Interspeech 2022
18-22 September 2022, Incheon, Korea

Speech Emotion Recognition in the Wild using Multi-task and Adversarial Learning

Jack Parry, Eric DeMattos, Anita Klementiev, Axel Ind, Daniela Morse-Kopp, Georgia Clarke and Dimitri Palaz

Speech Graphics Ltd, Edinburgh, United Kingdom

{jack.parry, eric.demattos, anita.klementiev, axel.ind, daniela, g.clarke, dpalaz}@speech-graphics.com

Abstract

Speech Emotion Recognition (SER) is an important and challenging task, especially when deploying systems in the wild i.e. on unseen data, as they tend to generalise poorly. One promising approach to improve the generalisation capabilities of SER systems is to incorporate attributes of the speech signal, such as corpus or speaker information, which can be a source of overfitting or confusion for the model. In this paper, we investigate using multi-task learning, where attribute prediction is given as an auxiliary task to the model, and adversarial learning, where the model is explicitly trained to incorrectly predict attributes. We compare two adversarial learning approaches: gradient reversal and an adversarial discriminator. We evaluate these approaches in a cross-corpus training setting using two unseen corpora as test sets. We use four attributes – corpus, speaker, gender and language – and evaluate all possible combinations of these attributes. We show that both multi-task learning and adversarial learning improve SER performance in the wild, with the gradient reversal approach being the most consistent across attributes and test sets.

Index Terms: speech emotion recognition, multi-task learning, adversarial learning

1. Introduction

Speech Emotion Recognition (SER) has seen a growing number of applications in recent years. For such applications to be successful, SER models must perform well *in the wild*, i.e. they need to generalise well to unseen data with varying characteristics. In recent years, deep learning approaches have been shown to yield significant performance improvements compared to traditional approaches on the SER task [1], however, they still seem to generalise poorly on unseen data when trained on one corpus as standard [2]. To address this issue, one promising approach may be to force the model to be independent of some often overlooked *attributes* of the speech signal. We define an attribute as a characteristic of speech that has an impact on the acoustic characteristics of the signal. These may be problematic when training an SER model for one of two reasons: the attribute could be a source of overfitting, because there is not enough variation in the training data; or it could be a source of confusion, as the acoustic characteristics associated with different attributes can be similar to those associated with different emotions. In this paper we focus on four attributes: corpus, gender, speaker and language. The corpus attribute is a major source of overfitting [3, 2] as recording conditions typically do not vary within a given corpus. The gender attribute, which here refers to the average pitch of the speaker, may also lead to overfitting as the model could learn frequency-specific filters. The speaker attribute may be a source of confusion for the model as prosody varies widely between speakers but is also an important indicator of emotion. Finally, the language could be a further source of overfitting as this attribute encompasses cultural and linguistic information relevant for detecting emotions [4].

To address these issues, several approaches have been developed including cross-corpus training, multi-task learning (MTL) and adversarial learning (AL). Cross-corpus training consists of aggregating several corpora with the goal of creating a diverse training set comprising a variety of attributes and thereby mitigating overfitting. This approach was initially shown to display poor generalisation capabilities on out-of-domain data [3], but recent works have shown more promising results [2, 5]. The MTL approach [6] uses attributes as auxiliary labels and has been shown to improve SER performance using attributes like gender [7] and corpus [8]. The AL approach, popularised by the GAN [9], also incorporates auxiliary labels, but unlike MTL the model is trained to *incorrectly* classify these labels, with the aim of being more independent of these tasks. This approach has been implemented for domain transfer [10] and successfully applied to SER [11, 8]. It is however not yet clear if the attribute-based methods are helpful to the cross-corpus training approach and which attributes are the most useful.

In this paper, we present a study on the potential of multi-task and adversarial learning approaches to improve the generalisation capabilities of speech emotion recognition models. We combine seven corpora to train a widely used CNN-LSTM model in a cross-corpus training setting, which we then evaluate using two out-of-domain corpora and an in-domain test set. We trained two different AL models, gradient reversal (GR) and adversarial discriminator (AD), which are compared with the MTL approach and a single-task baseline. The four attributes – corpus, gender, speaker and language – are first studied separately, before being combined in multi-attribute models. We show that the three approaches using these attributes can improve performance on out-of-domain data compared to a single-task model. We show that the GR approach achieves the best performance and is the most consistent across attributes and corpora, while the MTL and AD approaches can be beneficial but are inconsistent.

