ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10447167

BALANCING SPEAKER-RATER FAIRNESS FOR GENDER-NEUTRAL SPEECH EMOTION RECOGNITION

Woan-Shiuan Chien, Shreya G. Upadhyay, Chi-Chun Lee

Department of Electrical Engineering, National Tsing Hua University, Taiwan

ABSTRACT

Speech emotion recognition (SER) adds to the humane aspects of voice technologies to enhance user experiences. The ground truth emotion annotations provided by human raters and attributes related to the speakers themselves arise a compounded fairness issue in SER. While there exist works in fair SER, our work presents one of the first studies in addressing the unique joint speaker-rater (two-sided) bias, focusing on the issue of gender fairness. Our cross-reference evaluation demonstrates that the SER fair model, which merely mitigates one-sided bias introduces biases when examining from another viewpoint. Furthermore, in order to handle model stability when optimizing for these compounded speaker-rater constraints, we introduce a flexible controlled mechanism that dynamically balances the contribution of each viewpoint. Our analyses show the efficacy of our approach in achieving a fair SER that meets the dual speaker-rater gender neutrality criterion.

Index Terms— speech emotion recognition, fairness, gender neutrality, speaker-rater biases

1. INTRODUCTION

Emotion AI powered by speech emotion recognition (SER) is rapidly shaping our next-generation voice technology [1]. Particularly, the humane aspect of SER profoundly transforms user experiences into something more relatable and personal. Yet, as SER becomes an integral part of our daily life and even impacts many decision-making processes, ensuring fairness of SER is critical for advancing responsible-AI applications [2]. A typical SER model is constructed by learning on datasets comprised of human *speakers* engaging in spoken dialogs with human *raters* providing ground truth labels [3]. SER is thus *compoundly-biased* as learning happens on speech samples generated and rated by humans, inspiring researchers to begin actively tackling the fairness issue in SER (summarized in a recent review [4]).

These emerging works handle fairness issues of SER in an “one-sided” manner, that is to mitigate biases arise due to attributes of the speaker or rater only. For instance, Gorrostieta et al. propose an adversarial invariant network to alleviate SER biases attributed to speaker’s age and gender [5]. Similarly, Gu et al. use an attribute predictor along with adversarial training to ensure the gender neutrality of speaker’s acoustic embedding [6]. Our most recent work [7] is the first that addresses SER biases induced by rater’s gender, shifting from a common focus of speaker attributes to much less-studied rater attributes. Despite these advancements, taking an one-sided approach falls into another pitfall as a single viewpoint (e.g., speaker) of fairness ignores and even induces biases when examining from the other viewpoint (e.g., rater) [8].

While this one-sided issue is known, past literature has pointed out a critical challenge in handling “two-sided” fairness optimization, i.e., joint training naïvely by adding the two constraints induces model conflict [9, 10]. This is likely caused by the vast differences in the dual constraints, e.g., in the case of SER, one constraint caters to raters (attributes of multiple raters applying to each data sample) and the other to speakers (attributes of speakers encompassing multiple speech samples). The use of various invariant strategies to achieve fairness under these dual constraints often pulls the model in opposing directions, complicating the process toward balanced and stable training [11]. Consequently, developing algorithms to achieve a fair SER system extends beyond merely a single bias elimination but also generalization across constraints. It is imperative to design a strategy that carefully navigates and balances fairness when examining attributes from speakers and raters simultaneously.

In this work, we focus on the study of gender-neutral SER by jointly considering speaker-rater fairness. First, by adopting the use of cross-evaluations scheme [12] (termed as inter-fairness and intra-fairness [13]), our analysis shows that the one-sided fair SER model does not generalize well across different viewpoints, and further direct joint training on these two one-sided fairness constraints destabilizes the SER model. With these insights, we propose a balancing speaker-rater fairness mechanism toward realizing gender neutrality. It works by explicitly calculating the distributional distances between latent embeddings derived from a speaker fair model and a rater fair model as a flexible weight, which helps dynamically adjust the contribution of the two one-sided fair constraints at joint learning. Our experimental results reveal that our balancing mechanism stands as a robust arbitrator, effectively mediating the nuances between speaker and rater fairness constraints.

