electronics

MDPI

Article

Gender Neutralisation for Unbiased Speech Synthesising

Davit Rizhinashvili [1], Abdallah Hussein Sham [2,*] and Gholamreza Anbarjafari [1,3,4,5]

[1] iCV Lab, University of Tartu, Narva mnt 18, 51009 Tartu, Estonia; rizhinashvilidavid@gmail.com (D.R.); shb@ut.ee (G.A.)
[2] Enactive Virtuality Lab, Narva mnt 25, 10120 Tallinn, Estonia
[3] iVCV OÜ, Purpuri tn 12-2, 51011 Tartu, Estonia
[4] PwC Advisory, Itämerentori 2, 00180 Helsinki, Finland
[5] Institute of Higher Education, Yildiz Technical University, Yildiz, Beşiktaş District, Istanbul 34349, Turkey
* Correspondence: ahsham@tlu.ee; Tel.: +372-5677-0200

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 labelling, inaccurate representation of different population groups in training datasets, and chosen model structures and optimization methods. Our paper 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 to validate the main claim of this research work, which is the 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: responsible AI; speech analysis; emotion recognition; gender bias

check for updates

Citation: Rizhinashvili, D.; Sham, A.H.; Anbarjafari, G. Gender Neutralisation for Unbiased Speech Synthesising. *Electronics* **2022**, *11*, 1594. https://doi.org/10.3390/electronics11101594

Academic Editor: George A. Papakostas

Received: 15 April 2022
Accepted: 15 May 2022
Published: 17 May 2022

1. 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 [1–3].

Most machine learning models are designed to optimize only one performance metric, such as accuracy. Such designs inadvertently have consequences [4], having discriminatory results based on sensitive features, such as gender, and are considered to be 'biased' or 'unfair'. Examples of gender-based unfairness in speech processing applications are abundant. For instance, until recently, speech synthesis and speech recognition favoured lower pitch voices [5], typically present in adult males. As a result, speech recognition produced higher error rate scores for children and adult females.

Several factors can contribute to producing negative bias in ML models. One significant cause is incomplete training data 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 labelled data comes from people. Despite the effects of the dataset, since people are unintentionally 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 [6,7]. Similarly, as mentioned, humans might be gender-biased as well. While there are conventional methods for de-biasing AI, like modification of loss function to minimize the bias towards protected variable [8], they can not be applied to humans.

According to [9], one of the definitions of fairness can be considered “Fairness through unawareness”, which states that the model achieves fairness towards a certain attribute, if such attribute is not utilized to make predictions [10]. 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 and AI.

One of the crucial elements of gender-neutral speech synthesis is the speech gender recognition (SGR) network, which must be highly accurate and robust towards accents, languages, and noise that the sound may carry. Such a network has to validate that perceived gender is hard to distinguish. Generally, it is believed that perceived gender depends mainly on the pitch, the male being in the range of 85 to 155 Hz and female 165 to 255 Hz [11], though, in reality, factors like higher tier frequencies, pitch fluctuations, and overall pitch contour shape have a significant effect on it. Further, both male and female pitches can go out of those boundaries in many cases. Figure 1 shows the sorted distribution of average pitch for 720 female and male speakers, respectively. Figure 2 shows the pitch contours for female and male speakers expressing happy emotions over the same sentence.

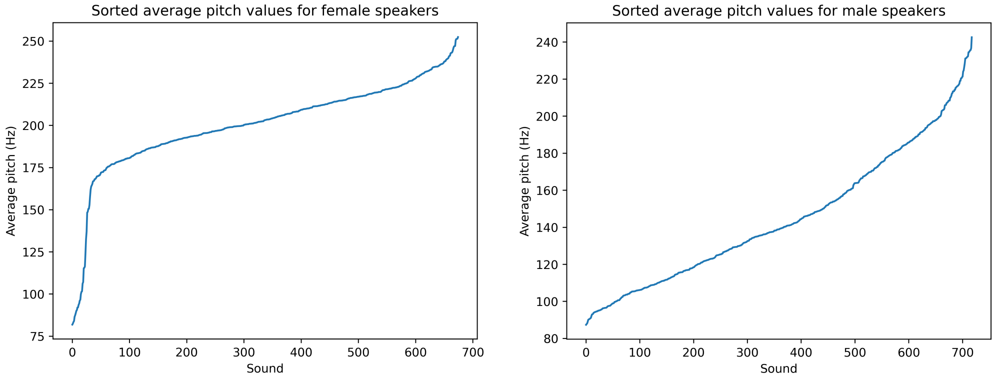

Figure 1. Sorted average pitch values for 1440 utterances by female and male speakers separately. Sounds from each gender category were analysed for average pitch and then sorted for illustrative purposes.