The key contributions of this paper are: (1) we present a comprehensive study of attributes as auxiliary tasks in a cross-corpus setting for out-of-domain generalisation of which 9 out of 15 combinations are novel to the best of our knowledge; (2) we compare three attribute-based training approaches and show that they can be beneficial for SER in the wild, with the GR approach being the most promising; and (3) we provide insight into the differences between representations learned by MTL and AL models.

10.21437/Interspeech.2022-10581

Table 1: *Summary of MTL and AL approaches using the (C)orpus, (G)ender, (S)peaker and (L)anguage attributes. ID refers to in-domain and OOD to out-of-domain. The last two rows of the table correspond to the scope of this paper.*

	Attribute			
Training setup	C	G	S	L
Single corpus	N/A	[21, 18, 28, 7]	[21]	–
Cross-corpus				
↳ ID eval.	[8]	[26]	[26, 22]	[23]
↳ OOD eval. (*this paper*)				
↳ MTL	[25]	[19, 25, 26, 20]	[26]	[20]
↳ AL	[11]	–	–	–

2. Related Work

In this section, we present the related work for cross-corpus SER, and MTL and AL approaches. For more details on other approaches in SER see [1, 12].

2.1. Cross-corpus Speech Emotion Recognition

The cross-corpus problem in SER refers to the performance mismatch between corpora, as models trained on a single corpus tend to overfit [13]. Alongside cross-corpus training [2, 5], other approaches have been proposed to combat this problem such as speaker-based z-normalisation [3]. Data augmentation approaches have also been used, for example using a CycleGAN [14]. Furthermore, [15] explored unsupervised domain adaptation. Adversarial approaches using cross-corpus training have also been proposed for domain shift adaptation [16] and learning common representations between corpora [17].

2.2. Attribute-based Multi-task and Adversarial Learning

MTL has been applied to SER with a variety of attributes as auxiliary tasks. Adding the auxiliary task of gender recognition has been shown to improve models trained and evaluated on a single corpus [7, 18] and on out-of-domain data using cross-corpus training [19, 20]. Speaker recognition has also been used on single corpus [21] and cross-corpus training [22]. Language recognition has also been explored in a cross-corpus training setting [23, 20], where this attribute has not been proven to be beneficial. Discrimination between acted and spontaneous speech has been shown to improve SER [24]. [25] studied the influence of corpus and mode (speech or singing) in addition to gender using cross-corpus training. The adversarial autoencoder approach was also incorporated into an MTL framework using gender and speaker as attributes to help learn discriminative features [26].

Attribute-based AL was originally developed for the corpus attribute by Ganin et al. [10]. This approach was applied to SER using cross-corpus training in the context of domain adaptation [11] and corpus-independent emotion encoding [8]. It was also investigated in the context of mitigating biases in machine learning models [27] and applied to SER for gender de-biasing [28], where it was shown that gender bias can be mitigated at the cost of lower performance on the task at hand.

Compared to this paper, most of these works focus on in-domain evaluation and do not provide an evaluation on out-of-domain data. Table 1 provides a summary of the above related works with regards to the attributes used, the training and evaluation setup and the type of approach.

3. Methodology

3.1. Multi-task Learning

The MTL model consists of three parts: the shared layers θ_f, the emotion classifier layers θ_{emo} and the attribute classifier layers θ_{a_i}, for attributes a_i, $i \in 0, ..., N$ where N denotes the number of attributes. The emotion classifier C and attribute classifier A_i parameters are (θ_f, θ_{emo}) and (θ_f, θ_{a_i}) respectively. The model is trained by minimising the loss $\mathcal{L}_{MTL}$:

$$\mathcal{L}_{MTL} = (1-\alpha)\mathcal{L}_{emo} + \alpha\left(\frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_{a_i}\right) \quad (1)$$

where $\mathcal{L}_{emo}$ is the cross-entropy classification loss between the predicted emotion $\hat{y}$ and the emotion label y and $\mathcal{L}_{a_i}$ is the classification loss between the predicted attribute $\hat{z}_i$ and the attribute label z_i for attribute i. The α term determines the weight of the different losses.

