UNIVERSITY OF TARTU

Faculty of Science and Technology

Institute of Technology

Davit Rizhinashvili

Gender Neutralisation for Unbiased Speech Synthesising

Bachelor's Thesis (12 ECTS)

Curriculum Science and Technology

Supervisors:

Abdallah Hussein Sham, MSc

Prof. Dr. Gholamreza Anbarjafari

Tartu 2022

Gender Neutralisation for Unbiased Speech Synthesising

Abstract:

Machine learning can encode and amplify negative biases or stereotypes already present in humans, resulting in high-profile cases. There can be multiple sources encoding the negative bias in these algorithms, like errors from human labeling, inaccurate representation of different population groups in training datasets, chosen model structures and optimization methods. This thesis proposes a novel approach to speech processing that can resolve the gender bias problem by eliminating the gender parameter. Therefore, we devised a system that transforms the input sound (speech of a person) into a neutralized voice to the point where the gender of the speaker becomes indistinguishable by both humans and AI. Wav2Vec based network has been utilised to conduct speech gender recognition in order to validate that main claim of this research work, which is neutralisation of gender from the speech. Such a system can be used as a batch pre-processing layer for training models, thus making associated gender bias irrelevant. Further, such a system can also find its application where speaker gender bias by humans is also prominent, as the listener will not be able to judge the gender from speech.

Keywords: Signal processing, Responsible AI, Speech analysis, Emotion recognition, Gender bias

CERCS: P175 Informatics, systems theory, T121 Signal processing

Sooline neutraliseerimine erapooletuks kõnesünteesiks

Lühikokkuvõte:

Masinõpe algoritmi treenimisel võib tekkida või võimeneda olemasolevad negatiivseid stereotüüpsed eelarvamused, mille tulemuseks on kõrgetasemelised juhtumid. Sellised negatiivsed kallutused võivad kodeerida mitmed allikad, nagu inimeste märgistamise vead, ebatäpne elanikkonnarühmade kaasamine treening andmetes, valitud mudeli struktuur ja optimeerimismeetodid. See artikkel pakub välja uudse lähenemise kõnetöötlusele, mis võimaldab soolise eelarvamuse probleemi lahendada, kõrvaldades soo parameetri. Me töötasime välja süsteemi, mis moondab sisend heli, et väljundi sugu oleks inimeste või tehisintellekti jaoks eristamatu. Seda süsteemi saab kasutada kui eeltöötlus kihina tehisintellekti treenimise jaoks, muutes sellega seotud soolise kallutatuse ebaoluliseks.

Lisaks võib seda süsteemi rakendada kõneleja soo maskeerimiseks kus vastasel juhul võib tekkida sooline eelarvamus. Sünteesitud helis sooneutraalsuse kinnitamiseks, kasutame uudset ja märkimisväärselt täpset kõne baasil töötavat sootuvastus meetodit.

Võtmesõnad: Signaali töötlemine,Vastutustundlik AI, Kõne analüüs, Emotsioonide äratundmine,Sooline eelarvamus

CERCS: P175 Informaatika, süsteemiteooria, T121 Signaalitöötlus

Contents

1 TERMS, ABBREVIATIONS AND NOTATIONS

ML - Machine Learning

AI - Artificial Intelligence

NN - Neural Network

ANN - Artificial Neural Network

CNN - Convolutional Neural Network

SER - Speech Emotion Recognition

SGR - Speech Gender Recognition

NLP - Natural Language Processing

MLP - Multi Layer Perseptron

MFCC - Mel-frequency cepstral coefficients

Chroma - In sound processing, Chroma represents the tonal content of a musical audio signal in a condensed form

Tonnetz - A conceptual lattice diagram representing tonal space

Formant - In sound processing, a formant is the broad spectral maximum that results from an acoustic resonance of the human vocal tract.

2 Introduction

In recent times, research in artificial intelligence (AI) and machine learning (ML) techniques have led to significant improvements in computer vision, speech processing, and language technologies, among others. Consequently, with these advances has come an inadvertent focus on the ethics of such ML models [42, 21].

Due to certain design choices or other reasons, ML models can end up having discriminatory results based on sensitive features and are considered to be ’biased’ or ’unfair’ [49]. Several factors can contribute to producing negative bias in ML models. One significant cause is incomplete training data [55] that lack sensitive information like gender or is unbalanced. Most models used in modern technology applications are based on supervised learning, and much of the labeled data comes from people. Despite the effects of the dataset, since people are inherently biased and models are estimates of people’s impressions, this bias will be passed on and implicitly encoded in the algorithms. As a result, there is the real risk that these systems can inadvertently perpetuate or even amplify bias contained in the label data [27, 59].

According to [35], one of the definitions of fairness can be considered "Fairness through unawareness," which states that the model achieves fairness towards certain attribute, if such attribute is not utilized to make predictions [25]. Approaches using such definition can be generalized not only for AI but also for humans. Therefore, we propose eliminating the gender parameter from speech processing altogether by pre-processing the sound, so that perceived gender is indistinguishable by both humans or AI.

In this work, we employ sound modification techniques in combination with ML models to create gender neutral speech. To accurately judge about gender neutrality, we have developed a new method for Speech Gender Recognition (SGR), that reaches state-of-the-art accuracy and even surpasses it in some scenarios like multi-lingual or cross dataset tests. We then assembled the system for speaker gender neutralization and tested it on variety of sounds. Finally, we assessed the quality of resynthesized samples and validated that no unwanted artifacts were added to it

3 Literature Review

3.1 Machine Learning

Machine learning (ML) is the technology of developing computer algorithms that try to mimic human intelligence [23]. It draws on ideas from different disciplines such as artificial intelligence(AI), probability, statistics, computer science, information theory, control theory, and philosophy [11, 7, 10]. A ML algorithm is a computational process which uses set data to achieve given task, without being "hard coded" to do so [51]. These algorithms alter or adapt their architecture for given task through a process called training, which is the "learning" part of the discipline. The models themselves are structurally similar to human brain, in which they have nodes (neurons) and connections organised in layers. A typical ML model has an input layer for the data, hidden layers for processing the data and an output layer, which are connected between each other [38] (see figure 1). Such architecture is termed as Artificial Neural Network (ANN) [63].

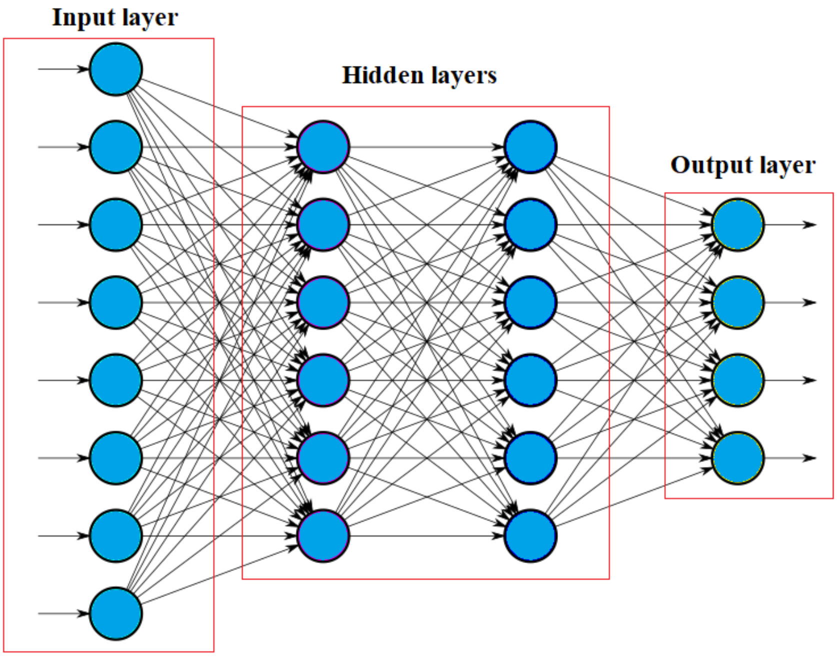

Figure 1. Example Structure of typical fully connected artificial neural network. Adapted from [60].

Each connection in ANN has an associated weight, which is a multiplier for the value it caries. Hence, the input of the given neuron can be determined by summing up all the multiples of connected neuron values and corresponding connection weights, given by the formula 1 [60].

$$N_j = \sum_{i=1}^{T} W_i P_i \tag{1}$$

Where $\mathrm{N_j}$ is the total input value of the neuron j, T is the number of the neurons that are connected to its input, $\mathrm{P_i}$ is the output value of the $\mathrm{i_{th}}$ neuron and $\mathrm{W_i}$ is the weight of the connection.

After calculating the total input, it passes through the "activation function" of the neuron, which determines its output. The most commonly used activation function are "hyperbolic tangent"(tanh) or "sigmoid" [48], which output 1 if the input is above certain level and 0 otherwise.

