Gradually Changing Gender Attribution of Speech Recordings Using Interpolated Filters

Mira Kjellin

Master's thesis
2022:E59

Lund University

Faculty of Engineering
Centre for Mathematical Sciences
Mathematical Statistics

Master’s Theses in Mathematical Sciences 2022:E59
ISSN 1404-6342

LUTFMS-3455-2022

Mathematical Statistics
Centre for Mathematical Sciences
Lund University
Box 118, SE-221 00 Lund, Sweden

http://www.maths.lu.se/

Abstract

When we listen to human speech, one of the first characteristics we assess is the gender of the speaker. For individuals who suffer from gender dysphoria, this may cause them to be negatively impacted by their voice not matching their gender identity. Therefore, some persons attempt to change their voices with a speech-language therapist. Differences between the average female and male voice have been studied extensively, and the findings are used in therapy to appropriately modify patients' voices. By applying this knowledge to digitally alter patients' voice recordings to sound more like their respective target voices, treatment could be made easier and more effective. This thesis explores the use of interpolated all-pole filters and TD-PSOLA to transform voice recordings of the vowel /a/ to be perceived as more feminine or masculine, while simultaneously attempting to preserve the qualities that make voices sound natural. Additionally, methods of measuring the distance between speech signals using the 2-Wasserstein metric are investigated. An online survey is conducted to evaluate the perceived gender and naturalness of 15 transformations. Results from the survey indicate that the gender attribution of the recordings changes when they are transformed and that the average gender scores correlate with transformation goals. It is found that five out of eleven transformed speech signals were rated as natural by more than 50% of listeners. Furthermore, the ratings imply that several of the transformed signals were as natural sounding as unmodified ones. In conclusion, this method of voice transformation shows promise, but additional research is required before real-world applications can be made.

Contents

1 Introduction

1.1 Background & Motivation

Speech is arguably one of the most important communication tools humans have and has had a major impact on our evolution as a species. We use the information conveyed in voices to understand more about the speaker's physical characteristics, such as age, sex, and health status, as well as their psychological and social characteristics [Laver, 2009]. For instance, listeners can learn about the speaker's personality, occupation, and social status. One attribute that is typically easy and fast to asses from listening to a speaker is their gender [Kreiman and Sidtis, 2011]. When we are born, we are assigned a gender based on our physical appearance: either female or male. However, some people find that their own experience of their gender does not match their assigned gender and therefore identify as transgender, trans for short. The American Psychology Association reports that several transgender persons experience gender dysphoria before or during their transition. They define gender dysphoria as the feeling of unhappiness arising from an inconsistency between how an individual views their own gender and how it appears to others [Association, 2013]. A report by the Swedish National Board of Health and Welfare from 2018 states that approximately 6000 people, or 0.06 percent of the population, were diagnosed with gender dysphoria in Sweden and that the number of diagnoses had increased consistently over the last decades [Socialstyrelsen, 2020].

Some studies have shown that having a voice that does not match the gender identity can have a negative impact on quality of life and personal safety as well as attract unwanted societal attention and potentially reveal birth-assigned gender [Oates and Dacakis, 2015]. For these reasons, some people choose to seek treatment with a speech-language pathologist. Oates and Dackis report that the majority of gender-nonconforming individuals who are seeking treatment are transgender women. This could be because hormone treatments lower the pitch of transmen's voices sufficiently for them not to need further treatment. Nonetheless, even without physical alterations, a study from 2004 found that transwomen were able to increase their fundamental frequency successfully and in a healthy manner [Söderpalm et al., 2004]. There is not a lot of evidence for the effectiveness of vocal treatment, most likely because of a lack of studies due to it being a recent field of research. However, some small studies that have been conducted show promise [Oates and Dacakis, 2015].

Certain digital tools are recommended by therapists for at-home practice, such as pitch measuring programs to use for controlling fundamental frequency [Davies et al., 2015]. One way to make the treatment easier and more successful could be to create more advanced digital tools. For instance, one that lets the patient make voice recordings and transform them in a manner such that the perceived gender of the speaker changes, and have the patient use it to practice with. By making gradual transformations from a source voice to a target voice, the transformations can mimic the patients' natural progress in therapy and serve as a guide for the voice realignment. The method of using linear predictive speech coding, interpolated filters using line spectral frequencies, and pitch-shifting algorithms needed to achieve such transformations of vowel sounds is the focus of this paper.

1.1.1 Previous work

There is still research done on differences between female and male voices and what as well as to what degree acoustic properties affect the gender attribution of speakers [Hardy et al., 2020, Chen et al., 2021]. However, many available tools that change the gender of a speech recording, for instance, Praat [Boersma and Weenink, 2022], modify only certain aspects of the speech signal, such as fundamental and formant frequencies. Hagelborn and Hulme Geber used machine learning and interpolated filters in "Interpolation of perceived gender in speech signals" where two voice signals were morphed using interpolated LPC filters and the pitch shifting algorithm PSOLA [Hagelborn and Hulme Geber, 2020]. However, the resulting interpolations were perceived as unnatural.

1.2 Goals & Purpose

This project aims to explore how transformed voice recordings can be made to sound like the natural progression from a typically female to a typically male voice and vice versa. The overarching goals were to gain a deeper understanding of which aspects that make morphed speech sound synthetic, and use that knowledge to improve the naturalness of the transformations, as well as to define a distance measurement between voiced speech signals to be used to quantify changes or improvements. This was achieved by focusing on the following research questions:

- Can interpolated autoregressive filters together with PSOLA be used to transform voice recordings of the sound /a/ so that the perceived gender of the speaker changes (while maintaining a natural sounding voice)?
- Can the distance between two voice recordings be measured to objectively gauge how the voice progresses throughout treatment?

2 The Human Voice

The following section focuses on human speech – how it's produced, perceived and what acoustic properties it has. This will serve as a foundation for the modeling section ahead.

2.1 Speech Production & Perception

All sound is created by an energy source that generates waves in the surrounding material [Fry, 1979]. For human speech, that energy source is the lungs expelling air [Kreiman and Sidtis, 2011]. Some of the sounds produced while speaking – called voiced sounds – are formed when the vocal folds in the larynx open and close and thereby periodically interrupt and allow the airflow. The rate of the vocal fold vibration depends on the size and stiffness of the vocal folds, where longer and thicker ones naturally vibrate at a lower frequency than smaller ones. Other types of speech sounds are created without vocal fold vibration and are therefore called voiceless or unvoiced. The airflow from the larynx, either modulated by the vocal folds or not, is then passed through and shaped by the acoustic properties of the vocal tract. The vocal tract is shown in Figure 1. By changing the shape of the mouth, for instance by protruding the lips or lowering the jaw, one changes the acoustics of the vocal tract which in turn changes the sound. This is the process with which we can form different vocal sounds known as phonemes.

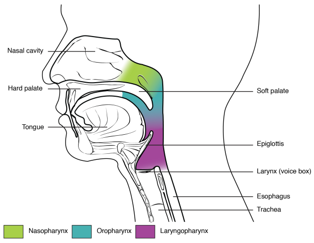


Figure 1: Illustration of the vocal tract [Openstax, 2013]

One popular theory for speech production is the source-filter model [Fant, 1960]. It describes how the vocal folds vibrating create a source signal which is passed through the vocal tract that acts as a resonator, meaning that it naturally oscillates at some frequencies referred to as formant frequencies. The collection of formant frequencies can be called the vocal tract transfer function as it dictates how the energy from the source signal is transferred to the air. The final aspect of the theory is a model for how the sound radiates from the mouth. This separation between the source and filter is a simplification and therefore not a perfect representation of how sound is made, but it is suitable for modeling and conceptual understanding. [Kreiman and Sidtis, 2011] Figure 2 shows the relationship between the source and radiating output according to the model.

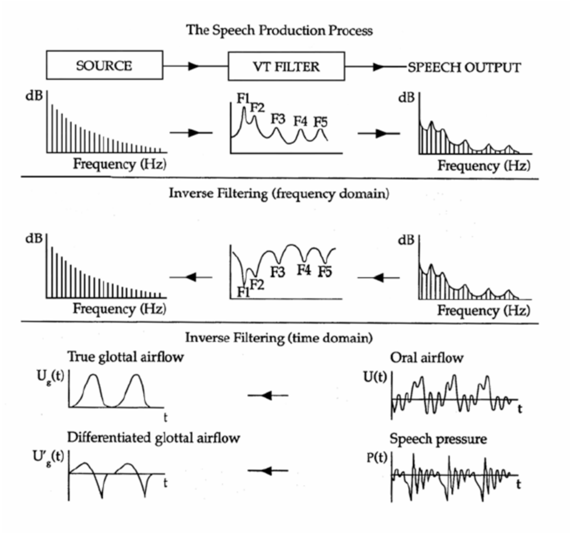


Figure 2: Source-filter decomposition [Hardcastle et al., 2010]

Three main aspects affect how speech sounds: fundamental frequency (F0), which is the vocal fold vibration rate, loudness, and quality. Humans can hear sounds ranging approximately from 20 to 20 000 Hz but are better at hearing differences between low-frequency sounds. Most of the important information for speech sounds has lower frequencies. For instance, F0 generally ranges from 80 to 260 Hz for adults and listeners can easily detect changes as small as 2% of it. Sound intensity, amplitude, and loudness refer to the same thing but are measured differently, where loudness means the subjective perception of the intensity and is commonly measured by comparing intensity to standard thresholds. The concept of quality is discussed in more detail below. [Kreiman and Sidtis, 2011]

2.2 Acoustic Properties

Speech signals can be analyzed both in the time and frequency domain, which can provide different kinds of insights. In the time domain, the signal is depicted as a periodic wave representing the pressure variation in the air over time. An example of this is shown in Figure 3. Therefore, time-domain features are for example signal duration, F0, and variation of period length and amplitude of the vocal fold vibration. On the other hand, the frequency domain focuses on the frequency contents of the signal and can give information about the formants, relative energy in different parts of the spectra, and harmonics. [Kreiman and Sidtis, 2011]

In the frequency domain, the signal can be represented using a power spectrum or a spectrogram. The spectrogram shows how the spectrum changes over time and gives insight into, for example, how the pitch and formants move during the speech. In Figure 4 the spectrogram for the Swedish word "munnen" (in English: "the mouth") is displayed, with visibly separated phonemes. The spectrum gives more details about the frequency contents in each frame, including the finer structure present in voices speech corresponding to the partials – the F0 and its overtones. An example of a power spectrum for the vowel /a/ is shown in Figure 5.