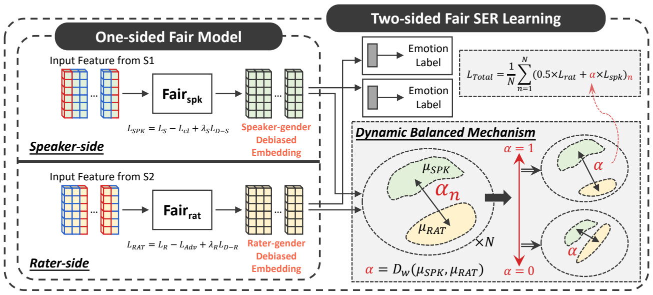

Two-sided Fair SER Learning
One-sided Fair Model
Input Feature from S1
Fair$_{spk}$
$L_{SPK} = L_S - L_{cl} + \lambda_S L_{D-S}$
Speaker-gender Debiased Embedding
Speaker-side
Input Feature from S2
Fair$_{rat}$
$L_{RAT} = L_R - L_{Adv} + \lambda_R L_{D-R}$
Rater-gender Debiased Embedding
Rater-side
Emotion Label
Emotion Label
$L_{Total} = \frac{1}{N}\sum_{n=1}^{N}(0.5\times L_{rat} + \alpha\times L_{spk})_n$
Dynamic Balanced Mechanism
μ_{SPK}
α_n
μ_{RAT}
$\times N$
$\alpha = D_w(\mu_{SPK}, \mu_{RAT})$
$\alpha = 1$
$\alpha = 0$
α
α

Fig. 1: Overview of the fair speech emotion recognition (SER) architecture using both one-sided and our proposed two-sided learning frameworks. The one-sided debiased embedding is first trained independently, then we compute the WD distance between these embeddings per batch to derive the flexible control parameter α, which adjusts the other side's contribution dynamically. Finally, the two-sided fair SER model is batch-wise optimized by the L_{Total}.

Table 1: Emotion distribution in the bias set.

	Overall	Neutral	Happiness	Anger	Sadness	Frustration
S1	7362	1706	1633	1099	1080	1844
S2	4324	1323	446	628	628	1299

2. FAIRNESS ANALYSIS: ONE-SIDED MODEL

2.1. Dataset

The IEMOCAP dataset [14] is a benchmark SER corpus with a gender balance (one male and one female) in each of its five dyadic spoken interaction sessions. There are six unique raters (two males and four females) that rate the emotions. The consensus labels are obtained based on the plurality rule. Aligning with most conventional SER research, our work treats emotion detectors as the primary task and specifically targets samples labeled with five emotion categories: Neutral, Happiness, Anger, Sadness, and Frustration.

2.1.1. Study Sets

There is a total of 7362 utterances comprising the five primary categorical emotions. In this work, we focus our study on two different sets of data: **S1** (the whole dataset), which includes the entire speaker set in the IEMOCAP (a total of 7362 samples included). This set is used for the speaker-side gender fairness study; **S2** (the rater-gender biased set), which includes those samples where the ground truth labels align with the emotion annotation given by either the male rater or female rater only, indicating rater biases. This set is used for the rater-side gender fairness study, which includes a total of 4323 samples. S2 is identical to the set used to study the gender bias of raters in our previous work [7]. Table 1 presents the emotion label distribution of S1 and S2 used in our study.

2.2. Cross Evaluation of One-sided Models

In this section, we present our analysis on cross referencing one-sided gender-neutral fair model to examine the generalizability of these models. We first describe the construction of the corresponding one-sided fair SER, then examine the fairness of these models in a cross-evaluation scheme including two experiments: intra-fairness and inter-fairness analysis.

2.2.1. Acoustic Features

We use the Huggingface framework [15] to extract 768-dimensional latent wav2vec 2.0 [16] vectors as the acoustic features. This pre-trained audio encoder can directly embed information from raw audio, transforming the waveform into the embedding. All features undergo speaker-wise z-normalization.

2.2.2. One-sided Fair SER Model

We construct two one-sided fair models, one for rater and another for speaker viewpoint. In terms of debiasing *rater-side* gender attribute, we utilize the model most recently proposed [7]. This model, **Fair**$_{\text{rat}}$, produces gender-debiased representations by using three loss functions: one for predicting ground truth emotional labels (L_{R}), another for minimizing the distance between gender class in the feature space ($L_{\text{D-R}}$), and the third for detecting gender from embedding (L_{Adv}). In summary, the **Fair**$_{\text{rat}}$ parameters are optimized using the following loss function with a hyper-parameter λ_{R}:

$$L_{\text{RAT}} = L_{\text{R}} - L_{\text{Adv}} + \lambda_{\text{R}} L_{\text{D-R}}, \tag{1}$$