The signal is transmitted from the first input layer to the last output layer through those neurons and connections. Therefore, the training process for ANN is essentially fine-tuning above mentioned weights to achieve certain task. In other words, it needs to learn the correlation between the data and desired output. This process starts by providing a model with training data, which it uses to guide the adjustment through iterative process called back-propagation [36]. In other words, the model tries to minimize the "cost function", which is calculated based on the difference between expected and actual results [31]. Following this, model can be provided with new data to make predictions on it.

3.2 Bias in machine learning

Most ML methods are tasked to optimize only one performance metric, such as accuracy, which will inadvertently have consequences [50]. More often, it is facilitated in having discriminatory results based on sensitive features, such as gender, and such models are considered to be ’unfair’ or ’biased’. With the widespread use of AI and its applications in our everyday lives, accounting for fairness has gained significant importance in designing such systems. AI systems are often used in many sensitive environments to make important and life-changing judgments, such as interviews [16], hiring [47] and personality analysis [29]. Therefore, it is essential to ensure that these decisions do not reflect discriminatory behavior towards certain groups or populations. More recently, some work has been developed in classical ML and deep learning (DL) techniques that address these challenges in different sub-domains, which can include different aspects of possible bias sources, be it gender, race, age, or others.

The majority of AI algorithms are driven by data on which they are trained, making them tightly reliant on its quality. In cases where data contains biases, algorithms that are trained on them will inherit those biases and possibly convey them in their predictions. Algorithms themselves can further amplify those biases due to particular design choices. Moreover, even if the data is considered 'fair', algorithms can still produce biases independently. Similarly, those biased outcomes can carry on into real-world systems and affect users' judgments, thus fueling more biased data production [40]. As an example, one can imagine how this could affect the common search engine, that puts some specific results at the top. As usual, users tend to interact with top most results and pay significantly less attention to ones below [37]. These interactions will be collected by the search engine and it will make decisions based on which results were more popular. As an outcome, results at the top will become more popular over time, because of the poor algorithm, not because of their actual content [37]. The common loop illustrating the life-cycle of AI based systems is shown on figure 2. The following subsections will briefly discuss potential sources and outcomes for each type of bias.

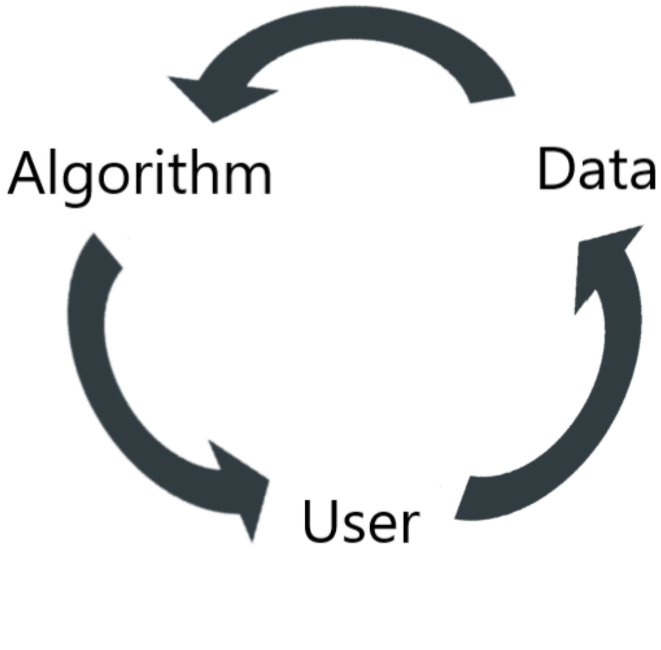

Figure 2. Common life-cycle of AI systems, where each interaction can produce or amplify bias.

3.2.1 Bias from user generated data

Considerable amount of data for training ML models are human-generated, thus any bias that they exhibit will be passed on to that data. Additionally, if algorithm used for this process is affected by some sort of bias, subsequent data generation will be affected in that manner as well. Following is the list of some possible types of bias sources that come from users.

1. **Population bias** - *arises when statistics, demographics, representatives, and user characteristics are different in the user population of the platform from the original target population* [44]. Population bias yields non representative data, where distribution of sensitive attributes, like gender, within the data is not uniform. An example would be data from user demographics on different social networks, such as men being more active on platforms like Twitter while women being more likely to use platforms like Instagram or Facebook [33].

2. **Social Bias** - *is observed when users actions are affected by others* [5]. An example of this can be seen when one is not satisfied with highly rated product. Even though one would like to give a low score to the product, seeing others high ratings can influence their decision.

3. **Behavioral Bias** - *arises from different user behavior across contexts, platforms, or even different datasets* [44]. For example, authors of [41] highlight that differences in emoji representations between platforms can cause different behavior or reaction from users and even lead to miscommunication.

4. **Content Production Bias -** *arises from lexical, structural and semantic differences in the contents generated by users* [44]. For instance, in [43], authors discuss how differences in use of language across different age and gender groups can cause miss-interpretation and subsequent bias.

3.2.2 Algorithm generated bias.

Algorithms can affect user's actions, hence any bias in them will be passed on to user's behavior. The following list will outline some of possible biases as a result of algorithm outcome.

1. **Algorithmic Bias -** *is evident when bias is not present in training data and is added solely by the algorithm itself* [5]. Number of things can cause algorithm to be bias generating, for instance, design choices, optimization functions, regularizations and use of statistically biased estimators [19].

2. **Popularity Bias.** *More popular items tend to be more exposed. Nevertheless, popularity can be manipulated artificially, for instance by fake reviewers or bots* [14]. This can be seen in some search engines' recommendation systems, which promote popular products. Nonetheless, their popularity does not always stem from high quality and instead could be a result of other biased factors.

3.2.3 Data generated bias

As discussed, bias present in training data will inadvertently be passed onto the algorithm and affect its fairness. Following list will discuss some of the possible sources of bias in data.

1. **Measurement Bias -** *arises from how we choose, measure or report particular features* [57]. For instance, if poorly trained group is collecting data about deaths and they include ones outside the intended time period, this would lead to overestimation of mortality rate.

2. **Omitted Variable Bias -** *occurs when one or more significant variables are not considered in the model* [15]. As an example, lets say we build a model that predicts the price of the house in some region, given its square footage, with relatively high accuracy. When we try to apply this model to a different region we might find that the actual price is very different from calculated. This happens because the location parameter was left out from the model, rendering it biased towards one region, on which it was trained on.

3. **Representation Bias -** *arises from sampling strategy used when collecting data* [57]. Commonly, non-representative datasets lack the diversity of different populations and sometimes are even missing some subgroups. Lack of Geographical diversity in facial processing datasets will inadvertently cause bias against less represented demographic members. A canonical example of this can be seen in

software **COMPAS**[1] used by United States court to make pretrial detention and release decisions. COMPAS measures the risk of a person to recommit another crime and based on that judges decide to release or keep the offender. However, an investigation has found that this software is biased against African-Americans, meaning it is more likely to produce false positive outcomes on African-Americans than Caucasians [2].

3.3 Gender bias in machine learning methods and humans

There is an ample amount of examples for gender-based unfairness in ML applications, especially speech processing. For instance, it was reported that, speech synthesis and speech recognition performed better on lower pitch voices [34], usually present in adult males. As a result, speech recognition yielded higher error rates for children and female adults. In [21], authors show how the gender of the subject affects its emotion recognition and compare gender bias of different models. They also assess the extent of said bias by measuring the accuracy gap in emotion recognition between male and female test sets and observing which types of emotions are better classified for each gender. In [30] authors demonstrate that the activation model for Speech Emotion Recognition (SER) is negatively biased towards females, meaning that these models consistently underwhelm activation for women compared men. .

While there are conventional methods for bias mitigation in AI, they can not be simply applied to humans, which might be biased as well. Many studies have been conducted by psychologists in which they highlighted correlation between culture [18], race [17], gender [26], and cultural differences in emotions. For example, In [46], authors have shown that participants believed women experienced and expressed the majority of the 19 emotions studied more often than men. In another study, it was shown that women were rated sadder and less angry than men.

Natural language processing (NLP), speech-to-text, and general sound recognition can also be the subject of gender bias. Authors of [6] mention that current state-of-the-art speech recognition is 13% more accurate for men than it is for women. Further, they go on to state that Dialects also affect accuracy. For example, Indian English has a 78% accuracy rate, and Scottish English has a 53% accuracy rate. Authors of [52] highlight

[1] Correctional Offender Management Profiling for Alternative Sanctions

[2] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

the said bias for machine translation, which facilitates itself by creating social gender expectations. for instance, translating engineers as masculine and nurses as feminine. Similar stereotypical gender bias can be observed in humans as well. Authors of [22] utilized a word association test to assess gender stereotypes in texts and found that bias scores correlate well with bias in the real world.