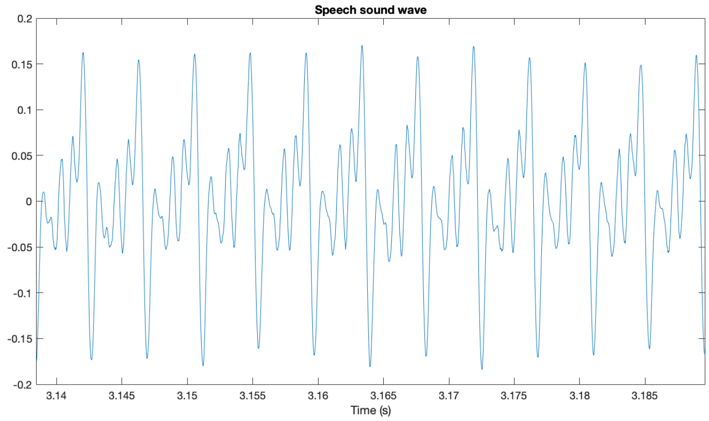


Figure 3: Speech sound wave of the vowel /a/

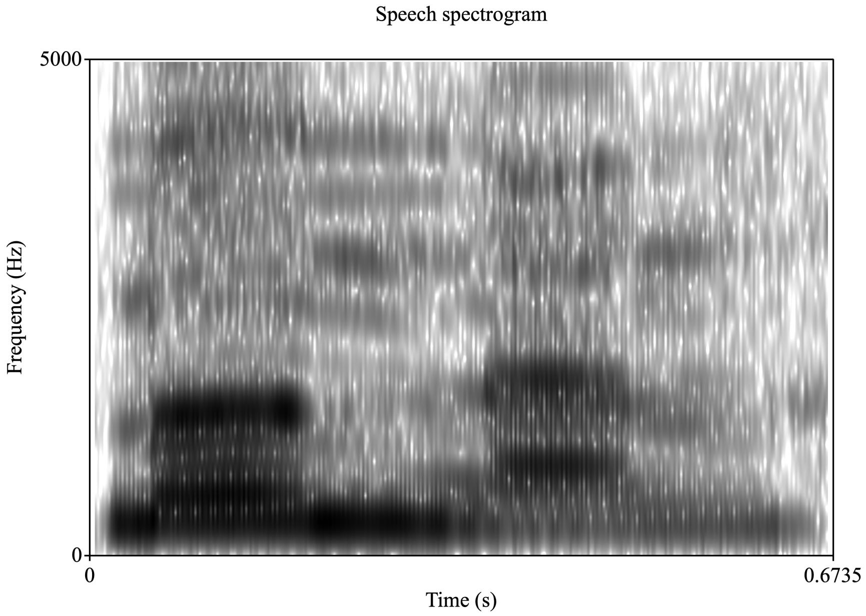


Figure 4: Spectrogram of the Swedish word ”munnen”

2.3 Voice Quality

Depending on how tense or forcefully tight the vocal folds are, how symmetric their vibration is or how much subglottal pressure there is, the voice can perceptually sound in different ways typically called phonation types. The most common phonation used is modal, but there are other nonmodal ones also used, such as falsetto, vocal fry, and breathy phonation. Vocal fry entails that the vocal folds close quickly and remain closed for longer than normal and is commonly associated with creaky

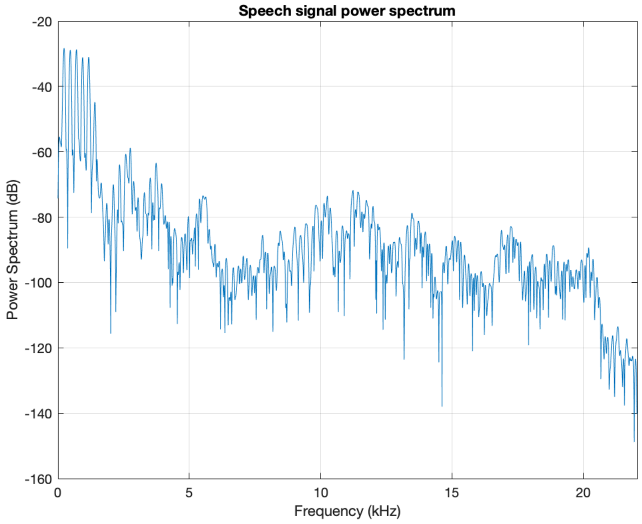


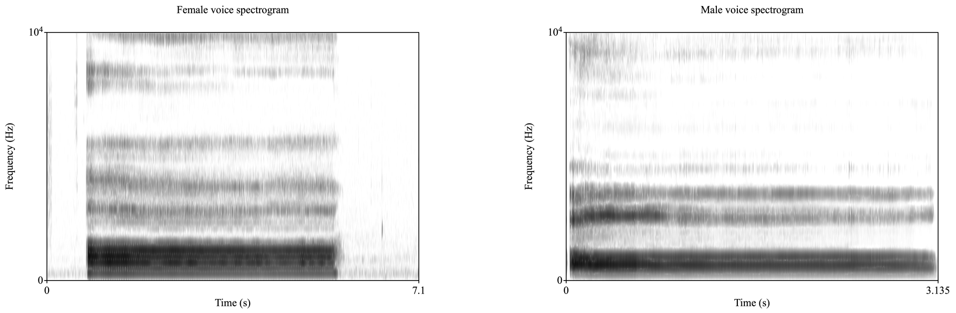
Figure 5: Spectrum of the vowel /a/

voice, while in breathy voices the vocal folds do not close completely allowing airflow to pass through introducing noise into the voice. [Kreiman and Sidtis, 2011]
Other common quality aspects are roughness – where the sound wave amplitude fluctuates, smoothness, brightness, stillness, and vigor, which all can be seen as continuous scales [Colton and Estill, 1981]. Furthermore, the formants, especially the high-frequency ones, contain much information about the speaker and can be associated with so-called personal quality [Kreiman and Sidtis, 2011].

2.4 Difference Between Female and Male Voices

One of the most important factors in gender attribution from voices is the F0 [Hardy et al., 2020]. The F0 of female voices is on average 220 Hz and generally ranges from 150 to 260 Hz while for male voices the average is 115 Hz and ranges from 80 to 170 Hz [Kreiman and Sidtis, 2011]. Another factor that contributes to the perceived gender is vowel formant frequencies, which are commonly lower for men than for women. Two spectrograms showing a comparison between a female and a male voice are displayed in Figure 6. The fundamental formants are visibly lower in the male voice.

Additionally, qualitative measures and how they vary in female and male voices have been studied. For instance, breathiness is commonly associated with female speakers while roughness or vocal fry is usually linked to masculine voices. Intonation, rate of speech, and loudness have also been proposed to differ between women and men. [Hardy et al., 2020]
Oates and Dacakis report that some common goals when trans persons are adjusting their voices in therapy are changing the pitch, formant frequencies, and F0 variability together with adjusting levels of breathiness, sound pressure level, and glottal closure [Oates and Dacakis, 2015].



(a) Female voice spectrogram

(b) Male voice spectrogram

Figure 6: Spectrogram of two speech signals

3 Speech Modeling & Processing

This section focuses on the speech recording transformation and what that process looks like. The goal of the speech modification is to take two recordings and transform one of them to sound more like the other. This is achieved by firstly modeling the resonance characteristics of both voices. Afterward, the filters are interpolated to find a middle ground between the two voices dependent on some interpolation factor between zero and one. Then, the source signal is inverse filtered using the source filters to get a residual corresponding to a glottal source. Next, the residual is pitch-shifted to reach a more similar F0 as the target signal. Lastly, the pitch-corrected signal is filtered using the interpolated filter to mimic the formant characteristics of the target voice. Each step is described further below.

3.1 Speech as a Stationary Stochastic Process

The stationary stochastic process is a useful tool for analyzing and modeling time series. Stochastic models describe processes that contain some degree of randomness, and stationary ones have statistical properties that do not vary over time [Lindgren et al., 2013]. Speech signals are often assumed to be piecewise stationary stochastic processes, meaning they are stationary during short segments, approximately 20 to 30 ms long [Tyagi et al., 2006], and can therefore be processed as such. The length during which the speech is stationary or quasi-stationary depends on what sound it is. Vowels, for instance, can be stationary for 30 – 80 ms while plosives, such as “p”, only are for less than 20 ms [Tyagi et al., 2006]. Based on the source-filter model for speech production, speech can be approximated by estimating the vocal tract resonance properties with linear filters.

3.2 Linear Predictive Coding

Linear predictive coding analysis (LPC) is a common method for modeling speech. The technique encodes speech waves through several parameters forming an all-pole filter [Atal and Hanauer, 1971]. Poles of the transfer function represent resonances of the vocal tract and zeros represent anti-resonances. As the poles are considered more perceptually important it is often not necessary to include zeros in the filter. The goal of LPC for speech signals is to model the spectral envelope – a smooth function encapsulating the peaks of the spectrum, and thereby separating the vocal tract characteristics from the glottal source. There are some drawbacks of using LPC for envelope estimation, such as difficulty defining the proper model order and distortion of the low-frequency part of the spectrum [Villavicencio et al., 2006]. An alternative method is to first estimate an envelope, using methods such as the True [Villavicencio et al., 2006] or CheapTrick Envelope [Morise, 2015], and use that envelope to estimate a filter. Both of these envelope estimation methods are based on cepstral smoothing of the amplitude spectrum. The cepstrum is the inverse Fourier transformation of the log power spectrum [Noll, 1967] and can be used to identify periodicities in the frequency domain, for instance, harmonics present in voiced speech. To retrieve a smoothed spectral shape from the cepstrum, a window function, referred to as a lifter function, is applied to the cepstrum, followed by a Fourier transform. The CheapTrick envelope uses a rectangular lifter function (in the frequency domain), defined as $L_s(n) = \mathrm{sinc}(f_0 n) = \frac{\sin(\pi f_0 n)}{\pi f_0 n}$ [Morise, 2015]. Because the lifter function is valued at 0 at multiples of the F0, it cancels the harmonic component of the spectrum and thereby smoothing it out.

$$\begin{aligned} S(w) &= \mathcal{F}[s(t)] \\ c(n) &= \mathcal{F}^{-1}[log(S(w))] \\ S_{\text{smooth}}(w) &= \mathcal{F}[c(n)L_s(n)] \end{aligned}$$

Because Hagelborn and Hulme Geber found that the CheapTrick envelope performed the best of all tested smoothing-techniques [Hagelborn and Hulme Geber, 2020], it was applied here as well.

3.2.1 Estimating Filters

The speech signal is segmented into short overlapping windows, of circa 50-100 ms, assumed to be quasi-stationary. For each segment, the CheapTrick Envelope is estimated as described above by liftering the cepstrum using an F0-adaptive window function, i.e. a sinc function periodic with the mean pitch period. By marking each pitch period in the speech signal, the F0 can be estimated using the average distance between the peaks. The inverse Fourier transform is then applied to the spectral envelope to attain the biased autocorrelation function. This is later employed to estimate the autoregressive filter coefficients using Levinson-Durbin recursion. Figure 7 shows the comparison between the power spectrum, CheapTrick envelope, and frequency response of filter for a signal segment.

A model order of 50 is used for the filters because it returned the best results. The source signal is inverse filtered and later regularly filtered in overlapping segments and added back together using an overlap-add procedure with triangular window functions.

3.3 Interpolating Filters

Interpolation of LPC parameters is a practice commonly used for speech coding and transmission of speech, where it enables the use of slower frame rates [Paliwal, 1995]. Since slow rates can lead to large differences in LPC parameters in neighboring frames the technique is used to close the spectral gaps by inserting short frames with interpolated parameters. The interpolation can be done for LPC parameters in different representations, such as LPC coefficients, reflection coefficients, and line spectral frequencies, LSF. A comparison between interpolations of eight such representations found that LSF led to the least amount of spectral distortion [Paliwal, 1995]. Similar results were established by Islam [Islam, 2000].

