Interspeech 2024
1-5 September 2024, Kos, Greece

An Investigation of *Group* versus *Individual* Fairness in Perceptually Fair Speech Emotion Recognition

Woan-Shiuan Chien, Chi-Chun Lee

Department of Electrical Engineering, National Tsing Hua University, Taiwan
wschien@gapp.nthu.edu.tw, cclee@ee.nthu.edu.tw

Abstract

Speech emotion recognition (SER) has been extensively integrated into voice-centric applications. A unique fairness issue of SER stems from the naturally biased labels given by raters as ground truth. While existing efforts primarily aim to advance SER fairness through a group (i.e., gender) fairness standpoint, our analysis reveals that label biases arising from individual raters also persist and require equal attention. Our work presents a systematic analysis to determine the effect of enhanced group (gender) fairness on individual fairness. Specifically, by evaluating two datasets we demonstrate that there exists a trade-off between group and individual fairness when removing group information. Moreover, our results indicate that achieving group fairness results in diminished individual fairness, particularly when the attribute distributions of the two groups are significantly distant. This work brings initial insights into issues of group and individual fairness in the SER systems.
Index Terms: speech emotion recognition, group fairness, individual fairness, perceptual bias

1. Introduction

Emotion AI manifested through Speech Emotion Recognition (SER) heralds a transformative era by seamlessly weaving humane and intimately personal nuances into the voice-enabled technologies [1]. With the growing integration of SER into our daily routines and its application in diverse areas, securing its fairness is essential not only for developing responsible-AI systems but also for enhancing user trust and inclusivity [2]. Particularly, SER models fundamentally rely on *human raters* for emotion interpretation and labeling, which inherently embeds perceptual biases into the systems [3]. Such biases erode the precision and fairness of emotion recognition, which in turn affects user trust and leads to perceptions of the system as unfair. Recognizing the complexity of this issue on perceptions, researchers have intensified their focus on dissecting and addressing these biases and fairness issues in rater perceptions, aiming to enhance the reliability and fairness of SER across different user groups [4–6].

Current research predominantly navigates fairness concerns in the SER system through a *group fairness* perspective, a concept focused on achieving equitable outcomes across groups (predefined attributes) by satisfying statistical parity criteria [7]. One notable effort to tackle the perceptual biases in SER is done by Chien et al. [8, 9], which effectively confronts these gender biases through a group fairness perspective, exhibiting promising results in mitigating such gender-based biases. However, these labeling differences are profoundly shaped by the diversity of human attributes (e.g., culture, gender, age) and the subjectivity inherent in emotion perception [10, 11]. Being one of the inputs to SER learning may lead to a generalized view that fails to acknowledge the unique ways in which individuals are perceived [12]. While this approach to mitigating rating bias forms a critical foundation for group fairness, it often overlooks the rating differences among individuals that exist within and between these groups.

Nevertheless, *individual fairness* stands in contrast to group fairness, which aims for equity across broader demographic groups. It seeks to ensure that individuals with similar representations would receive similar predictions from the system [13]. Despite the critical importance of both concepts, extensive research has revealed that achieving either group or individual fairness alone may not be sufficient for comprehensive fairness due to the distinct nature of these fairness concepts [14]. Moreover, efforts to uphold one can inadvertently compromise the other, as this distinction introduces inherent conflicts between the two fairness paradigms [15]. For instance, while statistical parity aims to balance outcomes across groups, it may neglect the nuanced differences between individuals, potentially leading to disparities in treatment and outcomes [16]. Consequently, the quest for a perceptually fair SER system extends beyond merely a single view of fairness but also entails understanding the impact between different fairness aspects. To gain insights into group versus individual fairness perspectives on SER learning, we formulate the following research questions:

- RQ1: How would the rating biases arising from group or individual perspectives manifest within emotional corpora?
- RQ2: How would individual fairness be affected when we improve group fairness?

To respond to these questions, we investigate *group* versus *individual* fairness on two speech corpora that provide raters' information, namely IEMOCAP [17] and BIIC-Podcast [18]. First, by examining labeling differences from both group and individual perspectives, our analysis indicates that even though rating biases manifest differently across various corpora, such rating biases indeed exist in both emotional corpora. With this insight, we adopt a perceptually fair SER model toward realizing group fairness. We systematically focus on the impact of group fairness constraints on individual fairness when constructing the SER model. Our results suggest that: 1) there is a clear trade-off between group fairness and individual fairness when removing partial group information. 2) satisfying group fairness decreases the level of individual fairness with large Wasserstein distance (WD) between attribute distributions of two groups. These findings in the SER system also align with the expected trends reported in the fairness literature, providing initial evidence that both group and individual fairness issues coexist in SER systems and that the need for a more comprehensive approach to achieving fairness.