3.2. Adversarial Learning

3.2.1. Gradient Reversal

Gradient reversal refers to the technique of reversing the sign of the gradient at a specific point in the model. The GR model consists of the same three parts as the MTL model, but differs in that the sign of the gradients calculated from classifiers A_i is reversed for the shared layers. Based on [10], the gradient reversal layer is defined as $R(x) = x$ with $dR/dx = -I$, where I is the identity matrix. This can then be incorporated into the final loss:

$$\mathcal{L}_{GR} = (1-\alpha)\mathcal{L}_{emo} + \alpha R\left(\frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_{a_i}\right). \quad (2)$$

3.2.2. Adversarial Discriminator

The adversarial discriminator approach separates the classifier models so that there are no shared layers. In addition to the classifier C from the above, a separate attribute classifier model D_i is also used, with parameters θ_{d_i}. To train this model, at each time step we first pass the input $\boldsymbol{x}$ through C to return the emotion prediction $\hat{y}$. We then pass $\hat{y}$[1] through D_i to return the attribute prediction $\hat{z}_i$. D_i is trained by minimising the cross-entropy loss $\mathcal{L}_{d_i}$ between $\hat{z}_i$ and the attribute label z_i. C is trained by minimising the loss $\mathcal{L}_{AD}$:

$$\mathcal{L}_{AD} = (1-\alpha)\mathcal{L}_{emo} - \alpha\left(\frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_{d_i}\right). \quad (3)$$

As with the other models, $\mathcal{L}_{emo}$ is the classification loss between $\hat{y}$ and y, and α is the task weighting. This approach is based on [27], however, we omit the projection term of the loss as we are not attempting to de-bias the models, merely improve SER performance. We also found that training without this term produced more stable models, less prone to exploding gradients.

3.3. Models

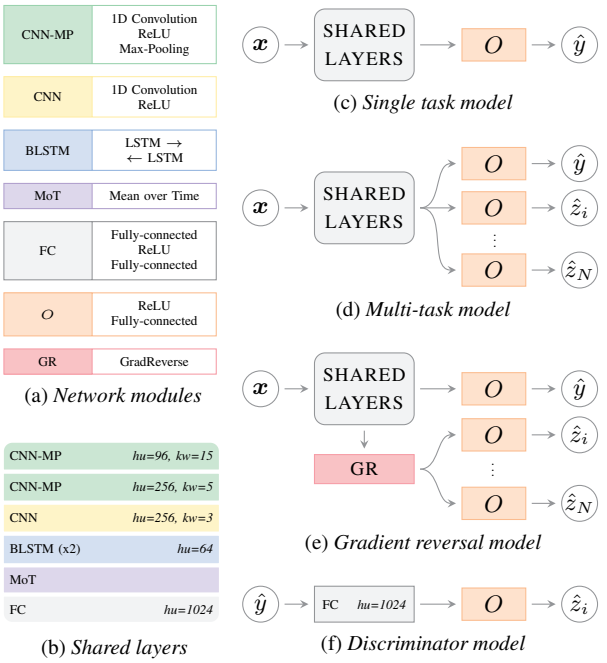
All models use the CNN-LSTM architecture described in [2] and shown in Figure 1, with varying output layers depending on the training approach. The baseline is a single-task model trained only on emotion. The MTL model returns one output for emotion and one output for the relevant attribute(s) as shown in

[1] In line with [27], the input to D is the softmax output vector.

Figure 1(d). The GR model is similar, with a gradient reversal layer used between the shared layers and the output layers of the relevant attribute(s). For the AD approach, two models are used: the first model is identical to the single-task model as shown in Figure 1(c) and the discriminator model D is composed of fully-connected layers as shown in Figure 1(f). For each of the three approaches we train a model for every possible combination of attributes using one to four attributes, resulting in fifteen models for each approach.