Line spectral frequencies, known also as line spectral pairs, are representations of LPC filter parameters with properties, such as well-behaved filter stability preservation, that make them robust for quantizations [Soong and Juang, 1984]. The representation is based on the m^{th} order LPC filter:

$$A_m(z) = 1 + a_1 z^{-1} + \ldots + a_m z^{-m} \tag{1}$$

The symmetric and anti-symmetric polynomials P(z) and Q(z) are defined as:

$$\begin{aligned} P(z) = A_m(z) + z^{-(m+1)} A_m(z^{-1}) \\ Q(z) = A_m(z) - z^{-(m+1)} A_m(z^{-1}) \end{aligned}$$

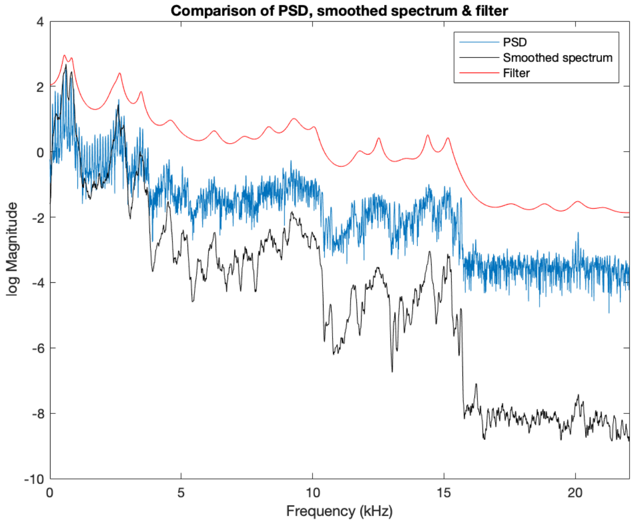


Figure 7: Example of power spectrum, smoothed spectrum using CheapTrick envelope and filter frequency response.

where

$$A_m(z) = \frac{P(z) + Q(z)}{2} \tag{2}$$

This is done by extending it to a $(m+1)$ order filter and letting the $(m+1)$th reflection coefficients be 1 or -1. An interpretation of $P(z)$ and $Q(z)$ is that they represent the vocal tract when the glottis is open versus closed (Sugamura & Itakura, 1986). All roots of $P(z)$ and $Q(z)$ are on the unit circle and can therefore be expressed as frequencies by letting $z = e^{j\omega}$ where the angles $\omega, 0 < \omega_i < \pi, \forall i = 1, ..., m/2$, are known as line spectrum pair frequencies. Because $P(z)$ and $Q(z)$ have real coefficients they have complex conjugate roots which makes the computations easier and you only need to find roots in the upper half unit circle. If $1/A(z)$ is stable, meaning that all poles lie within the unit circle, the line spectrum frequencies are ordered and alternate between being the angle of pole of $P(z)$ and $Q(z)$. Two sets of line spectrum frequencies ω^s and ω^t, representing the LPC parameters of the source and target filters, can be interpolated using factor $\tau \in [0, 1]$ as:

$$\Omega^\tau = (1-\tau)\omega^s + \tau\omega^t$$

This set of line spectral pair frequencies Ω^τ can then be converted back to updated polynomials $P(z)$ and $Q(z)$ using

$$P(z) = \left(1 - z^{-1}\right) \prod_{k=2,4,\ldots,p} \left(1 - 2z^{-1}\cos\Omega_k^\tau + z^{-2}\right) \tag{3}$$

$$Q(z) = \left(1 - z^{-1}\right) \prod_{k=1,3,\ldots,p-1} \left(1 - 2z^{-1}\cos\Omega_k^\tau + z^{-2}\right) \tag{4}$$

respectively, and thereafter to $A(z)$ using (2).

In Figures 8 and 9, are examples of the poles and frequency responses of several interpolated filters between a source and target filter.

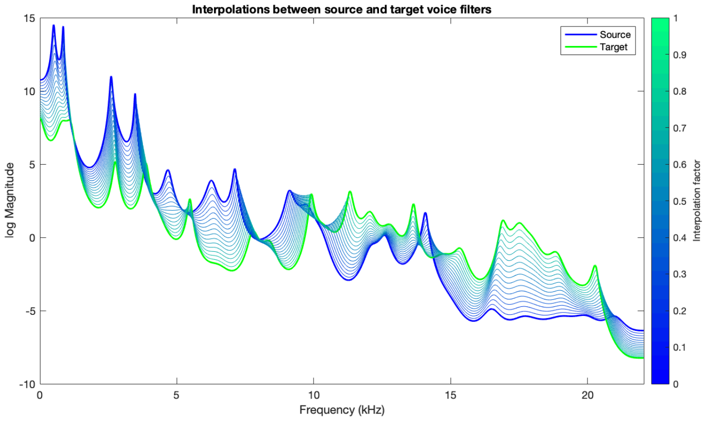


Figure 8: Interpolated filters between source and target filters, depicted in blue and green.

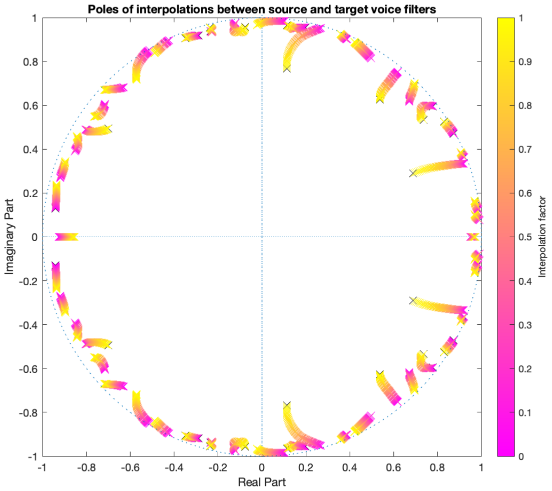


Figure 9: Poles of interpolated filters between source and target filters

3.4 Time Domain Pitch Synchronous overlap-add

To shift the pitch of the speech signal, the time domain pitch synchronous overlap-add (TD-PSOLA) method is applied. It is a popular and widely used algorithm because of its simplicity, quality and computational efficiency [Toma et al., 2010]. The algorithm works by excising a sequence of pitch synchronous overlapping frames from the waveform, then modifying this stream of short-term signals and finally adding them back together using an overlap-add synthesis procedure [Hamon et al., 1989]. Since it operates on pitch periods, it is only performed on voiced sections of the speech signal. If the method is executed correctly the pitch-shifting does not affect the spectral envelope and can therefore maintain the original vocal characteristics. One drawback of TD-PSOLA is that it can introduce perceptible distortion of the signal and that this distortion increases with the level of pitch modification [Longster, 2003]. Some studies indicate that pitch shifting of up to 50% can produce acceptable results without losing significant quality and naturalness [Longster, 2003]. Because the algorithm moves all frequency content when changing the pitch, it was found that substantial intervals of the highest frequencies were lost when decreasing it. It was naturally most noticeable when there was a large pitch difference.

3.4.1 Process

Firstly, each pitch period is marked by a pitch marker in the original speech signal waveform. Then the signal is decomposed into short, overlapping segments centered around each pitch marker, and these are subsequently modified to fit a set of target pitch markers, corresponding to the desired pitch. A mapping function synchronizes the original and target pitch markers and resampling is used to modify the length of each segment to fit in the synthetic signal. Additionally, some window function is used in the overlap-add process, for example, Hamming or triangular windows. [Hamon et al., 1989]

3.5 Filter and Interpolation Evaluation

To evaluate the performance and issues with the transformation pipeline, interpolated signals were compared to both source and target speech signals using power spectrums and spectrograms. Furthermore, the F0 variance, shimmer - relating to variation of sound wave amplitude, jitter – relating to variation of sound wave length, and harmonic-to-noise ratio, HNR, were used to analyze the signals. Local shimmer can be calculated as in 5, local jitter as in 6 and HNR as in 7 [Teixeira et al., 2013]. Simple sine waves were transformed to identify problem areas with regards to introducing artefacts such as buzz and ringing in both the filtering and PSOLA processes.

$$\text{shimmer } = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}\left|A_i - A_{i+1}\right|}{\frac{1}{N}\sum_{i=1}^{N}A_i} \times 100 \tag{5}$$

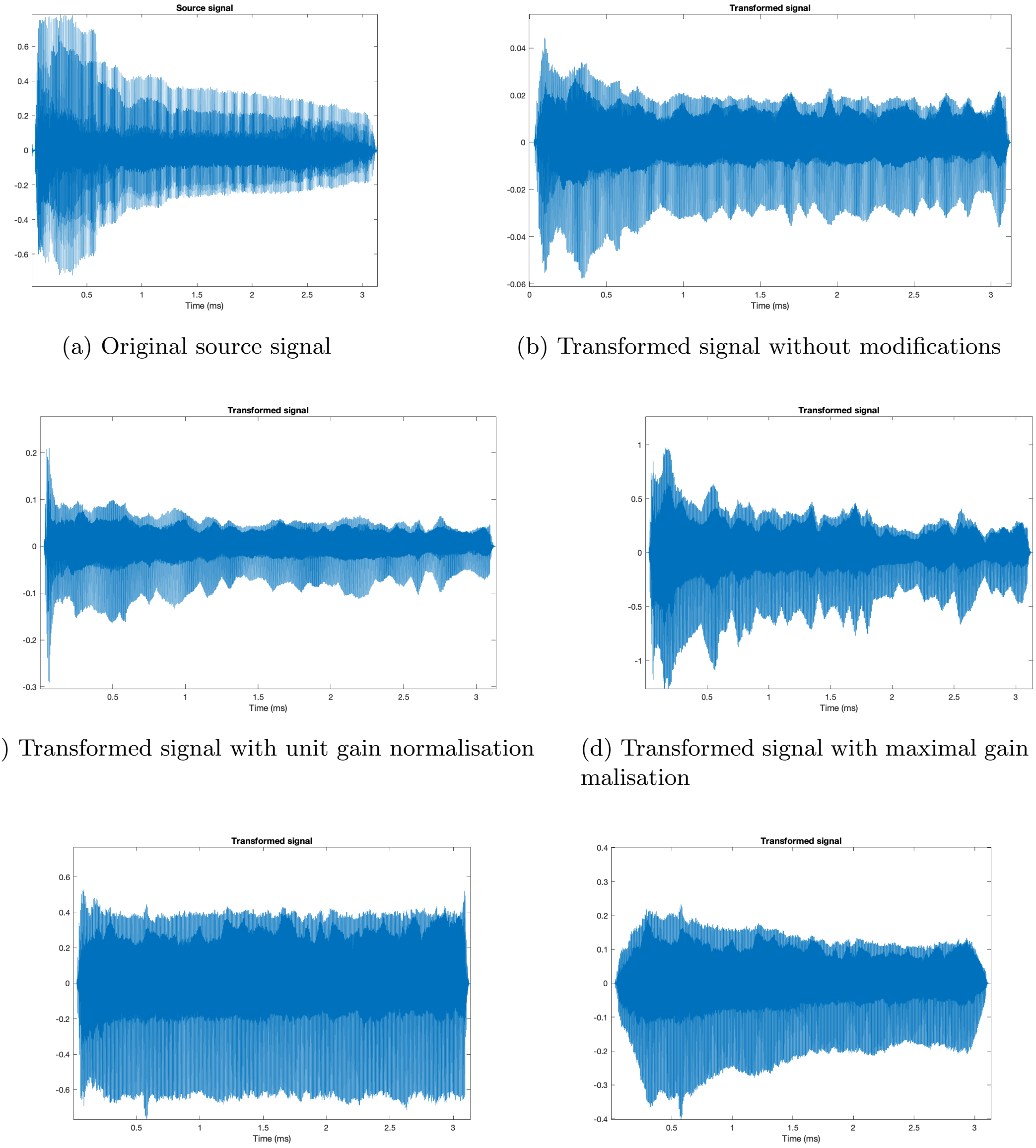
$$\text{jitter } = \frac{1}{N-1}\sum_{i=1}^{N-1}\left|T_i - T_{i-1}\right| \tag{6}$$

$$\text{HNR}(dB) = 10 \times \log_{10}\frac{r_x'\left(\tau_{\max}\right)}{1 - r_x'\left(\tau_{\max}\right)} \tag{7}$$

Two of the most significant findings were that the volume of the interpolations varied more than natural recordings and that the F0, as well as a few overtones, were suppressed in the synthetic signals. These issues are discussed further below.

3.5.1 Sound Wave Amplitude Variation

The signal was filtered in longer segments to reduce the “buzz” arising from the varying energy levels of each filter, creating segments with varying amplitude. This variation of sound wave amplitude of consecutive is called shimmer and is expected to a certain level in a healthy voice [Kreiman and Sidtis, 2011]. It is therefore also expected in the source or target speech recordings which can influence the resulting synthetic signal.