For *speaker-side*, we construct a similar framework as Fair$_{\text{rat}}$ by using a fairness constraint contrastive framework to train the gender debiasing model **Fair**$_{\text{spk}}$. Thus, Fair$_{\text{spk}}$ integrates the same three types of loss functions as well. The first loss is L_{S}, which corresponds to the standard cross entropy loss associated with the SER task. We further measure the Wasserstein Distance (WD) [17] between male speaker's features and female speaker's features as the optimization objective ($L_{\text{D-S}}$). To achieve a gender-neutral embedding, we employ a contrastive loss [18] that captures gender information into the embeddings. Given x_i as the speaker-bias embedding of the inputs with the encoder, we identify positive samples as those embeddings sharing the same gender as x_i and negative samples as the set of embeddings with the opposite gender. With the aim of eliminating the gender-specific information in the speaker-bias embedding, this loss L_{cl} is subtracted from the emotion detection network. Hence, the parameters of **Fair**$_{\text{spk}}$ are trained by minimizing the following loss function with the hyper-parameter λ_{S}:

$$L_{\text{SPK}} = L_{\text{S}} - L_{\text{cl}} + \lambda_{\text{S}} L_{\text{D-S}}, \tag{2}$$

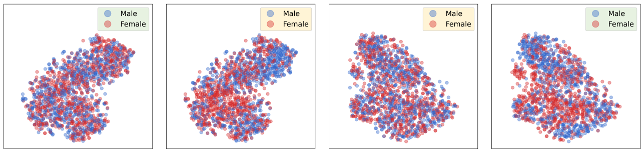
We train five binary emotion detectors for each of these two "one-sided" fair models. All of them are trained by setting the decaying factor at 0.001, the dropout is set to 0.2. We use the Adam optimizer with a 0.001 learning rate to optimize the parameters and train models for 500 epochs with a 32-fixed batch size.

2.2.3. Cross Reference Analysis

We conduct two experiments to evaluate the effectiveness of the one-sided gender-neutral models. Note that for all ex-

Table 2: Results of the recognition performances and fairness metrics on the one-sided model. Green backgrounds represent fairness evaluation from the speaker-side gender perspectives, while yellow highlights rater-side gender fairness evaluations.

	Neutral			Happiness			Anger			Sadness			Frustration		
	F1(%)	ΔSP_{spk}	ΔSP_{rat}	F1(%)	ΔSP_{spk}	ΔSP_{rat}	F1(%)	ΔSP_{spk}	ΔSP_{rat}	F1(%)	ΔSP_{spk}	ΔSP_{rat}	F1(%)	ΔSP_{spk}	ΔSP_{rat}
DNN	77.73	0.452	0.649	70.00	0.511	0.428	76.44	0.378	0.389	82.28	0.359	0.169	63.54	0.385	0.626
Fair_{spk}	70.68	0.226	0.488	65.80	0.380	0.366	73.26	0.234	0.379	75.50	0.260	0.208	63.34	0.301	0.550
Fair_{rat}	68.80	0.403	0.352	65.14	0.691	0.126	75.68	0.372	0.189	76.84	0.291	0.088	70.22	0.411	0.448



(a) Fair_{spk} / speaker-gender (b) Fair_{spk} / rater-gender (c) Fair_{rat} / rater-gender (d) Fair_{rat} / speaker-gender

Fig. 2: The t-SNE scatter plot of Fair_{spk} (a)(b) and Fair_{rat} (c)(d) from Anger detector. We color the points according to gender distribution from either the speaker-side or rater-side.

periments, a session-independent cross-validation scheme is applied. We present the recognition performances as well as fairness metrics for the two one-sided fair models (Fair_{spk} and Fair_{rat}). Our target emotion labels are derived from voted ground truths, and the emotion recognition performance on S1 is evaluated by the weighted F1-score. The fairness metric, statistical parity score [19] (ideal = 0), is evaluated on both study sets, S1 for *speaker-side* and S2 for *rater-side*, which is further denoted as ΔSP_{spk} and ΔSP_{rat} respectively. The parity metric quantifies whether our model favors one gender category over another when performing prediction. The cross-reference analysis involves two distinct evaluation schemes to assess the generalizability in the fairness of the models:

- **Intra-Fairness:** We evaluate the one-sided gender-neutral fairness in their own corresponding viewpoint, i.e., using ΔSP_{spk} for Fair_{spk} and ΔSP_{rat} for Fair_{rat}.
- **Inter-Fairness:** Based on the one-sided fair models, we evaluate the fairness metric of one-sided using the model of the other. This means using ΔSP_{spk} for Fair_{rat} and ΔSP_{rat} for Fair_{spk}.

