Interspeech 2024
1-5 September 2024, Kos, Greece

An Inter-Speaker Fairness-Aware Speech Emotion Regression Framework

Hsing-Hang Chou, Woan-Shiuan Chien, Ya-Tse Wu, Chi-Chun Lee

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

Abstract

Speech emotion recognition (SER) helps to achieve better human-to-machine interactions in voice technologies. Recent studies have pointed out critical fairness issues in the SER. While there are efforts in building fair SER, most of the works focus on fairness between demographic groups and rely on these broad categorical attributes to build a fair SER. In this paper, we instead focus on the fairness learning among individual speakers, which is rarely discussed yet much more intuitively appealing in constructing a *fair* SER model. To reduce the reliance on knowing speaker IDs, we perform unsupervised clustering on the utterance embeddings from a pre-trained speaker verification model that puts utterances with different characteristics into clusters that roughly represent the true speaker index. Our evaluation demonstrates that with these cluster IDs, we can construct a fairness-aware SER model at an *individual speaker*-level without knowing speaker IDs upfront.

Index Terms: Some keywords **Index Terms**: speech emotion recognition, fairness, privacy

1. Introduction

Speech emotion recognition (SER) helps enable machines with emotional intelligence in voice technology [1]. One key factor is to add the human aspect of SER into the system, which makes machine-human interaction more relatable. Along with the development of SER and the involvement and integration of this system in our daily life and many decision-making processes, it is now becoming critical for advancing AI-based applications to ensure fairness [2]. A SER model is often constructed by learning on datasets composed with sliced or recorded audio and ground truth labels about emotion provided by human raters [3]. Learning on these speech samples generated by humans may overlook diversities, equality, and inclusion elements, affecting not just performances but also fairness [4]; for example, emotion attributes annotated by humans also cause bias like affect priming [5]. In consequence, fairness constraint is important when developing the SER model in the current era.

Numerous efforts have endeavored to address fairness concerns within SER. For instance, Gorrostieta et al. tackled gender bias in SER by implementing an adversarial invariant strategy and model penalization [6]. Wagner et al. [7] demonstrated the potential of transformer-based SSL in achieving comparable performance across gender groups. These works mainly mitigate bias stemming from differences in emotional expression between broad demographic groups categorized by attributes (e.g., gender). However, individual speakers exhibit distinct emotional expressions influenced by personal attributes shaped over time through self-awareness and socialization processes [8]. Addressing bias at the group level may neglect fairness at the individual level [7]. Thus, in the pursuit of a fair SER model, it's crucial to consider inter-speaker fairness specifically.

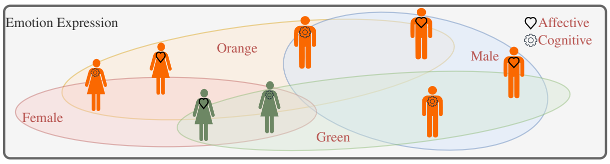


Figure 1: *Emotion expression not only differs between demographic groups but also differs at the individual level with different personal attributes*

While it's acknowledged that individuals express emotions differently, current research predominantly concentrates on enhancing overall performance by incorporating personal identifiers (e.g., speaker ID and representations) into SER models. For instance, Li et al. employed adversarial training to render model representation speaker-invariant [9]. Li et al. devised a graph integrating the speaker similarity between embeddings to capture inter-speaker relation [10]. However, pursuing performance improvement alone, devoid of explicit bias control, doesn't guarantee inter-speaker fairness.

Previous works to mitigate bias in SER predominantly aimed at achieving group fairness, whereas personalized SER that accounts for individual differences neglected the necessity for fairness. We argue that SER should encompass both aspects to ensure inter-speaker fairness. Moreover, discussions surrounding fairness in emotion regression are infrequent. To the best of our knowledge, our proposal marks the inaugural attempt to introduce direct inter-speaker fairness constraints in speech emotion regression.

In this paper, we introduce an SER model that prioritizes inter-speaker fairness. Moreover, in the realization of real-world application, we delve into automatic speaker clustering for datasets lacking explicit speaker identification. Initially, we leverage utterance-level embeddings to encode the distinctive features of each sample. Subsequently, we employ a clustering algorithm to organize these features into speaker clusters that roughly represent speaker ID. We then evaluate various fair SER models, each penalized for fairness concerning individual speakers, clusters, and gender, respectively. Additionally, we conduct comparative analyses between learning from an in-the-wild dataset, encompassing a diverse array of speakers, and a lab dataset containing both scripted and spontaneous recordings. Our results demonstrate that the proposed method enhances inter-speaker fairness while maintaining moderate performance levels. Importantly, it can be effectively generalized to in-the-wild datasets without prior knowledge of speaker IDs.