(a) Original source signal

(b) Transformed signal without modifications

(c) Transformed signal with unit gain normalisation

(d) Transformed signal with maximal gain normalisation

(e) Transformed signal with sound wave standardisation

(f) Transformed signal with sound wave standardisation and scaled volume

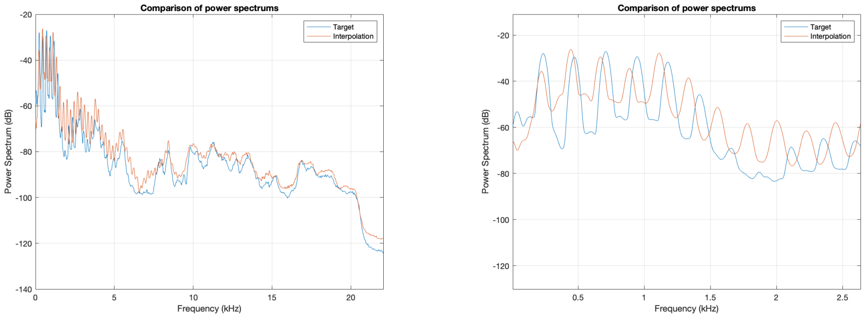
Figure 10: Source and transformed signal sound waves

However, it was found that the volume then contained a perceptible amount of slow amplitude variation exceeding the levels present in the original signals, causing the speech to sound ringing. Figure 10 depicts such a transformed sound wave compared to the source signal. Several attempts were made to combat this: unit gain at zero frequency, maximum peak height at the same power, and sound wave standardization of each segment. First, each filter was normalized to have unit gain at zero frequency, but this proved to have little to no effect. See Figure 10 (b) for reference. Normalizing the filter by max value of its frequency response had a similar, or worse, result. Finally, each sound wave segment was standardized to have the same amplitude which gave the least volume variation. However, some variation is required for a natural sound and the signal was therefore reshaped to mimic the volume of the original speaker. This was done by comparing the standard deviation of short segments of the signals and scaling the synthetic one accordingly. The resulting transformations from all of the aforementioned methods are displayed in Figure 10. The amplitude-standardized transformation in (f) seems to have the volume variation closest to the original sound wave.

Further investigations would be required to identify the underlying issue with the filters and why they have varying gains to such a high degree. The problem could potentially be solved by altering the filter estimation and spectrum smoothing processes. However, the ultimate goal of the transformation is to transform more complex speech sounds, such as complete words or phrases. In those speech signals, each phoneme is only sustained for short periods and a variation might not be as evident as in the case of long vowel sounds.

3.5.2 Low-Pass Filters and Augmentation of Partials

Additionally, it was found when comparing the transformed and target signals' spectrums that the first few partials were negatively impacted by the filtering, as seen in Figure 11 (b). Narrow bandpass filters were used to amplify the first four harmonics individually to better match the target voice signal. The difference is shown in Figure 12. Also visible in Figure 11 is that, especially for male-to-female voice transformations, the spectrum of the resulting signal had slightly higher intensities for frequencies above 2 kHz. To combat this, two separate lowpass filters – one from approximately 2 kHz and one from 20 kHz – were used to decrease the high-frequency contents with a few decibels. These values were chosen specifically to address each transformation and therefore varied slightly for different signals. The spectrums for the filtered signal, using both narrow bandpass and lowpass filters, and the target signal are compared in Figure 13.



(a) Spectrum comparison

(b) Spectrum comparison for low frequencies

Figure 11: Comparison between spectrum for transformed and target signal. Transformation done from a male to female voice using interpolation factor $\tau = 0.9$.

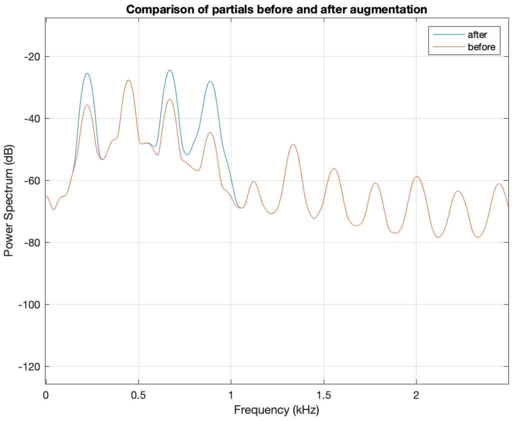
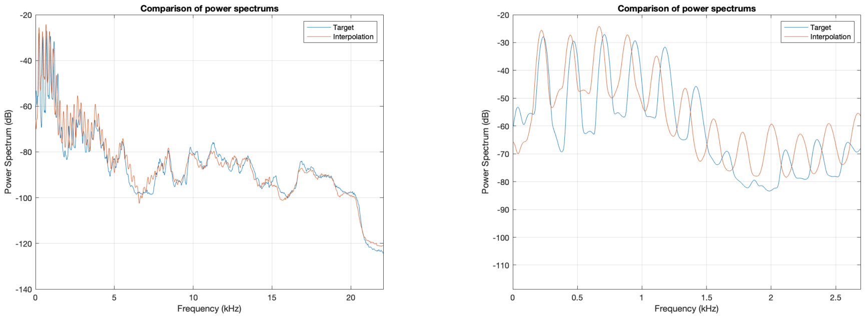


Figure 12: Augmented partials of synthetic signal, before and after filtering with bandpass filters.



(a) Spectrum comparison for filtered signals

(b) Spectrum comparison for low frequencies for filtered signals

Figure 13: Comparison between spectrum for transformed and target signal. Transformation done from a male to female voice using interpolation factor $\tau = 0.9$ and filtered using bandpass and lowpass filters.

4 Spectral Distance Measurement

Previous sections have dealt with the first research question, regarding the speech transformation, and the second, concerning the measurement of vocal differences, will be treated below. Several acoustic aspects of the voice contribute to its characteristic sound. Some of the most important ones are fundamental and formant frequencies, making it appropriate to compare the signals' spectra to measure their differences. However, simply comparing two spectra using a common distance measurement, such as the Euclidean norm, does not produce meaningful results because it only compares powers for the same frequencies. This does not translate well to the power spectrum since the harmonics and energy bands are located in different positions for, and during, all speech sounds. Therefore a different way of measuring distances must be implemented. As a basis for quantifying the distance used here is the concept called optimal mass transport, with many previous applications in signal processing and machine learning.

4.1 Optimal Mass Transport

Also known as Earth Mover's Distance, optimal mass transport is used to compare signals and images in regards to both intensity and spatial information. The main purpose of OMT is to find the most efficient transport plan or map, given a certain cost function, from one mass distribution to another. Hence, the name earth mover's distance, since it describes how one pile of earth can be moved into another pile most efficiently. [Kolouri et al., 2017]

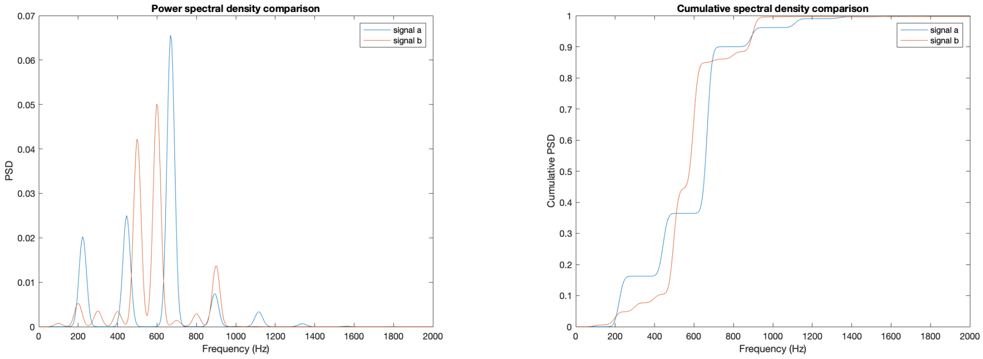
4.2 Wasserstein Distance

While OMT defines the best way to arrange one distribution to another, the p-Wasserstein metric measures the distance between distributions [Kolouri et al., 2017].

The normalised power spectral densities, I_0 and I_1, of two one-dimensional speech recordings, S_0 and S_1, are defined on Ω_0 and Ω_1. The cumulative densities for both of them, $F_0(x) \geq 0$ and $F_1(x) \geq 0$, are defined on $[0, 1]$. The p-Wasserstein distance then has a closed form solution:

$$W_p(I_0, I_1) = \left(\int_0^1 \left| F_0^{-1}(z) - F_1^{-1}(z) \right|^p dz \right)^{1/p}$$

Where $F_i^{-1}(z) = \inf\{x \in \Omega : F_0(x) \geq z\} \forall i$ is the pseudoinverse (inverse) to $F_i(x)$. In Figure 14 two power spectrums and cumulative densities corresponding to two different signals are shown.



(a) Comparison of power spectral densities for signal A and B

(b) Comparison of cumulative spectral densities for signal A and B

Figure 14: The normalised PSD and CDF for two signals

The difference between measuring distances using the p-Wasserstein metric as compared to, for example, the Euclidean norm is that it takes the spatial distance into account as well. This can be demonstrated using a simple example of two Gaussian distributions where the mean values start out being the same but one of them is shifted further and further to the right. For each step, the distance between the distributions is measured using 2-Wasserstein and the Euclidean norm. The plotted distributions and distance measurements are displayed in Figure 15. As the distributions move farther apart along the x-axis, the 2-Wasserstein metric increases linearly, because each mass unit has to be moved further, while the Euclidean norm stabilizes as soon as the distributions no longer have mass in the same x-coordinates.

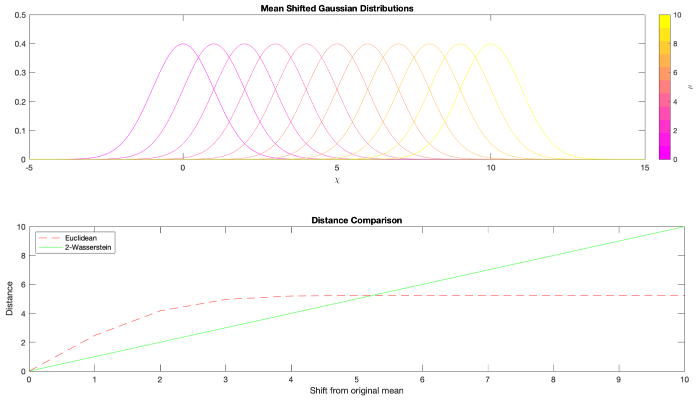
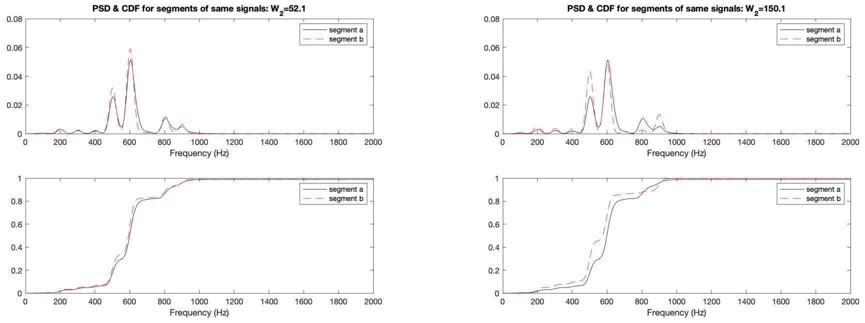


Figure 15: Distance comparison of mean-shifted Gaussian distributions between 2-Wasserstein and Euclidean norm

4.3 2-Wasserstein for Speech Signals

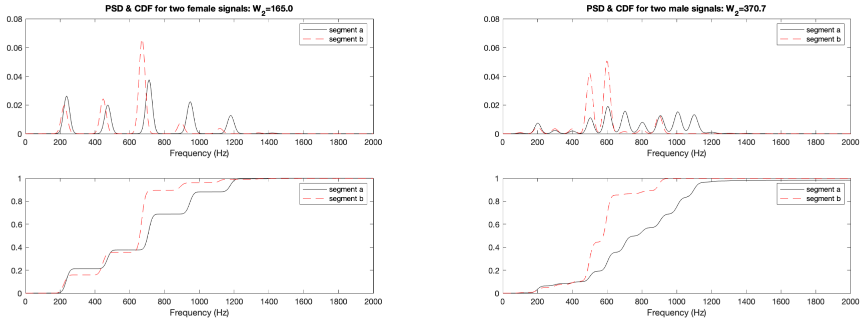
To understand how the distance measurement works on speech signals, a few different examples were tested. Figure 16 shows a comparison of power spectrums for two segments from the same speech signal. In (a), the segments are overlapping, and should therefore be very similar, while the segments in (b) have no overlap and might be more different. This is reflected both visually and in the 2-Wasserstein distance, which is larger for (b). Further, in Figure 17 two speech signals from speakers of the same gender with similar fundamental frequencies are compared in the same manner. For the two female voices in (a), the distance is quite small, which is not surprising given the speakers are sisters of similar age. The two male speech signals in (b) have visibly different spectrums and also have a larger 2-Wasserstein distance. Lastly, Figure 18 displays a comparison of distances between vowels from the same speaker. In (a) speech recordings of the same vowel - a - are compared, and the distance is evidently quite small. The plots in (b) on the other hand, compare recordings of vowels a and e, and exhibit a larger spectral distance.



