Elucidate Gender Fairness in Singing Voice Transcription

Xiangming Gu
Integrative Sciences and Engineering Programme, NUS Graduate School, National University of Singapore
Singapore
xiangming@u.nus.edu

Wei Zeng
Integrative Sciences and Engineering Programme, NUS Graduate School, National University of Singapore
Singapore
w.zeng@u.nus.edu

Ye Wang
School of Computing, National University of Singapore
Singapore
wangye@comp.nus.edu.sg

ABSTRACT

It is widely known that males and females typically possess different sound characteristics when singing, such as timbre and pitch, but it has never been explored whether these gender-based characteristics lead to a performance disparity in singing voice transcription (SVT), whose target includes pitch. Such a disparity could cause fairness issues and severely affect the user experience of downstream SVT applications. Motivated by this, we first demonstrate the female superiority of SVT systems, which is observed across different models and datasets. We find that different pitch distributions, rather than gender data imbalance, contribute to this disparity. To address this issue, we propose using an attribute predictor to predict gender labels and adversarially training the SVT system to enforce the gender-invariance of acoustic representations. Leveraging the prior knowledge that pitch distributions may contribute to the gender bias, we propose conditionally aligning acoustic representations between demographic groups by feeding note events to the attribute predictor. Empirical experiments on multiple benchmark SVT datasets show that our method significantly reduces gender bias (up to more than 50%) with negligible degradation of overall SVT performance, on both in-domain and out-of-domain singing data, thus offering a better fairness-utility trade-off.

CCS CONCEPTS

• **Applied computing** → **Sound and music computing**; • **Information systems** → *Music retrieval*; *Speech / audio search*; • **Social and professional topics** → **Gender**; • **Networks** → *Network reliability*.

KEYWORDS

fairness, singing voice transcription, pitch, adversarial learning, bias, fairness-utility trade-off

ACM Reference Format:
Xiangming Gu, Wei Zeng, and Ye Wang. 2023. Elucidate Gender Fairness in Singing Voice Transcription. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3581783.3612272

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
MM '23, October 29-November 3, 2023, Ottawa, ON, Canada
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3612272

1 INTRODUCTION

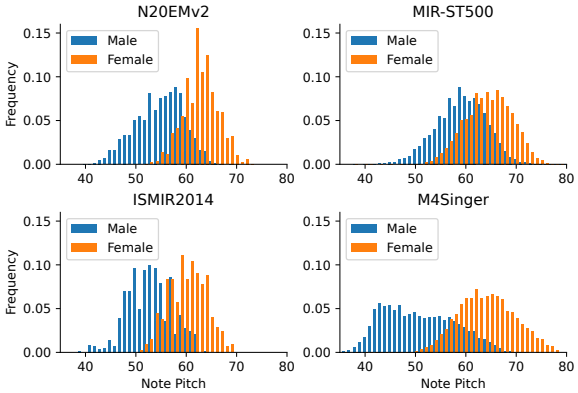

Figure 1: Pitch distributions of various singing voice transcription datasets.

Singing is a rich audio signal that consists of information from two aspects: textual and musical. The textual modality pertains to the lyrics, while the musical modality encompasses note events. Automatic lyric transcription (ALT) can be used to retrieve lyrical/textual data, as evidenced by prior literature [7–9, 17, 18, 23, 50, 74]. On the other hand, singing voice transcription (SVT) is utilized to retrieve note events, which include onsets, offsets, and pitches [15, 24, 29, 33, 42, 67, 68]. Transcribed information can enable the development of singing voice synthesis [38, 55] and aid in education [26, 46, 72] and therapy [60]. It is important to note that males and females have distinct sound characteristics in their singing voices [35], e.g. timbre. Additionally, compared to lyrics, the note events tend to be more subject to explicit gender-related biases, particularly for pitch, as shown in prior literature [52, 58] that males tend to have lower average pitch than females. This is also consistent with our analysis of four SVT datasets, namely N20EMv2 [24], MIR-ST500 [67], ISMIR2014 [47], and M4Singer [75]. As shown in Fig. 1, we observe that the pitch range of females is generally higher than that of males across these four SVT datasets. Besides, the proportion of males and females for each pitch value is also different. Consequently, a critical **fairness** question arises: since SVT systems target pitches, will the performance of these systems favor one gender over the other? Before we delve into this question, it is imperative to elucidate why this question holds substantial significance.

The rapid progress in machine learning techniques has facilitated their successful integration into various downstream applications, streamlining decision-making processes and reducing the need for repeated human efforts. However, the presence of bias in machine

learning systems can lead to the discriminatory treatment of certain groups with sensitive attributes such as gender, age, and race, resulting in unfair decisions. Moreover, general usage of biased machine learning systems can reinforce the stereotypes and exclude certain groups from the opportunities. This phenomenon is not only limited to traditional decision-making scenarios such as loan applications, hiring, legal proceedings, and policy-making [73], but also frequently appears in recent deep learning applications, including visual recognition [56, 62, 69], natural language processing [11, 25], speech processing and recognition [5, 10, 13, 22, 37, 54, 66], recommendation systems [36, 59], and generative models [6, 14]. Consequently, the issue of fairness has gained prominence in the machine learning community due to its pervasive nature and potential societal consequences. Returning to our SVT task, the unfairness of systems could directly lead to user inconvenience and negatively impact their experience. Applying biased SVT systems to downstream applications could cause more fairness issues. We can consider an automatic sight-singing exercise [72], where an SVT system could serve as an intermediate stage to transcribe the singing voice into musical notes. However, if a gender-biased SVT system is utilized, it could result in an unsatisfying user experience for certain groups, as well as reinforce gender stereotypes, which would be unfair to individuals of different genders.

In this study, we elucidate the fairness issue in the field of SVT. Our investigation demonstrates that SVT systems tend to perform better on females compared to males, raising questions about their fairness. Then, we assume that this gender bias is attributable to the differences of singing voices, especially in terms of pitch distributions, in different demographic groups. We utilize self-supervised-learning (SSL)-based SVT systems [24], which represent the current state-of-the-art in this field, to develop a bias mitigation approach. In contrast to previous research, which normally focuses on the in-domain fairness, we also pay attention to the out-of-domain fairness as user data is sometimes sampled from out-of-domain distribution. To implement this, our approach adopts the adversarial learning framework to "unlearn" gender-related information in the acoustic representations. Considering the effects of pitch distributions, we propose a note-conditional attribute predictor to conditionally align the representations between female and male groups by conditioning on note events. Empirical results from various SVT datasets confirm the effectiveness of our approach in mitigating gender bias. Our contributions are summarized as below:

- We provide evidence and analysis of the prevalence and source of gender bias in SVT systems. To the best of our knowledge, this is the first attempt at fairness in singing-centric deep learning.
- We first introduce a note-conditioned adversarial learning approach to achieve fair representation learning in audio modality, resulting in a significant reduction in the performance gap between the two gender groups on various benchmark SVT datasets while maintaining a good fairness-utility trade-off. Our method is effective in reducing biases on both in-domain and out-of-domain data.
- We demonstrate the superiority of our note-conditioned adversarial learning approach through comparisons with baseline adversarial learning and domain-independent training.

2 RELATED WORK

2.1 Singing Voice Transcription

Singing voice transcription (SVT) involves various sub-tasks, including pitch estimation and onset/offset detection. Earlier approaches [42, 43, 49, 71] typically relied on statistical models, such as Bayesian models and Hidden Markov Models (HMM), to predict fundamental frequency (F0) and note segmentation. In contrast, more recent SVT methods [15, 29, 33] have predominantly employed deep learning techniques, such as CNN and LSTM, and have demonstrated superior performance. Despite the promising SVT performance achieved by these methods, the intrinsic difficulty of curating large-scale, high-quality SVT datasets presents an obstacle to further improvements. To address this challenge, several approaches have been proposed. Among them, VOCANO [29] employed the Virtual Adversarial Training (VAT) [45] to train the note segment network on both labeled and unlabeled data. In [33], pseudo labels are obtained by quantizing frame-level pitch contours for training on unlabeled audio data. MusicYOLO [68] utilized the object detection model YOLOX [20], which has been trained on the image domain, to locate notes in the audio spectrogram. Recently, [24] adapted self-supervised learning (SSL) models from the speech domain to the SVT task, thus alleviating the label insufficiency as well as achieving state-of-the-art SVT performance.

