INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Gender de-biasing in speech emotion recognition

Cristina Gorrostieta, Reza Lotfian, Kye Taylor, Richard Brutti, John Kane

Cogito Corporation
{cgorrostieta,rlotfian,ktaylor,rbrutti,jkane}
@cogitocorp.com

Abstract

Machine learning can unintentionally encode and amplify negative bias and stereotypes present in humans, be they conscious or unconscious. This has led to high-profile cases where machine learning systems have been found to exhibit bias towards gender, race, and ethnicity, among other demographic categories. Negative bias can be encoded in these algorithms based on: the representation of different population categories in the training data; bias arising from manual human labeling of these data; as well as modeling types and optimisation approaches. In this paper we assess the effect of gender bias in speech emotion recognition and find that emotional activation model accuracy is consistently lower for female compared to male audio samples. Further, we demonstrate that a fairer and more consistent model accuracy can be achieved by applying a simple de-biasing training technique.

1. Introduction

In recent times, machine learning via deep neural networks has led to rapid improvements in computer vision, language technologies, and medical applications, among others. In tandem with these advances has come an increased focus on bias and ethics in machine learning [1, 2]. In machine learning, discrimination itself is not always negative: in fact, the objective of training these algorithms is to find patterns and characteristics in the data that allow you to effectively separate the categories you care about (e.g., recognising classes of emotion). However, this should not be at the cost of introducing negative bias.

There have been some high profile cases of bias including computer vision systems reportedly having a higher error rate for recognising dark skinned women [3]. Other research has found that Black and Hispanic borrowers were treated far less favourably than white borrowers by new credit worthiness technology [4, 5].

Language technology has also come under considerable scrutiny for gender bias. For many modern natural language processing models, words are converted to numerical representations known as word-embeddings. The embeddings encode semantic information by assessing the context (or neighbourhood) in which a word occurs. These representations result in words with similar semantics being close to each other in a mathematical sense and enable assessment of the relationships between them. For instance, you can easily use word-embeddings to predict the word "queen" from an input "man is to king, as woman is to?". However, a 2016 NeurIPS paper [6] found that such encodings led to biased and sexist output. For example, the input "man is to computer programmer as woman is to?" resulted in the output "home-maker". Recent studies show that word-embeddings capture common stereotypes because these implicit biases typically exist in large text databases [7, 8]. Discrimination has also been reported in systems advertisement delivery [9], object classification [10] and image search results for occupations [11] .

There are several factors which can contribute to producing negative bias in machine learning models. One major cause is incomplete or skewed training data. If certain demographic categories are missing from the training data, models developed on this can fail to generalise when applied to new data containing those missing categories. The majority of models deployed in modern technology applications are based on supervised machine learning and much of the labeled data comes from people. Labels can include movie reviews, hotel ratings, image descriptions, audio transcriptions, and perceptual ratings of emotion. As people are inherently biased, and because models are an estimate of people's impressions, it follows that this bias will be implicitly encoded into these algorithms. As a result, there is the real risk that these systems can inadvertently perpetuate or even amplify bias contained in the label data [12].

Given a large, appropriately sampled dataset and labels with a minimal degree of bias, there is still the risk of introducing bias from the features and modeling techniques used. This problem is not new and in fact has been present in the area of speech technology for many decades. Until recently, speech synthesis and speech recognition favoured lower pitch voices, typically present in adult males. As a result, speech recognisers produced higher word-error rate scores for children and adult females [13]. One reason for this is fact that early speech analysis and measurement techniques were developed mainly with low-pitched male voices in mind. But such acoustic characterisation of speech production did not sufficiently generalise to voices with a higher pitch, where the spacing of the harmonics in the audio spectrum is much wider and the assumptions of a linear time-invariant vocal system become seriously undermined.

Few studies have assessed the affect of gender bias in speech with respect to emotion. [14] assessed the perception of vocalisations and did not find any effects of negative bias towards gender. However, the stimuli consisted of acted speech samples and so the findings may not generalise to natural conversational speech. A later study [15] did in fact observe different identification accuracy rates in the perception of emotion for male compared with female speakers. Some recent research has approached the problem of mitigating negative bias in machine learning generally, with [16] proposing an adversarial training technique to attenuate different formulations of bias. A related paper [17] demonstrated how autoencoder models can be used to learn representations free of "nuisance variables" (e.g., gender) by incorporating orthogonal components to the network. Several publications have investigated gender de-biasing in natural language processing (see for instance [10, 18]).