(a) Comparison of PSD and CDF for two segments with overlap from the same signal

(b) Comparison of PSD and CDF for two segments with no overlap from the same signal

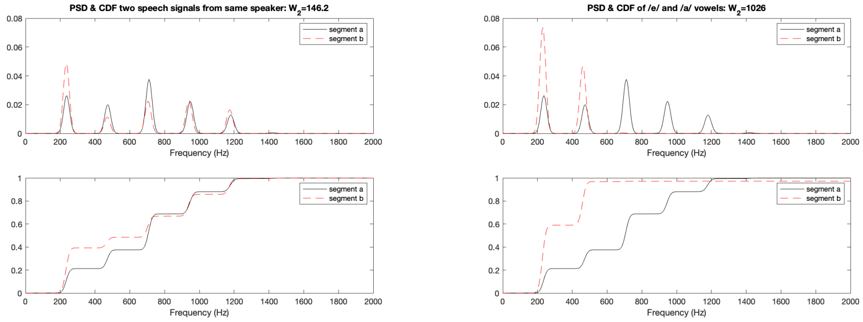
Figure 16: Power spectrum and cumulative density as well as 2-Wasserstein distance for segments from same signal.



(a) Comparison of PSD and CDF for two female voices

(b) Comparison of PSD and CDF for two male voices

Figure 17: Power spectrum and cumulative density as well as 2-Wasserstein distance between two female and two male voices.



(a) Comparison of PSD and CDF for vowel sounds /a/ from the same speaker

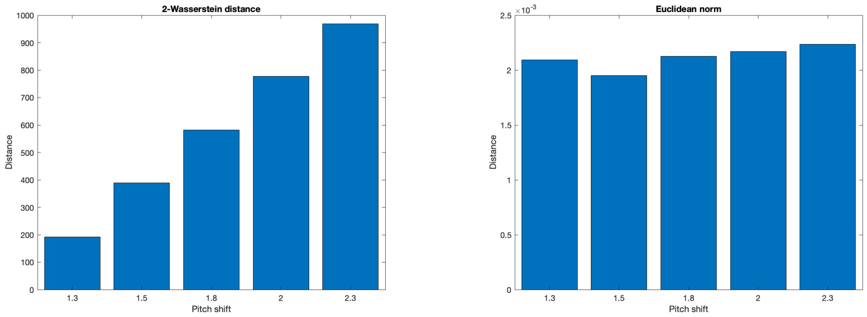
(b) Comparison of PSD and CDF for vowel sounds /a/ and /e/ from the same speaker

Figure 18: Power spectrum and cumulative density as well as 2-Wasserstein distance between two recordings from the same speaker.

4.4 Distance Between Interpolated Signals

One interesting application of this distance measurement for this project was the interpolated speech signals. Since the ultimate goal of the distance metric is to track progress in speech-and-language therapy, it was relevant to find out how the distance behaved for progressively transformed signals. Therefore, how the distance between the transformed signal and the source and target signals changed for various interpolation factors was tracked.

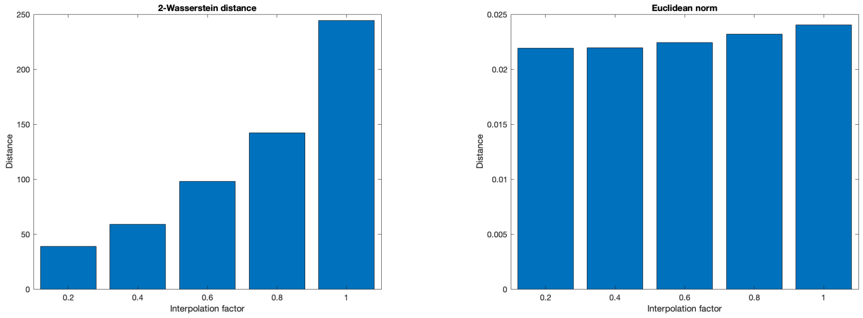
First, the two main modifications – pitch-shifting and filtering – were tested in isolation. Figures 19 and 20 display the distances from a modified signal to the original signal, using both 2-Wasserstein and Euclidean norm. The filters were achieved by interpolating the original signal's filters with target filters using increasing interpolation factors. Clearly, the 2-Wasserstein metric returns increasing distances for both pitch-shifted and filtered signals, while the Euclidean norm does not provide definite differences between signals. F0 seems to influence the distance the most since the distances increase more for each pitch shift compared to the filtering.



(a) 2-Wasserstein distance between original and pitch-shifted signal

(b) Euclidean norm of difference between original and pitch-shifted signal

Figure 19: Comparison of 2-Wasserstein metric and Euclidean norm for pitch-shifted speech signals



(a) 2-Wasserstein distance between original and filtered signal using different interpolation factors

(b) Euclidean norm of difference between original and filtered signal using different interpolation factors

Figure 20: Comparison of 2-Wasserstein metric and Euclidean norm for filtered speech signals

Secondly, full transformations using increasing interpolation factors were compared to the source and target speech signals. The resulting bar graphs showing distances to source and target signals for three separate sets of interpolations are displayed in Figures 21, 22 and 23. The distances do not reflect interpolations linearly transforming from the source signal to the target signal, which would entail decreasing distances from the target signal and increasing distances from the source signal. On the contrary, the distance between the source and target signal is for all examples exceeded by distances to some interpolations. As previously mentioned, the F0 and first harmonics were affected, in some cases distorted, by filters and pitch-shifting. Because the low-frequency content - especially the first harmonics – is emphasized in the power spectrum for speech signals, it also has the most effect on the distance measurement. Therefore, if the first harmonics are altered it could have a significant impact on the measured distance, which could potentially be a reason for the large distances.

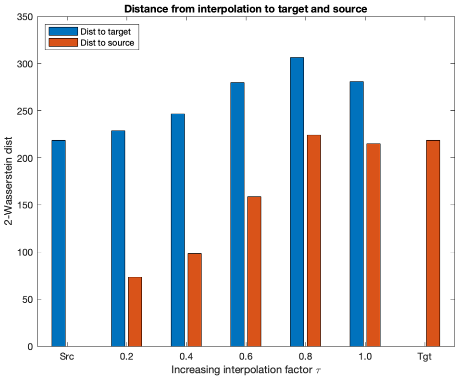


Figure 21: 2-Wasserstein distances from interpolations to source and target signals for interpolation factors 0.2, 0.4, 0.6, 0.8 and 1.0

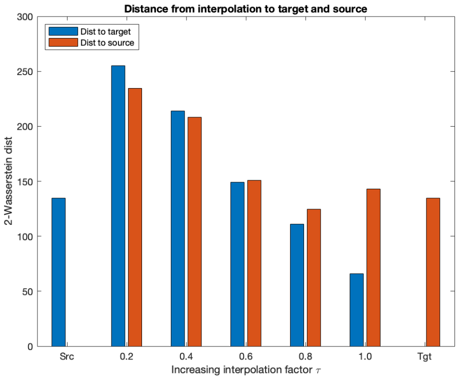


Figure 22: 2-Wasserstein distances from interpolations to source and target signals for interpolation factors 0.2, 0.4, 0.6, 0.8 and 1.0

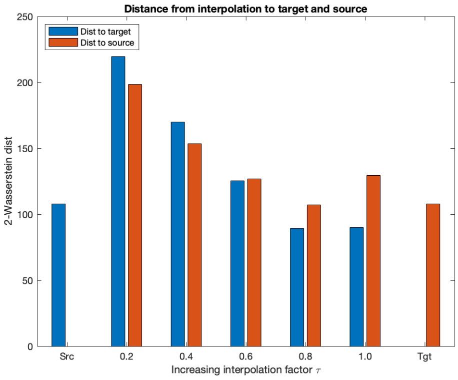


Figure 23: 2-Wasserstein distances from interpolations to source and target signals for interpolation factors 0.2, 0.4, 0.6, 0.8 and 1.0

5 Speech Transformation Evaluation: Survey

Because perception of gender in speech is complex and subjective the recordings needed to be reviewed by listeners to evaluate the result of the transformations. Furthermore, listeners were also able to evaluate the perceived naturalness of the recordings.

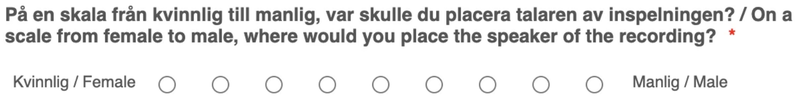
5.1 Survey Process

An online survey was conducted using the website Jotform [Jotform, 2022] featuring 15 speech recordings derived from seven different speakers, three women, and four men. Some of them were not modified, to serve as an anchor for the naturalness ratings. Furthermore, three of the recordings were included twice to measure the internal consistency. All recordings included are displayed in Table 1. The survey was directed at persons older than 18 and with self-proclaimed normal hearing.

Table 1: Original recordings used for transformations in survey

Source speaker	**Target speaker**	**Interpolation factor**
Female 1	Male 1	0.2
Female 1	Male 1	0.6
Female 1	Male 1	0.8
Male 2	Female 1	0.4
Male 2	Female 1	0.6
Male 2	Female 1	1.0
Male 3	Female 2	0.2
Male 3	Female 2	0.4
Male 3	Female 2	0.8
Female 3	-	0
Male 1	-	0
Male 4	-	0

For each recording, the listener was asked to rate the speaker on a digital VAS scale from female to male with nine different levels, and to answer if they experienced the recording as synthetic or natural. If they answered synthetic, they were prompted to describe the recording in words from a list including robotic, ringing, nasal, noisy, and other, with an option to insert a word. A question that checked whether the participant was wearing headphones and if they experienced any technical issues were included to avoid erroneous replies. 55 people had answered the survey at the time of collecting all responses.



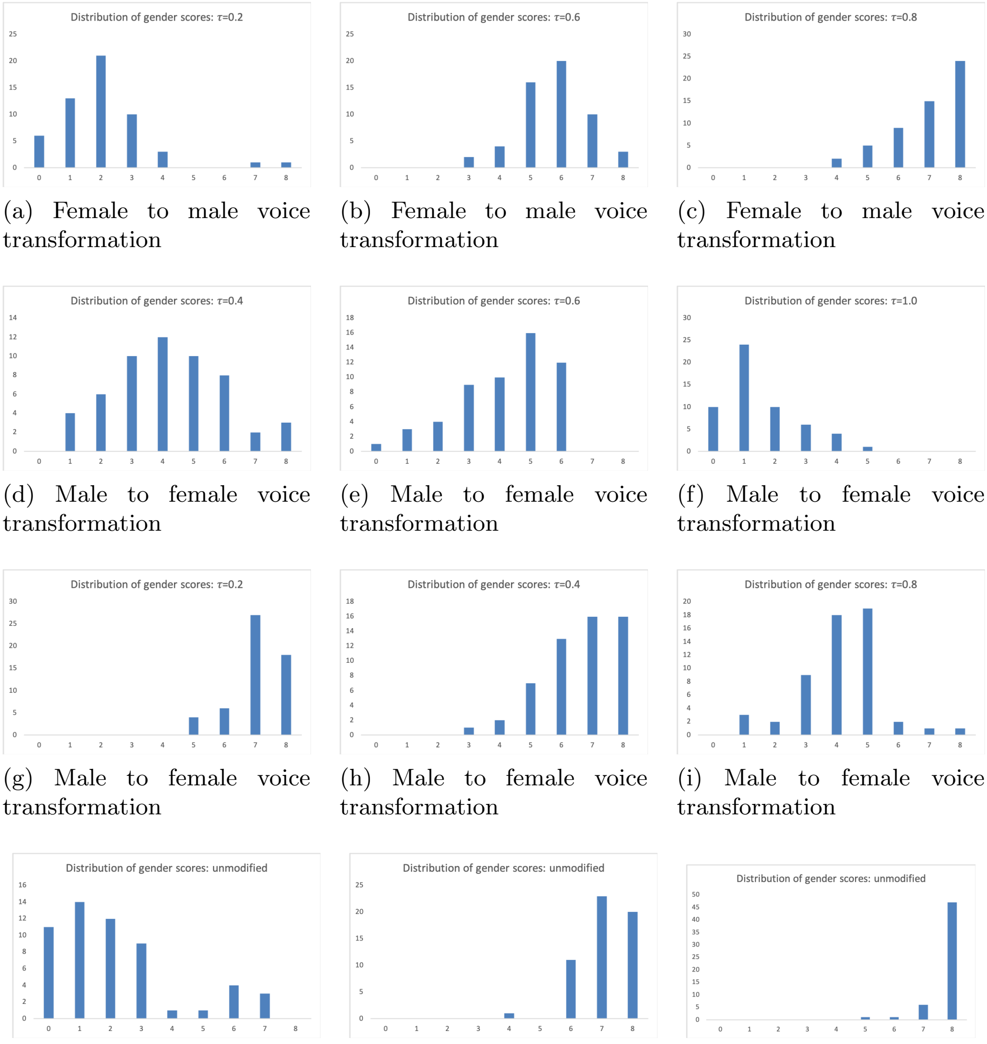
På en skala från kvinnlig till manlig, var skulle du placera talaren av inspelningen? / On a scale from female to male, where would you place the speaker of the recording? *