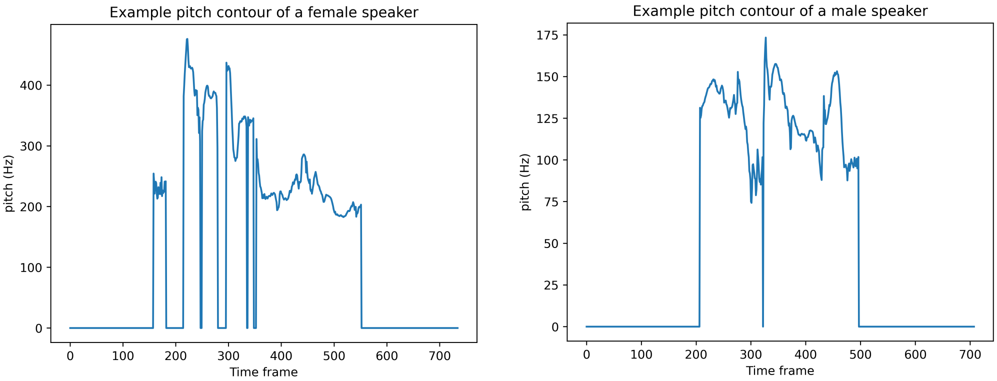

Figure 2. Pitch contour for female and male speakers on example utterance.

Most of the previous researches [12,13] 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, and its main contributions are as follows:

- We assembled and trained a network for SGR, with previously unseen accuracy and robustness towards a variety of sounds.

- We created a pipeline for eliminating gender parameters from the speech.
- We demonstrated that said method does not create unwanted artefacts in the sound, showing that speech-to-text performance and conveyed emotion stay the same.

The remaining of the paper is organized as follows: Section 2 gives an overview of the related works. Section 3 describes the used datasets, the experimental methods, and the training. Section 4 presents the experimental results and the discussion. Lastly, Section 5 concludes the paper.

2. Related Work

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 can be used in many sensitive environments to make important and life-changing decisions, such as interviews [14], hiring [15] and personality analysis [16]. Thus, it is crucial to ensure that these decisions do not reflect discriminatory behaviour towards certain groups or populations.

More recently, some work has been developed in traditional ML and deep learning (DL) techniques that address such challenges in different sub-domains [17]. These sub-domains can include different aspects of possible bias sources, be it gender, race, age, and others. For example, in [2], authors show which models are gender-biased, which are not, and how the gender of the subject affects its emotion recognition. They also describe the extent of this 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 men and women. In [8], authors demonstrate that the activation model for Speaker Emotion Recognition (SER) is negatively biased towards women, meaning that said models consistently display lower activation for women than men.

Furthermore, psychologists did many studies and found that there was a correlation between culture [18], race [19], gender [20], and cultural differences in emotions. For example, in [21], 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.

Better performance has been reported when separate acoustic models are employed for males and females [22], thus rendering SGR very significant. Many methods for SGR have been described in [23–27] and while most show relatively high accuracy rates, their robustness towards different aspects of input sound, such as quality, noise levels, language, and others can be questioned. Moreover, most of these methods rely on feature extraction, such as pitch and multiple spectrograms, for which multiple methods exist, and they ultimately differ in their output.

Natural language processing (NLP), speech-to-text, and general sound recognition can also be the subject of gender bias. Authors of [28] mention that current state-of-the-art speech recognition is 13% more accurate for men than it is for women. Furthermore, 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 [29] highlight the said bias for machine translation, which facilitates itself by creating social gender expectations—for example, translating engineers as masculine and nurses as feminine. Similar stereotypical gender bias can be observed in humans as well. For instance, authors of [30] utilized a word association test to assess gender stereotypes in texts and found that bias scores correlate well with bias in the real world.

On another note, we can discuss the possible adjustments needed for AI fairness. In [31], the 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 [32] explain how to use design methodology to create a responsible and fair AI. Authors of [10] provide a summary of formal definitions for AI fairness, those being 'Fairness through unawareness' [9], 'Counterfactual fairness' [9], 'Statistical Parity Difference' [9], 'Equal Opportunity Difference' [33], 'Average Odds Difference' [34] and 'Disparate Impact' [35]. 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 [36], 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.

Finally, we can look at the methods and approaches that try to overcome bias in AI. As discussed, most approaches try to target and optimize the network, training, or datasets; for example, in their comprehensive review paper, authors of [12] 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 [8] 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 [13,37–39] against gender bias in AI.

3. Methodology

This section gives an overview and details about the resources and methods used. Section 3.1 describes the datasets that are used for the project and outlines the reason why each of them was chosen. Section 3.2 describes the method for speech gender recognition. Lastly, Section 3.3 portrays the whole sound resynthesis process.

3.1. Databases

There are many speech datasets (https://github.com/coqui-ai/open-speech-corpora) (accessed on 15 December 2021) 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	7442	Male + Female	6	English
DEMoS	9365	Male + Female	6	Italian
TESS	2800	only female	7	English
AudioMNIST	30,000	Male + Female	–	English
EMO-DB	535	Male + Female	7	German
LibriSpeech	280,000	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 16 kHz sampling frequencies. For pre-processing, we set the intensity of each sound to 70 dB using Praat (https://www.fon.hum.uva.nl/praat/) (accessed on 5 January 2022). The following subsections describe the datasets we used.

3.1.1. RAVDESS

Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) (https://smartlaboratory.org/ravdess/) (accessed on 15 December 2021) the database contains 1440 audio files of 24 professional actors (12 female, 12 male), vocalizing two lexicallymatched statements in a neutral North American accent. Speech includes calm, happy, sad, angry, fearful, surprised, and disgusted 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.

3.1.2. CREMA-D

CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset) (https://github.com/CheyneyComputerScience/CREMA-D) (accessed on 15 December 2021) is a data set of 7442 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 being quite noisy. Labels include the gender of the speaker and convey emotion.