2.2 Fairness and Bias Mitigation

The notion of fairness in machine learning systems is defined as the absence of any discrimination based on sensitive attributes when making decisions. It can be categorized into group fairness [27, 76] and individual fairness [12, 34]. The former requires that there are no disparities among different demographic groups [76], while the latter requires that similar individuals receive similar predictions [34]. In our work, we focus on group fairness, which can be assessed by criteria, including independence, separation, and sufficiency [3, 57], along with metrics such as demographic parity, equalized odds, equal opportunity, and accuracy parity [44, 65, 76].

Numerous approaches have been proposed to mitigate bias in machine learning systems. Among them, adversarial learning has emerged as a powerful technique for removing sensitive attributes in the representations used for prediction. For instance, [40] repurposed the framework of generative adversarial networks (GANs) [21] to satisfy demographic parity. [41] used an adversary to predict sensitive attributes from latent representations and a decoder to reconstruct the input data from the latent representations and predicted attributes in adversarial learning. Additionally, they proposed different adversarial objectives according to the target group fairness criteria. [31] proposed using a bias prediction network to minimize mutual information between latent representations and bias through adversarial learning. In contrast to these approaches, [76] proposed conditional alignment of latent representations to strike a better balance between fairness and utility. In addition to adversarial learning, alternative methods have been proposed to mitigate bias. [69] advocated domain independent training, where different classifiers are trained for different demographic groups. [56, 62] disentangled the latent representations for task predictions and sensitive attributes, respectively, to reduce the influence of the sensitive attributes on the model's decision-making process.

2.3 Adversarial Learning for Invariance

In addition to the adversarial learning framework used in the fairness community, our work also shares similarities with unsupervised domain adaptation, which aims to learn domain-invariant representations. [16] proposed a domain classifier to learn the representations that are invariant to the domain shift, with an adversary implemented by a gradient reverse layer. [63] used a discriminator to align the distributions of source domain and target domain in the representation space, following the training of GANs [21]. Similar to [76], [39, 77] proposed a conditional discriminator that takes into account the multimodal nature of feature distributions. One key difference compared to the fairness scenario is that labels in the target domain are not available in domain adaptation. Therefore, the conditions in [39, 77] are the label predictions, rather than the ground-truth labels used in [76].

3 PRELIMINARY FOR SVT

We recap the formulation and solutions of singing voice transcription (SVT), which is defined according to the framework introduced in [24, 67]. The input to an SVT system is the waveform $\boldsymbol{x}$ while the output $\boldsymbol{y} = [(o_1, f_1, p_1), ..., (o_n, f_n, p_n), ..., (o_N, f_N, p_N)]$ is a sequence of note events, where o_n, f_n, p_n represent the onset/offset/pitch of each note, respectively. Consequently, the task of SVT can be regarded as a sequence-to-sequence problem. Since it is challenging to supervise the entire model using the ground truth note events directly, frame-level labels $\boldsymbol{O}, \boldsymbol{S}, \boldsymbol{V}, \boldsymbol{P}$ are constructed to mark the onset/silence/octave class/pitch class of each frame. Specifically, $\boldsymbol{O}, \boldsymbol{S}$ comprise binary classes, whereas $\boldsymbol{V}$ and $\boldsymbol{P}$ have multiple classes. The number of categories of pitch classes is fixed to 12, while the number of categories of octaves are chosen based on the pitch range. For example, the octave class of C4 is 4 and the pitch class is C. We add an additional octave/pitch class to represent the pitch of silence. The loss function is formulated to minimize the empirical risk as follows:

$$\mathcal{L}_{\text{SVT}}(\hat{\boldsymbol{y}}, \boldsymbol{y}) = \frac{1}{T}\sum_{t=1}^{T}[l_{\text{BCE}}(\hat{O}_t, O_t) + l_{\text{BCE}}(\hat{S}_t, S_t) + l_{\text{CE}}(\hat{V}_t, V_t) + l_{\text{CE}}(\hat{P}_t, P_t)], \tag{1}$$

where T refers to the number of frames, $\hat{O}_t, \hat{S}_t, \hat{V}_t, \hat{P}_t$ denote the frame-level predictions while $\hat{\boldsymbol{y}}$ represents note-level predictions. After post-processing, $\hat{O}_t, \hat{S}_t, \hat{V}_t, \hat{P}_t$ are transformed to $\hat{\boldsymbol{y}}$. Interested readers can refer to [24, 67] for more detailed information. An SVT system consists of an acoustic encoder and a note predictor. The note predictor is a simple linear layer, while the design of acoustic encoder can vary. For instance, [67] employed EfficientNet [61], while [24] adapted self-supervised-learning models, e.g. wav2vec 2.0 as the acoustic encoder. To evaluate SVT systems, f1-scores of COnPOff (correct onset, pitch, and offset), COnP (correct onset and pitch), and COn (correct onset) are commonly used. These metrics were proposed in [47] and have since been widely adopted. The first two metrics are related to the accuracy of pitch estimation. We use the default tolerances as implemented in the python package *mir_eval* [53] for the evaluation of SVT systems in this work.

4 FAIRNESS ANALYSIS FOR SVT

4.1 Female Superiority in SVT Performance: Evidence from Multiple Datasets

We evaluate the performance of state-of-the-art singing voice transcription (SVT) systems from [24] on three benchmark datasets, including MIR-ST500 [67], N20EMv2 [24], and ISMIR2014 [47]. The results for two gender groups are presented in Table 1, where "model1" is trained on the MIR-ST500 training split, "model2" on the N20EMv2 training split, and "model3" on both training sets. These models were built based on wav2vec 2.0 [2]. The ISMIR2014 dataset already has the gender label for each song. The gender labels for the N20EMv2 and MIR-ST500 can be obtained by directly listening to the audio recordings. For double confirmation, we also check the video modality provided in N20EMv2 and the original Youtube links in MIR-ST500. Across all datasets, we observe that the SVT performances of three models on females consistently outperform that on males in terms of COnPOff and COnP f1-scores, which are the metrics related to pitch estimation. However, the COn performance does not demonstrate the consistent female superiority. We compute the performance gap as the difference between male and female metrics, i.e., $\text{metric}_{\text{gap}} = \text{metric}_{\text{male}} - \text{metric}_{\text{female}}$. We find that the performance gaps between gender groups are significant for pitch-related metrics on the N20EMv2 and ISMIR2014 datasets. Even though the performance gap is comparably small for MIR-ST500, female superiority in SVT performance remains valid.

Apart from the datasets evaluated in [24], we conduct experiments on a recent Mandarin singing dataset called M4Singer [75]. This dataset comprises data from 20 singers across four main voice types (soprano, alto, tenor, and bass). Since M4Singer lacks official training-test splits, we manually partition the data, selecting data from two male and two female singers representing the above four voice types for the test split and the data from the remaining singers for the training split. To ensure a fair test split, we make sure that the total duration of selected female data in the test split is almost equal to that of male data. We follow the training configuration in [24] to train our SVT system from scratch using wav2vec 2.0, making minimal modifications to achieve high SVT performance. We add an additional octave category to the classifier to account for the larger pitch range of M4Singer. The M4Singer dataset already provides the gender labels, similar to ISMIR2014. The results in Table 1 show that the SVT performance of females is still better than that of males, with a significant performance gap.