10.21437/Interspeech.2024-461

Table 1: *Data distribution within the study sets of each corpus and preliminary analyses of perceptual differences on group (gender-based) and individual perspectives in study sets.*

		IEMOCAP					BIIC-Podcast				
		Overall	Neu.	Hap.	Ang.	Sad.	Overall	Neu.	Hap.	Ang.	Sad.
Data Distribution (Numbers)											
	S_C	2593	383	1187	471	552	30733	11828	13122	2293	3490
	S_{NC}	3025	1323	446	628	628	30736	12726	10888	4035	3087
Label Similarity (%)											
Group (Male)	All Data	80.66	90.04	91.73	90.81	85.83	63.56	77.22	73.65	86.55	72.02
	S_{NC}	67.72	87.30	69.73	85.03	77.55	56.22	51.65	42.17	60.22	42.60
Group (Female)	All Data	59.85	34.82	80.96	50.77	53.80	70.03	68.58	88.29	73.21	80.77
	S_{NC}	32.28	12.70	30.27	14.97	22.45	43.78	48.35	57.83	39.78	57.40
Inter-Annotator Agreement (κ)											
Individual	All Data	0.446	0.328	0.306	0.294	0.312	0.421	0.226	0.247	0.218	0.224
Group-level (Male)	All Data	0.467	0.348	0.360	0.402	0.316	0.372	0.212	0.218	0.194	0.226
Group-level (Female)	All Data	0.434	0.305	0.342	0.318	0.288	0.413	0.231	0.210	0.220	0.216

2. Emotional Corpora

In this study, we utilize datasets of two different scales, with both including detailed rater information crucial for our further analysis. Consistent with conventional SER research practices, our focus is primarily on emotion detection, targeting samples categorized into four emotional states: Neutral, Happiness, Anger, and Sadness. Several details are listed below:

- **IEMOCAP dataset** [17] stands as a widely recognized benchmark for SER research that contains five sessions of dyadic spoken interactions, featuring one male and one female actor per session. Emotion ratings within this dataset are provided by six unique raters (2 males and 4 females) with consensus labels established using the plurality rule.
- **BIIC-Podcast dataset** (v1.01) [18] contains 170 hours of emotional Taiwanese-Mandarin, sourced from audio-sharing websites. This dataset is uniquely characterized by its diversity in raters, with 89 individuals (30 males and 59 females) contributing to the emotional labeling. The number of emotional annotations ranges from 3-7 per sample, providing a more varied labeling context. Similar to typical emotion corpora, consensus labels are derived using the plurality rule.

2.1. Study Sets

There is a total of 5618 utterances and 61469 utterances comprising the four primary categorical emotions in IEMOCAP and BIIC-Podcast respectively. In this work, we take *gender* as a representative which is the dominant point of view in studying a "group" manner. We follow the guidelines of splitting study sets suggested by [8], that is to divide the sets according to examining those samples where the consensus among male and female raters for each utterance results in the same emotional ratings (Consensus Data) or differing emotional ratings (Non-Consensus Data). Thus, two subsets are formed for further experiments: $\mathbf{S_C}$ (the gender-wise labeling-**unbiased** set), both males and females have the same emotion ratings to the ground truth labels. $\mathbf{S_{NC}}$ (the gender-wise labeling-**biased** set), the ground truth labels have either identical emotion rating as males or females. The emotion label distribution of S_C and S_{NC} in both corpora used in our study is presented in Table 1.

3. Differences in Rater Labeling

To investigate the manifestations of group and individual fairness within emotional corpora, this section introduces targeted measures to assess labeling biases arising from differences in gender-based or individual ratings. For all analyses, we apply the measures to study sets defined in Sec. 2.1 and summarize the results in Table 1.