Although some previous research has investigated modeling emotion and gender separately [19, 20], to the best of our knowledge the present paper is the first to investigate gender de-biasing directly in speech emotion recognition, and its main

http://dx.doi.org/10.21437/Interspeech.2019-1708

contributions are as follows:

1. We assess the effect of gender bias in a standard speech emotion recognition model on a large, naturalistic and well-known dataset
2. We evaluate the efficacy of two de-biasing training techniques to mitigate the effect of gender bias
3. We highlight challenges associated with training with de-biasing

2. De-biasing emotion recognition models

2.1. Defining fairness

Before experimenting with techniques to mitigate the effects of unwanted bias, it is important to start by quantifying fairness. We adopt the definitions of fairness with respect to machine learning models outlined in [5] and [21]. The definitions we use are described below. We use the term "protected variable" to refer to some categorisation of the population (e.g., age, gender, demographics, ethnicity etc.).

"Demographic parity" requires that the prediction of a model be independent of the protected variable (e.g., gender). In [5] the authors give the example of a loan application and state that for demographic parity to exist in this case the decision on whether or not to grant a loan should have no correlation with being a member of certain protected variable class (e.g., being of some specific ethnicity). [5], however, highlights major flaws with this particular definition of fairness. Taking again the loan example. forcing demographic parity can result in applications being inappropriately accepted for certain classes of the protected variable.

Another definition of fairness is "Equality of odds". In order for this definition to hold, a model prediction (for example loan application acceptance) should have the same true positive rate, with respect to the ground truth, for all elements of the protected variable. So if the protected variable is gender, and we make the assumption that gender is strictly binary (which admittedly may not be an entirely valid assumption), the model true positive rate for males is the same as that for females. Adopting this approach forces modeling techniques away from just being accurate for a majority class or a typically more favourably treated class.

In the present study we focus on "Equality of odds", and this condition is held if for all values of the true label Y (in our case the binarised emotional activation label) the probability of the predicted label, $\hat{Y}$, being a particular value is the same for all values of the gender protected variable Z, i.e.

$$P(\hat{Y} = \hat{y}|Y = y) = P(\hat{Y} = \hat{y}|Z = z, Y = y) \quad (1)$$

2.2. De-biasing

In order to mitigate unwanted bias, we investigate two different techniques.

The first is an an adversarial learning approach, similar to that described in [16], with the objective of achieving "Equality of odds". The technique (inspired by the adversarial training of generative models proposed by [22]), in summary, looks to jointly minimise the primary objective (i.e. reduce error between the model prediction and the groundtruth) and maximise a metric which captures "Equality of odds". This technique can be applied to any machine learning model and to any protected variable (i.e. any category in the population which is at risk for negative bias).

The approach involves jointly training two models: a regression predictor and an adversary. The adversary is a low-complexity model which takes the continuous scalar output of the predictor and the binary label variable as inputs and is trained to optimally classify the binary protected variable. In this study we need to binarise the continuous emotional activation label in order to be compatible with this adversarial technique for achieving "Equality of odds". More formally, the adversary can be written as:

$$\hat{z} = w_2\big[s, sy_{bin}, s(1 - y_{bin})\big] + b \quad s = \sigma(1 + |c|)\hat{y} \quad (2)$$

where $\hat{y}$ is the regression predictor output prior to any transformation, c and b learnable scalars, w_2 a learnable vector, $\sigma()$ indicates the sigmoid activation function, and y_{bin} is the binarised activation label:

$$y_{bin} = \begin{cases} 0, & \text{if } y \leq \theta \\ 1, & \text{otherwise} \end{cases} \quad (3)$$

with θ being a constant threshold (set in this study to 3.5, which is the centre of the target variable's range). The adversary model is trained to minimise the loss, L_A.

The regression predictor model is trained to predict y given an input feature matrix, x. The loss term associated with predicting the primary emotional activation target, y, is L_P. The model is trained by optimizing its prediction of y while simultaneously being penalised as the adversary gets better at predicting the protected variable. The key idea here is to avoid transferring any information which is useful for predicting the protected variable beyond what is already provided in the binarised emotional activation ground truth. Hence, the loss function is a weighted combination of L_P and L_A and the loss function is updated with the following gradient:

$$\nabla_W L_P - \alpha_0 \nabla_W L_A \quad (4)$$

where W is the set of model predictor weights and α_0 a is hyperparameter to be tuned during training. As is noted in [16], moving in the direction of this gradient could actually help the adversary, especially for cases when $\nabla_W L_P$ and $\nabla_W L_A$ are correlated. To avoid this they suggest modifying the gradient by including a projection term, $\text{proj}_{\nabla_W L_A} \nabla_W L_P$ to modify parameters according to the expression

$$\nabla_W L_P - \text{proj}_{\nabla_W L_A} \nabla_W L_P - \alpha_0 \nabla_W L_A \quad (5)$$

We adopt this modified gradient approach for optimizing the regression predictor.