4. Experimental setup

The input to all models is Mel filterbank coefficients. We compute 40 coefficients over a 25ms window with a 10ms shift and do not add any delta features. We normalise utterances to have zero-mean and unit variance. We apply SpecAugment [29] to the features with a time mask width of 30 frames and a frequency mask width of 3. We use stochastic gradient descent with 0.9 momentum with a learning rate of 0.002 for all models including the discriminator. We use dropout at a rate of 0.5 before the FC layer. All models are trained for 150 epochs with early stopping. The shared layers as described in Figure 1 for all models have approximately 1.8 million parameters and each discriminator model totals about 1 million parameters. We evaluate model performance using Unweighted Accuracy (UA), which is the average of each individual class accuracy. For the MTL and AL models we set α to 0.1 and 0.5 respectively for the entire duration of training. For the gradient reversal model we set $\alpha = 0$ at the beginning of training and then increment it linearly with each epoch, ending at 0.5, similar to [11]. For fair comparison, we keep hyper-parameters the same for each model. We use PyTorch [30] to run experiments.



(a) *Network modules*

(b) *Shared layers*

(c) *Single task model*

(d) *Multi-task model*

(e) *Gradient reversal model*

(f) *Discriminator model*

Figure 1: *Model architectures. kw refers to kernel width and hu is the number of hidden units. Max-pooling layers have $kw = 3$ and stride $= 2$.*

Table 2: *Results on the TESS and CREMA-D corpora using the UA metric [%] for single-attribute experiments. The TESS baseline is from [2].*

	Corpus	Gender	Speaker	Language
TESS - baseline: 49.48				
MTL	48.90	35.13	**52.25**	**54.73**
GR	**54.71**	45.04	**55.17**	**50.10**
AD	40.73	39.35	36.98	33.29
CREMA-D - baseline: 52.21				
MTL	49.64	50.13	49.50	50.39
GR	51.82	**53.99**	47.60	**53.92**
AD	51.92	**53.33**	**53.23**	50.89

4.1. Corpora

For this cross-corpus study, the following corpora were used: CREMA-D [31]; EMOVO [32]; Emo-DB [33]; IEMOCAP [34]; EPST [35]; RAVDESS [36]; SAVEE [37]; and TESS [38]. All corpora apart from CREMA-D were also included in our work in 2019 [2]. For more information about these corpora, see Table 1 in that paper. All corpora excluding TESS and CREMA-D are aggregated, where each corpus is split between training (80%), validation (10%) and test sets (10%), yielding 11 hours 45 minutes for training and 1 hour 30 minutes each for validation and testing. Speakers in the validation and test sets do not appear in the training set. TESS (2 speakers, 1 hour 36 minutes) and CREMA-D (91 speakers, 5 hours 15 minutes) are used for out-of-domain (OOD) testing.

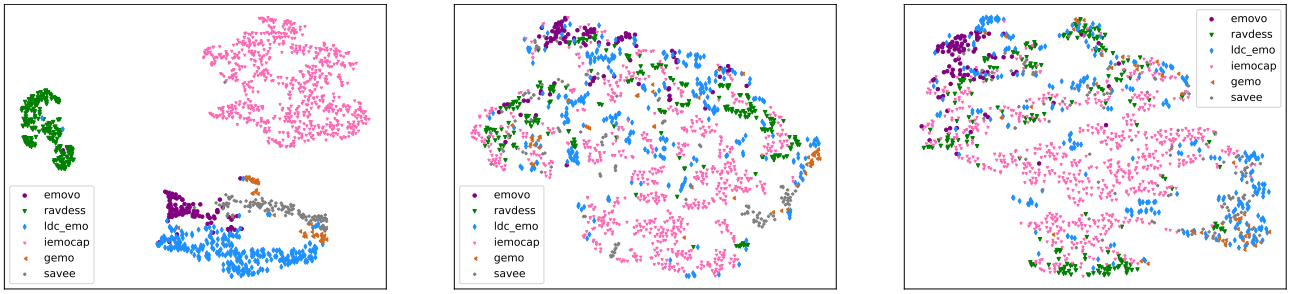
To avoid discarding data, the emotion labels for each corpus are mapped to three classes: *positive*, *negative* and *neutral*. For more information about this mapping, see Table 4 in [2]. The label set for CREMA-D consists of happiness, mapped to *positive*; sadness, anger, fear and disgust, mapped to *negative*; and neutral, mapped to *neutral*.