3.1.3. EMO-DB

EMO-DB (Berlin Database of Emotional Speech) (http://emodb.bilderbar.info/) (accessed on 15 December 2021) is a data set of 535 audio clips from 10 actors (5 male and 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.

3.1.4. LibriSpeech

LibriSpeech (https://www.openslr.org/12/) (accessed on 20 January 2022) is a corpus of approximately 1000 h 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 s 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 performance programs on unmodified and modified sounds to prove the absence of any unwanted artefacts after the "neutralization" step.

3.2. 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 [23]. Though many described models achieve almost perfect accuracy, some of them may be prone to over-fitting and perform purely 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 sample to sample. We propose a non-traditional yet quite simple method that is robust to the aforementioned problems.

We propose the following model for recognizing gender from speech, and it 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 [40] 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 25 ms frame in the sound and comprise a vector with 512 values.

For classification, the output of the Wav2Vec network over the input sound is averaged over time, yielding a vector with 512 values. In other words, for 100 ms sound, the output of Wav2Vec will be a 512 × 4 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 3.

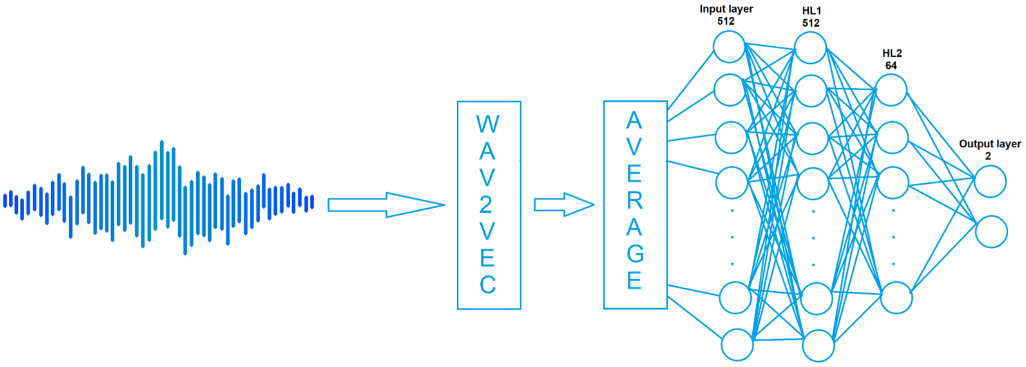

Figure 3. Structure of our SGR network.

3.3. Neutralization Process

As previously discussed, multiple types of alterations can be applied to the sound to change its perceived gender, including frequency shifting, filtering, time stretching and more. In line with this point, we have found that the combination of correctly chosen pitch and formant shifts (we refer to these 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 the above-mentioned goal. For modifying and resynthesizing sounds, we use *Praat* and its implementation in python called parselMouth (https://pypi.org/project/praat-parselmouth/) (accessed on 5 January 2022). 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.

To check that the speech is indeed gender-neutral, we use the SGR network described in Section 3.2. As the output of SGR is accompanied by the certainty of a 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. 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, the system transforms the speech with every possible combination of parameters (we use steps of 3 Hz in the range of 50 Hz to 240 Hz for pitch and steps of 0.01 in the 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 (1)). Such a search is quite time consuming and mostly unnecessary. For example, increasing the frequencies of the female voice will most certainly make it sound more feminine, thus our system should not waste time checking all possible combinations.

$$|P_{male} - P_{female}| < 10\% \tag{1}$$

where P is the certainty of given gender prediction.

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, the 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 4 for male samples and Figure 5 for female samples. Based on these correlation figures, we were able to devise the corresponding Equations (2) and (3) for males and Equations (4) and (5) for females.

For male:

$$P_{male} = 150 + \frac{sp}{1.63} \tag{2}$$

$$fs = 1.15 - \frac{p}{1500} \tag{3}$$

For female:

$$P_{female} = 140 - \frac{sp}{1.96} \tag{4}$$

$$fs = 0.8 + \frac{20}{p} \tag{5}$$

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 the pitch from the initial sound and using the 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, the search can take up an additional 100 iterations, while, on average with a coarse neutralization step, the process takes 11 iterations to find the correct transformation parameters.