10.21437/Interspeech.2024-471

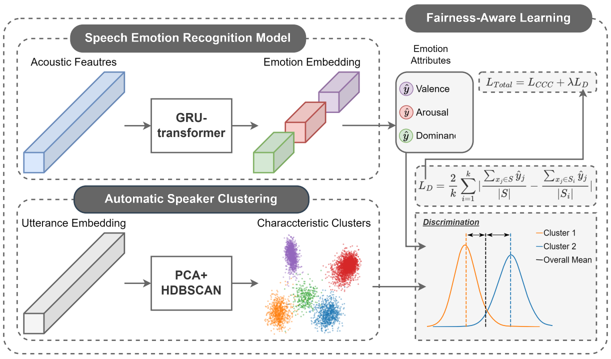


Figure 2: *Overview of the speech characteristic fairness aware speech emotion recognition (SER) architecture.*

2. Research Methodology

2.1. Dataset

2.1.1. MSP-Podcast

MSP-Podcast corpus version 1.10 [11] contains real podcast recordings (16kHz, 1ch) segmented in utterances. There are 104,267 utterances for a total of 106 hours in the dataset, including 1433 speakers. Each utterance is annotated with valence, arousal, and dominance as the emotion attributes by at least 5 annotators. A seven-point Likert scale is used to evaluate valence (very negative versus very positive), arousal (very calm versus very active), and dominance (very weak versus very strong). We follow the split provided by the authors for training, validation, and testing datasets, and the speakers in each set are independent of the others. In this research, we choose 515 labeled speakers with at least 32 utterances to avoid evaluating fairness among speakers with an insufficient amount of data to represent one speaker. The number of speakers and utterances used in this work is summarized in 1.

2.1.2. IEMOCAP

The IEMOCAP dataset [12] is a benchmark SER corpus with a gender balance (one male and one female) in each of its five dyadic spoken interaction sessions and results in 10 speakers with approximately equal number of utterances. There are 10039 utterances for a total of 12 hours in the dataset, and the emotion of the utterances is rated by six unique raters (two males and four females). The self-assessment manikins (SAMs) are used to evaluate the corpus in terms of the attributes valence [1-negative, 5-possitive], arousal [1-calm, 5-excited], and dominance [1-weak, 5-strong]. We split the utterances of each speaker with the ratio of 7:1:2 to form training, validation, and testing datasets. The number of speakers and utterances is summarized in 1.

2.2. Computational Framework

2.2.1. Speech Emotion Regression Model

To recognize the emotion attributes, we use an emotion recognition model with a structure similar to Wu et al. [13]. It includes HuBERT [14] and a GRU-transformer. The 768 dimensions acoustic features from HuBERT are first reduced to 16 dimensions by a linear fully connected layer, followed by stacking one GRU layer and one transformer encoder layer with two heads. Average pooling over time is applied before it passes to three independent two-stacked fully connected layers each ending with one unit corresponding to the prediction of arousal, valence, and dominance for learning emotion embedding that represents each emotion domain separately. Except for the last layer, all hidden dimensions of layers are consistent.

Table 1: *A summary of speakers and utterance distribution in two datasets*

	MSP-Podast			IEMOCAP		
	Train	Valid	Test	Train	Valid	Test
speaker	364	44	107	10	10	10
utterance	38513	10988	23809	7022	1004	2013

2.2.2. Fairness-Aware Learning

To mitigate the bias, we train the model by minimizing the following loss function with fairness constraint for each emotion attribute (i.e., arousal, valence, and dominance):

$$L_{\text{Total}} = L_{\text{CCC}} + \lambda L_{\text{D}}, \tag{1}$$

where L_{CCC} has been used in various works for regression task of SER [6, 7], which is the negative concordance correlation coefficient (CCC) as follows:

$$L_{\text{CCC}} = -\frac{2\rho\sigma_y\sigma_{\hat{y}}}{\sigma_y^2 + \sigma_{\hat{y}}^2(\mu_y - \mu_{\hat{y}})^2}, \tag{2}$$

where ρ is the Pearson correlation coefficient and σ_x and μ_x are the variance and the mean over a batch of training samples. L_{D} is the fairness criterion that is used to evaluate fairness and penalizes the model concerning bias between sampled groups with an equal number of samples over batch, with λ controlling the level of trade-off between performance and fairness. We discuss the fairness criterion in our interest in the following section.