4.2 Female Superiority in SVT Performance: Evidence from Multiple Models

The models evaluated in Sec. 4.1 are all based on wav2vec 2.0 [2]. However, it is important to investigate whether the observed female superiority is limited to this particular model choice. To explore this question, we replace the wav2vec 2.0 trained on M4Singer with other self-supervised-learning (SSL) models, e.g. Hubert [30], wavLM [4], and data2vec [1]. Although these models have different SSL objectives and slightly different model architectures, we find that the SVT performance of the female group is still better than the male group with significant margins, as shown in Table 2. To strengthen our findings, we further evaluate the performance of

Table 1: COnPOff/COnP/COn F1-score (%) of state-of-the-art SVT systems in [24] on multiple datasets. We using bold face to highlight the gender group with better performance and red bold face to mark the results with large bias.

Dataset	Model	COnPOff (%)				COnP (%)				COn (%)			
		total ↑	female ↑	male ↑	gap	total ↑	female ↑	male ↑	gap	total ↑	female ↑	male ↑	gap
MIR-ST500	model1	52.39	**53.11**	51.47	-1.64	70.73	**72.54**	68.42	-4.12	78.32	**79.36**	77.00	-2.36
	model2	34.55	**35.29**	33.60	-2.55	51.64	**52.76**	50.21	-2.55	71.33	**72.07**	70.39	-1.68
	model3	52.84	**54.02**	51.33	-2.69	70.00	**71.17**	67.85	-3.32	78.05	**78.84**	77.05	-1.79
N20EMv2	model1	55.20	**60.96**	51.55	**-9.41**	72.03	**79.76**	67.11	**-12.65**	88.51	**90.63**	87.16	-3.47
	model2	68.62	**74.82**	64.68	**-10.14**	75.69	**81.27**	72.14	**-9.12**	92.83	**94.33**	91.88	-2.44
	model3	73.06	**78.34**	69.69	**-8.65**	79.56	**84.38**	76.49	**-7.89**	93.66	**94.96**	92.83	-2.13
ISMIR2014	model1	52.58	**61.22**	45.28	**-15.94**	67.75	**74.90**	61.70	**-13.20**	92.13	91.93	**92.30**	+0.37
	model2	57.35	**62.42**	53.06	**-9.36**	72.15	**79.35**	66.06	**-13.29**	91.53	**92.25**	90.92	-1.33
	model3	59.95	**65.61**	55.16	**-10.45**	73.85	**80.55**	68.19	**-12.36**	92.80	**93.18**	92.49	-0.69
M4Singer	wav2vec2	53.66	**57.27**	49.93	**-7.34**	61.95	**66.36**	57.38	**-8.98**	82.60	81.66	**83.58**	+1.91
	Hubert	55.11	**58.39**	51.73	**-6.66**	64.17	**68.10**	60.10	**-8.00**	82.13	81.53	**82.76**	+1.23
	wavLM	57.06	**60.33**	53.68	**-6.66**	65.40	**68.87**	61.82	**-7.05**	82.73	82.29	**83.19**	+0.90
	data2vec	53.98	**57.45**	50.40	**-7.05**	63.04	**67.09**	58.86	**-8.23**	82.30	82.13	**82.48**	+0.35

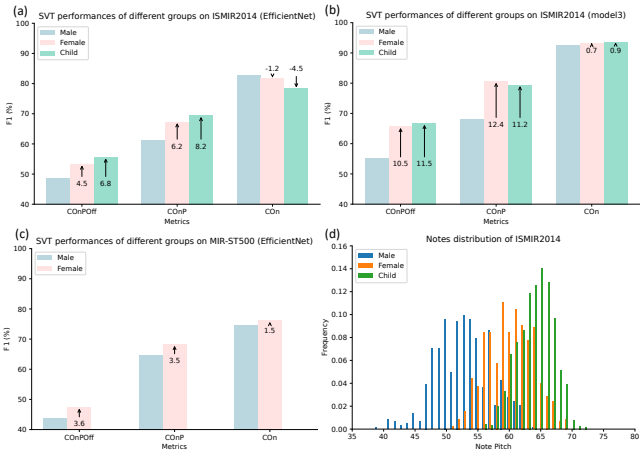

Figure 2: (a) SVT performance of EfficientNet on ISMIR2014. (b) SVT performance of model3 on ISMIR2014. (c) SVT performance of EfficientNet on MIR-ST500. (d) Pitch distributions of male/female/child groups on ISMIR2014.

EfficientNet-based SVT system in [67]. As presented in Fig. 2 (a) and (c), we note that on the MIR-ST500 and ISMIR2014 datasets, the SVT performance of female group still outperforms that of the male group in terms of COnPOff and COnP f1-scores, which is consistent with our earlier conclusion.

4.3 Possible Reasons for Female Superiority

From Sec. 4.1 and Sec. 4.2, we conclude that female superiority in SVT performance is valid across different datasets and model choices. To explore possible reasons behind this phenomenon, we first examine the statistics of the SVT datasets to investigate whether the singing datasets typically possess the property of gender data imbalance. As presented in Table 2, we include the statistics of two gender groups in the training splits of MIR-ST500, N20EMv2, and M4Singer (ISMIR2014 is only used for evaluation).

Table 2: Demographic statistics of SVT training splits.

Dataset	Songs Num			Duration (h)		
	total	female	male	total	female	male
MIR-ST500	400	**221**	179	27.62	**15.13**	12.49
N20EMv2	123	52	**71**	6.44	2.74	**3.70**
M4Singer	521	246	**275**	22.71	10.27	**12.44**

While MIR-ST500 has a larger proportion of female data, N20EMv2 and M4Singer have larger proportions of male data. Nevertheless, all the SVT systems trained on these datasets favored females, indicating that the source of bias is not merely the data imbalance. We hypothesize that the gender bias in SVT performance is attributed to the differences of sound characteristics across different demographic groups. Specifically, we assume that pitch distributions make substantial contributions to the female superiority. As shown in Fig. 1, we find that (1) the female group generally has higher pitch range than the male group; (2) the proportion of male and female labels for each specific pitch is different.

To further support our assumption, we perform an additional evaluation on the ISMIR2014 dataset, which includes the child group. Typically, children have different inherent properties in sound voices compared to female adults and male adults [48, 70]. For instance, the pitch distribution of the child group is different from both the female group and male group, as present in Fig. 2 (d). Consequently, we find that both the EfficientNet-based SVT system and wav2vec 2.0-based SVT system (model3) demonstrate performance disparity among the three groups. As shown in Fig. 2 (a) and (b), the SVT performance of the child group significantly outperforms that of the male group while is close to that of the female group. To interpret this, the pitch distribution difference between the male group and the child group is large while the difference between the female group and the child group is comparatively subtle. As there are no existing SVT annotations for child training data, we narrow down the scope of this work to gender fairness and leave the discussion on age fairness to future work. Similarly, as present

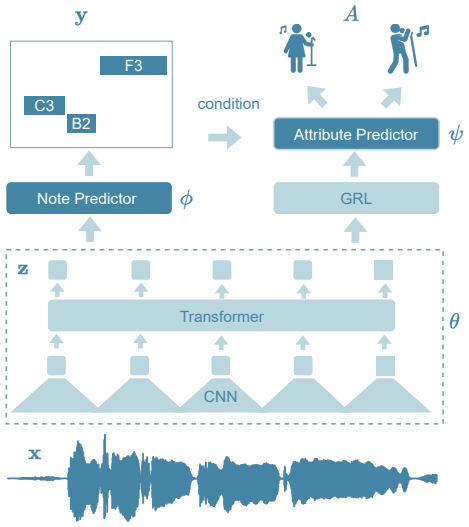

Figure 3: Framework of note-conditioned adversarial learning for singing voice transcription.

in Table 1, the performance gap on the MIR-ST500 dataset is not as large as the other three datasets. We assume the reason is that the singing recordings in MIR-ST500 were performed by professional singers, resulting in smaller pitch distribution difference between two gender groups.