The second de-biasing technique does not apply adversarial training but rather solely focuses on training a regression predictor model. Training involves minimising a loss function, L_N:

$$L_N = L_P + \alpha_1 L_{eq} \quad (6)$$

which additionally includes a weighted term, L_{eq}, which penalises the model for producing inconsistency in recall across classes of the protected variable. This additional term can be written as:

$$\begin{aligned} L_{eq} = &|\text{r}_{0,m}(\hat{y}_{bin}, y_{bin}) - \text{r}_{0,f}(\hat{y}_{bin}, y_{bin})| + \\ &|\text{r}_{1,m}(\hat{y}_{bin}, y_{bin}) - \text{r}_{1,f}(\hat{y}_{bin}, y_{bin})| \end{aligned} \quad (7)$$

where $r_{a,g}()$ is a function which computes recall for activation class, a (low is 0, high is 1), and gender class, g (f is female, m is male), with α_1 being a constant weight. The binary variable, $\hat{y}_{bin}$, is determined by thresholding the regression predictor output, $\hat{y}$, in the same way as in Equation 3.

3. Experimental protocol

3.1. Audio data and labels

Our research is concerned with automatically recognising naturally occurring emotion in conversations, rather than acted or feigned speech productions. As a result, our experimental evaluation is conducted on the version 1.3 of the MSP-Podcast corpus [23] (see also [24]). The version 1.3 of the corpus contains 33,439 utterances collected from podcasts available in audio-sharing websites. These naturalistic conversations are segmented into speaking turns with duration between 2.7 sec and 11 sec. The segments from the podcasts are annotated with emotional labels using a crowdsourcing framework which enables access to a diverse pool of annotators. The utterances are annotated by at least five annotators for the emotional attributes (activation/arousal, valence and dominance) and categorical emotions (e.g., happy, sad, angry). For 28,266 utterances, the dataset provides the identity and gender of the speakers which are collected by searching the names of the speakers either mentioned in the podcast recording or shared as part of the description of the podcast. The dataset is split into train, validation and test sets, with speaker independent partitions. The number of utterances is 15,132, 4,098 and 9,036, split over training, validation and test sets, respectively.

Note that we treat the protected variable, gender, as being binary in this study, despite this characterisation not being entirely valid. As our application interest is in processing call centre audio, we downsample audio to 8 kHz with16-bit precision.

Our modeling target for this study is continuous emotional activation and this target is created by averaging the activation label over multiple annotators who rated their perception of activation on a 7 point Likert scale. We also use a binarised version of this target variable, achieved by thresholding the continuous variable at 3.5, to be used with the gender de-biasing training techniques. In terms of the distribution of the protected variable there are overall 38 % female samples, with 18 % female in the training partition, 47 % in validation and 67 % in test.

3.2. Features and model architecture

Given that our objectives with this research are to assess gender bias and to mitigate its effects, we use a fairly standard speech emotion recognition modeling approach, with a similar architecture to some of the models described in [25]. For features, we use Mel frequency Cepstral Coefficients (MFCCs), which are derived by computing the magnitude spectrum from Hamming windowed 40 ms frames of audio (with 16 ms shift), and then applying a set of 40 triangular filters linearly spaced on the Mel scale. We then apply a discrete cosine transform (DCT) and energy scaling. We use the first 24 coefficients as features, excluding the 0th. Although feature decorrelation is not required for the modeling approach we use, it does produces energy independent features which can help with generalisability to new data recorded in different acoustic environments.

Z-normalisation is applied to the 24 features, using mean and standard deviation statistics computed on the training set only. We then create frames of roughly 4.5 seconds (279 time steps), which are computed every 160 ms on the input audio data.

24x279 matrices are then fed into a neural network model consisting of sequences of 2-dimensional convolutional layers, with rectified linear (ReLU) activation function. The output of the final convolutional layer is flattened to a 1-dimensional vector which is then fed into a sequence of fully-connected layers again with ReLU activation before a final linear output layer.