2.2.3. Fairness Definition

In the regression problem of SER, there is currently no clear definition for measuring fairness. Gorrostieta et al. transferred the continuous arousal label into the binary label (calm and active) and focused on "Equality of odds", which requires the model prediction to have the same true positive rate, conditioned on the ground truth, for all elements of the protected attributes [6]. On the other hand, Wanger et al. evaluated the performance difference between male and female groups [7]. While most of these works focus on achieving fairness considering the ground truth label, model prediction should be fair among groups regardless of the ground truth label beforehand.

In consequence, we use the discrimination of the model defined by Zemel et al. [15], which measures the average difference between the average prediction for each attribute value as the fairness criterion. Since the original definition is limited to only binary attributes, we follow the extended version for k-way categorical attributes by Raft et al. [16], which is done by re-formulating discrimination to consider the sub-population differences from the global mean as follows:

$$Discrimination = \frac{2}{k}\sum_{i=1}^{k}|\frac{\sum_{x_j \in S}\hat{y}_j}{|S|} - \frac{\sum_{x_j \in S_i}\hat{y}_j}{|S_i|}| \tag{3}$$

where x_i and $\hat{y}_i$ are the input feature and the prediction of the model, and $S_i \in S$ are k attribute values.

Table 2: *A summary of experimental performance and fairness concerning different groups over emotion attributes V (valence), A (arousal), and D (dominance) in each testing set.*

MSP-Podcast	Performance (CCC)			Discrimination (speaker)			Discrimination (cluster)			Discrimination (gender)		
Emotion	V	A	D	V	A	D	V	A	D	V	A	D
human rated				0.1722	0.2497	0.1898	0.1511	0.2258	0.1729	0.0129	0.0255	0.0056
baseline	0.2594	**0.5112**	0.3857	0.2851	0.3444	0.3310	0.2643	0.3207	0.3098	**0.0089**	0.0607	**0.0231**
Fair$_{\text{gender}}$	0.2548	0.5093	0.3844	0.2831	0.3424	0.3312	0.2604	0.3190	0.3100	0.0215	**0.0571**	0.0265
Fair$_{\text{speaker}}$	**0.2664**	0.5019	**0.3891**	**0.1793**	**0.2532**	**0.2263**	**0.1595**	**0.2335**	**0.2085**	0.0284	0.0578	0.0267
Fair$_{\text{cluster}}$	0.2598	0.5012	0.3865	0.1896	0.2612	0.2358	0.1682	0.2417	0.2171	0.0273	0.0594	0.0288
IEMOCAP	Performance (CCC)			Discrimination (speaker)			Discrimination (cluster)			Discrimination (gender)		
Emotion	V	A	D	V	A	D	V	A	D	V	A	D
human rated				0.0640	0.0303	0.0999	0.0651	0.0965	0.1058	0.0248	0.0002	0.0026
baseline	**0.5505**	**0.6980**	**0.5572**	0.0925	0.0483	0.0776	0.1038	**0.1252**	**0.1248**	**0.0203**	**0.0017**	0.0050
Fair$_{\text{gender}}$	0.5492	0.6962	0.5558	0.1240	0.0727	0.1169	0.1485	0.1845	0.1837	0.0232	0.0042	0.0075
Fair$_{\text{speaker}}$	0.5266	0.6907	0.5405	**0.0680**	**0.0430**	**0.0659**	0.1196	0.1686	0.1604	0.0217	0.0019	0.0077
Fair$_{\text{cluster}}$	0.5265	0.6864	0.5379	0.0765	0.0613	0.0841	**0.0892**	0.1454	0.1371	0.0237	0.0021	**0.0038**

In this study, our primary aim is to address inter-speaker fairness, particularly concerning individuals exhibiting diverse emotional expressions, as discerned through speaker ID discrimination. However, in-the-wild datasets often lack comprehensive speaker information, including speaker IDs for all samples, and may contain insufficient data for certain speakers to adequately capture the breadth of emotional variability. To enhance the relevance of our approach in real-world settings, our objective is to derive speaker groups directly from speech signals, leveraging individual differences to approximate speaker ID representation while combining speakers with similar characteristics into larger groups.