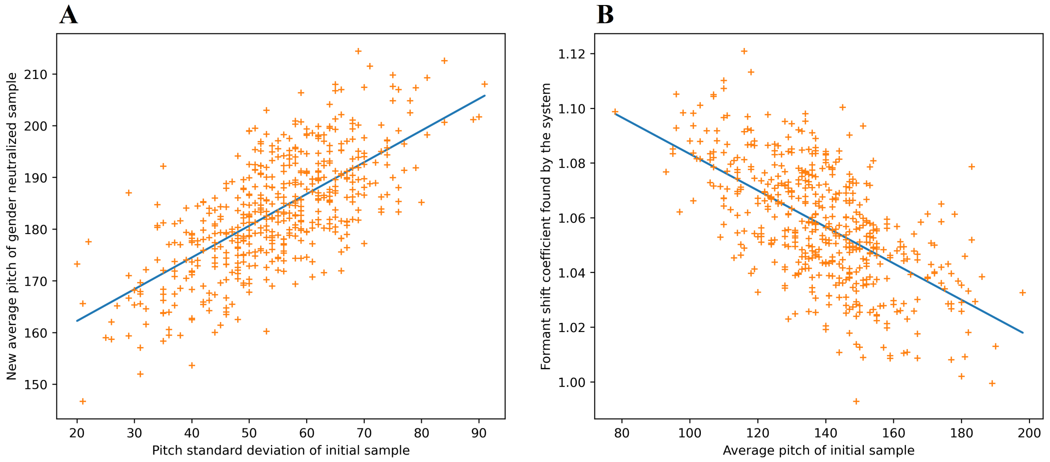

Figure 4. Correlation of initial sound features and correct transformation parameters for gender neutralization for male samples. Orange '+' markers correspond to each sound sample and blue line is representing equation of corresponding correlation. (**A**) Correlation of initial pitch standard deviation and new pitch value of neutralized sound. (**B**) Correlation of initial average pitch to the formant shift coefficient in neutralized sound.

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 (https://vokaturi.com/) (accessed on 3 February 2022) software's Python API to measure the emotion of the initial sample and during parameter search, prioritize the transformations where emotion was not affected. In other words, if the system finds a pitch and formant shift combination for which SGR outputs certainties close to 50% but the measure emotion differs from the initial, such combination will be discarded.

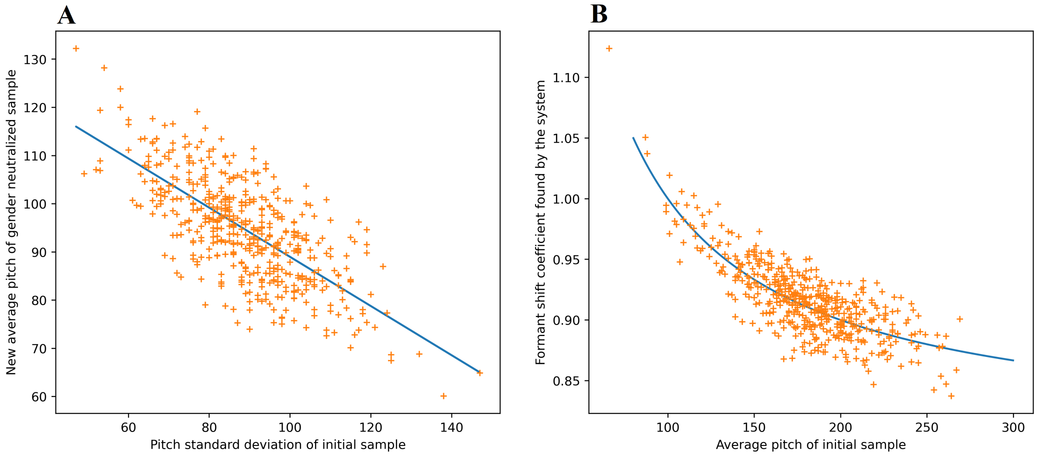

Figure 5. Correlation of initial sound features and correct transformation parameters for gender neutralization for female samples. Orange '+' markers correspond to each sound sample and blue line is representing equation of corresponding correlation. (**A**) Correlation of initial pitch standard deviation and new pitch value of neutralized sound. (**B**) Correlation of initial average pitch to the formant shift coefficient in neutralized sound.

All in all, our gender neutralization system can be split into two parts. The first is coarse neutralization, where we extract the features from the initial speech sample and use them to find a probable location of correct transformation parameters. And second, a feedback loop comprising sound resynthesis, SGR and SER networks. The loop starts from the combination of transformation parameters calculated by the coarse neutralization step. Then for each iteration, it slightly changes the 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 (1)) provided that there is no modification in the carried emotion from the given transformation. After such a pair is found, the loop is halted and the corresponding sound is saved. A block diagram of the gender neutralization system can be seen in Figure 6.

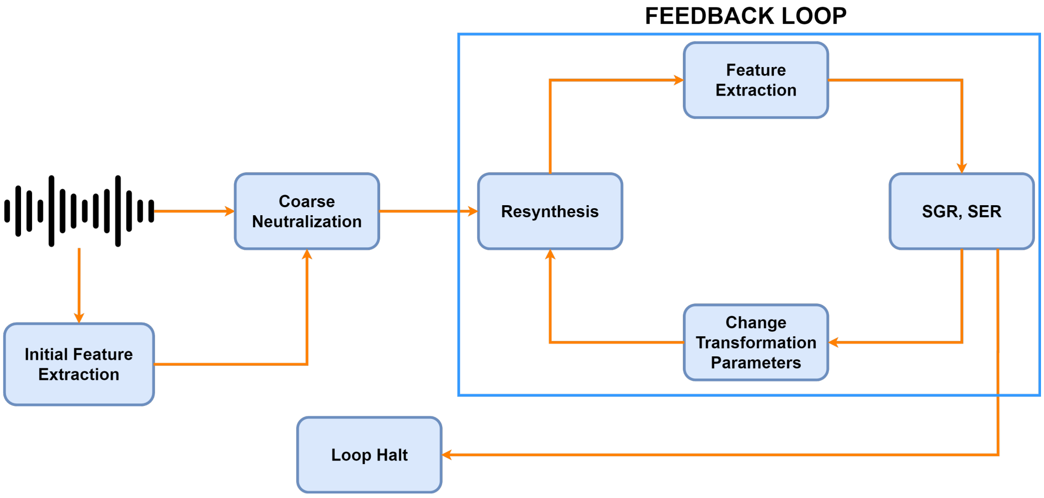

Figure 6. Diagram of our proposed gender neutralization system.

