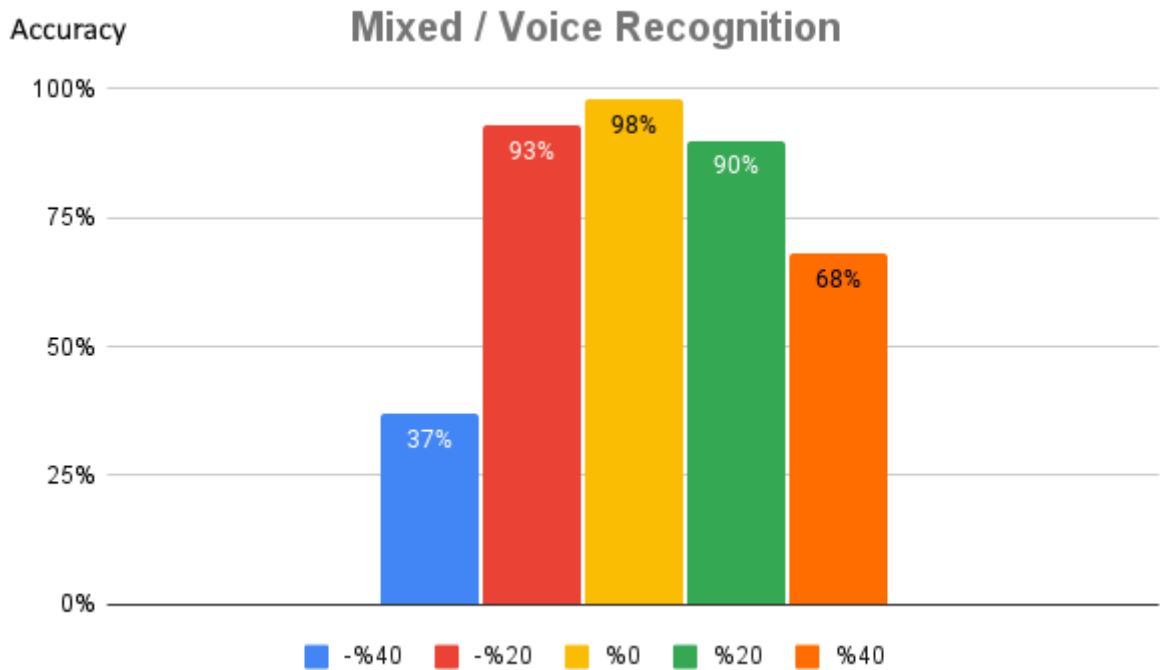


Figure 5.6. (-%40,%-20,%20%40 Pitch Shifts) Audacity Mixed / Voice Recognition(Lexical) Statistics

5.2. Identity Recognisability Scale

This section solely focuses on recognisability statistics of the two models obtained from objective testing by our 6 test subjects. To confidently compare the models we are using we need to know if the identification of the person after the alterations is possible. After all, being able to produce a result that still has the linguistic properties without altering the characteristics is not what both of the methods desire to do so. Following figures compare the 10 variations of the Stargan-VC model and 3 variations of the Audacity Pitch Shifting that contains 5 male target voices and 5 female target voices for the Stargan-VC and %40, %20, %20 pitch shifted voices for the audacity pitch shifting method. Since there are 6 test subjects thus 6 test answers, the baseline performance for this section is about %16 percent.

Looking at the recognisability results from the Audacity Pitch Shifting, it is fair to say that this method works better than not using anything but is not a valid option for our goal of choice with %44.64 correctness rate and cannot be used as an identity masker between the speech recognition API and the user. In the instance of Stargan-VC, it achieves astonishing results with an average accuracy rate of %27, which is fairly low given the experiment's baseline of roughly 16 percent. This shows us that the subjects are not able to identify the speaker on Stargan-VC and it can act as an identity masking layer between the speech recognition API and the user.