3.1. Analysis of Labeling Bias

3.1.1. Gender-based Rating Differences

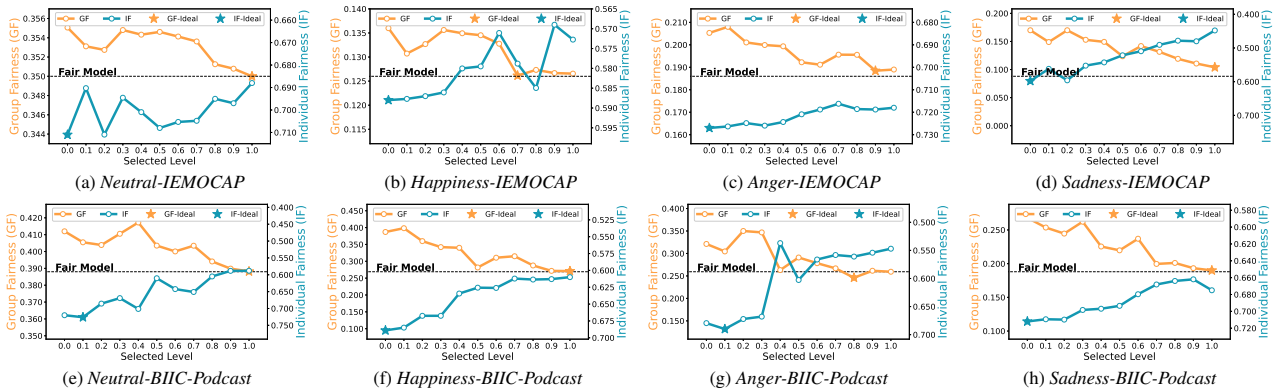
We calculate the Label Similarity to assess the matching percentage between ratings from a group perspective and the voted ground truth for each utterance. Specifically, This metric aims to measure the consistency between the consensus ratings by male and female raters against the established ground truth labels. Table 1 shows the label similarity results from a group perspective. We observe that there is a significant gender influence in the IEMOCAP dataset, where the voted ground truth labels tend to align more with male perspectives, indicating a decisive impact on male raters. This suggests that the consensus in this dataset may be skewed towards a male viewpoint. In contrast, the BIIC-Podcast dataset exhibits a more balanced scenario between male and female raters with the matching percentage differences being less than 20%. This is likely due to the larger base number of raters in BIIC-Podcast, which dilutes the impact of gender biases.

3.1.2. Individual Rating Differences

We shift focus to the individual perspective by analyzing the Inter-Annotator Agreement (IAA), recognized in literature as a measure of individual fairness [19]. For evaluating the consistency among raters' ratings, especially for categorical emotions, we employ Fleiss' Kappa (κ) [20] statistics. Typically, a lower $kappa$ value means a comparative disagreement between raters. However, as indicated by the values in Table 1, both datasets exhibit fair agreement (κ values ranging from 0.2 to 0.4) for each emotional category.

3.2. Assessment of Bias Dimensions

Noting the moderate outcomes from individual assessments, we extend the individual bias analysis to explore group-level consistency, specifically examining the rating consistency between male and female raters. This expansion aims to understand whether gender-based or individual differences play a more significant role in influencing ratings. We summarize the results in the last two rows of Table 1. The two databases show contrasting results when comparing individual κ value to group-level individual κ value. In IEMOCAP, a slightly increasing κ value as compared to the κ values in all data suggests that consensus among male raters aligns more closely, implying that gender factors within the group are the primary cause of bias. Conversely, the BIIC-Podcast exhibits little variation in κ values compared to the κ values in all data, signaling a possibility of heightened individual biases present in this particular dataset.



(a) *Neutral-IEMOCAP* (b) *Happiness-IEMOCAP* (c) *Anger-IEMOCAP* (d) *Sadness-IEMOCAP*

(e) *Neutral-BIIC-Podcast* (f) *Happiness-BIIC-Podcast* (g) *Anger-BIIC-Podcast* (h) *Sadness-BIIC-Podcast*

Figure 1: *Trade-off between group fairness and individual fairness while eliminating partial group information. The dotted grey line indicates the group and individual fairness value of the perceptually fair SER model as a reference. "⋆" represents the ideal value.*

In summary, our analysis reveals distinct patterns of labeling bias within the IEMOCAP and BIIC-Podcast datasets. Specifically, the IEMOCAP dataset exhibits a more pronounced gender influence due to fewer raters, thereby reducing the visibility of individual differences. Conversely, the BIIC-Podcast dataset, characterized by a larger pool of raters, tends to highlight individual variances more clearly. These contrasting outcomes showcase the presence of labeling biases at varying levels across different emotional corpora. Our findings are also corroborated by past literature [21], suggesting that biases can become more pronounced as the volume of data increases. While the influence of group and individual perspectives differs, neither should be neglected in the pursuit of constructing a fairer SER model.