4. Experimental Results and Discussion

To start, we evaluate the robustness and accuracy of our SGR method. For comparison purposes, we have employed the approach described in [23], namely, pre-processing, feature (MFCC, chroma, mel, and tonnetz) extraction, and classification using MLP. We then trained it using the RAVDESS dataset, and the network reached an accuracy of 98.4%

on the validation set. When we cross-checked the pre-trained network on a different dataset (EMO-DB), the accuracy was 56.9%, suggesting that this method is prone to over-fitting. The confusion matrix of such cross dataset test can be seen in Figure 7A.

On the other hand, our proposed network, trained on LibriSpeech (see Section 3.1.4, trained using 80% of the dataset and tested using 20%), shows similar accuracy of 99.87% on the validation set as seen in Figure 7B, while also excelling in cross dataset tests. Figure 7C,D show the confusion matrices of our method tested against RAVDESS and EMO-DB, reaching 96.9% and 97.3% respectively.

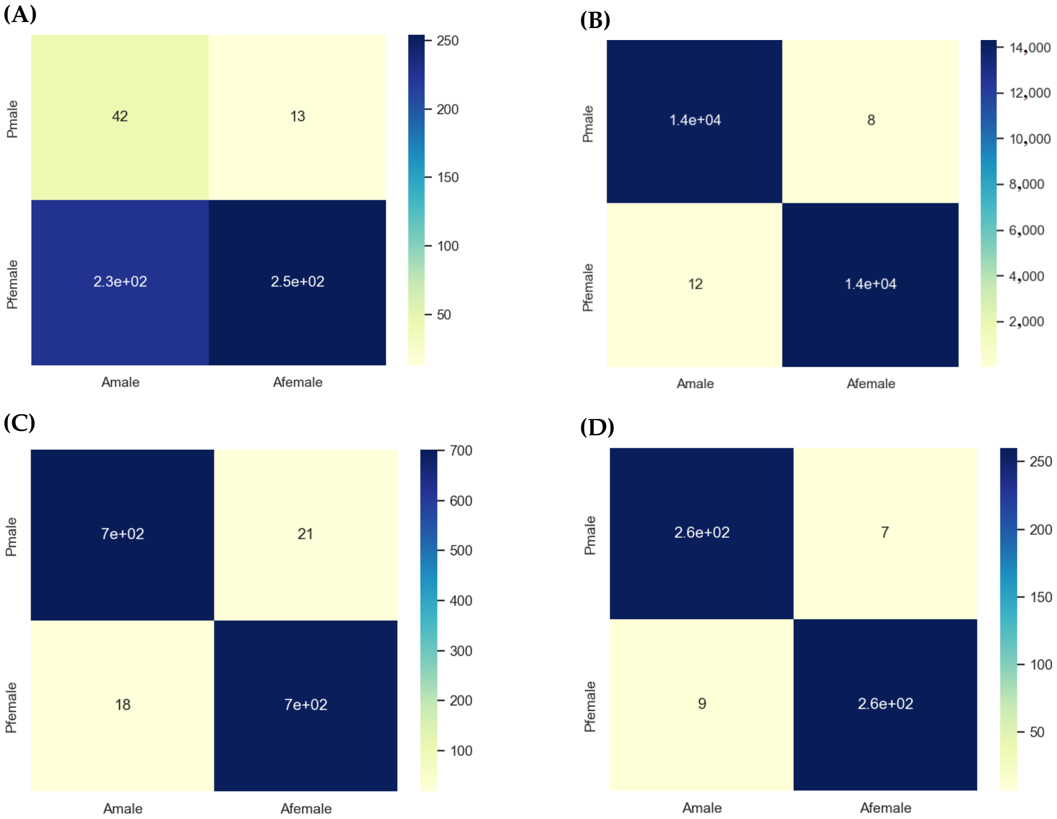

Figure 7. Confusion matrices of SGR methods tested on different datasets (**A**) Confusion matrix of SGR method from [23] trained on RAVDESS and tested on EMO-DB. (**B**) Confusion matrix of our SGR method trained on LibriSpeech. (**C**) Confusion matrix of our SGR method trained on LibriSpeech and tested on RAVDESS. (**D**) Confusion matrix of our SGR method trained on LibriSpeech and tested on EMO-DB.

After the proposed system described in Section 3.3 was assembled, we needed to check emotion integrity and the absence of the added sound artefacts. 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% match between original and transformed files, see Figure 8. 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.

To verify the absence of any unwanted artefacts, 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 (https://cloud.google.com/speech-to-text/) (accessed on 20 February 2022). To compare transcriptions, we used FuzzyWuzzy library (https://github.com/seatgeek/fuzzywuzzy) (accessed on 20 February 2022), 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 and Cross-checking transcriptions yielded 98.7% match, see Figure 9. 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.

Finally, we inspected the transformed sounds and compared them to the original. From Figures 10 and 11 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 behaviour is expected as the female voice is associated with having more high-frequency components. Most importantly, as far as the authors' impression is concerned, by playing and listening to transformed sounds, it was very hard to imply on the gender of the speaker.