3.4 Ways to overcome bias in AI

Unfair nature of most ML models has been known for long time, and there is a great amount of research targeting this problem. In [28] authors highlight the requirements and obstacles for responsible AI concerning two intertwined objectives: efforts toward socially beneficial applications and human and social dangers of AI systems. They also mentioned several reported bias cases in different fields due to a lack of transparency, intelligibility, and biased training data. Authors of [9] explain how to use design methodology to create a responsible and fair AI. Authors of [25] provide a summary of formal definitions for AI fairness, those being ’Fairness through unawareness’ [35], ’Counterfactual fairness’ [35], ’Statistical Parity Difference’ [35], ’Equal Opportunity Difference’ [32], ’Average Odds Difference’ [8] and ’Disparate Impact’ [24]. With correctly chosen technical tools, such as Meta’s Fairness Flow tool, IBM’s Fairness 360 toolkit or even Accenture’s AI Fairness tool, one can detect bias in sensitive datasets and even see correlations in said datasets. In [12], the authors highlight the importance of an appropriately diverse dataset to achieve fairness. They then propose a maximum entropy-based approach for data pre-processing, ultimately leading to bias mitigation.

Most approaches try to target and optimize the network, training, or used datasets [49]; for example, in their comprehensive review paper, authors of [56] focus on NLP and discuss existing methods for recognizing and mitigating gender bias, such as data manipulation and algorithm adjustment. They also outline the advantages and drawbacks of each. Authors of [30] propose two methods for AI de-biasing; while they focus on gender and gender bias, such approaches can be applied to any protected parameter. First, through an adversarial learning approach to achieve "Equality of odds" towards gender. The approach involves jointly training two models: a regression predictor and an adversary. The adversary is a low complexity model that takes the continuous scalar output of the predictor and the binary label variable as inputs and is trained to classify the binary protected variable(gender) optimally. The authors propose that the second method solely focuses on training a regression predictor model. Training involves minimizing a loss function which additionally includes a weighted term, which penalizes the model for producing inconsistency in recall across classes of the protected variable. These methodologies are being actively employed [61, 62, 58, 20] against gender bias in AI.

3.5 Speech gender recognition

Employing different acoustic models for males and females has been reported to yield better performance in speech processing and recognition tasks [54]. This has rendered the performance of Speech Gender Recognition (SGR) networks very significant. Generally, it is accepted that the pitch of the speaker can dictate their perceived gender, the male being in the range of 85 to 155 Hz and female 165 to 255 Hz [45], though, in reality, factors like higher tier frequencies, pitch contour shape and pitch fluctuations have a significant effect on it as well [49]. Further, both male and female pitch can go out of those boundaries quite frequently. Figure 3 displays the sorted distribution of average pitch for 720 female and male speakers, respectively. Figure 4 displays the pitch contours for female and male speakers while expressing happy emotions saying the same sentence.

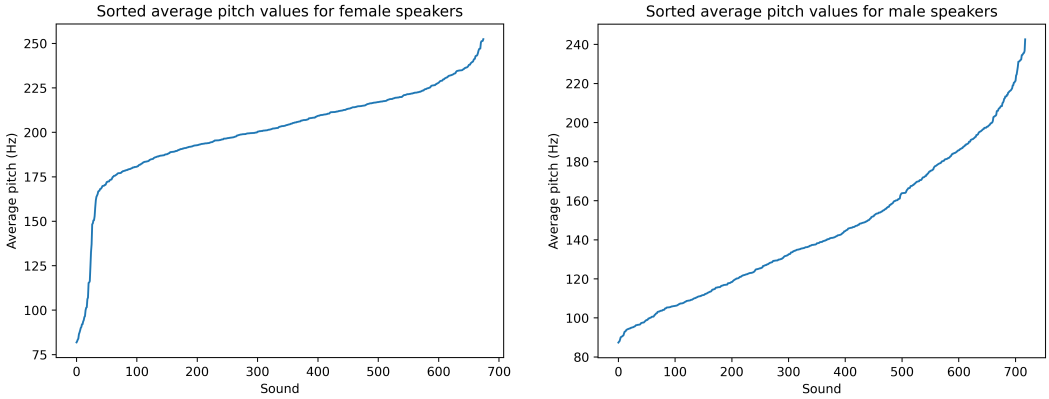

Figure 3. Sorted average pitch values for 1440 utterances by female and male speakers. Sounds from each gender category were analysed for average pitch and then sorted for illustrative purposes.

Many methods for SGR have been described in [3, 13, 1, 39, 2]. The most common approach is to extract features like pitch and different spectrograms from the sound and feed those to ANN for classification. Although it has been reported to have reached almost

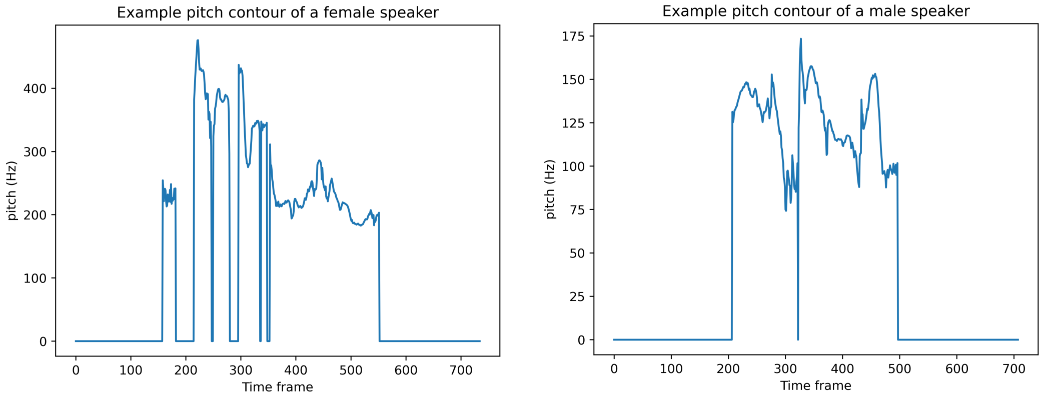

Figure 4. Pitch contour for female and male speakers on example utterance.

perfect accuracies on multitude of datasets, such approaches have many caveats. As different speech sample may differ in language, dialect, noise levels, sampling frequencies or overall quality, features extracted from them may be impacted significantly. Further, as some spectrogram extractions are based on dividing the sound sample into small frames, low sampling frequency will have a negative effect on its correctness.

4 The aims of the thesis

Gender bias in ML models, especially ones focusing on speech processing, is a well known and quite serious problem. While the most of the previous researches [56, 62] have targeted model or dataset optimizations for gender bias mitigation, to the best of our knowledge, our proposed method is the first to investigate gender de-biasing directly by parameter elimination. Therefore, the aims of the thesis are:

- To develop highly accurate method for SGR which does not discriminate between languages, dialects or other artifacts of the sound.
- To assemble a pipeline for speech gender neutralization.
- To test that such method works as de-biasing tool for gender and that no unwanted artifacts are created in original sound.

5 Experimental Part

5.1 Materials and Methods

5.1.1 Speech datasets

There are many speech datasets[3] that differ in their content, labels, language, access level, and more. We presented some of the datasets in table 1.

Table 1. Examples of different speech datasets, their content and description.

Database	Samples	Genders	Emotions	Language
CREMA-D	7,442	Male+Female	6	English
DEMoS	9,365	Male+Female	6	Italian
TESS	2800	only female	7	English
AudioMNIST	30000	Male+Female	–	English
EMO-DB	535	Male+Female	7	German
LibriSpeech	280000	Male+Female	–	English
RAVDESS	1440	Male+Female	8	English

For training and experimentation, we used 4 Speech databases from table 1. All sound samples used were 16 bit and 16KHZ sampling frequencies. For pre-processing, we set the intensity of each sound to 70db using Praat[4]. The following subsections describe the datasets we used.

5.1.1.1 RAVDESS

Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)[5] database contains 1440 audio files of 24 professional actors (12 female, 12 male), vocalizing two lexically-matched statements in a neutral North American accent. Speech includes calm, happy, sad, angry, fearful, surprise, and disgust expressions. This set was primarily used for testing emotional integrity after the speech was resynthesized. As the dataset contains relatively noiseless voices and equal gender distribution, the results of SER will be more comprehensive and accurate. Its labels include the gender of the speaker and conveyed emotion.

[3] https://github.com/coqui-ai/open-speech-corpora

[4] https://www.fon.hum.uva.nl/praat/

[5] https://smartlaboratory.org/ravdess/

5.1.1.2 CREMA-D

CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset)[6] is a data set of 7,442 original clips from 91 actors. These clips were from 48 male and 43 female actors between the ages of 20 and 74 coming from various races and ethnicities (African America, Asian, Caucasian, Hispanic, and Unspecified). This set was primarily used to test the robustness of our gender recognition network, as there is a wide range of speakers and sounds themselves are quite noisy. Labels include the gender of the speaker and conveyed emotion.