Kvinnlig / Female ○ ○ ○ ○ ○ ○ ○ ○ ○ Manlig / Male

Figure 24: Question prompting listener to rate gender attribution for each recording

5.2 Gender Score Distributions

The scale from female to male speaker corresponds to a scale of values from 0 to 8, where 0 represents a female-sounding voice and 8 a male-sounding voice. The distributions of gender scores for all (unique) recordings are shown in Figure 25, where they are grouped by the voices that were used as source and target signals. For (a) to (c), which are transformations from a female to a male voice, it is clear that the distributions move from more female to more male which corresponds to the interpolation factors increasing. In (d) to (f), the trend is not as evident but the variation is larger for interpolation factors 0.4 and 0.6 than for 1.0 where most listeners rated it as on the female spectrum. For the final transformation set, (g) to (i), which is also from a male to a female voice, the distributions move towards the female end of the scale as the interpolation factor increases. The average gender scores of the transformed speech recordings are shown in Figure 26. For the unmodified recordings in (j) to (l), listeners perceived them as similar to the true gender of the speakers. However, the variation in ratings of the female and first male voice indicates that it could have been difficult to assign gender in this setting.



(a) Female to male voice transformation

(b) Female to male voice transformation

(c) Female to male voice transformation

(d) Male to female voice transformation

(e) Male to female voice transformation

(f) Male to female voice transformation

(g) Male to female voice transformation

(h) Male to female voice transformation

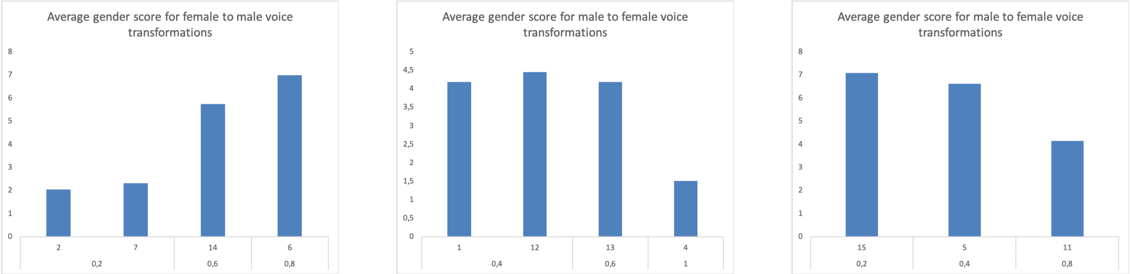
(i) Male to female voice transformation

(j) Female original voice

(k) Male original voice

(l) Male original voice

Figure 25: Gender scores for all recordings grouped by transformation set. τ equates the interpolation factor used in the transformation.



(a) Female to male voice transformation average

(b) Male to female voice transformation average

(c) Male to female voice transformation average

Figure 26: Average gender scores for all recordings grouped by transformation set. τ equates the interpolation factor used in the transformation.

5.2.1 Interpolation target

To define a value based on the interpolation factor that is comparable to the gender scores, the interpolation target is defined as:

$$T_\tau = \begin{cases} \tau \times 8 & \text{if source signal is female} \\ (1 - \tau) \times 8 & \text{if source signal is male} \end{cases} \tag{8}$$

The average gender score is plotted against this variable in Figure 27. This plot indicates that there may be a linear correlation between the two.

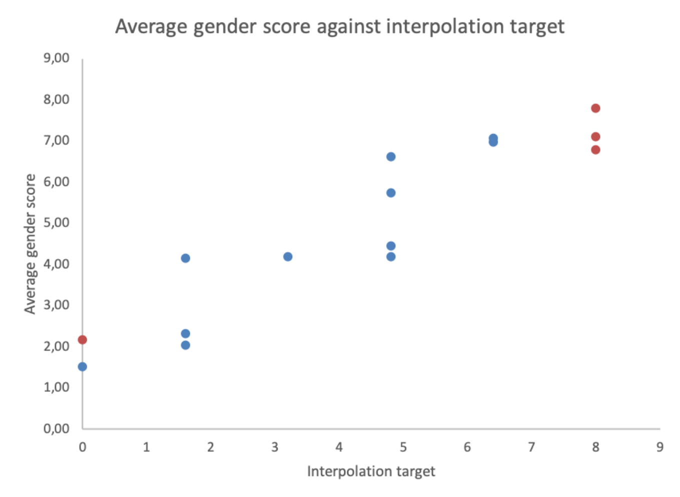


Figure 27: Average gender score against interpolation target. Unmodified recordings marked in red.

5.2.2 Normalised Gender Score

One interpretation of the interpolation factor is that it should – if assumed that the transformations are linear interpolations between two voices – correspond to the degree of how female or male a recording is perceived. Therefore, if the interpolation targets are used to normalize all gender scores you could expect them to be centered around zero. A histogram depicting the distribution of normalized gender scores is shown in Figure 28. The distribution appears to be slightly skewed toward the right, indicating that the recordings were perceived as more male than required.

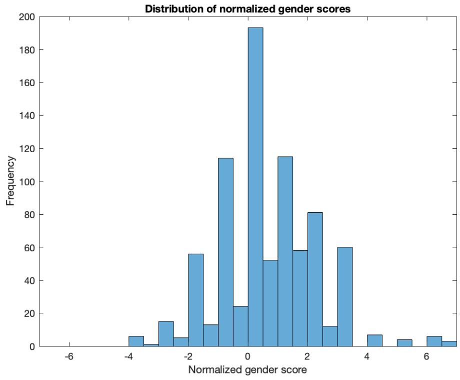


Figure 28: Histogram depicting the distribution of normalised gender scores

5.3 Naturalness Score

The perceived naturalness for all recordings is displayed in Figure 29. Out of the eleven transformed recordings, six were rated as synthetic and five were rated as natural by a majority of listeners. Not even the unmodified recordings were perceived as natural by all listeners, and some transformed recordings, such as number 5, have similar scores as them. A few of them stand out with very low naturalness ratings, especially numbers 6, 12, and 14. These were explored further to identify which aspects made them sound synthetic.

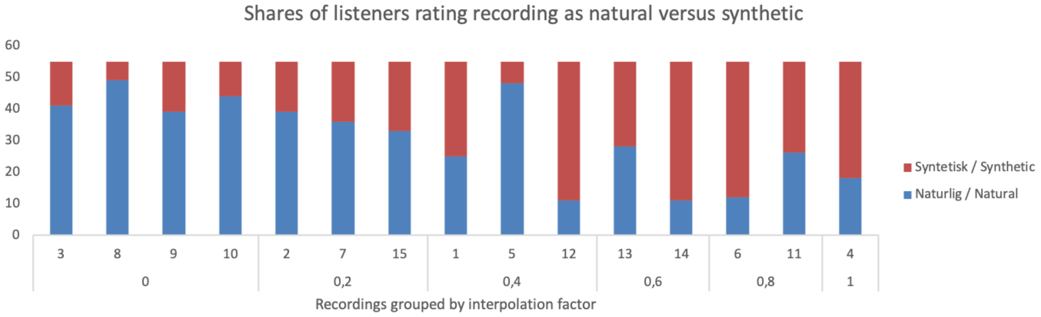
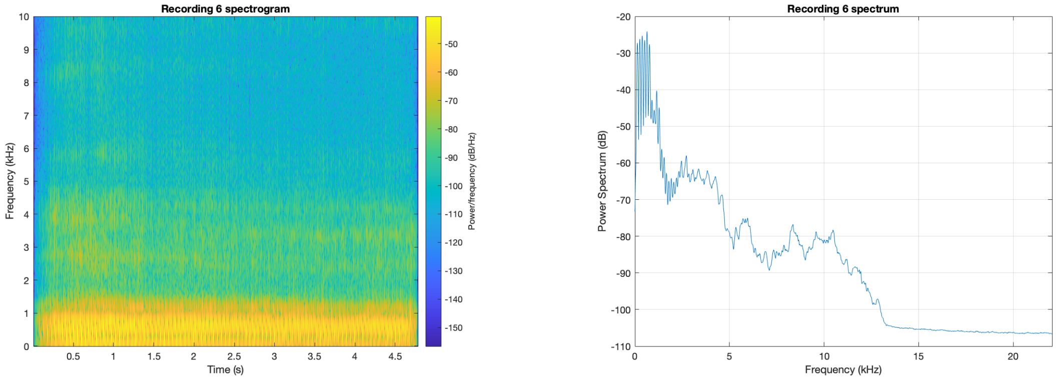


Figure 29: Bar graphs for each recording showing the share of listeners that perceived it as natural and synthetic ordered by increasing interpolation factor.

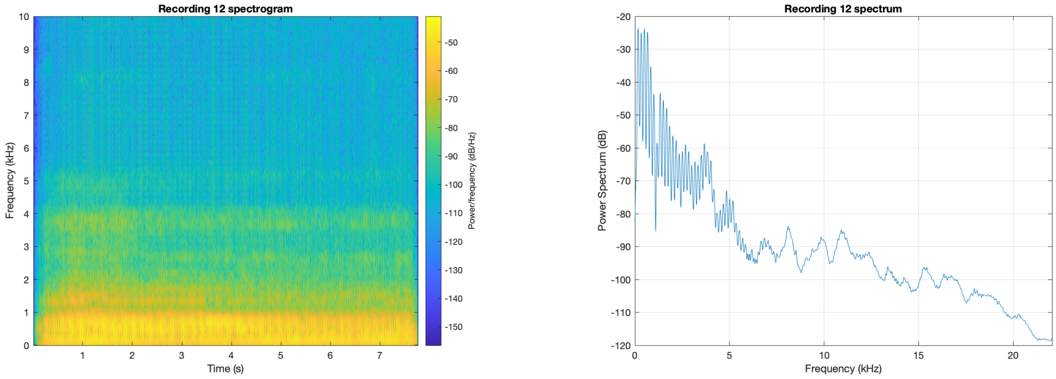
Listeners had the opportunity to give descriptors for the recordings they perceived as synthetic, and these offer some clues as to what the major issues were. The most common characterizations of recording 6 were robotic and noisy. In the spectrogram and spectrum, showed in Figure 30 it is revealed that the voice is lacking high-frequency contents and that the formants are faint and uneven. It has an 8.2 % shimmer. Similarly, recording 12 was perceived as noisy, robotic, and nasal and it had a 2.6 % shimmer. As evident from the spectrogram and spectrum, in Figure 31 the signal is also mostly comprised of low frequencies. The formants do not appear to be well-defined and stable. Lastly, recording 14 was described as noisy, breathy, monotonous, and reminiscent of the sound of an engine. It has a shimmer of 7 %. Looking at the spectrogram, the formants appear undefined and noisy. In the spectrum, it is clear that the formants have low energy and that the first harmonics are the strongest frequency contents. The spectrogram and spectrum for recording 14 are displayed in Figure 32 In comparison, the spectrogram and spectrum of recording 5, showed in Figure 33 – which was perceived as natural to the same extent as the unmodified voices – show that the signal has distinct and even formants and more high-frequency contents in general. Recording 5 has a shimmer value of 0.0297.