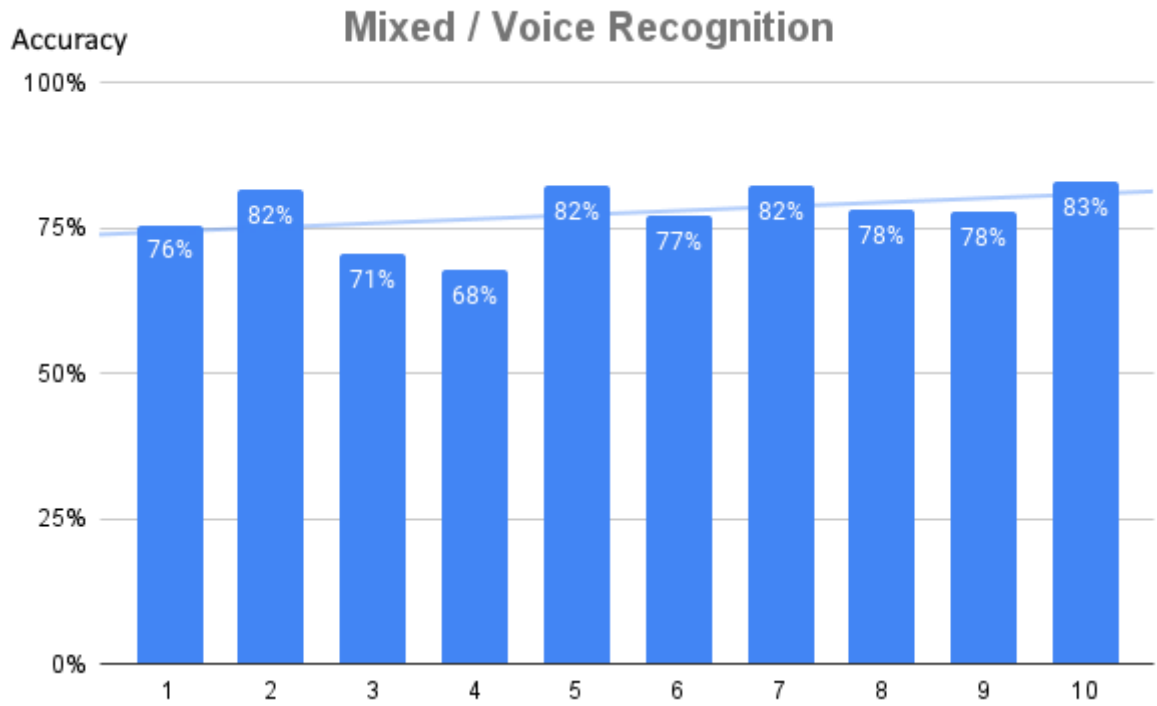


Figure 5.7. (Target Voices 1 to 10) Stargan-VC Mixed / Voice Recognition(Lexical) Statistics