This predictor regression model output together with the binary groundtruth label make up the 1-dimensional, 2-element vector fed into the adversarial model (see Equation 2).

3.3. Model training procedure

Our experimental procedure proceeds in two phases. First we look to find the training settings and model hyperparameters that achieve good accuracy on the validation set. We evaluate results based on our primary accuracy metric, CCC, and in terms of metrics related to "Equality of odds" with respect to gender. We quantify this by measuring the difference in true positive rate (TPR) for low and high activation classes, and further by summing those two quantities (note absolute difference is not used):

$$\begin{aligned} \text{TPR-low}_{diff} &= \text{r}_{0,f}(\hat{y}, y) - \text{r}_{0,m}(\hat{y}, y) \\ \text{TPR-high}_{diff} &= \text{r}_{1,f}(\hat{y}, y) - \text{r}_{1,m}(\hat{y}, y) \\ \text{TPR}_{diff} &= \text{TPR-low}_{diff} + \text{TPR-high}_{diff} \end{aligned} \tag{8}$$

We also compute the across-gender difference in CCC (CCC_{diff}). For both evaluation metrics negative values indicate poorer results for female samples compared to male.

The first phase is executed without applying the any debiasing techniques. For the second phase we use the best settings from the first, and we look to maintain the primary accuracy level but better achieve "Equality of odds", by applying the debiasing methods described in Section 2.2. We search over settings including the learning rate of the optimisation algorithm applied to the adversarial model and the weighting of the adversarial term, α_0 in equation 5.

We define the L_P loss function as:

$$L_P = -\frac{2\rho\sigma_y\sigma_{\hat{y}}}{\sigma_y^2 + \sigma_{\hat{y}}^2(\mu_y - \mu_{\hat{y}})^2} \tag{9}$$

which is the negative CCC, with ρ being the Pearson correlation coefficient, and σ_x and μ_x being variance and mean statistics computed over a batch of training samples. We define the adversary loss function, L_A, as:

$$L_A = -\frac{1}{N}\sum_{n=0}^{N-1}\sum_{i=0}^{K-1} z_i^{(n)} log(\hat{z}_i^{(n)}) \tag{10}$$

which is the cross entropy between the adversary model prediction, $\hat{z}$, and the ground truth gender label, z, and K is the number of output elements (in our case 2) averaged over a batch of N training samples, and where the superscript (n) is used to indicate index of the sample in the batch.

For the adversarial de-biasing method two optimisers are applied, one for the regression predictor (with gradients as per Equation 5) and another for the adversary (which seeks to minimise L_A), both with different learning rates. For the non-adversarial method, a single optimiser is used to minimise L_N (see Equation 6). The Adam optimization algorithm is used throughout [26].

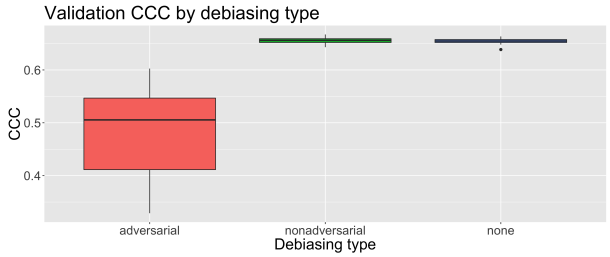

Figure 1: *Validation set CCC plotted as a function of de-biasing approach*

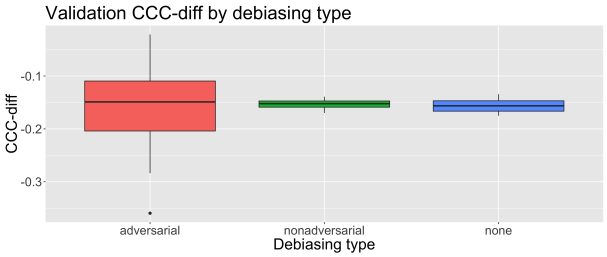

Figure 2: *Validation set CCC-diff plotted as a function of de-biasing approach*

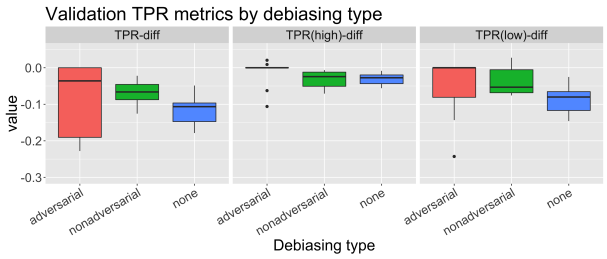

Figure 3: *Validation set TPR metrics plotted as a function of de-biasing approach*