4. Impact on Perceptually Fair Model

Given the analyses described in the previous section, we realize that labeling biases are present from group and individual perspectives. In this section, our goal is to investigate how *individual fairness* is affected when we improve *group fairness*. We first construct an SER model that achieves perceptual fairness predicated on the foundation of attaining gender-based group fairness. Then, we examine the impact on individual fairness through techniques aimed at mitigating gender bias: (1) employing a domain-invariant classifier and (2) utilizing WD measures as constraints.

4.1. Perceptually Fair Model Construction

In addressing the elimination of group bias, particularly with respect to gender, we utilize the recently proposed model for perceptual fairness [8]. This model intends to produce gender-debiased representations. There are two main operations that aim to mitigate the group distribution to achieve fairer distribution. The first is constructing a domain-invariant classifier for detecting gender from embedding. Another one is for minimizing the distance between gender classes in the feature space. Hence, as the model presented by [8], we also train the perceptually fair model by optimizing the following total loss functions with a hyper-parameter λ:

$$L_{\text{Total}} = L_{\text{R}} - L_{\text{Adv}} + \lambda L_{\text{D}}, \tag{1}$$

where L_{R} is the standard cross-entropy loss for predicting ground truth emotional labels, L_{D} measures how close the distributions over groups of the rater gender attribute are in the feature space, and L_{Adv} is the gender information loss term for evaluating how much gender can be detected from embedding.

4.2. Experimental Setup

We train four binary emotion detectors for each of these models on both corpora. We first employ the Huggingface framework [22] to derive 768-dimensional latent wav2vec 2.0 [23] vectors as the acoustic features, then apply speaker-wise z-normalization to all extracted features for standardization purposes. For all experiments, we implement a session-independent cross-validation strategy in IEMOCAP. For BIIC-Podcast, we consider all emotional samples with pre-defined train-valid-test sets of the corpus. All of them are configured with a learning rate and decay factor of 0.001, and the drop out is set at 0.2. We set the batch size to 32, limit the maximum number of epochs to 500, and employ Adam as the optimizer.

4.2.1. Evaluation Schemes

The target emotion labels are derived from voted ground truth labels. Then, we further consider fairness metrics from two different perspectives.

- **Group Fairness**: statistical parity score [24] (ideal = 0) is evaluated on S_{NC}, which is satisfied if performing prediction is independent of the gender attribute. That is, the proportion of individuals in any group receiving an emotional outcome is equal to the proportion of the population as a whole [25].
- **Individual Fairness**: we adopt consistency score [25] on all data as the individual fairness, which evaluates the consistency between the embedding and raters within a k-nearest neighbor set ($k = 20$). It aims to guarantee that similar individuals should be treated similarly.

4.3. Trade-off between Group and Individual Fairness

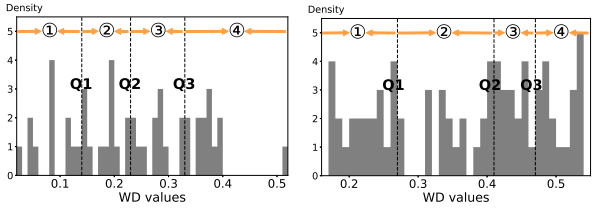
In this section, we aim to shed light on the impact of a perceptually fair model on individual fairness when optimizing for group fairness. To evaluate the extent to which operations for group fairness affect individual fairness, we conduct analyses focusing on two perspectives to investigate the trade-off between these two fairness: 1) effects of removing group information on fairness metrics (Sec. 4.3.1) and 2) influence when satisfying group fairness through Wasserstein Distance (WD) measures (Sec. 4.3.2).

4.3.1. Effects of Partial Group Information Elimination

As mentioned in Sec. 4.1, one intuitive approach to achieving group fairness involves the removal of gender attribute information. However, we recognize that while simply eliminating the gender attribute would yield a model that is fairer in terms of

Table 2: *A summary of experimental results of group fairness and individual fairness on different subsets which is to satisfy group fairness based on WD values. The bold numbers represent the ideal fairness performance under subset comparisons. The underlined numbers indicate the least optimal performance in comparison.*