2.2.4. Automatic Speaker Clustering

To delineate distinct speech characteristics, we aim to devise an unsupervised clustering procedure leveraging representative features extracted from each utterance. Initially, we harness the power of state-of-the-art ECAPA-TDNN speaker verification model [17], accessible via SpeechBrain [18], to derive utterance-level embeddings comprising 192 dimensions. The model is pre-trained on vast datasets encompassing over one million utterances from Voxceleb 1 [19] (featuring 1,211 speakers) and Voxceleb 2 [20] (encompassing 5,994 speakers), predominantly comprising English speakers from the US. Notably, this model attains a commendable 0.8% equal error rate on the Voxceleb 1 testing dataset, affirming its efficacy in capturing speaker characteristics.

Moreover, we employ a Principal Component Analysis (PCA) model for dimensionality reduction, retaining 80% of the original data's variance to mitigate computational costs and potential noise. Subsequently, the data undergo clustering utilizing the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) algorithm [21] without a predetermined number of clusters to form groups of utterances with similar characteristics. HDBSCAN is an unsupervised algorithm known for its ability to identify clusters of varying densities and handle noise robustly, particularly valuable in scenarios with imbalanced data distributions across speakers in in-the-wild datasets, potentially resulting in sparser distributions in the embedding space. This method effectively classifies certain points as noise data, which do not align with any clusters.

To leverage these noise data, we assign them labels based on the cluster exhibiting the highest cosine similarity between its centroid and the noise data. This clustering procedure is independently conducted on the training, validation, and testing datasets. Subsequently, armed with these characteristic clusters, we proceed to develop a SER model that conscientiously considers fairness concerning these pseudo-speaker ID clusters and improves inter-speaker fairness as well.

3. Experimental Setup and Results

3.1. Experimental Setup

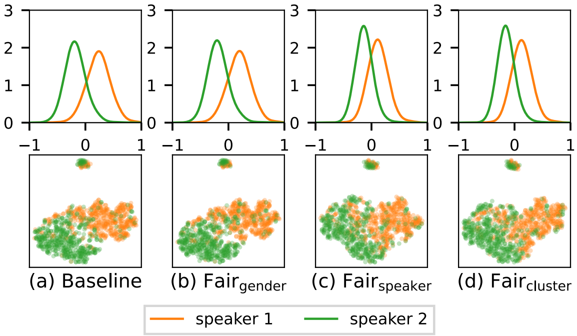
The experiment is run on the speech emotion regression over three emotion attributes (i.e., arousal, valence, and dominance) on MSP-Podcast and IEMOCAP. For both datasets, the emotion attributes of each utterance are derived by the average values among annotators and are scaled to $[-1, 1]$ to ensure the result is comparable between datasets. We experiment with four different SER models as follows:

- Baseline: SER model without any fairness constraint.
- Fair$_{\text{gender}}$: Fariness-Aware SER model respect to gender.
- Fair$_{\text{speaker}}$: Fariness-Aware SER model respect to speaker index.
- Fair$_{\text{cluster}}$: Fariness-Aware SER model respect to clusters.

We employ the CCC, with an ideal value of 1, to assess model performance. Fairness is evaluated using discrimination, where a lower value is desirable, ideally approaching 0. To gauge discrimination within the datasets, we incorporate the analysis of discrimination between ground truth labels annotated by humans. For each fair SER method, we utilize early stopping with patience of 20 epochs to identify the model exhibiting the lowest average discrimination of emotion attributes concerning their target groups on the validation dataset, ensuring that it maintains at least 90% of the best performance over epochs. Additionally, we determine the trade-off coefficient (λ) through linear searching within the range [0.05, 0.5] with a step size of 0.05, adhering to the same criteria.

All experiments are conducted using PyTorch 1.12.1 [22], with model parameters initialized using the default settings. Throughout the experiments, the decaying factor is set to 0.001, the dropout rate to 0.3, and the batch size remains fixed at 128. We employ the Adam optimizer with a learning rate of 1e-5 for parameter optimization, spanning 200 epochs. Training each method on an Nvidia GeForce RTX 1080Ti GPU typically requires 3 to 4 hours. Our implementation is publicly available on GitHub[1]. Detailed performance metrics and fairness evaluations of the test dataset are provided in Table 2.