We train on hardware with a single GPU and four CPUs. TensorFlow v1.12 is used as the machine learning framework and the dataset API settings were carefully chosen (as per https://www.tensorflow.org/guide/performance/datasets) to effectively utilise GPU and CPUs and ensure experimental efficiency. All stochastic components of the models (e.g., weight initialisation, dropout randomisation) as well as training (e.g., dataset shuffling) are seeded to ensure each individual experiment is fully reproducible. Once we have determined the optimal hyperparameters and settings for a particular modeling approach, we then re-run training using 10 different random seeds to produce a distribution of metrics for the validation set. Then only the best single selected model from each variant is subjected to the test set to ensure that a fair assessment of model generalisability.

4. Results

Following the first phase of experimentation we find best validation set accuracy (in terms of CCC) when using two layers of 2-dimensional convolutions (with 16 then 32 filters, 4x4 stride, and 8x8 filter size) without max-pooling followed by two layers of fully connected layers, each with 128 hidden units. Batch normalisation is not found to help, dropout keep probability is selected as 0.7, and a training batch size of 200 samples produced the highest CCC. The distribution of CCC on the validation (by varying the random seed) can be seen in Figure 1. The best performing model on the test set achieves 0.684 CCC (see Table 1) which is marginally lower than the best reported results on this dataset [24].

Considering the two de-biasing techniques, we see in Figure 1 that whereas the non-adversarial approach maintains around the same levels of CCC, the adversarial approach consistently produces lower accuracy. Additionally, it introduces considerable instability in the training and we observe high sensitivity to settings like the adversarial learning rate and the α_0 weight. This instability is apparent in Figure 1 where even variation in the random seed produces a wide variance in CCC.

From Figures 2 and 3 it can be seen that the adversarial training approach has the potential to achieve "Equality of odds" (as indicated by the TPR metrics) and more consistent accuracy across the two gender classes (as indicated by CCC-diff). However, the distributions of these metrics on the validation are large making model selection precarious. The non-adversarial approach, on the other hand, achieves much better consistency with all metrics and produces substantially better TPR-diff than the approach without de-biasing on the validation set.

The test set results (see Table 1) highlight the challenge of using the adversarial approach where although substantial improvements in across-gender consistency can be achieved this is at the risk of a significant reduction in overall accuracy. The simple non-adversarial approach produces improvements in fairness with just a marginal reduction in overall accuracy.

Table 1: *Test set results with a single model for each de-biasing variant chosen based on performance on the validation set*

De-biasing	CCC	CCC-diff	TPR-diff
None	0.684	-0.234	-0.067
Adversarial	0.518	0.072	0.038
Non-adversarial	0.661	-0.171	-0.016

5. Discussion & conclusion

Emotional activation models produce favourable accuracy for male samples compared to female throughout our experimentation, including in the many experiments that we ran which are not reported here. Although it is not completely clear why this is, a likely factor is the underrepresentation of female samples, in particular in the training partition. Another factor may be that the features, here MFCCs, are better configured for representing spectra with denser harmonic components associated with lower pitch voices. The two de-biasing approaches displayed qualities of attenuating this bias, however the adversarial technique introduces instability into model training, and better overall results are observed with the non-adversarial approach.

There remains considerable outstanding research needed to mitigate negative bias in machine learning generally but also in speech emotion recognition specifically. In this paper we focus on the modeling approach and optimisation techniques as well as sampling bias. The present research did not, for instance, assess the potential for negative bias to be introduced during annotation of the audio data. Our continuing research in this area will look to define a holistic approach for machine learning and emotion recognition, and provide policies and procedures to detect and reduce the sources and effects of negative bias.

6. References

[1] Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., Floridi. L., "The ethics of algorithms: Mapping the debate" *Big Data and Society*,2(2), 1-21, 2016.

[2] Char DS, Shah NH, Magnus D. "Implementing Machine Learning in Health Care - Addressing Ethical Challenges" *The New England Journal of Medicine*, 378(11), 981-983, 2018.

[3] Buolamwini, J., Gebru, T., "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification", *Proceedings of Machine Learning Research*, 81, 1-15, 2018.