	IEMOCAP								**BIIC-Podcast**							
	Group Fairness				Individual Fairness				Group Fairness				Individual Fairness			
	①	②	③	④	①	②	③	④	①	②	③	④	①	②	③	④
Neu.	0.358	0.357	0.357	**0.353**	0.672	0.677	0.670	0.632	0.426	0.418	0.418	**0.402**	0.675	0.628	0.505	0.508
Hap.	0.138	0.137	0.139	**0.137**	0.577	0.572	0.586	0.569	0.382	0.378	0.362	**0.354**	0.682	0.684	0.621	0.632
Ang.	0.211	0.202	0.213	**0.192**	0.702	0.670	0.683	0.624	0.354	0.360	0.322	**0.282**	0.576	0.558	0.531	0.529
Sad.	0.168	0.164	0.164	**0.142**	0.606	0.622	0.608	0.573	0.274	0.268	0.256	**0.252**	0.668	0.683	0.642	0.634



(a) *IEMOCAP* (b) *BIIC-Podcast*

Figure 2: *Histogram of WD values.*

group fairness definitions, it might exacerbate conditions of individual unfairness. Hence, we design the following experiment in which we randomly select data to remove gender information from those identified with perceptual biases (S_{NC}), aiming to weaken the domain-invariant classifier. Specifically, we randomly choose to train the domain-invariant classifier using $N\%$ of the S_{NC} sets in each corpus, where N varies from 0 to 100 in increments of 10. The remaining data are allocated together with the S_C sets for emotion learning purposes only.

Figure 1 presents the trade-off trend between group fairness and individual fairness. We provide the performance of fairness metrics for the perceptually fair model (dotted gray line) as a reference, where the left y-axis indicates group fairness values and the right y-axis corresponds to individual fairness values. It is noticeable that individual and group fairness metrics cannot align closely on the vertical axis; the stronger the removal of group fairness constraints, the poorer the individual fairness metric tends to be. This is particularly evident when over 70% of the S_{NC} dataset is used to eliminate gender information, resulting in a significant drop in the individual fairness score for emotional outcomes from both datasets (a higher value represents fairer conditions). Conversely, the values for group fairness appear to converge towards an ideal state under such operations. Additionally, the extent to which individual fairness deteriorates is noteworthy, with some emotions even faring worse than the perceptually fair model, such as Anger in BIIC-Podcast and Happiness and Sadness in IEMOCAP. These findings are especially aligned with the rating differences illustrated in Table 1; instances where original individual rating differences are large exhibit exacerbated decreases in individual fairness evaluations upon the group fairness constraints.

Moreover, the effect on individual fairness metrics differs between IEMOCAP and BIIC-Podcast datasets. Specifically, the discrepancy in IEMOCAP generally remains below 4%, whereas in the BIIC-Podcast dataset, the difference can reach around 20%. This finding correlates with the analyses in Section 3, attributed to the larger base of raters in the BIIC-Podcast that introduces greater individual diversity, which in turn significantly affects individual fairness within the framework of conventional optimal group fairness mechanisms. Our results resonate with most prevailing fairness theories [15, 26], indicating that under current group fairness algorithmic approaches, achieving group fairness while disregarding individual fairness is common. This is particularly true in our study, as we confirm this trend extends to rating biases to further influence the SER model fairness.

4.3.2. Influence of WD on Individual Fairness

Minimizing the WD is considered a key criterion of group fairness, as reflected in past research which commonly employs WD as a metric or constraint for group fairness [27]. Our goal is to understand the impact on individual fairness while adhering to this distance criterion. The following experiment is designed in that we compute the WD for each pair of samples in S_{NC}, with the resulting histograms presented in Figure 2. Then we partition each dataset into four equal parts, effectively calculating the quartiles based on the data quantity: ① for the first 25%, ② for Q1 to Q2, ③ for Q2 to Q3, and ④ for the upper 25%. Following this division, we focus on minimizing the distance between the embeddings for each set of data while observing its effects on the individual fairness metrics. We test the individual fairness metrics on the outcomes.

Table 2 provides a summary of the experimental results. We observe that for data below Q3 (lower 75%) in IEMOCAP, maintaining group fairness typically does not significantly reduce individual fairness. However, in models trained on the top 25% of data which is characterized by larger distances, there is a clear deterioration in individual fairness. Similarly, this shift is noticed at Q2. These observations are consistent with existing studies suggesting that higher WD within the dataset often correlates with a decrease in individual fairness when group fairness is prioritized [26]. Furthermore, literature posits that individual fairness implies group fairness only with minimal WD between groups [27]. This is supported by similar patterns observed across both datasets, with relatively larger WD values in BIIC-Podcast leading to a sharp decline in individual fairness past the Q2. A trend is consistent with IEMOCAP, where individual fairness declines past the Q3.