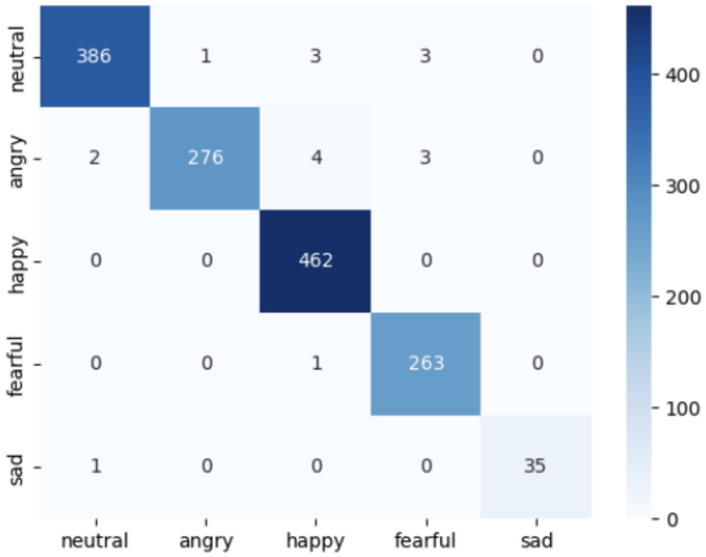

Figure 8. Emotion predictions on original and Transformed sounds from RAVDESS dataset.

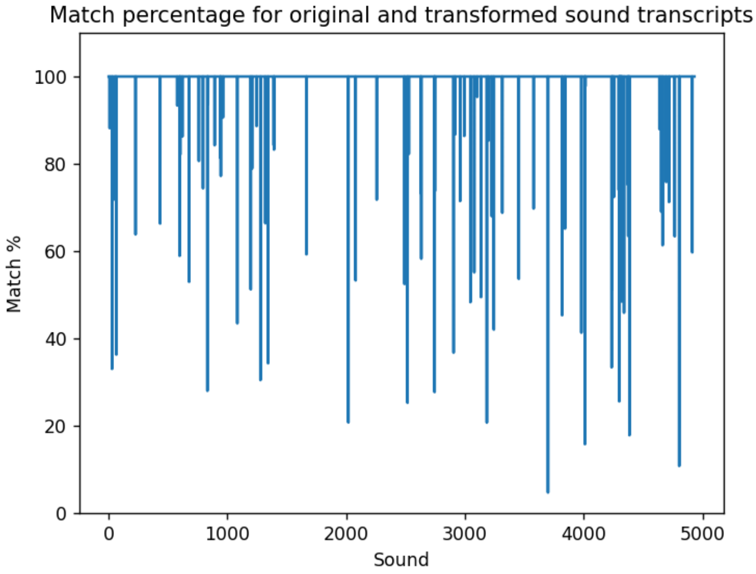

Figure 9. Transcription match pattern between original and neutralized sounds.

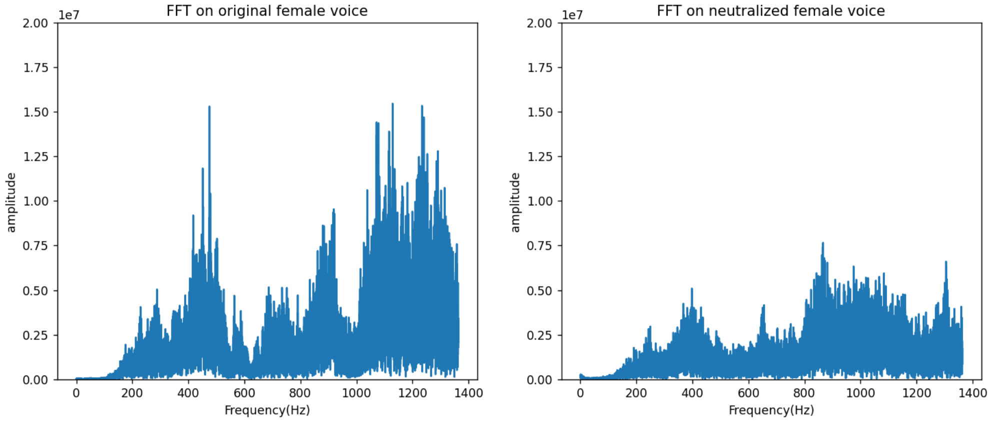

Figure 10. Frequency spectrum of female speech, before and after neutralization.

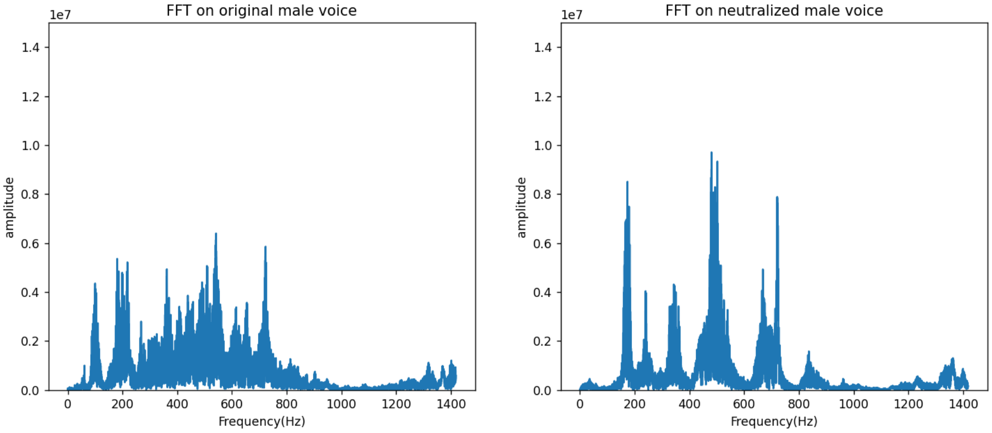

Figure 11. Frequency spectrum of male speech, before and after neutralization.