[1] `https://github.com/HenryChou36/FairEmo`



(a) Baseline (b) Fair$_{\text{gender}}$ (c) Fair$_{\text{speaker}}$ (d) Fair$_{\text{cluster}}$

Figure 3: *First row is probability density distribution of valence, while the second row is the corresponding model representation*

3.2. Analysis of Recognition Results

The best λ are $\lambda = 0.05$ for Fair$_{\text{gender}}$, $\lambda = 0.5$ for Fair$_{\text{speaker}}$ and $\lambda = 0.45$ for Fair$_{\text{cluster}}$ respectively on MSP-Podcast dataset, and the best λ are $\lambda = 0.5$ for all fair SER model on IEMOCAP dataset. The Fair$_{\text{speaker}}$ model achieves the lowest discrimination on both datasets while retaining more than 95% overall performance compared to the baseline model. On the other hand, on MSP-Podcast, the Fair$_{\text{cluster}}$ model achieves comparable performance and discrimination to Fair$_{\text{speaker}}$. However, the model on IEMOCAP performs worse in discrimination than the baseline model for emotion attributes other than valence.

Comparing results across datasets, the model on IEMOCAP shows generally less discrimination than on MSP-Podcast, even with human-annotated attributes. This discrepancy likely arises from the imbalance and limited emotional variety for each demographic group and individual speakers in the in-the-wild dataset compared to the balanced scripted and improvised recordings in the static dataset. Consequently, learning from biased data increases discrimination.

Regarding gender discrimination, we observe that gender discrimination is significantly lower than discrimination against other groups across both datasets. The baseline model already achieves relatively fair results, with the Fair$_{\text{gender}}$ model only enhancing fairness for specific attributes. However, achieving gender fairness does not necessarily ensure fairness concerning individual speakers.

Regarding cluster discrimination, in the MSP-Podcast dataset, improving fairness for either individual speakers or clusters generally benefits both. However, this correlation does not extend to the results from the IEMOCAP dataset. Fair$_{\text{clusterer}}$ only improves the fairness of valence concerning clusters compared to the baseline. This suggests that efforts to enhance fairness concerning cluster groups may primarily improve fairness among speakers in models trained with in-the-wild datasets.

3.3. Ablation Study of De-bias Technique

To examine the effect of mitigating discrimination in our model, we analyze the probability density distribution of valence predictions and the corresponding 16-dimensional hidden layer representations before the fully connected layers. This involves two randomly selected speakers from the MSP-Podcast dataset, visualized using t-distributed Stochastic Neighbor Embedding (t-SNE) for nonlinear dimension reduction in Figure 3.

As Figure 3 shows, utterances of speaker 1 are concentrated in space that is separated from the speaker 2 for baseline and the Fair$_{\text{gender}}$ model. On the other hand, for the model with lower discrimination between speakers like Fair$_{\text{speaker}}$ and Fair$_{\text{cluster}}$, the utterances of speaker 1 are scattered and overlapped more with the other, and the probability density function is closer to each other. This shows that by lowering the discrimination among speakers, we also make it harder to identify specific speakers from model representation.

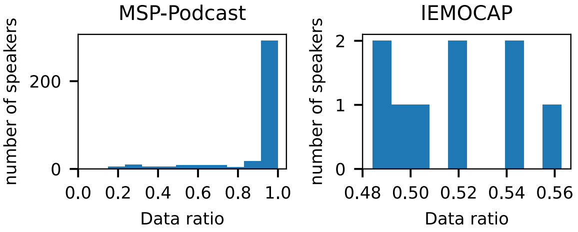


Figure 4: *Distribution of the data ratio of a speaker labeled with their majority vote cluster in training datasets.*

3.4. Analysis of Clusters

To figure out the differences in fairness results of Fair$_{\text{clusterer}}$ model between two datasets, we inspect the data ratio of each speaker labeled with the majority vote cluster that has the largest data ratio of a speaker among all the clusters they are labeled within the training datasets and summarize in Figure 4. For the MSP-Podcast dataset, above 80% of speakers have utterances being labeled with a single cluster for each of them. This shows that the clustering procedure roughly identifies each speaker as a whole based on the characteristics encoded by the utterance embedding. Thus, the Fair$_{\text{gender}}$ model that mitigates the bias concerning clusters can also reduce bias concerning individual speakers.

On the other hand, for the IEMOCAP dataset, only around 50% of the utterances of a speaker are in the same cluster, indicating that utterances of a speaker can be distributed in more than two clusters. Therefore with the clusters not corresponding to characteristics of speakers on the IEMOCAP dataset, constraining the Fair$_{\text{cluster}}$ model by fairness concerning it may not help in improving inter-speaker fairness.