5.1.1.3 EMO-DB

EMO-DB (Berlin Database of Emotional Speech)[7] is a data set of 535 audio clips from 10 actors(5 male 5 female) spoken in German. this set was primarily used to test the robustness of our gender recognition network toward the language. Using databases that contain speech in different languages, we can verify that our model is language invariant in a given task. Labels for EMO-DB include the gender and emotion of the speaker.

5.1.1.4 LibriSpeech

LibriSpeech[8] is a corpus of approximately 1000 hours of 16 kHz read English speech. The data is derived from reading audiobooks from the LibriVox project and has been carefully segmented and aligned. Each sound segment is 3-20 seconds long and contains labels for the gender of the speaker and a full transcript of what is read. We used this set for training our gender recognition network, primarily because of its size and quality. Transcripts were used to compare speech-to-text programs performance on unmodified and modified sounds to prove the absence of any unwanted artifacts after the "neutralization" step.

[6] https://github.com/CheyneyComputerScience/CREMA-D

[7] http://emodb.bilderbar.info/

[8] https://www.openslr.org/12/

5.1.2 Speaker gender recognition

The classical approach to SGR is extracting features from the sound and then feeding those to some NN for classification, as described in [4]. Though many described models achieve almost perfect accuracy, some of them may be prone to over-fitting and perform poorly on cross-dataset tests. Primarily, this is because the feature extraction part relies heavily on the quality of the sound itself, which may vary from sound to sound [49]. We propose a non-traditional yet quite simple method that is robust to the aforementioned problems.

As said, SGR can be split into two parts, feature extraction, and classification. For feature extraction, instead of relying on traditional methods like different spectrograms, we use a pre-trained Wav2Vec [53] network. Wav2vec is a convolutional neural network model that takes raw audio as input and computes a general representation that can be input to a speech recognition system. These representations are computed for each 25ms frame in the sound and comprise a vector with 512 values.

For classification, the output of Wav2Vec network over the input sound is averaged over time, yielding vector with 512 values. In other words, for 100ms sound, output of Wav2Vec will be a 512X4 matrix, which will then get flattened by averaging corresponding values for each time frame. We then feed those values to a simple MultiLayer Perceptron (MLP) classifier with two fully connected layers of size (512,64) and rectified linear unit (Relu) as activation function. The output layer contains two neurons, corresponding to each gender prediction. The structure diagram of described SGR model can be seen in figure 5.

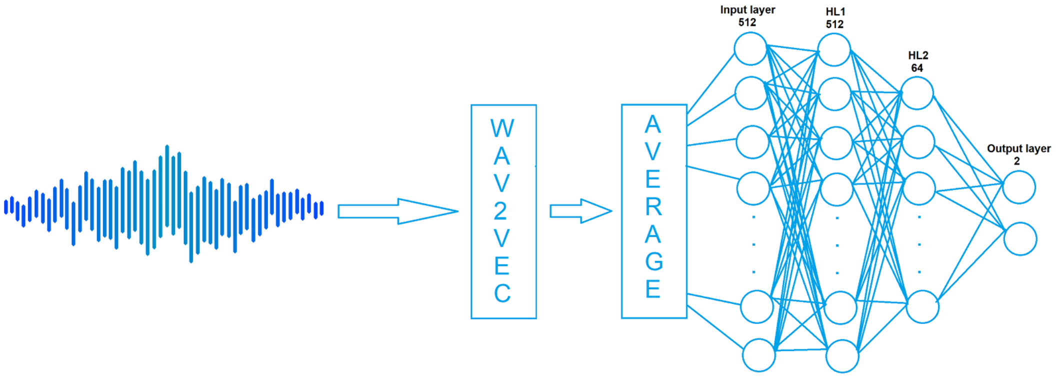

Figure 5. Structure of our SGR network

5.1.3 Neutralization process

As previously discussed, there are multiple types of alterations that can be applied to the sound in order to change its perceived gender, including frequency shifting, filtering, time stretching and more. In line of this point, we have found that the combination of correctly chosen pitch and formant shifts (we refer those as transformation parameters) is sufficient for achieving gender neutrality from any speech sample. Therefore, the task of our method is to find a correct combination of pitch and formant shift values for any given sample to achieve above mentioned goal. For modifying and resynthesizing sounds, we use *Praat* and its implementation in python called parselMouth[9]. Praat is a computer program that can analyze, synthesize, and manipulate speech. Most importantly, Praat has functions for pitch and formant manipulation that take one argument for each. These functions take arguments for new average pitch (50 Hz to 300 Hz) and formant shift coefficient (from 0.5 to 1.5, where values more than 1 would increase formant frequencies and the reverse for values below 1) and resynthesizes the sound accordingly.

In order to check that the speech is indeed gender neutral, we use the SGR network described in section 5.1.2. As the output of SGR is accompanied with the certainty of given prediction, at the point of gender neutrality, this certainty for each gender prediction would be close to 50%, implying that the model is not able to distinguish them accurately [49]. Therefore, we devised a system which searches through all transformation parameter combinations, until the SGR network outputs predictions with certainty close to 50%. In other words, system transforms the speech with every possible combination of parameters (we use steps of 3 Hz in range of 50 Hz to 240 Hz for pitch and steps of 0.01 in range of 0.75 to 1.25 for formant shift) and runs the output sound through SGR. This process continues until the absolute difference of certainties for each gender prediction falls below 10% (given by formula 2). Such search is quite time consuming and mostly unnecessary. For example, increasing frequencies of female voice will most certainly make it sound more feminine [49], thus our system should not waste time checking all possible combinations.

$$|P_{male} - P_{female}| < 10\% \tag{2}$$

Where P is the certainty of given gender prediction.

[9] https://pypi.org/project/praat-parselmouth/

To address this, we first neutralized 1000 different speech samples using the lengthy method described above and looked for possible relations between pitch, pitch standard deviation and gender of the initial sound and the transformation parameters found by the system. We have found that for each gender, initial pitch was correlated to the correct formant shift coefficient, and pitch standard deviation was correlated to the new pitch transformation found by the system. Said correlations are demonstrated on figure 6 for male samples and figure 7 for female samples. Based on these correlation figures, we were able to devise the corresponding equations 3,4 for males and equations 5,6 for females.

For male:

$$P_{male} = 150 + \frac{sp}{1.63} \tag{3}$$

$$fs = 1.15 - \frac{p}{1500} \tag{4}$$

For female:

$$P_{female} = 140 - \frac{sp}{1.96} \tag{5}$$

$$fs = 0.8 + \frac{20}{p} \tag{6}$$

Where P is the new average pitch, sp is the measured standard deviation of pitch, fs is the formant shift coefficient, and p is the initial measured pitch.

These findings enabled us to add an initial "*coarse neutralization*" step to the system. Here, we extract gender, pitch and standard deviation of pitch from the initial sound and using above correlation formulas, calculate the initial neutralization parameters. Based on these parameters, we do not need to start the searching cycle from scratch, hence saving time and processing power. Moreover, without this step, search can take up an additional 100 iterations, while, on average with coarse neutralization step, the process takes 11 iterations to find the correct transformation parameters.

Finally, as said it is very important that we do not change the emotion carried by speech during neutralization. To make sure this is the case, we use *Vokaturi*[10] software's python API to measure the emotion of initial sample and during parameter search, prioritize the transformations where emotion was not affected. In other words, if the

[10] https://vokaturi.com/

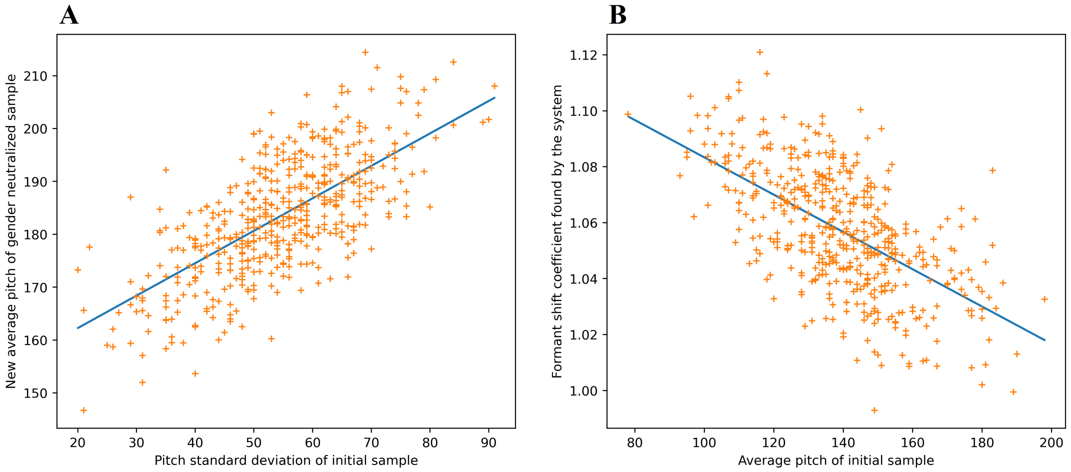

Figure 6. Correlation of initial sound features and correct transformation parameters for gender neutralization for male samples. Orange '+' markers correspond to each sound sample and blue line is representing equation of corresponding correlation. A) Correlation of initial pitch standard deviation and new pitch value of neutralized sound. B) Correlation of initial average pitch to the formant shift coefficient in neutralized sound.