5. Results

We present the results of the single-attribute experiments in Table 2 along with the single-task baseline. When evaluating TESS, MTL and GR models trained with corpus, speaker or language yield similar or better performance compared to the baseline, up to 6% UA improvement. However, training with gender as an attribute is noticeably worse. This discrepancy might be explained by the demographics of the corpora: TESS contains only two female speakers, while the training set is gender-balanced. Evaluating CREMA-D yields an improvement when training with gender using the AL approach, but overall does not improve much beyond the baseline, the MTL approach being consistently the worst.

A subset of the multi-attribute experiments is shown in Table 3 and the full set of results can be found in the Supplementary Materials. On TESS, most of the configurations yield higher performance than the baseline, with the GR and AL approaches broadly outperforming MTL. Interestingly, when evaluating CREMA-D most multi-attribute experiments yield lower performance than the baseline and the single attribute models, which is consistent with the literature [8]. Only the GR approach shows improvements on this corpus, the AD being consistently the worst.

When comparing the OOD performance across the two cor-



(a) *MTL* (b) *GR* (c) *Baseline*

Figure 2: *t-SNE plots coloured by corpus.*

pora, the three approaches are inconsistent: for the MTL and AD approaches, none of the configurations yield consistent performance on both corpora. The GR approach is more promising, having one single attribute model (L) and 5 multi-attribute models (C+L,S+L,G+L,C+S+G,C+S+L) of out 15 which are consistently better than the baseline across the two corpora.

On the in-domain corpora, the single-task baseline performance is 59.32% UA. As in OOD experiments, single-attribute AL approaches consistently outperform MTL, but only two models actually surpass the baseline: GR with speaker (59.98%) and MTL with gender and language (60.34%). The highest performing model when combining all four attributes is the GR model with 56.53% UA. Ultimately, training with attributes appears to have a negative effect on in-domain performance, clearly showing the importance of OOD evaluation of SER models.

Table 3: *Results on the TESS and CREMA-D corpora using the UA metric [%] for select multi-attribute experiments. The TESS baseline is from [2].*

	S+G	S+L	G+L	C+G+L	C+S+L	C+S+G+L
TESS - baseline: 49.48						
MTL	34.77	47.42	**51.85**	**52.90**	48.69	49.17
GR	**51.58**	**55.46**	**52.79**	49.40	**54.75**	**51.71**
AD	49.19	**50.40**	**52.40**	**51.92**	**50.08**	**52.48**
CREMA-D - baseline: 52.21						
MTL	**52.85**	50.26	52.26	51.53	48.95	51.54
GR	50.36	**53.19**	**52.92**	50.57	**53.71**	50.12
AD	50.11	46.47	49.80	45.07	46.32	47.15

6. Analysis

When comparing attributes, most of the combinations are beneficial to some extent. The best-performing attribute combinations across the two OOD corpora are G+L, S+L and C+S+L. This indicates that language is a useful attribute for cross-corpus training with multiple languages. This is a novel insight, as previous work indicated that language was not useful as auxiliary task [23]. The gender attribute yields the most inconsistent performance, as it yields the lowest performance on TESS and the highest on CREMA-D. This indicates that this attribute can be useful as reported in the literature (see Table 1) but should be used with careful consideration. Finally, one could expect that using all four attributes would lead to the best performing model, but it is not the case. This suggests that using more attributes does not necessarily lead to better generalisation capabilities.

When comparing approaches, all three approaches outperforms the baseline in some settings, the GR approach outperforms the MTL and AD approaches in most settings and is also more consistently across corpora. The AD and MTL approaches perform inconsistently across the two datasets for various attributes. Overall, it seems that all three approaches require careful hyper-parameter tuning, which will be part of our future work.

In both the MTL and GR approaches, the goal is to gain more independence with respect to an attribute, but through a completely different implementation. In the MTL approach, independence to an attribute is *implicit*, as the hidden representations learn to represent the attribute, but they are decoupled from the emotion representation to some extent to allow the model to perform well on both tasks. In the GR approach, independence is *explicit*, as the hidden representations learn to be as independent as possible through the loss function. These two approaches hence make the model learn very different hidden representations. To validate these differences, we apply the t-SNE algorithm [39] to the output of the shared layers for the baseline as well as the MTL and GR models trained using corpus as auxiliary task. Figure 2 presents the learned representation, coloured by corpus. We can clearly see that the representation learned by the MTL model are heavily clustered by corpus, where the representation learned by the GR and baseline models do not form any apparent clusters. This cannot be explained solely by performance, as the baseline and the MTL model yield similar performance (49.5% and 48.9% UA), and the GR outperforms both (54.7% UA). This analysis shows that the MTL and GR models have taken a different route to gain independence from the attribute. Understanding these differences further is an encouraging avenue of research for OOD generalisation and will be part of our future work.