(a) Spectrogram

(b) Spectrum

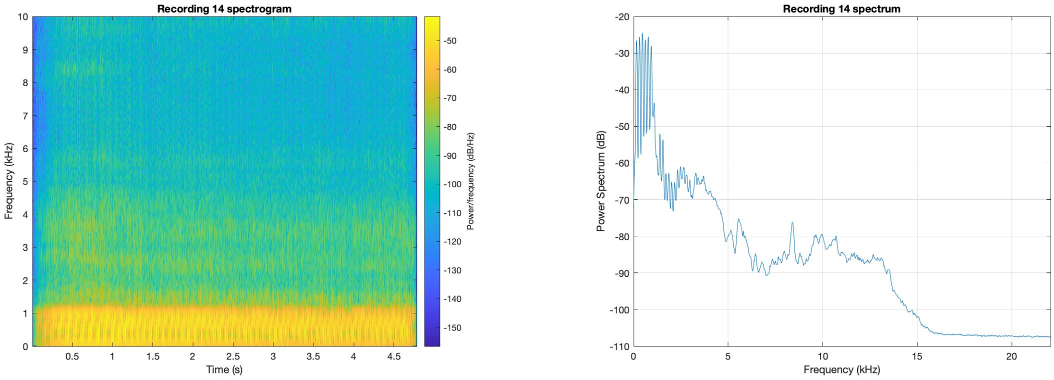
Figure 30: Spectrogram and spectrum for recording 6



(a) Spectrogram

(b) Spectrum

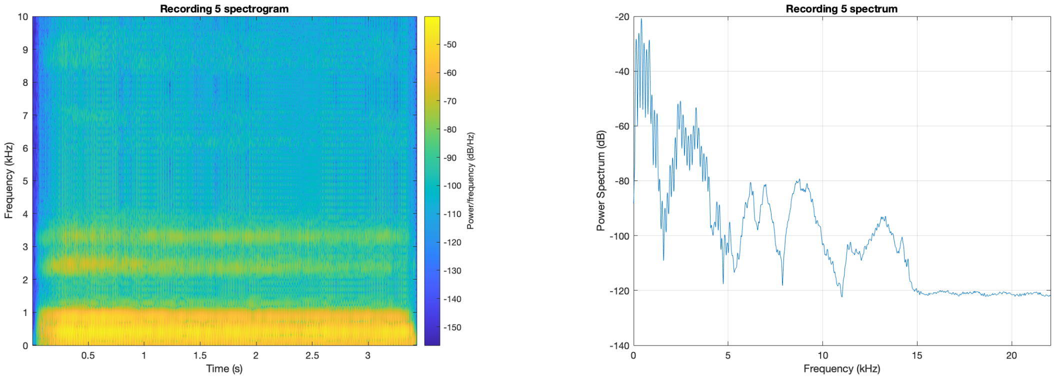
Figure 31: Spectrogram and spectrum for recording 12



(a) Spectrogram

(b) Spectrum

Figure 32: Spectrogram and spectrum for recording 14



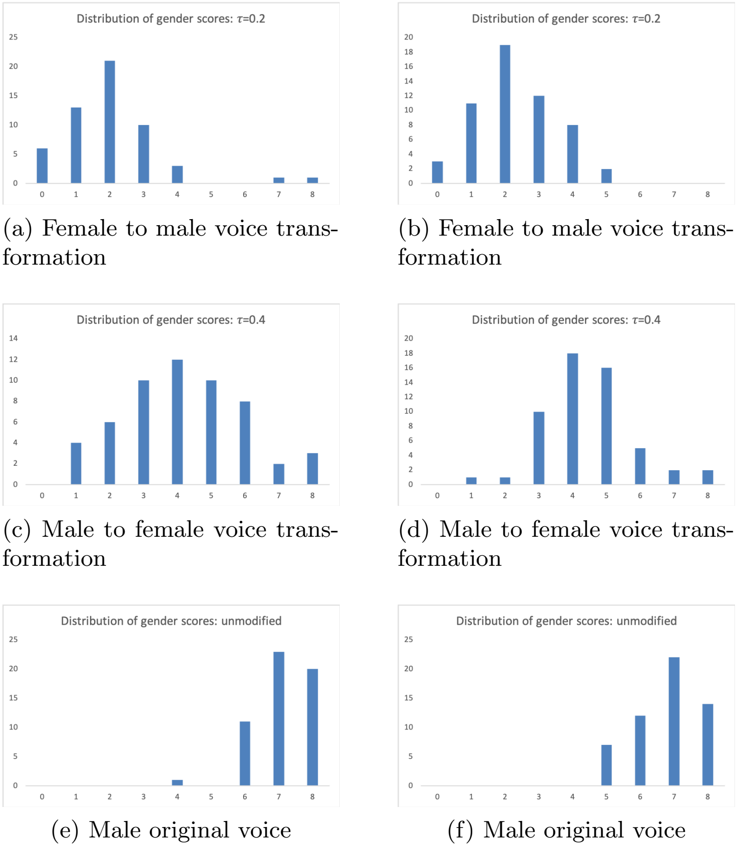
(a) Spectrogram

(b) Spectrum

Figure 33: Spectrogram and spectrum for recording 5

5.4 Internal Consistency

Three recordings were repeated in the survey to measure the internal consistency of the questions. By comparing the answers for different iterations of the questions, it can be studied how consistent listeners were in their ratings. Figure 34 displays the distributions of gender scores and Figure 35 shows the naturalness ratings for each repeated recording. The histograms all look quite similar for the same recordings and the type values are the same for both questions. As for the share of listeners that perceived the recordings as natural, these values are also in the same range in both iterations.



(a) Female to male voice transformation

(b) Female to male voice transformation

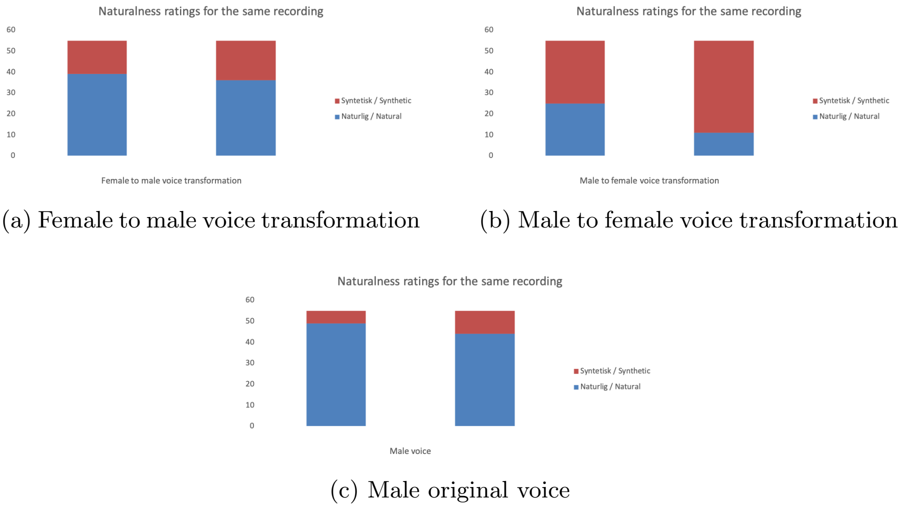
(c) Male to female voice transformation

(d) Male to female voice transformation

(e) Male original voice

(f) Male original voice

Figure 34: Comparison of gender scores for the same recordings



(a) Female to male voice transformation

(b) Male to female voice transformation

(c) Male original voice

Figure 35: Comparison of naturalness ratings for the same recordings

5.4.1 Cronbach's Alpha

The internal consistency can be measured using Cronbach's alpha [Cronbach, 1951]. This depends on both the variance of each item, σ_k^2, and the total variance, σ_{tot}^2, and is defined as:

$$\alpha_C = \frac{K}{K-1}(1 - \frac{\sum \sigma_k^2}{\sigma_{tot}^2}) \tag{9}$$

where K is the number of test items. The score ranges from 0 to 1, and higher scores indicate better internal consistency. It was calculated for all repeated questions and the results are displayed in Table 2. The alpha values are clearly low for the questions rating perceived gender for transformed voices, while it is rather high for the corresponding question for the unmodified voice. This may imply that it is more difficult to be consistent in attributing gender for the transformed recordings. Since both transformations are intended to sound androgynous, the alpha value might be indicative that the speakers' genders are indeed hard to specify. Moreover, the questions regarding naturalness have similar, low alpha values for all recordings. Because the score is low for all recordings, including the unmodified one, this could mean that listeners had a hard time assessing whether a voice sounded synthetic or natural.

Table 2: Cronbach's Alpha results for repeated recordings

Recording Type	**Question**	**Cronbach's Alpha**
Transformed: female to male voice	gender score	0.405
	naturalness	0.547
Transformed: male to female voice	gender score	0.421
	naturalness	0.526
Original: male voice	gender score	0.832
	naturalness	0.567

6 Discussion

6.1 Pitch-Changes in Interpolations

In the interpolated transformations from one perceived gender to another, the pitch and spectral envelope are here assumed to change with the same interpolation factor. In speech therapy, this corresponds to the patient making gradual changes to both F0 and formants simultaneously. However, this may not reflect the treatment process accurately. Because pitch is so strongly associated with gender attribution of speakers, it is an obvious starting point when changing one's voice to be more congruent with one's gender identity. Furthermore, if the patient has undergone surgical procedures or hormonal treatments that have affected the vocal fold vibration rate, the pitch would need little to no adjustment. Therefore, it may be more appropriate to immediately change the pitch to the target value and then make gradual changes to the formants. It depends on what is deemed appropriate by the therapist to establish a healthy and functional voice.

6.2 Acoustic Aspects Not Applicable To Sustained Vowel Sounds

The low consistency scores in the survey revealed that listeners found it somewhat challenging to identify the gender of speakers and gauge the naturalness of recordings. Several factors could have contributed to this, for instance, the fact that the recordings did not contain all information naturally found in spontaneous speech thought to be used to attribute genders, such as intonation and speech rate. The sustained vowel sound used in the survey is not naturally occurring which may have affected listeners. Because the speech signals lacked some attributes used to make assessments this might have made it more difficult, meaning that more meaningful results could be attained by using full words or sentences.

Further, the assumed gender of the speaker can influence the perception of naturalness. Meaning that if a voice is thought to come from a female speaker the listener expects to hear qualities commonly associated with feminine voices, such as breathiness, and rates it as synthetic if they do not. For recording 14, discussed in detail in section 5.3, it might have been perceived as unnatural because it was breathy but thought to come from a male speaker.

Lastly, the importance of natural-sounding transformations, given the intended purpose, can be questioned. Although a transformation should serve as a useful voice to mimic and practice with, and that it, therefore, should sound realistic, it is not by definition necessary that it should be indistinguishable from a human voice. It is therefore arguably acceptable that recordings contain some artifacts such as buzzing because it does not affect the way a person could recreate the sound of the recording when practicing.

6.3 Using Target Voices

The proposed method makes use of a target voice to modify the voice. This does not take the patient's physical limitation and personal goals into account, as opposed to being able to specify desired values for properties such as F0 and formant frequencies. Therefore, the therapist and patient are responsible for finding a target voice that fulfills their wishes and can serve as a realistic goal for the treatment.