To sum up, we demonstrated that our method for SGR performs with similarly high accuracy towards validation or completely different sets. We assembled the system for speaker gender neutralization using said SGR for validation, which also includes some optimizations to cut down processing time finally, we have shown that transformations used in this method are non-destructive for the sound and all key aspects of it stay almost untouched.

5. Conclusions

Our paper proposes a novel approach 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. 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 the correct combination of pitch and formant shifts is sufficient for the 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 artefacts 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.

Author Contributions: Conceptualization, D.R., A.H.S. and G.A.; methodology, D.R., A.H.S. and G.A.; software, D.R.; validation, D.R., A.H.S. and G.A.; formal analysis, D.R., A.H.S. and G.A.; investigation, D.R., A.H.S. and G.A.; resources, D.R., A.H.S. and G.A.; data curation, D.R., A.H.S. and G.A.; writing—original draft preparation, D.R., A.H.S. and G.A.; writing—review and editing, D.R., A.H.S. and G.A.; visualization, D.R., A.H.S. and G.A.; supervision, A.H.S. and G.A.; project administration, D.R., A.H.S. and G.A.; funding acquisition, G.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: This work has been partially supported by the EU Mobilitas Pluss grant (MOBTT90) (2017–2022).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Mittelstadt, B.; Allo, P.; Taddeo, M.; Wachter, S.; Floridi, L. The Ethics of Algorithms: Mapping the Debate. *Big Data Soc.* **2016**, *3*, 2053951716679679. [CrossRef]
2. Domnich, A.; Anbarjafari, G. Responsible AI: Gender bias assessment in emotion recognition. *arXiv* **2021**, arXiv:2103.11436.
3. Sham, A.H.; Aktas, K.; Rizhinashvili, D.; Kuklianov, D.; Alisinanoglu, F.; Ofodile, I.; Ozcinar, C.; Anbarjafari, G. Ethical AI in facial expression analysis: Racial bias. *Signal Image Video Process.* **2022**, 1–8. [CrossRef]
4. Rolf, E.; Simchowitz, M.; Dean, S.; Liu, L.T.; Bjorkegren, D.; Hardt, M.; Blumenstock, J. Balancing Competing Objectives with Noisy Data: Score-Based Classifiers for Welfare-Aware Machine Learning. In Proceedings of the 37th International Conference on Machine Learning, Virtual Event, 13–18 July 2020; Volume 119, pp. 8158–8168.
5. Jiang, Y.; Murphy, P. Voice Source Analysis for Pitch-Scale Modification of Speech Signals. In Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFX-01), Limerick, Ireland, 6–8 December 2001.
6. Fischer, A.H.; Kret, M.E.; Broekens, J. Gender differences in emotion perception and self-reported emotional intelligence: A test of the emotion sensitivity hypothesis. *PLoS ONE* **2018**, *13*, e0190712. [CrossRef] [PubMed]
7. Vallor, S. Artificial Intelligence and Public Trust. *Santa Clara Mag.* **2017**, *58*, 42–45.
8. Gorrostieta, C.; Lotfian, R.; Taylor, K.; Brutti, R.; Kane, J. Gender De-Biasing in Speech Emotion Recognition. In Proceedings of the Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, ISCA, Graz, Austria, 15–19 September 2019. [CrossRef]
9. Kusner, M.J.; Loftus, J.; Russell, C.; Silva, R. Counterfactual fairness. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30.
10. Feldman, T.; Peake, A. On the Basis of Sex: A Review of Gender Bias in Machine Learning Applications. *arXiv* **2021**, arXiv:2104.02532v1.
11. Pépiot, E. 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, Dublin, Ireland, 20–23 May 2014; pp. 305–309.
12. Sun, T.; Gaut, A.; Tang, S.; Huang, Y.; ElSherief, M.; Zhao, J.; Mirza, D.; Belding, E.; Chang, K.W.; Wang, W.Y. Mitigating gender bias in natural language processing: Literature review. *arXiv* **2019**, arXiv:1906.08976.
13. Wang, T.; Zhao, J.; Yatskar, M.; Chang, K.W.; Ordonez, V. 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), Seoul, Korea, 27 October–2 November 2019.
14. Cohen, L.; Lipton, Z.C.; Mansour, Y. Efficient candidate screening under multiple tests and implications for fairness. *arXiv* **2019**, arXiv:1905.11361.
15. Raghavan, M.; Barocas, S.; Kleinberg, J.; Levy, K. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, 27–30 January 2020; pp. 469–481.
16. Gorbova, J.; Avots, E.; Lüsi, I.; Fishel, M.; Escalera, S.; Anbarjafari, G. Integrating vision and language for first-impression personality analysis. *IEEE MultiMedia* **2018**, *25*, 24–33. [CrossRef]
17. Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; Galstyan, A. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.* **2021**, *54*, 1–35. [CrossRef]
18. Dailey, M.N.; Joyce, C.; Lyons, M.J.; Kamachi, M.; Ishi, H.; Gyoba, J.; Cottrell, G.W. Evidence and a computational explanation of cultural differences in facial expression recognition. *Emotion* **2010**, *10*, 874–893. [CrossRef] [PubMed]
19. Conley, M.I.; Dellarco, D.V.; Rubien-Thomas, E.; Cohen, A.O.; Cervera, A.; Tottenham, N.; Casey, B. The racially diverse affective expression (RADIATE) face stimulus set. *Psychiatry Res.* **2018**, *270*, 1059–1067. [CrossRef]