5 BIAS MITIGATION FOR SVT

5.1 Problem Formulation and Basic Framework

Suppose the training samples are drawn from the domain $\mathcal{D}_S = \{(\boldsymbol{x}^{(n)}, \boldsymbol{y}^{(n)}, A^{(n)})\}_{n=1}^{N_S}$, while the test data are sampled from the domain $\mathcal{D}_T = \{(\boldsymbol{x}^{(n)}, \boldsymbol{y}^{(n)}, A^{(n)})\}_{n=1}^{N_T}$. Here $\boldsymbol{x}^{(n)}$ represents the raw waveform of the singing audio, $\boldsymbol{y}^{(n)}$ represents the ground-truth note events, and $A^{(n)}$ is the sensitive attribute. In our study, $A = 0$ represents female singers, while $A = 1$ represents male singers. As explained in Sec. 3, the basic SVT system consists of two primary components: an acoustic encoder θ and a note predictor ϕ. We select self-supervised-learning (SSL) models as our acoustic encoder due to their state-the-of-art performance in SVT tasks [24].

Fig. 3 illustrates our proposed bias mitigation framework for the SVT task. The input singing waveform $\boldsymbol{x}$ is first processed by the acoustic encoder, which consists of a CNN and a transformer [64]. We refer readers to original papers of self-supervised-learning (SSL) models, e.g. wav2vec 2.0 [2], Hubert [30], wavLM [4], and data2vec [1] for details on these models, since they are not the focus of this paper. These models are first pre-trained on unlabeled speech data and then adapted to our singing data using a linear probing and then full fine-tuning approach [24]. After the acoustic encoder θ, we obtain the acoustic features $\boldsymbol{z} \in \mathcal{R}^{T \times D}$, where T is the number of frames and D is the number of feature dimensions. The note predictor ϕ is parameterized by a linear layer, and accepts the acoustic features $\boldsymbol{z}$ to predict the frame-level labels: $\hat{O}_t, \hat{S}_t, \hat{V}_t, \hat{P}_t = \phi(\boldsymbol{z}_t)$. The loss function for SVT task in domain $\mathcal{D}_S$ is:

$$\mathcal{L}_y = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{D}_S}\left[\mathcal{L}_{\text{SVT}}(\hat{\boldsymbol{y}}, \boldsymbol{y})\right], \quad \hat{\boldsymbol{y}} = \phi(\boldsymbol{z}) = \phi \circ \theta(\boldsymbol{x}), \tag{2}$$

where $\mathcal{L}_{\text{SVT}}$ is defined in Eq. 1. To evaluate the SVT performance, we compute the f1-scores of the COnPOff, COnP, and COn in domain $\mathcal{D}_T$. We refer to these metrics as utility metrics:

$$U = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{D}_T}\left[\text{metric}(\hat{\boldsymbol{y}}, \boldsymbol{y})\right], \tag{3}$$

where metric can be the f1-score of COnPOff or COnP. As our objective is to mitigate gender bias in SVT systems, we propose the following fairness metrics, which focus on the performance disparity between two demographic groups:

$$F = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y},A=1)\sim\mathcal{D}_T}\left[\text{metric}(\hat{\boldsymbol{y}}, \boldsymbol{y})\right] - \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y},A=0)\sim\mathcal{D}_T}\left[\text{metric}(\hat{\boldsymbol{y}}, \boldsymbol{y})\right], \tag{4}$$

This definition is similar to the accuracy parity in [76]: $P(\hat{\boldsymbol{y}} \neq \boldsymbol{y}|A = 0) = P(\hat{\boldsymbol{y}} \neq \boldsymbol{y}|A = 1)$ only except that we replace accuracy with f1-score as the evaluation metric for parity. To achieve bias mitigation, we formulate the problem as a bundle of optimization objectives:

$$\begin{cases} \max_{\theta,\phi} \min\{F(\theta, \phi), 0\} \\ \max_{\theta,\phi} U(\theta, \phi) \end{cases} \tag{5}$$

5.2 Adversarial Learning for Fairness

The optimization objectives in Eq. 5 cannot be directly optimized due to the unavailability of the test data $\mathcal{D}_T$ during the training of θ, ϕ. To mitigate the gender bias, we propose an adversarial learning framework to learn fair acoustic representations by assuming that the acoustic encoder cannot discriminate between two gender groups for SVT task. To achieve this, we employ an attribute predictor ψ that predicts the labels of sensitive attributes, such as gender. The goal is to eliminate the gender information in the acoustic features $\boldsymbol{z}$ while preserving the information necessary for predicting the note events. The binary cross-entropy between gender predictions and ground truth gender labels serves as the learning objective of the attribute predictor:

$$\mathcal{L}_A = \mathbb{E}_{(\boldsymbol{x},s)\sim\mathcal{D}_S}\left[\frac{1}{T}\sum_{t=1}^{T} l_{\text{BCE}}(\hat{A}_t, A)\right], \quad \hat{A} = \psi(\boldsymbol{z}) = \psi \circ \theta(\boldsymbol{x}). \tag{6}$$

To perform frame-level gender classification independently, each frame of acoustic features, denoted as $\boldsymbol{z}_t$, is annotated with the same attribute label A. We then average the frame-level classification loss to obtain the song-level or utterance-level loss. We assume that the temporal model structure in SSL models has sufficiently learned the gender representations and thus do not require a temporal attribute predictor. Our preliminary experiments show no empirical gains by including a temporal attribute predictor.

From a distribution alignment perspective, our objective is to achieve gender-invariant acoustic features, which requires that the distribution $P(\boldsymbol{z}|A = 0)$ and the distribution $P(\boldsymbol{z}|A = 1)$ be similar. The loss function in Eq. 6 serves as a proxy measure for the distance between the two distributions, and thus, the attribute predictor ψ must be trained to be powerful enough to distinguish between $P(\boldsymbol{z}|A = 0)$ and $P(\boldsymbol{z}|A = 1)$. Additionally, the acoustic encoder θ must be trained to deceive ψ. Therefore, the loss function for the acoustic encoder is formulated as $\mathcal{L}_y - \lambda\mathcal{L}_A$, where λ is a hyper-parameter that balances the two loss terms. Theoretically, this can be implemented by a gradient reverse layer (GRL) [16]. From the perspective of fairness criteria, the adversarial learning approach enforces that $\boldsymbol{z} \perp A$. Since $\hat{\boldsymbol{y}} = \phi(\boldsymbol{z})$, $\hat{\boldsymbol{y}} \perp A$ can be further enforced, which is the independence criteria [3].

5.3 Note-conditioned Adversarial Learning

We have observed that there is a difference in pitch distributions between males and females, which may contribute to the performance disparity. Motivated by this, we propose the conditional distribution alignment, which enforces that $P(\boldsymbol{z}|\boldsymbol{y}, A = 0) = P(\boldsymbol{z}|\boldsymbol{y}, A = 1)$. By incorporating note labels as an extra input to the attribute predictor ψ, we can eliminate the conditional dependencies between the acoustic features and the gender labels. Since the note labels are available for both demographic groups, we propose two variants for the design of condition. In variant 1, the ground-truth notes $\boldsymbol{y}$ are transformed into frame-level labels $\boldsymbol{O}, \boldsymbol{S}, \boldsymbol{V}, \boldsymbol{P}$, which are then fed into the attribute predictor. For variant 2, the logits of frame-level predictions $\hat{O}_t, \hat{S}_t, \hat{V}_t, \hat{P}_t$ are used instead.