7. Conclusion

In this paper, we presented an analysis of multi-task and adversarial learning approaches on four different attributes for SER in the wild. We trained these models in a cross-corpus setting and evaluated them on an in-domain test set as well as on two out-of-domain corpora. We showed that both approaches can improve performance compared to the baseline, the gradient reversal technique being the most consistent across attributes and test sets. While there is no obvious best combination of attributes, we showed that training on attributes is generally beneficial. For future work, we will focus on studying the MTL and AL approaches in more detail, such as studying different model architectures and exploring label mapping across multiple corpora.

8. References

[1] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech Emotion Recognition Using Deep Learning Techniques: A Review," *IEEE Access*, vol. 7, pp. 117 327–117 345, 2019.

[2] J. Parry, D. Palaz, G. Clarke, P. Lecomte, R. Mead, M. Berger, and G. Hofer, "Analysis of Deep Learning Architectures for Cross-Corpus Speech Emotion Recognition," *Proc. of Interspeech*, pp. 1656–1660, 2019.

[3] B. Schuller, B. Vlasenko, F. Eyben, M. Wöllmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll, "Cross-Corpus Acoustic Emotion Recognition: Variances and Strategies," *IEEE Transactions on Affective Computing*, vol. 1, no. 2, pp. 119–131, 2010.

[4] M. D. Pell, L. Monetta, S. Paulmann, and S. A. Kotz, "Recognizing Emotions in a Foreign Language," *Journal of Nonverbal Behavior*, vol. 33, no. 2, pp. 107–120, Jun. 2009.

[5] V. Dissanayake, H. Zhang, M. Billinghurst, and S. Nanayakkara, "Speech Emotion Recognition 'in the Wild' Using an Autoencoder," in *Proc. of Interspeech*, 2020, pp. 526–530.

[6] R. Caruana, "Multitask Learning," *Machine Learning*, vol. 28, no. 1, pp. 41–75, 1997.

[7] A. Nediyanchath, P. Paramasivam, and P. Yenigalla, "Multi-Head Attention for Speech Emotion Recognition with Auxiliary Learning of Gender Recognition," in *Proc. of ICASSP*, 2020, pp. 7179–7183.

[8] Z. Zhu and Y. Sato, "Reconciliation of Multiple Corpora for Speech Emotion Recognition by Multiple Classifiers with an Adversarial Corpus Discriminator," in *Proc. of Interspeech*, 2020, pp. 2342–2346.

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," *Proc. of NeurIPS*, vol. 27, 2014.

[10] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-Adversarial Training of Neural Networks," *Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[11] M. Abdelwahab and C. Busso, "Domain Adversarial for Acoustic Emotion Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 12, pp. 2423–2435, 2018.

[12] M. B. Akçay and K. Oğuz, "Speech Emotion Recognition: Emotional Models, Databases, Features, Preprocessing Methods, Supporting Modalities, and Classifiers," *Speech Communication*, vol. 116, pp. 56–76, 2020.

[13] B. Schuller, Z. Zhang, F. Weninger, and G. Rigoll, "Selecting Training Data for Cross-Corpus Speech Emotion Recognition: Prototypicality vs. Generalization," in *Proc. of Afeka-AVIOS Speech Processing Conference*, 2011.

[14] B.-H. Su and C.-C. Lee, "A Conditional Cycle Emotion Gan for Cross Corpus Speech Emotion Recognition," in *Proc of SLT*, Shenzhen, China, 2021, pp. 351–357.

[15] J. Deng, Z. Zhang, F. Eyben, and B. Schuller, "Autoencoder-based Unsupervised Domain Adaptation for Speech Emotion Recognition," *IEEE Signal Processing Letters*, vol. 21, no. 9, pp. 1068–1072, 2014.