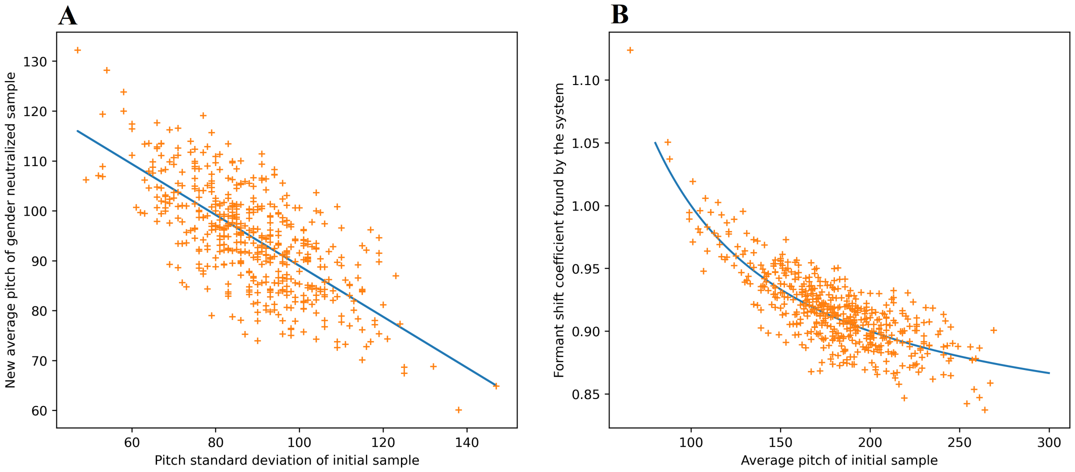

Figure 7. Correlation of initial sound features and correct transformation parameters for gender neutralization for female samples. Orange '+' markers correspond to each sound sample and blue line is representing equation of corresponding correlation. A) Correlation of initial pitch standard deviation and new pitch value of neutralized sound. B) Correlation of initial average pitch to the formant shift coefficient in neutralized sound.

system finds pitch and formant shift combination for which SGR outputs certainties close to 50% but the measured emotion differs from initial, such combination will be discarded.

All in all, our gender neutralization system can be split into two parts. First, coarse neutralization, where we extract the features from initial speech sample and use them to find a probable location of correct transformation parameters. And second, a feedback loop comprising of sound resynthesis, SGR and SER networks. The loop starts from the combination of transformation parameters calculated by coarse neutralization step. Then for each iteration, it slightly changes pitch or formant shift coefficient, resynthesizes the sound with new parameters and puts the sound through SGR and SER networks. We loop this process until the absolute difference of certainties for each gender prediction is below 10% (equation 2) provided that there is no modification in the carried emotion from the given transformation. After such pair is found, loop is halted and the corresponding sound is saved. A block diagram of gender neutralization system can be seen on figure 8.

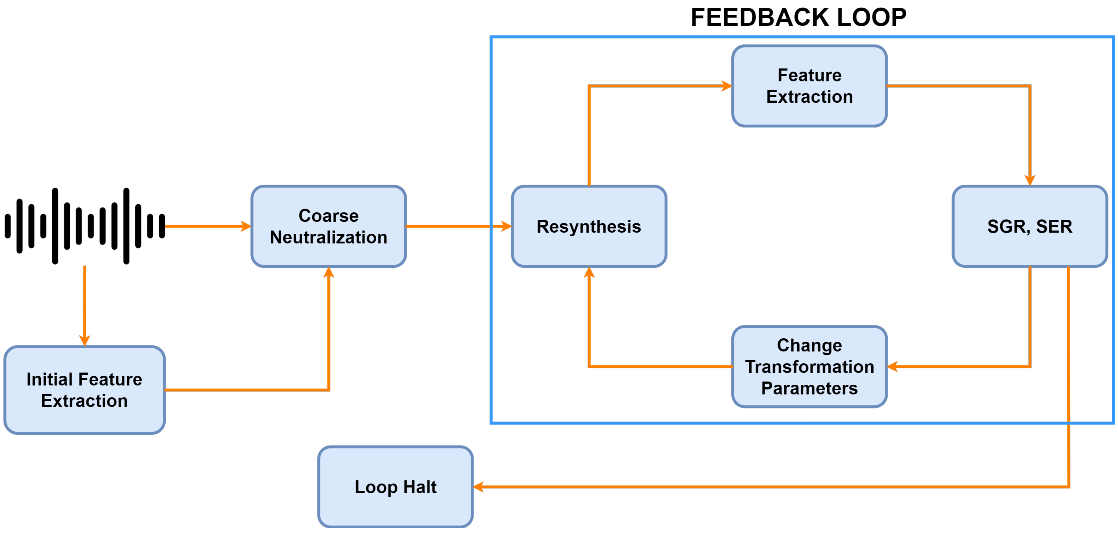

Figure 8. Diagram of our proposed gender neutralization system

5.2 Results

5.2.1 Speech Gender Recognition

To start, we evaluate the robustness and accuracy of our SGR method. For comparison purposes, we have employed the approach described in [3], namely, pre-processing, feature(MFCC, chroma, mel, and tonnetz) extraction, and classification using MLP. We then trained it using the LibriSpeech dataset, and the network reached an accuracy of 97.3% on the validation set. However, when tested against other datasets, this network produced very low accuracy, particularly, 58.3% on EMO-DB, 53.5% on CREMA-D and 68.5% on RAVDESS, suggesting that this method is prone to over-fitting. Confusion matrices of above tests can be seen on figures 9A,9B,9C and 9D (prefix ’A’ represents actual and ’P’ represents predicted).

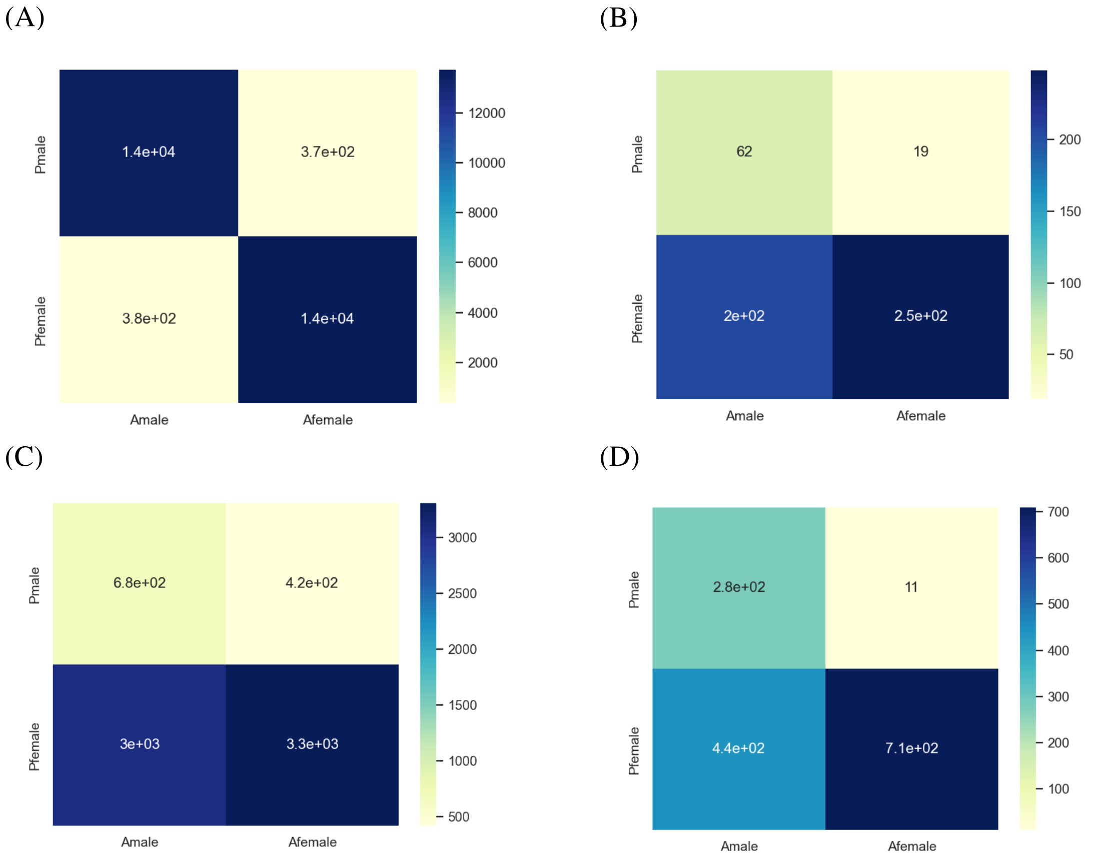

Figure 9. **Confusion matrices of SGR network from [3] trained on LibriSpeech and tested on:** A) Validation set, accuracy: 97.3%. B) EMO-DB, accuracy: 58.3%. C) CREMA-D, accuracy:53.5%. D) RAVDESS, accuracy: 68.5%.