From the perspective of fairness criteria, the variant 1 enforces $\boldsymbol{z} \perp A|\boldsymbol{y}$ while the variant 2 enforces $\boldsymbol{z} \perp A|\hat{\boldsymbol{y}}$. Therefore, the former can enforce $\hat{\boldsymbol{y}} \perp A|\boldsymbol{y}$, which is the separation criteria [3]. The formulation of variant 2 is motivated by the perspective of distribution alignment. The prediction $\hat{\boldsymbol{y}}$ contains prior knowledge about the classifier ϕ and similarities among different pitches (the predicted probabilities of other pitch values besides the true pitch value are not zeros, similar to dark knowledge in knowledge distillation [28]), while the ground truth $\boldsymbol{y}$ is one hot and has no information about other pitches. Considering that the pitch distributions of two gender groups are different, the proportion of male and female labels for each specific pitch value is also different. When conducting the conditional alignment for a specific pitch, the similarities contained in the pitch logits can assist in aligning other pitches, resulting in an improved and more efficient conditional alignment by the attribute predictor.

The framework of note-conditioned adversarial learning for SVT is depicted in Fig. 3, and its variant 2 is presented in Alg. 1. The algorithm of variant 1 can be similarly derived. To train the SVT system, we follow the linear-probing and full-finetuning strategy proposed in [24]. During the linear probing stage, only the label predictor ϕ and attribute predictor ψ are updated. This stage serves as a warm-up for these two models and preserves the pre-trained features of SSL models. Then in the full-finetuning stage, all three models θ, ϕ, ψ are updated using different learning rates. To parameterize the note-conditioned attribute predictor, we first use two linear layers to embed each frame of acoustic features $\boldsymbol{z}$ and notes $\boldsymbol{y}$ (or $\hat{\boldsymbol{y}}$), respectively. Then the two embeddings are concatenated and passed through two linear layers with ReLU activations. We note that by modifying the loss function in Eq. 6, our bias mitigation framework can be readily applied to other types of sensitive attributes, such as A is a multi-class attribute, or a continuous attribute, or a vector that encompasses multiple attributes.

5.4 Fairness-Utility Trade-off

Our task involves an inherent trade-off between fairness F and utility U, as noted in previous literature [32]. This is due to the fact that when the gender information is removed from the acoustic features, the representations used for the SVT task may be affected. Optimizing $\mathcal{L}_y - \lambda\mathcal{L}_A$ of the acoustic encoder θ poses a challenge as the the two loss terms have conflicting natures. Empirical experiments demonstrate that in some cases, improvements in the fairness metric F come at a cost of reduced utility metric U. Our goal is to improve the fairness metric U without significantly degrading the utility metric U, starting from the initial point (F_0, U_0) without mitigating bias. To achieve this, we introduce a tolerance hyper-parameter δ for U. We aim to increase the value of F as much as possible within the range of $U > U_0 - \delta$. Meanwhile, sacrificing utility beyond this range is not acceptable for real-world applications. This trade-off criterion facilitates the model selection. Typically, we set δ as 2% or 5% for f1-scores of COnPOff and COnP.

Algorithm 1 Note-conditioned adversarial learning for SVT

Require: Acoustic encoder $\theta^{(0)}$ pre-trained on speech data under SSL objective, randomly initialized label predictor $\phi^{(0)}$ and attribute predictor $\psi^{(0)}$, learning rates η_1, η_2, η_3 for θ, ϕ, ψ, training steps K_1, K_2 for linear probing and full finetuning.

for $k = 1$ **to** $K_1 + K_2$ **do**
  $\boldsymbol{z} = \theta^{(k-1)}(\boldsymbol{x})$, $\hat{\boldsymbol{y}} = \phi^{(k-1)}(\boldsymbol{z})$
  $\hat{A}_t = \psi^{(k-1)}(\boldsymbol{z}.\text{detach}(), \hat{\boldsymbol{y}}.\text{detach}())$, compute $\mathcal{L}_A$ in Eq. 6
  $\psi^{(k)} = \psi^{(k-1)} - \eta_3 \frac{\partial \mathcal{L}_A}{\partial \psi^{(k-1)}}$ ▷ Update ψ
  $\hat{A}_t = \psi^{(k)}(\boldsymbol{z}, \hat{\boldsymbol{y}})$, compute $\mathcal{L}_y, \mathcal{L}_A$ in Eq. 2 and 6
  $\phi^{(k)} = \phi^{(k-1)} - \eta_2 \frac{\partial \mathcal{L}_y}{\partial \phi^{(k-1)}}$ ▷ Update ϕ
  if $k \leq K_1$ **then** ▷ Update θ
    $\theta^{(k)} = \theta^{(k-1)}$
  else
    $\theta^{(k)} = \theta^{(k-1)} - \eta_1(\frac{\partial \mathcal{L}_y}{\partial \theta^{(k-1)}} - \lambda \frac{\partial \mathcal{L}_A}{\partial \theta^{(k-1)}})$
  end if
end for

6 EMPIRICAL EXPERIMENTS

6.1 Bias Mitigation Performance

To evaluate in-domain fairness, we adopt the proposed note-conditioned adversarial learning method on the M4Singer dataset. The SVT system is based on wav2vec 2.0 [2] and trained on the training split of M4Singer and evaluated on its test split. We set the learning rates for acoustic encoder and label predictor to be fixed at $\eta_1 = \eta_2 = 3 \times 10^{-4}$. To further evaluate out-of-domain fairness, we conduct experiments on model1 and model3, as displayed in Table 1. We set the learning rates to fixed values of $\eta_1 = 5 \times 10^{-5}$, $\eta_2 = 3 \times 10^{-4}$ following [24]. During the bias mitigation, we select the learning rate for the attribute predictor η_3 from the set $\{0.1, 0.01, 0.001, 0.0001\}$ and the balancing term λ from the set $\{0.2, 0.5, 1.0, 2.0\}$. We report the best results we can achieve. This hyper-parameter selection also applies to the baselines we compare with in the following Sec. 6.2. We find that further increasing the learning rate η_3 or the balancing term λ results in severe degradation in utility, even though the gender bias seems to be eliminated, as elaborated in Sec. 6.3. Additionally, when η_1 is large, we find that the best results are achieved when η_3 is also large. Given our framework is based on adversarial learning, aiming to achieve equilibrium between the acoustic encoder and the attribute predictor, a larger learning rate for the encoder necessitates a corresponding increase in the learning rate of the attribute predictor to attain equilibrium[1].

[1]We conducted our experiments using the open-sourced repo: https://github.com/guxm2021/SVT_SpeechBrain.

Table 3: Bias mitigation performance of note-conditioned adversarial learning.

Train set	Test set	Method	COnPOff (%)		COnP (%)	
			Utility (U) ↑	Fairness (F) ↑	Utility (U) ↑	Fairness (F) ↑
M4Singer	M4singer	ERM	53.66	- 7.34	61.95	- 8.98
		Ours	52.48 (-1.18)	- 3.61 (+3.73)	60.67 (-1.28)	- 4.21 (+4.77)
MIR-ST500	N20EMv2	ERM	55.20	- 9.41	72.03	-12.65
		Ours	53.29 (-1.91)	- 4.14 (+5.27)	72.71 (+0.68)	-10.82 (+1.83)
	ISMIR2014	ERM	52.58	-15.94	67.75	-13.20
		Ours	48.18 (-4.40)	- 8.29 (+7.65)	65.40 (-2.35)	- 9.08 (+4.12)
MIR-ST500 N20EMv2	N20EMv2	ERM	73.06	- 8.65	79.56	- 7.89
		Ours	72.43 (-0.63)	- 5.78 (+2.87)	78.47 (-1.09)	- 6.82 (+1.07)
	ISMIR2014	ERM	59.95	-10.45	73.85	-12.36
		Ours	59.57 (-0.38)	- 7.49 (+2.96)	73.33 (-0.52)	- 7.75 (+4.61)

Table 3 presents a comparison between the SVT systems trained with our note-conditioned adversarial learning (variant 2) and those trained using empirical risk minimization (ERM) in Eq. 2 without bias mitigation. Our experiments on the M4Singer dataset reveal that the fairness metrics improve by over 50% for COnPOff and COnP, respectively. At the same time, the utility metrics drop only by 1.18% and 1.28% for COnPOff and COnP, respectively. Apart from the in-domain fairness results, we evaluate the out-of-domain fairness results on model1 and model3 (mentioned in Sec. 4.1). Our experiments demonstrate that applying the bias mitigation method on model1 significantly improves its fairness. In particular, the performance disparity decreases by 50% on N20EMv2 and ISMIR2014, in terms of COnPOff. These results validate the effectiveness of our bias mitigation method on out-of-domain data. Furthermore, we achieve fairer SVT performance on both in-domain and out-of-domain scenarios with minor total performance degradation when using the state-of-the-art performing SVT system (model3).