Table 2 shows the results of the recognition and fairness performances. We also train a vanilla three-layer DNN model without any fairness constraint to provide a reference on the assessment of the intra and inter-fairness of Fair_{spk} and Fair_{rat}. We highlight the inter-fairness evaluation (yellow for rater's perspective and green for speaker's perspective). We observe that the fair model does not transfer well in addressing fairness issues from another perspective. Specifically, the speaker-side model, Fair_{spk}, exhibits a substantial increase in parity score of 0.136 for Neutral, 0.24 for Happiness, 0.19 for Anger, 0.12 for Sadness and 0.102 for Frustration in ΔSP_{rat} as compared to Fair_{rat}. Likewise, Fair_{rat} shows increases parity scores in $\Delta SPspk$ across all emotions relative to results of Fair_{spk}. Moreover, we visualize the model's embedding using t-SNE (Fig. 2) by randomly selecting 1500 samples from Anger-Fair_{spk} and Anger-Fair_{rat} detector by viewing from the speaker or rater perspective. For example, Fig. 2(b) plots Fair_{spk} embeddings with color indicate *rater's* gender class (blue is male, red is female). From both Fig. 2 and Table 2, we can see that although the intra-unfair issue can be addressed with one-sided fairness learning, the inter-fairness experiment shows that fairness is not maintained (sometimes even degrade) when viewing from another viewpoint.

3. TWO-SIDED FAIR SER LEARNING

3.1. Proposed Dynamic Balanced Mechanism

Given the analyses described in the previous section, we propose to address gender fairness learning through joint training to derive a "two-sided" model. A most basic approach is through straight optimizing the summation of L_{SPK} and L_{RAT}, which can be regarded as a simple multi-task learning (termed as **Basic**). In this work, we propose an improved strategy that dynamically balances the contribution of each fairness constraint while performing this joint learning, as illustrated in Fig. 1. Specifically, at every batch (n) during training N batches, we compute the WD distances (D_W) between speaker debiased embedding (μ_{SPK}) and rater debiased embedding (μ_{RAT}) derived from Fair_{spk} and Fair_{rat}. Then we derive a flexible control parameter (denoted as α) from the computed distributional distance: $D_W(\mu_{\text{SPK}}, \mu_{\text{RAT}})$. Then, by anchoring on a fixed one-sided model (e.g., Fair_{rat}), α values can dynamically determine the extent of the contribution at every batch from the other side (in this case Fair_{spk}). That is, a higher α value signifies a further distance between the two embedding sets, indicating a need for a stronger contribution of fairness constraint from the speaker-side, and vice versa for rater-side. The complete balancing strategy is optimized batch-wise using the following total loss function:

$$L_{\text{Total}} = \frac{1}{N} \sum\nolimits_{n=1}^{N} (0.5 \times L_{\text{RAT}} + \alpha \times L_{\text{SPK}})_n, \qquad (3)$$

α ranges from 0 to 1. No matter whether the viewpoint is from the rater or speaker, it is the *same* speech sample projecting through a network into a fair representation. As learning happens, the network should learn to move this embedding to a space where both fairness constraints meet. If the two viewpoints' embeddings are close, that indicates the two-sided model already encodes the speech sample into a jointly fair space; if they are further, the network should consider the fairness constraint more from that viewpoint to move the embedding. This optimizing strategy is adaptive to within each batch to alleviate the conflicts, leading to unstable learning. Our proposed balancing mechanism provides flexible and dynamic control during dual-constrained fairness learning.

Table 3: A summary of the recognition results on each emotion category by F1 (%) with SD and the fairness performance by ΔSP with SD. All numbers are performed based on the average of ten models.