On the other hand, our proposed network trained on LibriSpeech, while reaching almost perfect accuracy of 99.87% on validation set, also excels in cross dataset test. In particular, accuracies were 97.3% on EMO-DB, 96.7% on CREMA-D and 96.9% on RAVDESS. confusion matrices can be seen on figures 10A,10B,10C and 10D.

5.2.2 Speech Gender Neutralization Pipeline

As stated, the most important factor when dealing with sound re-synthesis is checking for absence of added artifacts. Particularly, when we assembled the system described in section 5.1.3, we needed to check that carried emotion stayed the same and there was no unwanted effect added to newly synthesized sound. Using the previously mentioned tool for SER, VokaTuri, we can measure the emotion before and after the transformation and then compare the two. Doing this for the whole RAVDESS dataset yielded a 98.75%

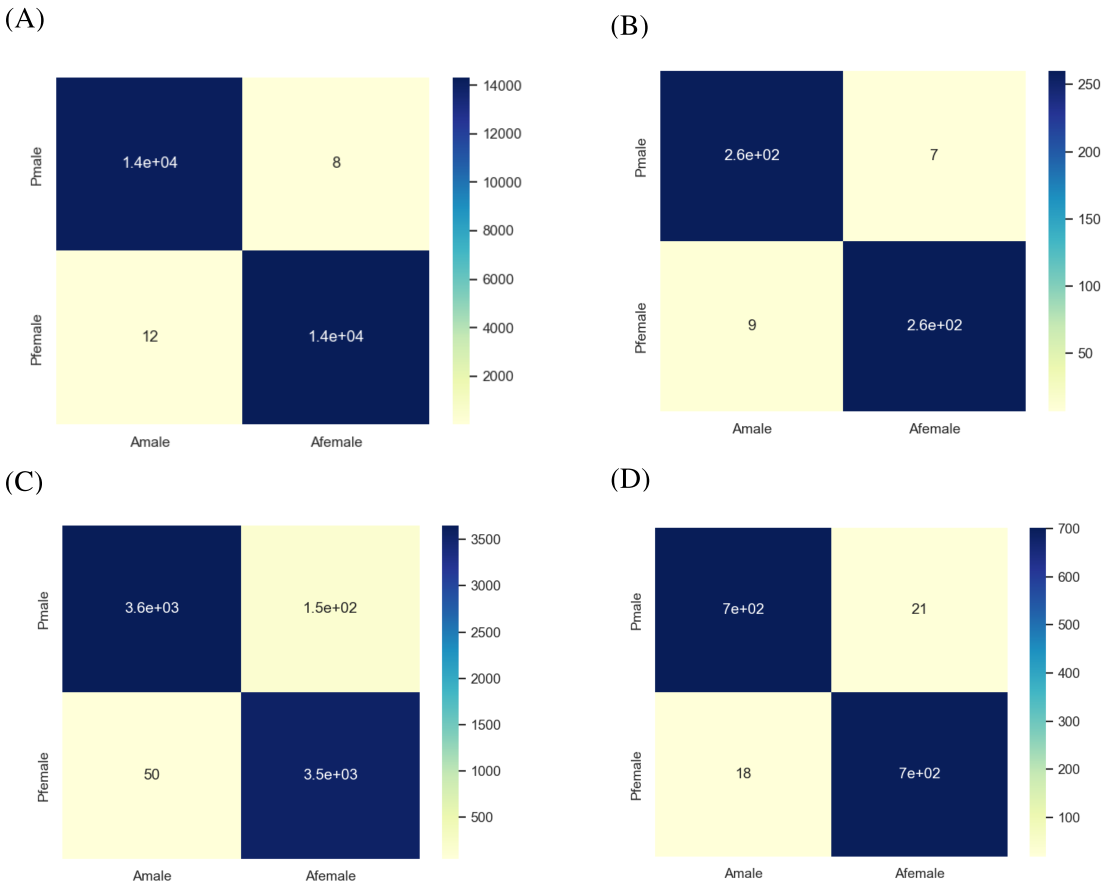

Figure 10. **Confusion matrices of our SGR network, trained on LibriSpeech and tested on:** A) Validation set, accuracy: 99.87%. B) EMO-DB, accuracy: 97.3%. C) CREMA-D, accuracy:96.7%. D) RAVDESS, accuracy: 96.9%.

match between original and transformed files. Figure 11 illustrates the matrix of emotion predictions, where horizontal axis represents original and vertical represents synthesized sounds. It is important to stress that we do not measure the accuracy of said SER method but rather make sure that the emotional prediction of a given sound does not change upon transformation.

Next, to verify the absence of any unwanted artifacts, we compare speech-to-text performance on original and neutralized sounds. We used 5000 speech samples from LibriSpeech, as it also comes with text transcriptions and google's speech-to-text API[11]. To compare transcripts, we used FuzzyWuzzy library[12], which basically compares two strings and measures the edit distance between them, then outputs a match score in percentages. Consequently, google's speech-to-text scored 86.12% on original set and 85.67% on Transformed sounds. Cross-checking transcriptions yielded 98.7% match.

[11] https://cloud.google.com/speech-to-text/

[12] https://github.com/seatgeek/fuzzywuzzy

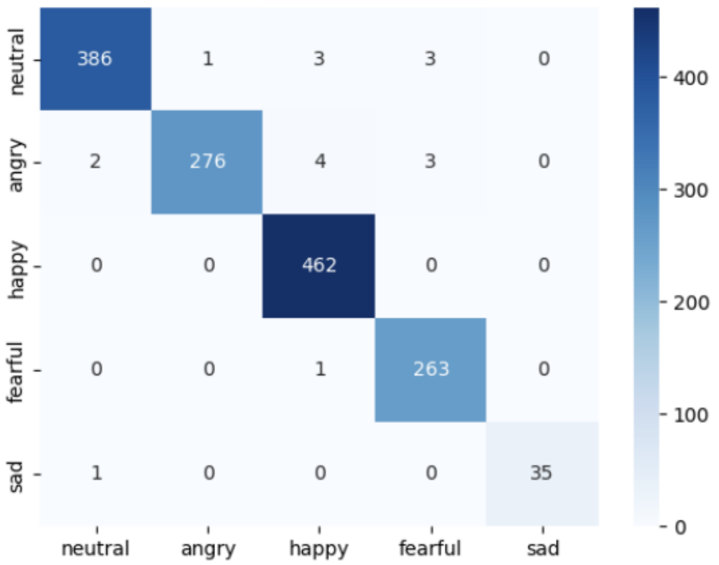

Figure 11. Emotion predictions on original and Transformed sounds from RAVDESS dataset

However, the said string matching algorithm does not take into account the fact that some words may sound similar but have different spellings, which could be one of the main contributors to the matching gap. Figure 12 shows the match percentage between transcripts on original and synthesized sounds.

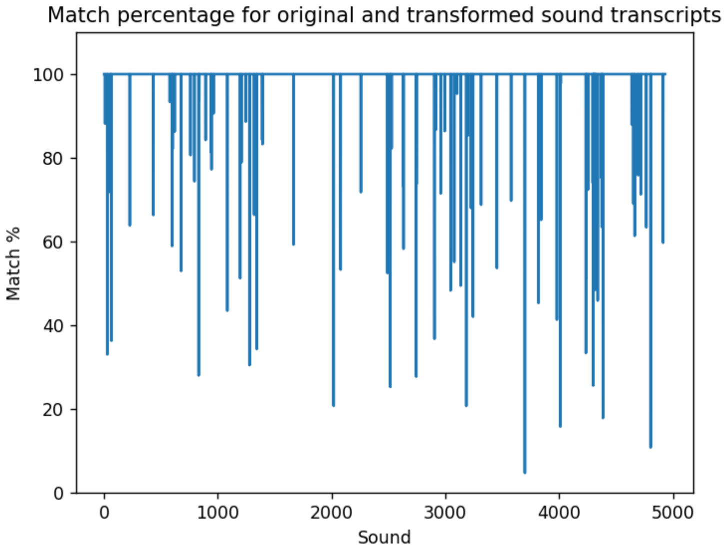

Figure 12. Transcription match pattern between original and neutralized sounds.

Finally, we inspected the transformed sounds and compared them to the original. From figures 13 and 14 you can see the frequency spectra of voices, before and after transformation. In the case of a female voice, frequencies in the vicinity of 400 Hz and over 800 Hz were drastically dampened. While in the case of male voice, some high-tier frequencies have increased in amplitude after transformations. Such behavior is expected as the female voice is associated with having more high-frequency components. Most importantly, by playing and listening to transformed sounds, we found that for human hearing, it was also tough to guess the gender of the speaker.