6.2 Comparisons with Baselines

We compare our note-conditioned adversarial learning framework with two baseline methods: adversarial learning (AL), and domain independent training proposed in [69]. We denote our method as "NCAL (variant 1)" and "NCAL (variant 2)". For AL, we keep the same configuration as our NCAL, except that we do not feed any condition into the attribute predictor. For domain independent training, we compare our method with two variants: calibrated inference and miscalibrated inference, which differ in whether the gender labels are used during the inference. These two inferences are abbreviated as " DInD (w/ calibr.)" and "DInD (w/o calibr.)", respectively. We refer readers to [69] for more technical details. We evaluate the comparisons on model3 and report the results on the N20EMv2 and ISMIR2014 datasets.

We present the Fairness-Utility trade-off of various bias mitigation methods in Fig. 4. According to the optimization bundle in Eq. 5, an upper-right point signifies a better trade-off compared to a lower-left point. Firstly, we observe that our two variants of NACL, "NCAL (variant 1)" and "NCAL (variant 2)", perform the best in the most cases in terms of fairness-utility trade-off. These two variants perform similarly on N20EMv2. On the ISMIR2014 dataset, NACL variant 2 performs better for COnPOff while variant

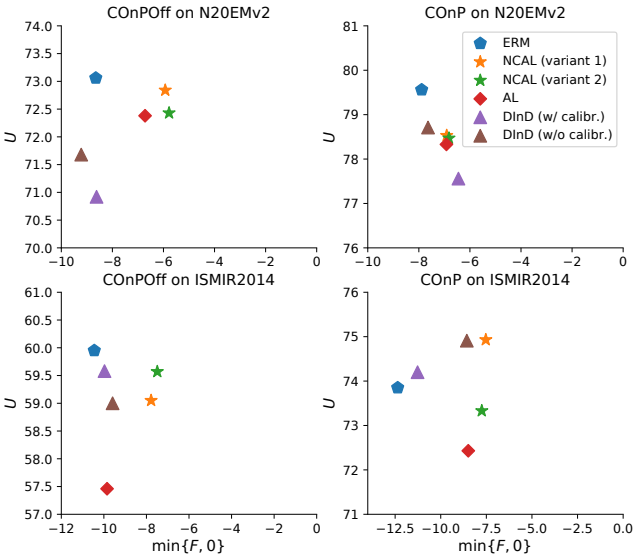

Figure 4: Comparisons among different bias mitigation methods of Fairness-Utility trade-off on N20EMv2 and ISMIR2014.

1 exhibits superiority for COnP. We then note that NCAL consistently outperforms the baseline AL. The performance of AL is similar to NACL only in terms of COnP on N20EMv2. However, in other cases, AL shows lower utility and less fairness. With respect to DInD, it shows better fairness than NCAL only in terms of COnP on N20EMv2 with calibration. However, its utility is more severely degraded than NCAL. In other cases, NACL consistently outperforms DInD. Although adversarial learning approaches are prone to instability during training compared to non-adversrial learning approaches, such as DInD, our proposed NCAL offers a better fairness-utility trade-off. Moreover, DInD cannot be easily applied to the scenarios where continuous sensitive attributes or multiple sensitive attributes are considered.

6.3 Further Empirical Analysis

In our experiments, we observe that setting a larger learning rate η_3 or the balancing term λ leads to near-perfect gender fairness but a drastic degradation in utility. As presented in Table 4, on M4Singer,

Table 4: More bias mitigation results.

Train set	Test set	Method	η_3	λ	COnPOff (%)		COnP (%)	
					Utility (U) ↑	Fairness (F) ↑	Utility (U) ↑	Fairness (F) ↑
M4Singer	M4singer	ERM	-	-	53.66	- 7.34	61.95	- 8.98
		AL	1.0	3.0	42.56 (-11.10)	+ 4.65	49.60 (-12.35)	+ 6.35
		NCAL	0.1	5.0	49.97 (- 3.69)	+ 0.39	57.66 (- 4.29)	+ 1.57
		NCAL	0.1	3.0	49.87 (- 3.79)	+ 1.29	58.53 (- 3.42)	+ 0.60
MIR-ST500	N20EMv2	ERM	-	-	55.20	- 9.41	72.03	-12.65
		NCAL	0.0001	5.0	45.78 (- 9.42)	+ 2.45	69.54 (- 2.49)	- 3.77
	ISMIR2014	ERM	-	-	52.58	-15.94	67.75	-13.20
		NCAL	0.0001	5.0	42.24 (-10.34)	+ 3.01	62.44 (- 5.31)	+ 1.22
MIR-ST500 N20EMv2	N20EMv2	ERM	-	-	73.06	- 8.65	79.56	- 7.89
		NACL	0.001	5.0	72.28 (- 0.78)	- 7.78	78.70 (- 0.86)	- 8.23
	ISMIR2014	ERM	-	-	59.95	-10.45	73.85	-12.36
		NACL	0.001	5.0	50.95 (- 9.00)	+ 2.49	65.87 (- 7.98)	+ 0.64

our NCAL method can achieve almost gender performance equality when λ is set to 3.0 or 5.0, but at the cost of more utility degradation than the results in Table 3. However, the baseline AL method fails to achieve such equality. We only observe a better performance for the male group with much more utility deterioration compared to NCAL when $\eta_3 = 1.0$ and $\lambda = 3.0$. This trend is consistently observed on model1 and model3 using NCAL for bias mitigation. We hypothesize that increasing either η_3 or λ enhances the discrimination ability of the attribute predictor, causing the models to focus more on the fairness while neglecting the utility. As explained in Sec. 5.4, such models may not be suitable for real-world applications where both fairness and utility are important.

The results in Table 3 and 4 can also validate our assumptions behind the performance disparity in SVT. Our baseline AL aims to ensure that the learnt acoustic representations contains as little gender information as possible. In this way, the effects of sound characteristics, which are related to gender, can be implicitly mitigated. The improvements in terms of fairness metrics brought by baseline AL provide evidence about our assumption that gender bias in SVT performance is attributed to the differences of sound characterises across different demographic groups. Additionally, when conditioning on the pitch information, our SVT systems could achieve nearly perfect fairness, as presented in Table 4, which further demonstrates the substantial contributions of pitch distribution difference to the performance disparity.

7 DISCUSSION AND FUTURE WORK

In addition to the primary focus on group fairness in this work, max-min fairness [51] is also raised in certain cases. Max-min fairness aims to minimize the worse-case error rates, offering an alternative perspective on fairness evaluation. Our motivation stems from the objective of enhancing user experience and mitigating potential discriminatory treatment when utilizing SVT systems and their downstream applications. Consequently, our main target is to strive for performance equalization across different demographic groups. Therefore, there is less discussion on max-min fairness. Despite this, we still observed improved performance for male data in most cases after applying our bias mitigation approach.