6.4 Losing High-Frequency Energy From PSOLA

As discussed in section 3.4, downshifting speech signals with the TD-PSOLA algorithm causes some loss of high-frequency energy. When testing the largest pitch decrease – a 50% reduction – the highest frequency was reduced to approximately 10 kHz. Research has historically focused on acoustic features of low-frequency energy in speech signals and studied how frequencies below 5-6 kHz influence intelligibility [Vitela et al., 2015, Monson et al., 2012]. This, together with the fact that band-limited speech signals in telephones are recognizable for normal hearing persons, would suggest that the energy loss exceeding 10 kHz is above the threshold for perceptual influence. However, two of the recordings with the highest synthetic ratings had down-shifted pitches and were perceived as nasal and monotonous, hinting that the lack of high-frequency contents affected the sound more than anticipated. Furthermore, other phonemes – especially consonants – have more high-frequency energy and may be impacted.

Increasing research on high-frequency energy in speech has revealed that it does in fact impact the intelligibility of speech.

6.5 The Distance Metric

The 2-Wasserstein metric returned expected and reasonable distances between different unmodified speech signals, such as larger distances between different vowels. When testing distances between an original signal and pitch-shifted as well as filtered versions, the distance increased with the degree of modification. This indicates that the metric was able to capture changes to both pitch and spectral envelope. However, for the gradual transformations from source to target signals the resulting distances did not indicate that they were interpolations of the speech spectrums. It cannot be concluded here whether this occurred because the transformation method or the distance measurement is not performing as intended.

A selection of transformations was also reviewed by listeners in the survey to evaluate if they fulfilled their purpose. It was then concluded that the transformations' gender attributions moved closer to the target voices', but that the naturalness on average was lower compared to unmodified recordings. Evidently, some aspects of the transformations' spectrums were altered which made them be perceived as unnatural. This difference could potentially be the underlying factor that made the distances from transformed signal to target and source signal larger. Since the survey only measured the perceived gender and naturalness of the recordings, while the distance metric measured the total spectral difference, it is not surprising that the results did not coincide.

6.6 Future Improvements

In order to move forward with the transformation method presented in this thesis, improvements and further work must be done. Firstly, the filter estimation process has to be re-evaluated to improve the filter gain variance and avoid volume fluctuation. Furthermore, aspects found to negatively affect the naturalness of transformations – such as undefined formants – should be investigated. For instance, efforts could be made to make sure formants are always prominent throughout the interpolated filters. Additionally, more attention could be paid to modifying other qualitative vocal characteristics that influence gender perception in speech, such as breathiness and roughness.

As mentioned above, naturalness could potentially be more accurately gauged by transforming full words or phrases, which would require a generalization of the method. To transform speech signals containing multiple different phonemes, each frame needs to be mapped to a frame in the other signal to ensure that correct filters are utilized. There are several ways to do this. Latsch and Netto suggest a method for prosody transplantation incorporated in TD-PSOLA where pitch marks in two signals are matched using dynamic time warping. Additionally, here the pitch is shifted to the same F0 for all frames in the signal. This would not be appropriate for more complex sounds since the pitch naturally changes during speech and would have to be addressed as well.

Finally, the distance metric must be tested further to understand how well it works for the intended purpose. By making distance comparisons on recordings from a successful treatment process, more insight could be attained on how it performs. Furthermore, speech signals can be altered to test the influence of different aspects more thoroughly by, for example, adjusting single formants and shifting F0 for several signals.

7 Conclusion

In this thesis, two main research questions were constructed regarding transforming voices and measuring distances between them. The goal was to use these findings to assist patients in speech therapy find a voice more congruent with their gender identity.

The first question was: is it possible to achieve natural transformations that progressively change the perceived gender of a voice? From the survey, it was found that speech recordings could be transformed in such a way that the perceived gender changed. Additionally, a few transformed recordings were rated as equally natural as unmodified recordings, and 6 out of 11 transformations were perceived as natural by the majority of listeners. However, the naturalness of the recordings varied substantially and transformed signals were on average perceived as synthetic by more listeners compared to natural speech recordings. Therefore, it was found that natural results are possible, but that additional work is needed to achieve consistent results. Target gender scores, based on the interpolation factor used for transformations, correlated with average gender scores, indicating that the transformations fulfilled the requirements.

The second question dealt with measuring distances between speech signals. When testing the 2-Wasserstein metric on pitch-shifted and filtered signals, the results implied that it could measure differences to F0 and formant frequencies. The distance measurement did not perform as expected when comparing distances between transformations and the source and target signals. However, this may not be a fault of the distance metric, since interpolations at this stage do not equate to a natural progression from one voice to another. Therefore, it cannot be determined at this point whether 2-Wasserstein can be used to measure voice progression throughout treatment because that would require other data.

In conclusion, the methods for transforming speech signals in regards to their gender attribution and quantifying the distance between speech signals show promise and may one day serve as useful tools for transgender patients in speech therapy. Until then, they both need further development to produce consistent and reliable results and handle more general speech signals.

References

[Association, 2013] Association, A. P. (2013). *Diagnostic and Statistical Manual of Mental Disorders (DSM–5-TR).* American Psychiatric Association Publishing, 5 edition.

[Atal and Hanauer, 1971] Atal, B. S. and Hanauer, S. L. (1971). Speech Analysis and Synthesis by Linear Prediction of the Speech Wave. *The Journal of the Acoustical Society of America*, 50(2B):20.

[Boersma and Weenink, 2022] Boersma, P. and Weenink, D. (2022). Praat: doing phonetics by computer. `http://www.praat.org/`.

[Chen et al., 2021] Chen, F., Togneri, R., Maybery, M., and Tan, D. (2021). Voice Gender Scoring and Independent Acoustic Characterization of Perceived Masculinity and Femininity. *arXiv:2102.07982 [cs, eess].* arXiv: 2102.07982.

[Colton and Estill, 1981] Colton, R. H. and Estill, J. A. (1981). Elements of Voice Quality: Perceptual, Acoustic, and Physiologic Aspects. In *Speech and Language*, volume 5, pages 311–403. Elsevier.

[Cronbach, 1951] Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3):297–334.

[Davies et al., 2015] Davies, S., Papp, V. G., and Antoni, C. (2015). Voice and Communication Change for Gender Nonconforming Individuals: Giving Voice to the Person Inside. *International Journal of Transgenderism*, 16(3):117–159. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/15532739.2015.1075931.

[Fant, 1960] Fant, G. (1960). *Acoustic Theory Of Speech Production: with calculations based on X-ray studies of Russian articulations.* Number 2 in Description and analysis of contemporary standard Russian. Mouton, The Hague.

[Fry, 1979] Fry, D. B. (1979). *The Physics of Speech.* Cambridge Textbooks in Linguistics. Cambridge University Press, Cambridge.

[Hagelborn and Hulme Geber, 2020] Hagelborn, A. and Hulme Geber, J. (2020). Interpolation of perceived gender in speech signals. Student Paper.

[Hamon et al., 1989] Hamon, C., Mouline, E., and Charpentier, F. (1989). A diphone synthesis system based on time-domain prosodic modifications of speech. In *International Conference on Acoustics, Speech, and Signal Processing,*, pages 238–241 vol.1. ISSN: 1520-6149.

[Hardcastle et al., 2010] Hardcastle, W., Laver, J., and Gibbon, F. (2010). Voice source variation. In *The Handbook of Phonetic Sciences*, page 381. Blackwell, Oxford, 2 edition.

[Hardy et al., 2020] Hardy, T. L. D., Rieger, J. M., Wells, K., and Boliek, C. A. (2020). Acoustic Predictors of Gender Attribution, Masculinity–Femininity, and Vocal Naturalness Ratings Amongst Transgender and Cisgender Speakers. *Journal of Voice*, 34(2):300.e11–300.e26. Publisher: Elsevier.

[Islam, 2000] Islam, T. (2000). Interpolation of linear prediction coefficients for speech coding. Technical report.

[Jotform, 2022] Jotform (2022). Jotform Inc. – The Easiest Online Form Builder. `https://www.jotform.com/about/`.

[Kolouri et al., 2017] Kolouri, S., Park, S. R., Thorpe, M., Slepcev, D., and Rohde, G. K. (2017). Optimal Mass Transport - Signal processing and machinelearning applications. *IEEE Signal Processing Magazine.*

[Kreiman and Sidtis, 2011] Kreiman, J. and Sidtis, D. (2011). *Foundations of Voice Studies - An Interdisciplinary Approachto Voice Production and Perception.* John Wiley & Sons, Ltd, 1 edition. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781444395068.

[Laver, 2009] Laver, J. (2009). *The Phonetic Description of Voice Quality | Phonetics and phonology.* Number 31 in Cambridge Studies in Linguistics. Cambridge University Press.

[Lindgren et al., 2013] Lindgren, G., Rootzen, H., and Sandsten, M. (2013). *Stationary Stochastic Processes for Scientists and Engineers.* Chapman and Hall/CRC, New York.

[Longster, 2003] Longster, J. A. (2003). Concatenative Speech Synthesis: A Framework for Reducing Perceived Distortion when using the TD-PSOLA Algorithm. Technical report, Bournemouth.

[Monson et al., 2012] Monson, B. B., Lotto, A. J., and Story, B. H. (2012). Analysis of high-frequency energy in long-term average spectra of singing, speech, and voiceless fricatives. *The Journal of the Acoustical Society of America*, 132(3):1754–1764. Publisher: Acoustical Society of America.

[Morise, 2015] Morise, M. (2015). CheapTrick, a spectral envelope estimator for high-quality speech synthesis. *Speech Communication*, 67:1–7.

[Noll, 1967] Noll, A. M. (1967). Cepstrum Pitch Determination. *The Journal of the Acoustical Society of America*, 41(2):293–309.

[Oates and Dacakis, 2015] Oates, J. and Dacakis, G. (2015). Transgender Voice and Communication: Research Evidence Underpinning Voice Intervention for Male-to-Female Transsexual Women. *Perspectives on Voice and Voice Disorders*, 25(2):48–58. Publisher: American Speech-Language-Hearing Association.

[Openstax, 2013] Openstax (2013). Anatomy and Physiology.

[Paliwal, 1995] Paliwal, K. K. (1995). Interpolation Propertied Of Linear Prediction Parametric Representations. *Eurospeech*, pages 1029–1032.

[Socialstyrelsen, 2020] Socialstyrelsen (2020). Utvecklingen av diagnosen könsdysfori - Förekomst, samtidiga psykiatriska diagnoser och dödlighet i suicid. Technical Report 2020-2-6600.

[Soong and Juang, 1984] Soong, F. and Juang, B. (1984). Line spectrum pair (LSP) and speech data compression. In *ICASSP '84. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 9, pages 37–40.

[Söderpalm et al., 2004] Söderpalm, E., Larsson, A., and Almquist, S.-A. (2004). Evaluation of a consecutive group of transsexual individuals referred for vocal intervention in the west of Sweden. *Logopedics, Phoniatrics, Vocology*, 29(1):18–30.

[Teixeira et al., 2013] Teixeira, J. P., Oliveira, C., and Lopes, C. (2013). Vocal Acoustic Analysis - Jitter, Shimmer and HNR Parameters. *Procedia Technology*, 9:1112–1122.

[Toma et al., 2010] Toma, -A., Târşa, G.-I., Oancea, E., Munteanu, D.-P., Totir, F., and Anton, L. (2010). A TD-PSOLA based method for speech synthesis and compression. In *2010 8th International Conference on Communications*, pages 123–126.

[Tyagi et al., 2006] Tyagi, V., Bourlard, H., and Wellekens, C. (2006). On variable-scale piecewise stationary spectral analysis of speech signals for ASR. *Speech Communication*, 48(9):1182–1191. Publisher: Elsevier B.V.

[Villavicencio et al., 2006] Villavicencio, F., Robel, A., and Rodet, X. (2006). Improving Lpc Spectral Envelope Extraction Of Voiced Speech By True-Envelope Estimation. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 1, pages I–I. ISSN: 2379-190X.

[Vitela et al., 2015] Vitela, A. D., Monson, B. B., and Lotto, A. J. (2015). Phoneme categorization relying solely on high-frequency energy. *The Journal of the Acoustical Society of America*, 137(1):EL65–EL70.