[4] Corbett-Davies, S., Goel S., "The measure and mismeasure of fairness: A critical review of fair machine learning", *arXiv preprint arXiv:1808.00023*, 2018.

[5] Hardt, M., Price, E., Srebro, N. "Equality of Opportunity in Supervised Learning", *Advances in Neural Information Processing Systems*, 29, 3315-3323, 2016.

[6] Bolukbasi, T., Chang, K-W, Zou, J. Y., Saligrama V., Kalai, A. T, "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings", *Advances in Neural Information Processing Systems*, 29, 4349-4357, 2016.

[7] Caliskan, A., Bryson, J., Narayanan, A., "Semantics derived automatically from language corpora contain human-like biases" *American Association for the Advancement of Science*, 356(6334), 183-186, 2017.

[8] Garg, N., Schiebinger, L., Jurafsky, Dan., Zou, J., "Word embeddings quantify 100 years of gender and ethnic stereotypes" *Proceedings of the National Academy of Sciences*, 115(16), 3635-3644, 2018.

[9] Sweeney, L., "Discrimination in Online Ad Delivery" *Queue*, 11(3), 10-20, 2013.

[10] Zhao, J., Wang, T., Yatskar, M., Ordonez, V., Chang, K-W., "Men also like shopping: Reducing gender bias amplification using corpus-level constraints", *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2979-2989, 2017

[11] Kay, M., Matuszek, C., Munson, S., "Unequal Representation and Gender Stereotypes in Image Search Results for Occupations" *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 3819-3828, 2015.

[12] Vallor, S., "Artificial intelligence and public trust", *Santa Clara Magazine*, 58, 42-45, 2017.

[13] Das, S., Nix, D., Picheny, M., "Improvements in children speech recognition performance", *Proceedings of ICASSP*, 1, 433-436, 1998.

[14] Young, K. S., Parsons, C. E., LeBeau, R. T., Tabak, B. A., Sewart, A. R., Stein, A., Kringelbach, M. L., Craske, M. G., "Sensing Emotion in Voices: Negativity Bias and Gender Differences in a Validation Study of the Oxford Vocal ("OxVoc") Sounds Database", *Psychological Assessment*, 29(8), 967-977, 2017.

[15] Lausen, A., Schacht, A., "Gender differences in recognition of vocal emotions", *Frontiers in Psychology*, 9(882), 2018.

[16] Zhang, B. H., Lemoine, B., Mitchell, M., "Mitigating Unwanted Biases with Adversarial Learning", *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 335-340, 2018.

[17] Liu, L., Ghosh, S., Scherer, S., "Towards Learning Nuisance-Free Representations of Speech", *Proceedings of ICASSP*, 2018

[18] Zhao, J., Wang, T., Yatskar, M., Ordonez, V., Chang, K-W., "Gender bias in coreference resolution: Evaluation and debiasing methods", *Proceedings of NAACL*, 2018

[19] C., Myungsu, Kim, T-H., Shin, Y. H., Kim, J-W., Lee, S-Y., "End-to-end multimodal emotion and gender recognition with dynamic joint loss weights", *arXiv preprint arXiv:1809.00758*, 2018

[20] Wang, Z-Q., Tashev, I., "Learning utterance-level representations for speech emotion and age/gender recognition using deep neural networks,", *Proceedings of ICASSP*, 2017

[21] Beutel, A., Chen, J., Zhao, Z., Chi, E. H. "Data Decisions and Theoretical Implications when Adversarially Learning Fair Representations", *arXiv preprint arXiv:1707.00075*, 2017.

[22] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., "Generative adversarial nets', *Advances in neural information processing systems*, 2672-2680, 2014.

[23] Lotfian, R., Busso, C., "Building Naturalistic Emotionally Balanced Speech Corpus by Retrieving Emotional Speech From Existing Podcast Recordings", *IEEE Transactions on Affective Computing*, 3698-3702, 2017.

[24] Parthasarathy, S., Busso, C., "Ladder Networks for Emotion Recognition: Using Unsupervised Auxiliary Tasks to Improve Predictions of Emotional Attributes", *Proceedings of Interspeech*, 3698-3702, 2018.

[25] Fayek, H. M., Lech, M., Cavedon, L., "Evaluating deep learning architectures for Speech Emotion Recognition", *Neural Networks*, 92, 60-68, 2017.

[26] Kingma, D., Ba, J., "Adam: A method for stochastic optimization", *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.