In this work, we formulate our solution from the perspective of fairness. We think the perspective of signal processing may also be beneficial to further interpret our findings. By incorporating signal processing approaches, our adversarial learning framework may further enhance SVT performance in terms of both fairness and utility. We identify this as an avenue for future exploration. Furthermore, we think our approach could be extended to consider other sensitive attributes, such as age, race, language. Beyond demographic groups, we also recognize that different instruments generally exhibit distinct pitch distributions and timbre characteristics. Hence, our approach holds potential applicability in the domain of automatic music transcription [19], wherein musical notes are inferred from audio signals produced by diverse instruments.

8 CONCLUSION

This work represents the first attempt of fairness topic within the singing-centric deep learning community. We presented evidence that the performance of singing voice transcription (SVT) on female data surpasses that of male data, irrespective of the models or datasets employed. Our findings suggested that this performance disparity is attributed to the inherent differences between male and female singing voices, especially in pitch distribution. Given the significance of this fairness issue, we proposed a note-conditioned adversarial learning approach to mitigate gender bias in SVT. Specifically, our approach leveraged an attribute predictor to learn gender-invariant acoustic representations. By conditioning on the note events, we further achieved conditional alignment between acoustic features of different groups. Our results demonstrated the effectiveness of our bias mitigation method, as it significantly improves fairness metrics while maintaining utility metrics across both in-domain and out-of-domain data.

ACKNOWLEDGMENTS

We would like to thank anonymous reviewers for their valuable suggestions. We also appreciated Nicholas Wong's writing advice and Xudong Shen's comments on fairness. This project is funded in part by a research grant MOESOL-2021-0017 from the Ministry of Education in Singapore.

REFERENCES

[1] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. 2022. Data2vec: A general framework for self-supervised learning in speech, vision and language. *arXiv preprint arXiv:2202.03555* (2022).

[2] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems* 33 (2020), 12449–12460.

[3] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org. http://www.fairmlbook.org.

[4] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing* 16, 6 (2022), 1505–1518.

[5] Xingyu Chen, Zhengxiong Li, Srirangaraj Setlur, and Wenyao Xu. 2022. Exploring racial and gender disparities in voice biometrics. *Scientific Reports* 12, 1 (2022), 3723.

[6] Kristy Choi, Aditya Grover, Trisha Singh, Rui Shu, and Stefano Ermon. 2020. Fair generative modeling via weak supervision. In *International Conference on Machine Learning*. PMLR, 1887–1898.

[7] Gerardo Roa Dabike and Jon Barker. 2019. Automatic Lyric Transcription from Karaoke Vocal Tracks: Resources and a Baseline System.. In *Interspeech*. 579–583.

[8] Emir Demirel, Sven Ahlbäck, and Simon Dixon. 2020. Automatic lyrics transcription using dilated convolutional neural networks with self-attention. In *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[9] Emir Demirel, Sven Ahlbäck, and Simon Dixon. 2021. MSTRE-Net: Multistreaming acoustic modeling for automatic lyrics transcription. *arXiv preprint arXiv:2108.02625* (2021).

[10] Pranav Dheram, Murugesan Ramakrishnan, Anirudh Raju, I-Fan Chen, Brian King, Katherine Powell, Melissa Saboowala, Karan Shetty, and Andreas Stolcke. 2022. Toward Fairness in Speech Recognition: Discovery and mitigation of performance disparities. *arXiv preprint arXiv:2207.11345* (2022).

[11] Lei Ding, Dengdeng Yu, Jinhan Xie, Wenxing Guo, Shenggang Hu, Meichen Liu, Linglong Kong, Hongsheng Dai, Yanchun Bao, and Bei Jiang. 2022. Word embeddings via causal inference: Gender bias reducing and semantic information preserving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 11864–11872.

[12] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.

[13] Gianni Fenu, Hicham Lafhouli, and Mirko Marras. 2020. Exploring algorithmic fairness in deep speaker verification. In *Computational Science and Its Applications–ICCSA 2020: 20th International Conference, Cagliari, Italy, July 1–4, 2020, Proceedings, Part IV 20*. Springer, 77–93.

[14] Felix Friedrich, Patrick Schramowski, Manuel Brack, Lukas Struppek, Dominik Hintersdorf, Sasha Luccioni, and Kristian Kersting. 2023. Fair Diffusion: Instructing Text-to-Image Generation Models on Fairness. *arXiv preprint arXiv:2302.10893* (2023).

[15] Zih-Sing Fu and Li Su. 2019. Hierarchical classification networks for singing voice segmentation and transcription. In *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR 2019)*. 900–907.

[16] Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*. PMLR, 1180–1189.

[17] Xiaoxue Gao, Chitralekha Gupta, and Haizhou Li. 2022. Automatic lyrics transcription of polyphonic music with lyrics-chord multi-task learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30 (2022), 2280–2294.

[18] Xiaoxue Gao, Chitralekha Gupta, and Haizhou Li. 2022. Genre-Conditioned Acoustic Models for Automatic Lyrics Transcription of Polyphonic Music. In *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 791–795.

[19] Joshua P Gardner, Ian Simon, Ethan Manilow, Curtis Hawthorne, and Jesse Engel. 2022. MT3: Multi-Task Multitrack Music Transcription. In *International Conference on Learning Representations*.

[20] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. 2021. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430* (2021).

[21] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Commun. ACM* 63, 11 (2020), 139–144.

[22] Cristina Gorrostieta, Reza Lotfian, Kye Taylor, Richard Brutti, and John Kane. 2019. Gender De-Biasing in Speech Emotion Recognition.. In *Interspeech*. 2823–2827.

[23] Xiangming Gu, Longshen Ou, Danielle Ong, and Ye Wang. 2022. Mm-alt: A multimodal automatic lyric transcription system. In *Proceedings of the 30th ACM International Conference on Multimedia*. 3328–3337.

[24] Xiangming Gu, Wei Zeng, Jianan Zhang, Longshen Ou, and Ye Wang. 2023. Deep Audio-Visual Singing Voice Transcription based on Self-Supervised Learning Models. *arXiv preprint arXiv:2304.12082* (2023).

[25] Yue Guo, Yi Yang, and Ahmed Abbasi. 2022. Auto-debias: Debiasing masked language models with automated biased prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1012–1023.

[26] Chitralekha Gupta, Haizhou Li, and Ye Wang. 2017. Perceptual evaluation of singing quality. In *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 577–586.

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

[28] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* 2, 7 (2015).

[29] Jui-Yang Hsu and Li Su. 2021. VOCANO: A note transcription framework for singing voice in polyphonic music.. In *Proceedings of the 22th International Society for Music Information Retrieval Conference (ISMIR)*. 293–300.

[30] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 3451–3460.

[31] Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. 2019. Learning not to learn: Training deep neural networks with biased data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9012–9020.

[32] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016).

[33] Sangeun Kum, Jongpil Lee, Keunhyoung Luke Kim, Taehyoung Kim, and Juhan Nam. 2022. Pseudo-Label Transfer from Frame-Level to Note-Level in a Teacher-Student Framework for Singing Transcription from Polyphonic Music. In *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 796–800.

[34] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. *Advances in neural information processing systems* 30 (2017).

[35] Marianne Latinus and Margot J Taylor. 2012. Discriminating male and female voices: differentiating pitch and gender. *Brain topography* 25 (2012), 194–204.

[36] Yunqi Li, Hanxiong Chen, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2021. User-oriented fairness in recommendation. In *Proceedings of the Web Conference 2021*. 624–632.

[37] Chunxi Liu, Michael Picheny, Leda Sarı, Pooja Chitkara, Alex Xiao, Xiaohui Zhang, Mark Chou, Andres Alvarado, Caner Hazirbas, and Yatharth Saraf. 2022. Towards measuring fairness in speech recognition: casual conversations dataset transcriptions. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6162–6166.