20. Fischer, A.H.; Rodriguez Mosquera, P.M.; Van Vianen, A.E.; Manstead, A.S. Gender and culture differences in emotion. *Emotion* **2004**, *4*, 87–94. [CrossRef]
21. Plant, E.A.; Hyde, J.S.; Keltner, D.; Devine, P.G. The Gender Stereotyping of Emotions. *Psychol. Women Q.* **2000**, *24*, 81–92. [CrossRef]
22. Sedaaghi, M. A Comparative Study of Gender and Age Classification in Speech Signals. *Iran. J. Electr. Electron. Eng.* **2009**, *5*, 1–12.
23. Alkhawaldeh, R.S. DGR: Gender Recognition of Human Speech Using One-Dimensional Conventional Neural Network. *Sci. Program.* **2019**, *2019*, 7213717. [CrossRef]
24. Childers, D.G.; Wu, K. Gender recognition from speech. Part II: Fine analysis. *J. Acoust. Soc. Am.* **1991**, *90*, 1841–1856. [CrossRef] [PubMed]
25. Abdulsatar, A.A.; Davydov, V.V.; Yushkova, V.V.; Glinushkin, A.P.; Rud, V.Y. Age and gender recognition from speech signals. *J. Phys. Conf. Ser.* **2019**, *1410*, 012073. [CrossRef]
26. Levitan, S.; Mishra, T.; Bangalore, S. Automatic identification of gender from speech. In Proceedings of the Speech Prosody 2016, Boston, MA, USA, 31 May–3 June 2016; pp. 84–88. [CrossRef]
27. Ali, M.; Islam, M.; Hossain, M.A. Gender recognition system using speech signal. *Int. J. Comput. Sci. Eng. Inf. Technol. (IJCSEIT)* **2012**, *2*, 1–9.
28. Bajorek, J. Voice Recognition Still Has Significant Race and Gender Biases. Available online: 2019. https://hbr.org/2019/05/voice-recognition-still-has-significant-race-and-gender-biases (accessed on 10 January 2022).
29. Savoldi, B.; Gaido, M.; Bentivogli, L.; Negri, M.; Turchi, M. Gender Bias in Machine Translation. *Trans. Assoc. Comput. Linguist.* **2021**, *9*, 845–874. [CrossRef]
30. Du, Y.; Wu, Y.; Lan, M. 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, 3–7 November 2019; Association for Computational Linguistics: Hong Kong, China, 2019; pp. 6133–6143. [CrossRef]
31. Ghallab, M. Responsible AI: Requirements and challenges. *AI Perspect.* **2019**, *1*, 1–7. [CrossRef]
32. Benjamins, R.; Barbado, A.; Sierra, D. Responsible AI by design in practice. *arXiv* **2019**, arXiv:1909.12838.
33. Hardt, M.; Price, E.; Srebro, N. Equality of opportunity in supervised learning. In Proceedings of the Advances in Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; Volume 29.
34. Bellamy, R.K.E.; Dey, K.; Hind, M.; Hoffman, S.C.; Houde, S.; Kannan, K.; Lohia, P.; Martino, J.; Mehta, S.; Mojsilović, A.; et al. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM J. Res. Dev.* **2019**, *63*, 4:1–4:15. [CrossRef]
35. Feldman, M.; Friedler, S.A.; Moeller, J.; Scheidegger, C.; Venkatasubramanian, S. Certifying and removing disparate impact. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, Australia, 10–13 August 2015; pp. 259–268.
36. Celis, L.E.; Keswani, V.; Yildiz, O.; Vishnoi, N.K. Fair Distributions from Biased Samples: A Maximum Entropy Optimization Framework. *arXiv* **2019**, arXiv:1906.02164.
37. Wang, T.; Zhao, J.; Chang, K.W.; Yatskar, M.; Ordonez, V. Adversarial removal of gender from deep image representations. *arXiv* **2018**, arXiv:1811.08489.
38. Thong, W.; Snoek, C.G. Feature and Label Embedding Spaces Matter in Addressing Image Classifier Bias. *arXiv* **2021**, arXiv:2110.14336.
39. David, K.E.; Liu, Q.; Fong, R. Debiasing Convolutional Neural Networks via Meta Orthogonalization. *arXiv* **2020**, arXiv:2011.07453.
40. Schneider, S.; Baevski, A.; Collobert, R.; Auli, M. wav2vec: Unsupervised pre-training for speech recognition. *arXiv* **2019**, arXiv:1904.05862.