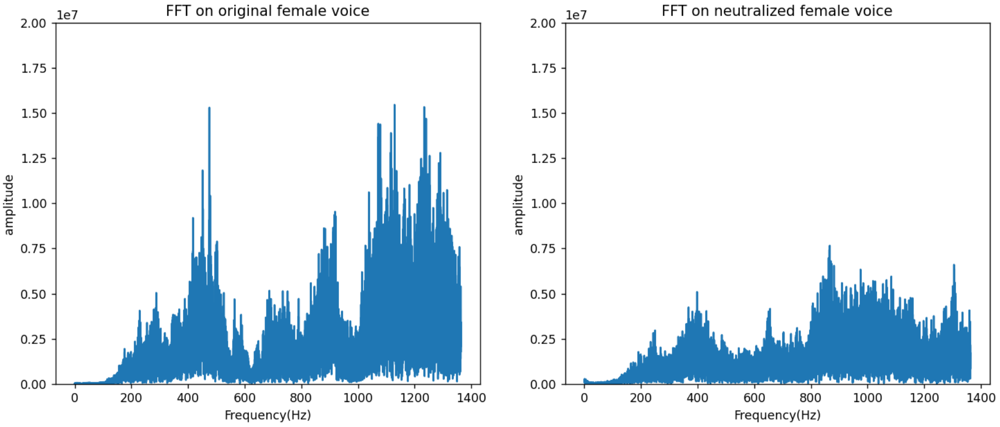

Figure 13. Frequency spectrum of female speech, before and after neutralization

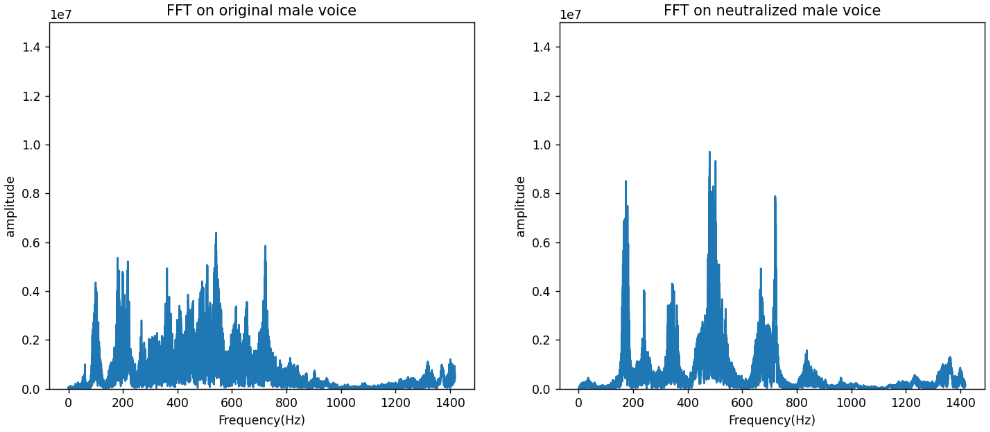

Figure 14. Frequency spectrum of male speech, before and after neutralization

6 Summary

The application of deep neural networks for speech and language processing has drastically increased during the last few years. Due to the advancements in these fields, ethical problems like gender bias have become inevitable. Such bias can cause sensitive decisions to be made inaccurately and cause a dispute by the affected class members. Gender bias can be found in most branches of speech processing and NLP, like emotion recognition, speech-to-text, speaker recognition, and machine translation. While multiple approaches exist that try to overcome it through dataset or network structure manipulations, most of them require more extensively labeled datasets, which mostly render them inapplicable for unsupervised learning or other real-world scenarios.

This thesis proposes a novel method for gender de-biasing in speech processing. Instead of focusing on dataset or model optimizations, our method implies the removal of gender parameters from speech data altogether. In order to achieve this, we employ the speech manipulation tools to transform the original sound to the point where gender becomes indistinguishable, thus rendering such parameters redundant for consideration in speech processing. Specifically, we have found that correct combination of pitch and formant shifts is sufficient for given task. As a benchmark for gender neutrality, we employ a Wav2Vec based speech gender recognition network, which demonstrated remarkable accuracy on validation, as well as on cross dataset tests.

Results of our gender neutralizing system have shown that transformations and validations are used to ensure that key aspects of the original sound, like carried emotion, stay the same and that there are no unwanted artifacts added. Furthermore, we have shown that our method for SGR excels in robustness towards sounds that differ in noise levels, language, and accents. Such a system can be used as a batch pre-processing tool for models in speech processing applications, where gender bias is an evident problem. By removing the gender factor from speech processing, we ultimately eliminate bias towards it as well.

It is essential to highlight that our implementation can be further optimized. For future work, we firmly believe that computation times for sound resynthesis can be cut down, and other types of transformations can be introduced as well. To further make the process close to real time, system can first "adapt" to particular speakers characteristics, and optimize the search for transformation parameters.

References

[1] Assim Ara Abdulsatar et al. “Age and gender recognition from speech signals”. In: *Journal of Physics: Conference Series* 1410.1 (Dec. 2019), p. 012073. DOI: 10.1088/1742-6596/1410/1/012073. URL: https://doi.org/10.1088/1742-6596/1410/1/012073.

[2] Md Ali, Md Islam, and Md Alamgir Hossain. “GENDER RECOGNITION SYSTEM USING SPEECH SIGNAL”. In: Vol.2 (Jan. 2012).

[3] Rami S. Alkhawaldeh. “DGR: Gender Recognition of Human Speech Using One-Dimensional Conventional Neural Network”. In: *Scientific Programming* 2019 (Sept. 2019), pp. 1–12. DOI: 10.1155/2019/7213717. URL: https://doi.org/10.1155/2019/7213717.

[4] Rami S. Alkhawaldeh. “DGR: Gender Recognition of Human Speech Using One-Dimensional Conventional Neural Network”. In: *Scientific Programming* Volume 2019 (2019). URL: https://doi.org/10.1155/2019/7213717.

[5] Ricardo Baeza-Yates. “Bias on the web”. In: *Communications of the ACM* 61.6 (2018), pp. 54–61.

[6] Joan Bajorek. “Voice Recognition Still Has Significant Race and Gender Biases”. In: *Harvard business review* (May 2019).

[7] Yalin Baştanlar and Mustafa Özuysal. “Introduction to machine learning”. In: *miRNomics: MicroRNA biology and computational analysis* (2014), pp. 105–128.

[8] R. K. E. Bellamy et al. “AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias”. In: *IBM Journal of Research and Development* 63.4/5 (2019), 4:1–4:15. DOI: 10.1147/JRD.2019.2942287.

[9] Richard Benjamins, Alberto Barbado, and Daniel Sierra. “Responsible AI by design in practice”. In: *arXiv preprint arXiv:1909.12838* (2019).

[10] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*. Vol. 4. 4. Springer, 2006.

[11] Jaime G Carbonell, Ryszard S Michalski, and Tom M Mitchell. “An overview of machine learning”. In: *Machine learning* (1983), pp. 3–23.

[12] L Elisa Celis et al. “Fair Distributions from Biased Samples: A Maximum Entropy Optimization Framework.” In: *arXiv preprint arXiv:1906.02164* (2019).

[13] D. G. Childers and Ke Wu. “Gender recognition from speech. Part II: Fine analysis”. In: *The Journal of the Acoustical Society of America* 90.4 (Oct. 1991), pp. 1841–1856. DOI: 10.1121/1.401664. URL: https://doi.org/10.1121/1.401664.

[14] Giovanni Luca Ciampaglia et al. “How algorithmic popularity bias hinders or promotes quality”. In: *Scientific reports* 8.1 (2018), pp. 1–7.

[15] Kevin A Clarke. “The phantom menace: Omitted variable bias in econometric research”. In: *Conflict management and peace science* 22.4 (2005), pp. 341–352.

[16] Lee Cohen, Zachary C Lipton, and Yishay Mansour. “Efficient candidate screening under multiple tests and implications for fairness.(2019)”. In: *arXiv preprint cs.LG/1905.11361* (2019).

[17] May I Conley et al. “The racially diverse affective expression (RADIATE) face stimulus set”. In: *Psychiatry research* 270 (2018), pp. 1059–1067.

[18] Matthew N Dailey et al. “Evidence and a computational explanation of cultural differences in facial expression recognition.” In: *Emotion* 10.6 (2010), p. 874.

[19] David Danks and Alex John London. “Algorithmic Bias in Autonomous Systems.” In: *IJCAI*. Vol. 17. 2017, pp. 4691–4697.

[20] Kurtis Evan David, Qiang Liu, and Ruth Fong. “Debiasing Convolutional Neural Networks via Meta Orthogonalization”. In: *arXiv preprint arXiv:2011.07453* (2020).