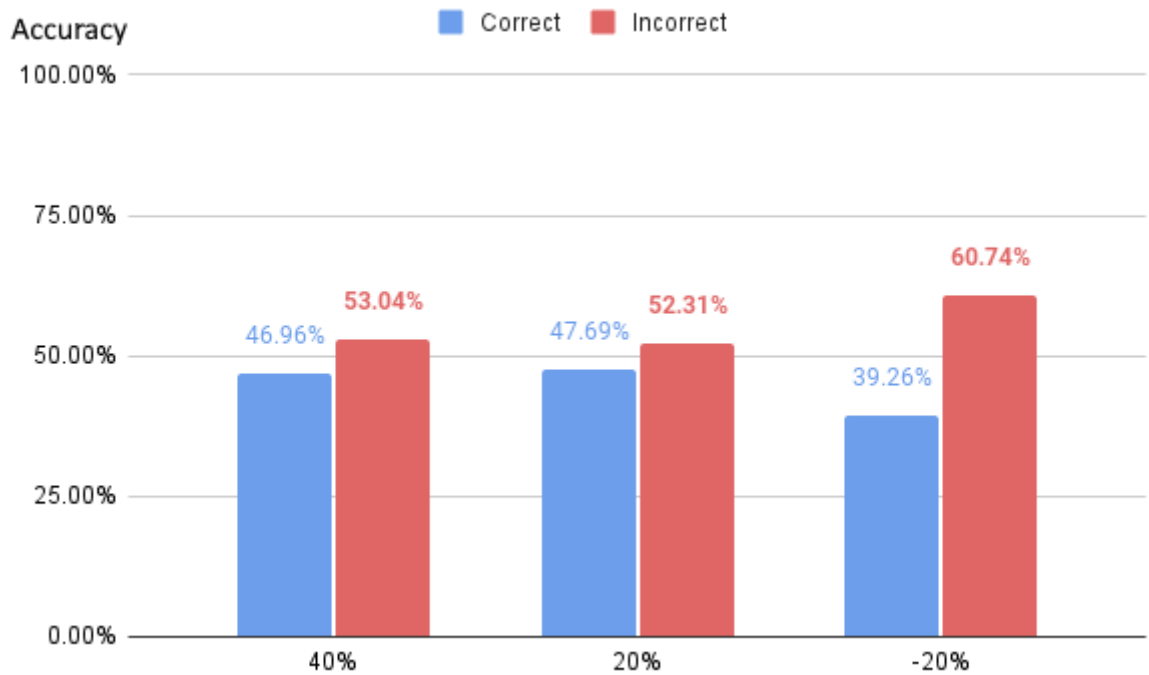


Figure 5.8. (%-20,%20%40 Pitch Shifts) Audacity / Identity Recognition Statistics