		Neutral	Happiness	Anger	Sadness	Frustration
F1(%)	Basic	66.26 ± 2.277	65.37 ± 1.987	69.70 ± 2.776	69.15 ± 2.406	66.10 ± 2.417
	Proposed	68.06 ± 1.241	67.67 ± 1.514	73.09 ± 1.546	71.35 ± 1.119	69.20 ± 1.617
ΔSP_{spk}	Basic	0.365 ± 0.082	0.566 ± 0.085	0.302 ± 0.032	0.318 ± 0.045	0.347 ± 0.026
	Proposed	0.269 ± 0.025	0.359 ± 0.064	0.254 ± 0.015	0.215 ± 0.027	0.310 ± 0.006
ΔSP_{rat}	Basic	0.525 ± 0.039	0.311 ± 0.071	0.257 ± 0.060	0.116 ± 0.028	0.521 ± 0.034
	Proposed	0.407 ± 0.043	0.214 ± 0.049	0.222 ± 0.026	0.112 ± 0.014	0.406 ± 0.009
$\Delta SP_{spk}\times\Delta SP_{rat}$	Basic	0.192	0.175	0.078	0.037	0.181
	Proposed	0.109	0.075	0.056	0.024	0.126
$\lvert\Delta SP_{spk}-\Delta SP_{rat}\rvert$	Basic	0.160	0.255	0.060	0.202	0.174
	Proposed	0.137	0.154	0.039	0.103	0.096

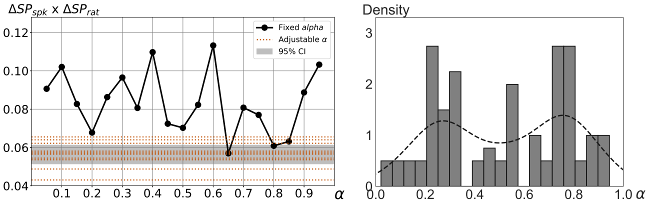
3.2. Results and Analyses

We run experiments ten times and also report the standard deviation (SD) to examine the stability of performances. We also analyze the balancing behavior from our proposed gender-neutral model to determine its learning trajectory.

3.2.1. Recognition Performances

Table 3 summarizes the average performance and SD values over the ten trials for **Basic** (the basic joint speaker-rater learning model) and **Proposed** (our proposed gender-neutral SER). Our goal is to examine the stability and effectiveness of our proposed SER model in terms of recognition and fairness performances. These performances are based on speaker-rater joint training, therefore, SP values can be regarded as intra-gender values. Moreover, to demonstrate that ΔSP_{spk} and ΔSP_{rat} are more likely to converge in a joint speaker-rater fair state, we also report the product and difference between ΔSP_{spk} and ΔSP_{rat} as an additional metric.

Several observations can be noted by looking at both Table 2 and Table 3. First, we can observe that: (1) Our Proposed SER can maintain a competitive emotion recognition performance to those of separate one-sided gender-neutral SER (Fair_{spk} or Fair_{rat}), with the most significant drop being only 2.41% in the Anger detector. (2) Our Proposed model shows an enhanced optimal balance in performance, with relative improvements of 1.8% for Neutral, 2.3% for Happiness, 1.39% for Anger, 2.2% for Sadness, and 3.1% for Frustration over the Basic. Second, while the performance of our Proposed model might be slightly lower than the intra-fairness of one-sided SER, it outperforms the inter-fairness value of the one-sided SER significantly. As for comparison with Basic model, ΔSP_{spk} and ΔSP_{rat} of our Proposed model performs well for all emotions. At last, we turn our attention to the stability of joint training. Without a balancing mechanism between two one-sided models, the Basic model results in obviously unstable performances in SD, both in terms of recognition and fairness performances (Proposed outperforms in every emotion). This is likely attributed to our balanced mechanism that offers opportunities for adapting the two one-sided models during the learning process.



(a) The product of ΔSP_{spk} and ΔSP_{rat} under different fixed α settings.

(b) Histogram of dynamic control parameter α.

Fig. 3: Illustration of balancing behavior in Proposed SER.

3.2.2. Balancing Behavior

To understand the balancing behavior in our proposed gender-neutral SER, we conduct two analyses on the *Anger* detector to understand our balanced learning mechanism. First, we examine the results obtained when using fixed *alpha* during our two-sided training. Specifically, we fixed the α weight, incrementing from 0.05 to 0.95 in steps of 0.05. Fig. 3a shows the results. Black solid line corresponds to fixed $alpha$, each of the dotted orange lines indicates a trial of our proposed SER model, and the grey interval indicates the 95% confidence interval of our approach. It is clear that our proposed approach obtains a much better fairness metric (lower values of $\Delta SP_{spk}\times\Delta SP_{rat}$). Second, we can identify which side has more influence on the fairness of SER. We extract the batch-wise α values from the *Anger* detector and present them in a histogram, as depicted in Fig. 3b. We first note that these values are diverse and signify the dynamic properties in controlling two-sided learning. Further, the two peaks of α values located near the 0.25 and 0.75 marks suggest that both the speaker bias and rater-side bias hold major influences on the learning toward achieving the model. In short, simply using a constant α is not adequate for this compounded fairness learning. Instead, the adaptability of the flexible dynamic balancing α plays a key role in the reported robustness.