[21] Artem Domnich and Gholamreza Anbarjafari. “Responsible AI: Gender bias assessment in emotion recognition”. In: *arXiv preprint arXiv:2103.11436* (2021).

[22] Yupei Du, Yuanbin Wu, and Man Lan. “Exploring Human Gender Stereotypes with Word Association Test”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 6133–6143. DOI: 10.18653/v1/D19-1635. URL: https://aclanthology.org/D19-1635.

[23] Issam El Naqa and Martin J Murphy. "What is machine learning?" In: *machine learning in radiation oncology*. Springer, 2015, pp. 3–11.

[24] Michael Feldman et al. "Certifying and removing disparate impact". In: *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 2015, pp. 259–268.

[25] Tal Feldman and Ashley Peake. "On the Basis of Sex: A Review of Gender Bias in Machine Learning Applications". In: *arXiv e-prints* (2021), arXiv–2104.

[26] Agneta H Fischer et al. "Gender and culture differences in emotion." In: *Emotion* 4.1 (2004), p. 87.

[27] Agneta H. Fischer, Mariska E. Kret, and Joost Broekens. "Gender differences in emotion perception and self-reported emotional intelligence: A test of the emotion sensitivity hypothesis". In: *PLOS ONE* 13.1 (Jan. 2018). Ed. by Gilles van Luijtelaar, e0190712. DOI: 10.1371/journal.pone.0190712. URL: https://doi.org/10.1371/journal.pone.0190712.

[28] Malik Ghallab. "Responsible AI: requirements and challenges". In: *AI Perspectives* 1.1 (2019), pp. 1–7.

[29] Jelena Gorbova et al. "Integrating vision and language for first-impression personality analysis". In: *IEEE MultiMedia* 25.2 (2018), pp. 24–33.

[30] Cristina Gorrostieta et al. "Gender De-Biasing in Speech Emotion Recognition". In: *Interspeech 2019*. ISCA, Sept. 2019. DOI: 10.21437/interspeech.2019-1708. URL: https://doi.org/10.21437/interspeech.2019-1708.

[31] Su-Hyun Han et al. "Artificial neural network: understanding the basic concepts without mathematics". In: *Dementia and Neurocognitive Disorders* 17.3 (2018), pp. 83–89.

[32] Moritz Hardt, Eric Price, and Nati Srebro. "Equality of opportunity in supervised learning". In: *Advances in neural information processing systems* 29 (2016).

[33] Eszter Hargittai. "Whose space? Differences among users and non-users of social network sites". In: *Journal of computer-mediated communication* 13.1 (2007), pp. 276–297.

[34] Yinglong Jiang and Peter Murphy. "Voice Source Analysis for Pitch-Scale Modification of Speech Signals". In: (Dec. 2001).

[35] Matt J Kusner et al. “Counterfactual fairness”. In: *Advances in neural information processing systems* 30 (2017).

[36] Yann LeCun et al. “A theoretical framework for back-propagation”. In: *Proceedings of the 1988 connectionist models summer school*. Vol. 1. 1988, pp. 21–28.

[37] Kristina Lerman and Tad Hogg. “Leveraging position bias to improve peer recommendation”. In: *PloS one* 9.6 (2014), e98914.

[38] Frank Hung-Fat Leung et al. “Tuning of the structure and parameters of a neural network using an improved genetic algorithm”. In: *IEEE Transactions on Neural networks* 14.1 (2003), pp. 79–88.

[39] Sarah Levitan, Taniya Mishra, and Srinivas Bangalore. “Automatic identification of gender from speech”. In: May 2016, pp. 84–88. DOI: 10.21437/SpeechProsody.2016-18.

[40] Ninareh Mehrabi et al. “A Survey on Bias and Fairness in Machine Learning”. In: *ACM Computing Surveys* 54.6 (July 2021), pp. 1–35. DOI: 10.1145/3457607. URL: https://doi.org/10.1145/3457607.

[41] Hannah Jean Miller et al. ““Blissfully Happy” or “Ready toFight”: Varying Interpretations of Emoji”. In: *Tenth international AAAI conference on web and social media*. 2016.

[42] Brent Mittelstadt et al. “The Ethics of Algorithms: Mapping the Debate”. In: *Big Data & Society* In press (Oct. 2016). DOI: 10.1177/2053951716679679.

[43] Dong Nguyen et al. “" How old do you think I am?" A study of language and age in Twitter”. In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 7. 1. 2013.

[44] Alexandra Olteanu et al. “Social data: Biases, methodological pitfalls, and ethical boundaries”. In: *Frontiers in Big Data* 2 (2019), p. 13.

[45] E. Pépiot. “Male and female speech: A study of mean f0, f0 range, phonation type and speech rate in parisian French and American English speakers”. In: *Proceedings of the International Conference on Speech Prosody* (Jan. 2014), pp. 305–309.

[46] E. Ashby Plant et al. “The Gender Stereotyping of Emotions”. In: *Psychology of Women Quarterly* 24.1 (Mar. 2000), pp. 81–92. DOI: 10.1111/j.1471-6402.2000.tb01024.x. URL: https://doi.org/10.1111/j.1471-6402.2000.tb01024.x.

[47] Manish Raghavan et al. “Mitigating bias in algorithmic hiring: Evaluating claims and practices”. In: *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 2020, pp. 469–481.

[48] Andrinandrasana David Rasamoelina, Fouzia Adjailia, and Peter Sinčák. “A review of activation function for artificial neural network”. In: *2020 IEEE 18th World Symposium on Applied Machine Intelligence and Informatics (SAMI)*. IEEE. 2020, pp. 281–286.

[49] Davit Rizhinashvili, Abdallah Hussein Sham, and Gholamreza Anbarjafari. “Gender Neutralisation for Unbiased Speech Synthesising”. In: *Electronics* 11.10 (2022), p. 1594.

[50] Esther Rolf et al. “Balancing Competing Objectives with Noisy Data: Score-Based Classifiers for Welfare-Aware Machine Learning”. In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, 13–18 Jul 2020, pp. 8158–8168. URL: https://proceedings.mlr.press/v119/rolf20a.html.

[51] Arthur L Samuel. “Some studies in machine learning using the game of checkers. II—Recent progress”. In: *IBM Journal of research and development* 11.6 (1967), pp. 601–617.

[52] Beatrice Savoldi et al. “Gender Bias in Machine Translation”. In: *Transactions of the Association for Computational Linguistics* 9 (2021), pp. 845–874. DOI: 10.1162/tacl_a_00401. URL: https://doi.org/10.1162/tacl_a_00401.

[53] Steffen Schneider et al. “wav2vec: Unsupervised pre-training for speech recognition”. In: *arXiv preprint arXiv:1904.05862* (2019).

[54] Mohammad Sedaaghi. “A Comparative Study of Gender and Age Classification in Speech Signals”. In: *Iranian Journal of Electrical & Electronic Engineering* 5 (Mar. 2009).

[55] Abdallah Hussein Sham et al. "Ethical AI in facial expression analysis: racial bias". In: *Signal, Image and Video Processing* (2022), pp. 1–8.

[56] Tony Sun et al. "Mitigating gender bias in natural language processing: Literature review". In: *arXiv preprint arXiv:1906.08976* (2019).

[57] Harini Suresh and John V Guttag. "A framework for understanding unintended consequences of machine learning". In: *arXiv preprint arXiv:1901.10002* 2 (2019).

[58] William Thong and Cees GM Snoek. "Feature and Label Embedding Spaces Matter in Addressing Image Classifier Bias". In: *arXiv preprint arXiv:2110.14336* (2021).

[59] S. Vallor. "Artificial Intelligence and Public Trust". In: 58(2) (2017), pp. 42–45.

[60] Sun-Chong Wang. "Artificial neural network". In: *Interdisciplinary computing in java programming*. Springer, 2003, pp. 81–100.

[61] Tianlu Wang et al. "Adversarial removal of gender from deep image representations". In: *arXiv preprint arXiv:1811.08489* 3 (2018).

[62] Tianlu Wang et al. "Balanced Datasets Are Not Enough: Estimating and Mitigating Gender Bias in Deep Image Representations". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2019.

[63] Bayya Yegnanarayana. *Artificial neural networks*. PHI Learning Pvt. Ltd., 2009.

Non-exclusive licence to reproduce thesis and make thesis public

I, **Davit Rizhinashvili**,

*(*author's name)

1. grant the University of Tartu a free permit (non-exclusive licence) to

   reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright, my thesis

   Gender Neutralisation for Unbiased Speech Synthesising,

   *(*title of thesis)

   supervised by Abdallah Hussein Sham and Gholamreza Anbarjafari.

   *(*supervisor's name)

2. I grant the University of Tartu a permit to make the thesis specified in point 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 4.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in points 1 and 2.

4. I confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Davit Rizhinashvili
25/05/2022