This difference between datasets can also be found in other research [23] that tried to identify protected groups from correlated features. In our case, since information like emotion states can also be disentangled from the embedding space [24], the method can be less effective for static datasets with a smaller variety of speakers than those in-the-wild datasets.

4. Conclusion

In this work, we proposed a method to tackle inter-speaker fairness issues residing in the SER model and further generalized without speaker ID upfront through automatic speaker clustering. Our results and analyses show three insights: 1) The inter-speaker fairness issue depends on the target dataset. 2) Relation between inter-speaker fairness and model representation. 3) effectiveness of the de-bias technique and the clustering method. Our current work is limited to one fairness criterion in the SER model, and we plan to extend to other criteria and de-biased methods to provide more insight into inter-speaker fairness. We also want to explore the speaker clustering method to make the procedure not limited to in-the-wild datasets.

5. References

[1] B. W. Schuller, “Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends,” *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, 2018.

[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*, vol. 111, no. 10, pp. 1142–1158, 2023.

[3] 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 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2023, pp. 1–8.

[4] T. Verhoef and E. Fosch-Villaronga, “Towards affective computing that works for everyone,” in *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2023, pp. 1–8.

[5] 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)*, 2023, pp. 1–8.

[6] C. Gorrostieta, R. Lotfian, K. Taylor, R. Brutti, and J. Kane, “Gender De-Biasing in Speech Emotion Recognition,” in *Proc. Interspeech 2019*, 2019, pp. 2823–2827.

[7] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B. W. Schuller, “Dawn of the transformer era in speech emotion recognition: closing the valence gap,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[8] J.-L. Li and C.-C. Lee, “Attentive to Individual: A Multimodal Emotion Recognition Network with Personalized Attention Profile,” in *Proc. Interspeech 2019*, 2019, pp. 211–215.

[9] H. Li, M. Tu, J. Huang, S. Narayanan, and P. Georgiou, “Speaker-invariant affective representation learning via adversarial training,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7144–7148.

[10] J.-L. Li and C.-C. Lee, “Using Speaker-Aligned Graph Memory Block in Multimodally Attentive Emotion Recognition Network,” in *Proc. Interspeech 2020*, 2020, pp. 389–393.

[11] R. Lotfian and C. Busso, “Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings,” *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, October-December 2019.

[12] 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, pp. 335–359, 2008.

[13] Y.-T. Wu and C.-C. Lee, “MetricAug: A Distortion Metric-Lead Augmentation Strategy for Training Noise-Robust Speech Emotion Recognizer,” in *Proc. INTERSPEECH 2023*, 2023, pp. 3587–3591.

[14] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, “Hubert: Self-supervised speech representation learning by masked prediction of hidden units,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[15] 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.

[16] E. Raff, J. Sylvester, and S. Mills, “Fair forests: Regularized tree induction to minimize model bias,” in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018, pp. 243–250.

[17] B. Desplanques, J. Thienpondt, and K. Demuynck, “ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification,” in *Interspeech 2020*, H. Meng, B. Xu, and T. F. Zheng, Eds. ISCA, 2020, pp. 3830–3834.

[18] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, “SpeechBrain: A general-purpose speech toolkit,” 2021, arXiv:2106.04624.

[19] A. Nagrani, J. S. Chung, and A. Zisserman, “Voxceleb: A large-scale speaker identification dataset,” in *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*, F. Lacerda, Ed. ISCA, 2017, pp. 2616–2620. [Online]. Available: https://doi.org/10.21437/Interspeech.2017-950

[20] J. S. Chung, A. Nagrani, and A. Zisserman, “VoxCeleb2: Deep Speaker Recognition,” in *Proc. Interspeech 2018*, 2018, pp. 1086–1090.

[21] L. McInnes and J. Healy, “Accelerated hierarchical density based clustering,” in *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, 2017, pp. 33–42.

[22] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, “Pytorch: An imperative style, high-performance deep learning library,” *Advances in neural information processing systems*, vol. 32, 2019.

[23] P. Lahoti, A. Beutel, J. Chen, K. Lee, F. Prost, N. Thain, X. Wang, and E. Chi, “Fairness without demographics through adversarially reweighted learning,” *Advances in neural information processing systems*, vol. 33, pp. 728–740, 2020.

[24] J. Williams and S. King, “Disentangling Style Factors from Speaker Representations,” in *Proc. Interspeech 2019*, 2019, pp. 3945–3949.