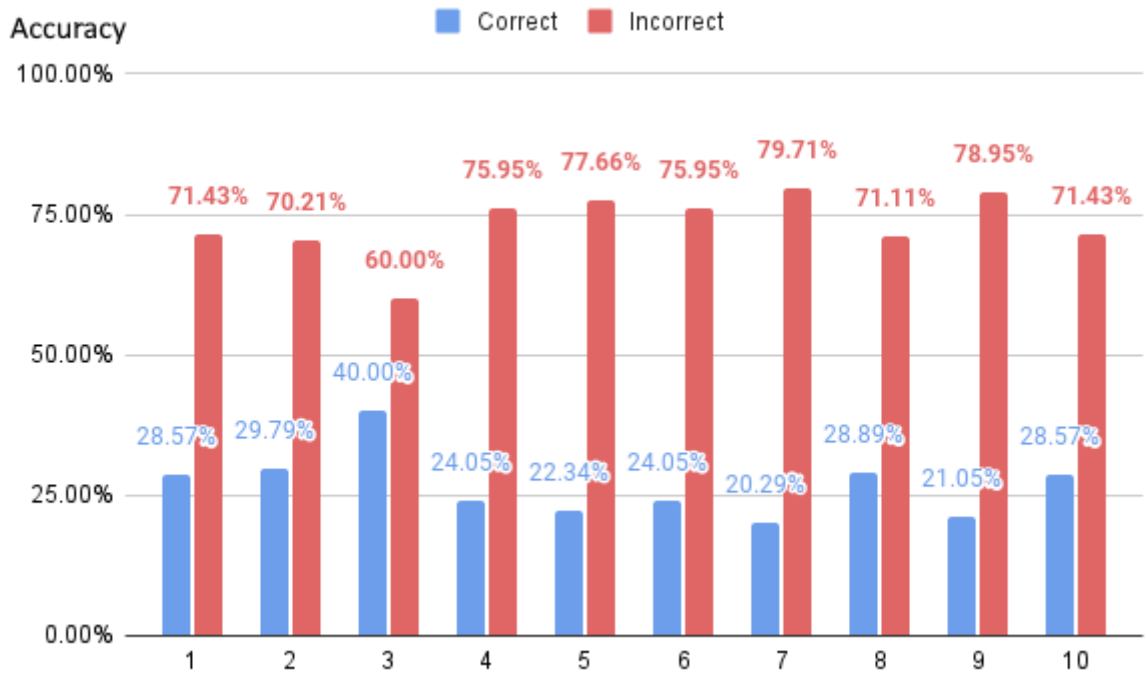


Figure 5.9. (Target Voices 1 to 10) Stargan / Identity Recognition Statistics

5.3. Best of Both Worlds

Looking at both the Voice Recognition Statistics and the Identity Recognition Statistics from our experiments, it is safe to say that Stargan-VC outperformed simple pitch shifting, with a comparable voice recognition rate from the Google API and a significantly lower Identity Recognizability rate conducted by the experiment.

6. Conclusion And Future Work

Looking back to the start of this report, we were challenged with privacy concerns in voice recognition and we knew that there were some working methods to this subject such as pitch shifting and GAN voice conversion. In the end we were able to get a good understanding of pitch shifting and Stargan-VC by comparing them in terms of voice recognition rate and recognizability rate with objective and subjective testings and filled a gap in the literature on the subject. Our results show that Stargan-VC outperforms Pitch Shifting in terms of both Voice Recognition(Lexical) Accuracy and Speaker Identity Recognition Accuracy. To improve the testing we could have focused on more than voice identity recognition such as testing sensitive word scrubbing used by other privacy methods on voice recognition, also talking to the test subjects about the experiment confirmed our theory on how people can pick up on accents to identify the voice of someone. To fix this issue on a certain level, we could have trained the voice conversion model on Turkish speech data-sets or make sure that the experiment participants had similar English accents. But limitations on time and data resources forced us to focus on what we have.

Bibliography

1. Ahmed, Shimaa, et al. "Preech: A System for Privacy-Preserving Speech Transcription." ArXiv:1909.04198 [Cs, Eess], July 2020. arXiv.org, <http://arxiv.org/abs/1909.04198>.
2. Mousa, Allam. (2011). Voice Conversion using Pitch Shifting Algorithm by Time Stretching with PSOLA and Re-Sampling. Journal of Electrical Engineering. 61. 2011. 10.2478/v10187-010-0008-5.
3. Kameoka, Hirokazu, et al. "Stargan-vc: Non-parallel many-to-many voice conversion using star generative adversarial networks." 2018 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2018.
4. Audacity - High Pass Filter / Low Pass Filter. (2021-11-16). Retrieved from https://manual.audacityteam.org/man/high_pass_filter.html https://manual.audacityteam.org/man/low_pass_filter.html
5. Audacity - Pitch Shifting. (2021-11-16). Retrieved from https://manual.audacityteam.org/man/change_pitch.html
6. Zen, Heiga, et al. "LibriTTS: A corpus derived from LibriSpeech for text-to-speech." arXiv preprint arXiv:1904.02882 (2019).
7. Google Speech Recognition (2022). Retrieved from <https://cloud.google.com/speech-to-text>
8. Boudreau, Tim, et al. NetBeans: the definitive guide: developing, debugging, and deploying Java code. " O'Reilly Media, Inc.", 2002.
9. Baken, Ronald J. "The aged voice: a new hypothesis." Journal of Voice 19.3 (2005): 317-325.
10. * Jianwei Qian, Haohua Du, Jiahui Hou, Linlin Chen, Taeho Jung, Xiang-Yang Li, Yu Wang, Yanbo Deng: "VoiceMask: Anonymize and Sanitize Voice Input on Mobile Devices", 2017; [<http://arxiv.org/abs/1711.11460> arXiv:1711.11460].

11. [Word Error Rate Estimation for Speech Recognition: e-WER](<https://aclanthology.org/P18-2004>) (Ali & Renals, ACL 2018)
12. Kobayashi, Kazuhiro, and Tomoki Toda. "sprocket: Open-Source Voice Conversion Software." Odyssey. 2018.