[38] Jinglin Liu, Chengxi Li, Yi Ren, Feiyang Chen, and Zhou Zhao. 2022. Diffsinger: Singing voice synthesis via shallow diffusion mechanism. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 11020–11028.

[39] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. 2018. Conditional adversarial domain adaptation. *Advances in neural information processing systems* 31 (2018).

[40] Gilles Louppe, Michael Kagan, and Kyle Cranmer. 2017. Learning to pivot with adversarial networks. *Advances in neural information processing systems* 30 (2017).

[41] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. 2018. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*. PMLR, 3384–3393.

[42] Matthias Mauch, Chris Cannam, Rachel Bittner, George Fazekas, Justin Salamon, Jiajie Dai, Juan Bello, and Simon Dixon. 2015. Computer-aided melody note transcription using the Tony software: Accuracy and efficiency. In *Proceedings of the 1st International Conference on Technologies for Music Notation and Representation*.

[43] Matthias Mauch and Simon Dixon. 2014. pYIN: A fundamental frequency estimator using probabilistic threshold distributions. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 659–663.

[44] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–35.

[45] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. 2018. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 8 (2018), 1979–1993.

[46] Emilio Molina, Isabel Barbancho, Emilia Gómez, Ana Maria Barbancho, and Lorenzo J Tardón. 2013. Fundamental frequency alignment vs. note-based melodic similarity for singing voice assessment. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 744–748.

[47] Emilio Molina, Ana Maria Barbancho-Perez, Lorenzo Jose Tardon-Garcia, Isabel Barbancho-Perez, et al. 2014. Evaluation framework for automatic singing transcription. *Proceedings of the 15th International Society for Music Information Retrieval Confence (ISMIR)* (2014).

[48] Randall S Moore. 1991. Comparison of children's and adults' vocal ranges and preferred tessituras in singing familiar songs. *Bulletin of the Council for Research in Music Education* (1991), 13–22.

[49] Ryo Nishikimi, Eita Nakamura, Masataka Goto, Katsutoshi Itoyama, and Kazuyoshi Yoshii. 2020. Bayesian singing transcription based on a hierarchical generative model of keys, musical notes, and f0 trajectories. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020), 1678–1691.

[50] Longshen Ou, Xiangming Gu, and Ye Wang. 2022. Transfer learning of wav2vec 2.0 for automatic lyric transcription. In *Proceedings of the 23rd International Society for Music Information Retrieval Conference (ISMIR)*.

[51] Thai-Hoang Pham, Xueru Zhang, and Ping Zhang. 2023. Fairness and accuracy under domain generalization. *arXiv preprint arXiv:2301.13323* (2023).

[52] David Andrew Puts, Carolyn R Hodges, Rodrigo A Cárdenas, and Steven JC Gaulin. 2007. Men's voices as dominance signals: vocal fundamental and formant frequencies influence dominance attributions among men. *Evolution and Human Behavior* 28, 5 (2007), 340–344.

[53] Colin Raffel, Brian McFee, Eric J Humphrey, Justin Salamon, Oriol Nieto, Dawen Liang, Daniel PW Ellis, and C Colin Raffel. 2014. mir_eval: A transparent implementation of common MIR metrics. In *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR)*.

[54] Sai Sathiesh Rajan, Sakshi Udeshi, and Sudipta Chattopadhyay. 2022. Aequevox: Automated fairness testing of speech recognition systems. In *Fundamental Approaches to Software Engineering: 25th International Conference, FASE 2022, Held as Part of the European Joint Conferences on Theory and Practice of Software, ETAPS 2022, Munich, Germany, April 2–7, 2022, Proceedings*. Springer International Publishing Cham, 245–267.

[55] Yi Ren, Xu Tan, Tao Qin, Jian Luan, Zhou Zhao, and Tie-Yan Liu. 2020. Deepsinger: Singing voice synthesis with data mined from the web. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1979–1989.

[56] Mhd Hasan Sarhan, Nassir Navab, Abouzar Eslami, and Shadi Albarqouni. 2020. Fairness by learning orthogonal disentangled representations. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*. Springer, 746–761.

[57] Xudong Shen, Yongkang Wong, and Mohan Kankanhalli. 2022. Fair representation: guaranteeing approximate multiple group fairness for unknown tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 1 (2022), 525–538.

[58] Adrian P Simpson. 2009. Phonetic differences between male and female speech. *Language and linguistics compass* 3, 2 (2009), 621–640.

[59] Nasim Sonboli, Jessie J Smith, Florencia Cabral Berenfus, Robin Burke, and Casey Fiesler. 2021. Fairness and transparency in recommendation: The users' perspective. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*. 274–279.

[60] Cynthia Tam, Heidi Schwellnus, Ceilidh Eaton, Yani Hamdani, Andrea Lamont, and Tom Chau. 2007. Movement-to-music computer technology: a developmental play experience for children with severe physical disabilities. *Occupational therapy international* 14, 2 (2007), 99–112.

[61] Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*. PMLR, 6105–6114.

[62] Enzo Tartaglione, Carlo Alberto Barbano, and Marco Grangetto. 2021. End: Entangling and disentangling deep representations for bias correction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 13508–13517.

[63] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. 2017. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7167–7176.

[64] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[65] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *Proceedings of the international workshop on software fairness*. 1–7.

[66] Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben, and Björn W Schuller. 2023. Dawn of the transformer era in speech emotion recognition: closing the valence gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).

[67] Jun-You Wang and Jyh-Shing Roger Jang. 2021. On the preparation and validation of a large-scale dataset of singing transcription. In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 276–280.

[68] Xianke Wang, Wei Xu, Weiming Yang, and Wenqing Cheng. 2022. Musicyolo: A Sight-Singing Onset/Offset Detection Framework Based on Object Detection Instead of Spectrum Frames. In *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 396–400.

[69] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. 2020. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8919–8928.

[70] Sandra P Whiteside and Carolyn Hodgson. 2000. Some acoustic characteristics in the voices of 6-to 10-year-old children and adults: a comparative sex and developmental perspective. *Logopedics Phoniatrics Vocology* 25, 3 (2000), 122–132.

[71] Luwei Yang, Akira Maezawa, Jordan BL Smith, and Elaine Chew. 2017. Probabilistic transcription of sung melody using a pitch dynamic model. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 301–305.

[72] Weiming Yang, Xianke Wang, Bowen Tian, Wei Xu, and Wenqing Cheng. 2022. A Multi-stage Automatic Evaluation System for Sight-singing. *IEEE Transactions on Multimedia* (2022).

[73] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *International conference on machine learning*. PMLR, 325–333.

[74] Chen Zhang, Jiaxing Yu, LuChin Chang, Xu Tan, Jiawei Chen, Tao Qin, and Kejun Zhang. 2022. PDAugment: Data Augmentation by Pitch and Duration Adjustments for Automatic Lyrics Transcription. In *Proceedings of the 23rd International Society for Music Information Retrieval Conference (ISMIR)*.

[75] Lichao Zhang, Ruiqi Li, Shoutong Wang, Liqun Deng, Jinglin Liu, Yi Ren, Jinzheng He, Rongjie Huang, Jieming Zhu, Xiao Chen, et al. 2022. M4Singer: A Multi-Style, Multi-Singer and Musical Score Provided Mandarin Singing Corpus. *Advances in Neural Information Processing Systems* 35 (2022), 6914–6926.

[76] Han Zhao, Amanda Coston, Tameem Adel, and Geoffrey J Gordon. 2019. Conditional learning of fair representations. *arXiv preprint arXiv:1910.07162* (2019).

[77] Mingmin Zhao, Shichao Yue, Dina Katabi, Tommi S Jaakkola, and Matt T Bianchi. 2017. Learning sleep stages from radio signals: A conditional adversarial architecture. In *International Conference on Machine Learning*. PMLR, 4100–4109.