[16] H. Zhou and K. Chen, "Transferable Positive/Negative Speech Emotion Recognition via Class-wise Adversarial Domain Adaptation," in *Proc. of ICASSP*, 2019, pp. 3732–3736.

[17] J. Gideon, M. G. McInnis, and E. M. Provost, "Improving Cross-Corpus Speech Emotion Recognition with Adversarial Discriminative Domain Generalization (ADDoG)," *IEEE Transactions on Affective Computing*, vol. 12, no. 4, pp. 1055–1068, 2021.

[18] Y. Li, T. Zhao, and T. Kawahara, "Improved End-to-End Speech Emotion Recognition Using Self Attention Mechanism and Multitask Learning." in *Proc. of Interspeech*, 2019, pp. 2803–2807.

[19] J. Kim, G. Englebienne, K. Truong, and V. Evers, "Towards Speech Emotion Recognition "in the Wild" Using Aggregated Corpora and Deep Multi-Task Learning," in *Proc. of Interspeech*, 2017.

[20] M. Sharma, "Multi-lingual multi-task speech emotion recognition using wav2vec 2.0," in *Proc. of ICASSP*, 2022, pp. 6907–6911.

[21] F. Tao and G. Liu, "Advanced LSTM: A Study About Better Time Dependency Modeling in Emotion Recognition," in *Proc. of ICASSP*, 2018, pp. 2906–2910.

[22] C. Fu, C. Liu, C. T. Ishi, and H. Ishiguro, "An End-to-end Multitask Learning Model to Improve Speech Emotion Recognition," in *Proc. of EUSIPCO*, 2021, pp. 1–5.

[23] S. Goel and H. Beigi, "Cross Lingual Cross Corpus Speech Emotion Recognition," *arXiv preprint arXiv:2003.07996*, 2020.

[24] H. Zhang, M. Mimura, T. Kawahara, and K. Ishizuka, "Selective multi-task learning for speech emotion recognition using corpora of different styles," in *Proc. of ICASSP*, 2022, pp. 7707–7711.

[25] B. Zhang, E. M. Provost, and G. Essl, "Cross-Corpus Acoustic Emotion Recognition with Multi-Task Learning: Seeking Common Ground While Preserving Differences," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 85–99, 2019.

[26] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Epps, and B. W. Schuller, "Multi-Task Semi-Supervised Adversarial Autoencoding for Speech Emotion Recognition," *IEEE Transactions on Affective Computing*, 2020.

[27] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating Unwanted Biases with Adversarial Learning," *arXiv:1801.07593 [cs]*, 2018, arXiv: 1801.07593.

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

[29] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A Simple Augmentation Method for Automatic Speech Recognition," in *Proc. of Interspeech*, 2019.

[30] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic Differentiation in PyTorch," in *Proc. of NIPS Workshop*, 2017.

[31] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "CREMA-D: Crowd-sourced emotional multimodal actors dataset," *IEEE Transactions on Affective Computing*, vol. 5, no. 4, pp. 377–390, 2014.

[32] G. Costantini, I. Iadarola, A. Paoloni, and M. Todisco, "EMOVO Corpus: an Italian Emotional Speech Database," *LREC*, pp. 3501–3504, 2014.

[33] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A Database of German Emotional Speech," in *Proc. of Interspeech*, 2005, pp. 1517–1520.

[34] 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.

[35] M. Liberman and al., "Emotional Prosody Speech and Transcripts LDC2002S28," *Web Download. Philadelphia: Linguistic Data Consortium*, 2002.

[36] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A Dynamic, Multimodal Set of Facial and Vocal Expressions in North American English," *PLOS ONE*, vol. 13, no. 5, 2018.

[37] S. Haq and P. Jackson, *Machine Audition: Principles, Algorithms and Systems*, W. Wang, Ed. Hershey PA: IGI Global, 2010.

[38] K. Dupuis and M. K. Pichora-Fuller, "Recognition of Emotional Speech for Younger and Older Talkers: Behavioural Findings From the Toronto Emotional Speech Set," *Canadian Acoustics*, vol. 39, no. 3, pp. 182–183, 2011.

[39] L. v. d. Maaten and G. Hinton, "Visualizing Data Using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.