5. Conclusion

In this work, we formulate two research questions to investigate *group fairness* versus *individual fairness* in perceptually fair SER system on two emotional corpora. Recognizing the presence of biases in both gender-based group and individual perspectives from our first analysis, we systematically study the impact of techniques designed to satisfy group fairness on individual fairness. Our analyses reveal interesting insights: 1) practical results from two datasets demonstrate a typical trade-off between group fairness and individual fairness, especially evident when group information is removed. 2) achieving group fairness can lead to diminished individual fairness when there are significant disparities in group attributes. By gaining a deeper understanding of the gap between group fairness and individual fairness, we can benefit from subsequent research and devise targeted strategies to achieve a fairer SER system.

6. Acknowledgements

This work was supported by the National Science and Technology Council (NSTC) under Grants 110-2221-E-007-067-MY3.

7. References

[1] S. Latif, A. Shahid, and J. Qadir, "Generative emotional ai for speech emotion recognition: The case for synthetic emotional speech augmentation," *Applied Acoustics*, vol. 210, p. 109425, 2023.

[2] C.-C. Lee, T. Chaspari, E. M. Provost, and S. 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] J. Kochan, "Subjectivity and emotion in scientific research," *Studies in History and Philosophy of Science Part A*, vol. 44, no. 3, pp. 354–362, 2013.

[4] S. G. Upadhyay, W.-S. Chien, B.-H. Su, and C.-C. Lee, "Learning with rater-expanded label space to improve speech emotion recognition," *IEEE Transactions on Affective Computing*, 2024.

[5] G. Chochlakis, G. Mahajan, S. Baruah, K. Burghardt, K. Lerman, and S. Narayanan, "Leveraging label correlations in a multi-label setting: A case study in emotion," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[6] L. Martinez-Lucas, A. Salman, S.-G. Leem, S. G. Upadhyay, C.-C. Lee, and C. Busso, "Analyzing the effect of affective priming on emotional annotations," in *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2023, pp. 1–8.

[7] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and machine learning: Limitations and opportunities*. MIT Press, 2023.

[8] W.-S. Chien and C.-C. 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.

[9] W.-S. Chien, S. G. Upadhyay, and C.-C. Lee, "Balancing speaker-rater fairness for gender-neutral speech emotion recognition," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 11 861–11 865.

[10] H. A. Elfenbein and N. Ambady, "On the universality and cultural specificity of emotion recognition: a meta-analysis." *Psychological bulletin*, vol. 128, no. 2, p. 203, 2002.

[11] A. Schirmer and R. Adolphs, "Emotion perception from face, voice, and touch: comparisons and convergence," *Trends in cognitive sciences*, vol. 21, no. 3, pp. 216–228, 2017.

[12] A. Batliner, S. Steidl, B. Schuller, D. Seppi, K. Laskowski, T. Vogt, L. Devillers, L. Vidrascu, N. Amir, L. Kessous *et al.*, "Combining efforts for improving automatic classification of emotional user states," pp. 240–245, 2006.

[13] I. Zliobaite, "A survey on measuring indirect discrimination in machine learning," *arXiv preprint arXiv:1511.00148*, 2015.

[14] S. Xu and T. Strohmer, "On the (in) compatibility between group fairness and individual fairness," *arXiv preprint arXiv:2401.07174*, 2024.

[15] R. Binns, "On the apparent conflict between individual and group fairness," in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020, pp. 514–524.

[16] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq, "Algorithmic decision making and the cost of fairness," in *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, 2017, pp. 797–806.

[17] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[18] S. G. Upadhyay, W.-S. Chien, B.-H. Su, L. Goncalves, Y.-T. Wu, A. N. Salman, C. Busso, and C.-C. 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.

[19] T. Räz, "Inter-rater reliability is individual fairness," *arXiv preprint arXiv:2308.05458*, 2023.

[20] J. L. Fleiss, "Measuring nominal scale agreement among many raters." *Psychological bulletin*, vol. 76, no. 5, p. 378, 1971.

[21] R. M. Kaplan, D. A. Chambers, and R. E. Glasgow, "Big data and large sample size: a cautionary note on the potential for bias," *Clinical and translational science*, vol. 7, no. 4, pp. 342–346, 2014.

[22] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. 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.

[23] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

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

[25] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *International conference on machine learning*. PMLR, 2013, pp. 325–333.

[26] W. Zhou, "Group vs. individual algorithmic fairness," Ph.D. dissertation, University of Southampton, 2022.

[27] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd innovations in theoretical computer science conference*, 2012, pp. 214–226.