4. CONCLUSION

In this work, we explore and address speaker-rater fairness issues in gender-neutral SER. We first construct the corresponding fair one-sided SER and propose a balancing speaker-rater fairness mechanism toward realizing gender neutrality in SER. Our results and analyses reveal interesting insights: 1) The one-sided fair SER model struggles to adapt effectively across different viewpoints. 2) The effectiveness and stability of the dynamic balancing mechanism in mitigating two-sided biases. We plan to extend our approach to other attributes beyond gender to continue advancing fair SER models.

5. REFERENCES

[1] Andrew McStay, "Emotional ai, soft biometrics and the surveillance of emotional life: An unusual consensus on privacy," *Big Data & Society*, vol. 7, no. 1, pp. 2053951720904386, 2020.

[2] Chi-Chun Lee, Theodora Chaspari, Emily Mower Provost, and Shrikanth S Narayanan, "An engineering view on emotions and speech: From analysis and predictive models to responsible human-centered applications," *Proceedings of the IEEE*, 2023.

[3] Shreya G Upadhyay, Woan-Shiuan Chien, Bo-Hao Su, Lucas Goncalves, Ya-Tse Wu, Ali N. Salman, Carlos Busso, and Chi-Chun Lee, "An intelligent infrastructure toward large scale naturalistic affective speech corpora collection," in *2023 10th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2023.

[4] Tiantian Feng, Rajat Hebbar, Nicholas Mehlman, Xuan Shi, Aditya Kommineni, Shrikanth Narayanan, et al., "A review of speech-centric trustworthy machine learning: Privacy, safety, and fairness," *APSIPA Transactions on Signal and Information Processing*, vol. 12, no. 3, 2023.

[5] Cristina Gorrostieta, Reza Lotfian, Kye Taylor, Richard Brutti, and John Kane, "Gender de-biasing in speech emotion recognition.," in *INTERSPEECH*, 2019, pp. 2823–2827.

[6] Xiangming Gu, Wei Zeng, and Ye Wang, "Elucidate gender fairness in singing voice transcription," *arXiv preprint arXiv:2308.02898*, 2023.

[7] Woan-Shiuan Chien and Chi-Chun Lee, "Achieving fair speech emotion recognition via perceptual fairness," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[8] Marco Morik, Ashudeep Singh, Jessica Hong, and Thorsten Joachims, "Controlling fairness and bias in dynamic learning-to-rank," in *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, 2020, pp. 429–438.

[9] Sungwon Han, Seungeon Lee, Fangzhao Wu, Sundong Kim, Chuhan Wu, Xiting Wang, Xing Xie, and Meeyoung Cha, "Dualfair: Fair representation learning at both group and individual levels via contrastive self-supervision," in *Proceedings of the ACM Web Conference 2023*, 2023, pp. 3766–3774.

[10] Lingxiao Huang and Nisheeth Vishnoi, "Stable and fair classification," in *International Conference on Machine Learning*. PMLR, 2019, pp. 2879–2890.

[11] Lequn Wang and Thorsten Joachims, "Uncertainty quantification for fairness in two-stage recommender systems," in *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, 2023, pp. 940–948.

[12] Quan Zhou, Jakub Mareček, and Robert Shorten, "Subgroup fairness in two-sided markets," *Plos one*, vol. 18, no. 2, pp. e0281443, 2023.

[13] Tom Sühr, Asia J Biega, Meike Zehlike, Krishna P Gummadi, and Abhijnan Chakraborty, "Two-sided fairness for repeated matchings in two-sided markets: A case study of a ride-hailing platform," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 3082–3092.

[14] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[15] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al., "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 2020, pp. 38–45.

[16] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12449–12460, 2020.

[17] Rui Feng, Yang Yang, Yuehan Lyu, Chenhao Tan, Yizhou Sun, and Chunping Wang, "Learning fair representations via an adversarial framework," *arXiv preprint arXiv:1904.13341*, 2019.

[18] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan, "Supervised contrastive learning," *Advances in neural information processing systems*, vol. 33, pp. 18661–18673, 2020.

[19] Ke Yang and Julia Stoyanovich, "Measuring fairness in ranked outputs," in *Proceedings of the 29th international conference on scientific and statistical database management*, 2017, pp. 1–6.