Interspeech 2022
18-22 September 2022, Incheon, Korea

# Multi-Corpus Speech Emotion Recognition for Unseen Corpus Using Corpus-Wise Weights in Classification Loss

*Youngdo Ahn*[1], *Sung Joo Lee*[2], *Jong Won Shin*[1]

[1]School of Electrical Engineering and Computer Science,
Gwangju Institute of Science and Technology, Gwangju, Korea
[2]Electronics and Telecommunications Research Institute, Daejeon, Korea
ayoungdo@gm.gist.ac.kr

## Abstract

Since each of the currently available emotional speech corpora is rather small to deal with personal or cultural diversity, multiple emotional speech corpora can be jointly used to train a speech emotion recognition (SER) model robust to unseen corpora. Each corpus has different characteristics, including whether acted or spontaneous, in which environment it was recorded, and what lexical contents it contains. Depending on the characteristics, the emotion recognition accuracy and time required to train a model for it are different. If we train the SER model utilizing multiple corpora equally, the classification performance for each training corpus would be different. The performance for unseen corpora may be enhanced if the model is trained to show similar recognition accuracy for each training corpus that covers different characteristics. In this study, we propose to adopt corpus-wise weights in the classification loss, which are functions of the recognition accuracy for each of the training corpus. We also adopt pseudo-emotion labels for the unlabeled speech corpus to further enhance the performance. Experimental results showed that the proposed method outperformed previously proposed approaches in the out-of-corpus SER using three emotional corpora for training and one corpus for evaluation.

**Index Terms**: speech emotion recognition, multi-corpus, corpus-wise weights, out-of-corpus

## 1. Introduction

Speech emotion recognition (SER), which classifies categorical emotions using spoken utterances, is useful for many applications [1, 2]. For real-world applications, the speech utterances that the SER model encounters may have different characteristics such as speakers, acoustics, and lexical contents, which were not present in the training data. To cope with the vulnerability of the model to the unseen factors, many researches focus on cross-corpus scenarios, in which limited information on the target corpus is available when the model is trained, such as a few labeled samples from the target corpus or unlabelled target samples [3, 4, 5, 6, 7, 8]. Another class of approaches is the out-of-corpus SER which does not utilize any information on the target corpus. In these approaches, multiple emotional speech corpora are jointly used to train an SER model robust to an unseen corpus [9, 10, 11].

Several approaches have been proposed to facilitate the generalization ability of the trained models, which may enhance the performance of the multi-corpus SER. [12, 13, 14, 15] constructed discriminative latent features with an autoencoder utilizing unlabeled speech corpus. [16] proposed using "soft labels" as the target of the classifier reflecting all annotators' opinions as probabilities instead of one-hot vectors obtained by majority voting of the annotators. [17] exploited two types of label smoothing in which the target label vector was a linear combination between a one-hot vector and a predefined distribution. The distribution was uniform for the uniform label smoothing and the class distribution in the training set for the unigram label smoothing. [18] adopted the focal loss [19] to focus more on utterances that are difficult to classify rather than easy samples. [20] proposed a confidence penalty that discourages the output of the neural networks to have low entropy. These methods have not yet been applied to out-of-corpus scenarios, but may potentially be useful.

Each corpus used in the training of the multi-corpus SER model represents different characteristics, including speakers, languages, lexical contents, recording environments, and whether the utterances are acted or spontaneous. For better generalization, these diverse characteristics should be incorporated into the trained model. If the SER model is trained with multiple corpora without further considering which corpus is more difficult to learn, the classification performance for each training corpus would be different. As each training corpus covers a different aspect of the emotional recordings, it may be beneficial to construct a model that performs similarly well for all the training corpora. In this study, for out-of-corpus SER utilizing multiple corpora, we propose to adopt corpus-wise weights in the classification loss, which are functions of the recognition accuracy for the individual training corpora. To further enhance the performance on the unseen corpus, we also exploited pseudo-emotion labels for unlabeled speech corpora. Experimental results showed that the proposed method outperformed other approaches to enhance generalization ability in the out-of-corpus SER utilizing three training corpora without any information on the target corpus.

## 2. Methods

In the multi-corpus SER, several emotional speech databases are utilized to train the SER model, while it is evaluated with another database. For the SER model, we exploit a neural network that brought advancement in the SER task. An input utterance is classified into one of the $C$ emotional classes by a neural-network-based emotion classifier $F$. One of the most popular choices for the loss function to train $F$ with $D$ training corpora is the cross-entropy (CE), which is defined as

$$\mathcal{L}_{\text{CE}}(F) = -\frac{1}{D}\sum_{d=1}^{D}\frac{1}{M^{(d)}}\sum_{i=1}^{M^{(d)}}\alpha_i^{(d)}y_i^{(d)}\cdot\log F(x_i^{(d)}) \quad (1)$$

10.21437/Interspeech.2022-111

where $x_i^{(d)}$ and $y_i^{(d)}$ are the $i$-th input feature and one-hot class label vector in the minibatch of size $M^{(d)}$ from the $d$-th training corpus, respectively, $\cdot$ represents an inner product, and $\alpha_i^{(d)}$ is the class-weight which is inverse proportional to the number of data in the same class with the $i$-th sample among $M^{(d)}$ samples in given minibatch. The class weight $\alpha_i^{(d)}$ is introduced to relieve the bias caused by class-imbalanced training data.

### 2.1. Corpus-wise weights in the classification loss

The recognition accuracy for each training corpus would be different when $F$ is trained to minimize the CE loss in (1), as some corpora are harder to classify than others. However, as each training corpus may cover different aspects of emotional expression, recording environments, or lexical contents, it may enhance the performance for unseen corpora to enforce the SER model to work equally well for all the training corpora.

To train a model to work similarly well for all training corpora, we propose to adopt corpus-wise weights (CWW) in the classification loss, which are functions of the recognition accuracy for given training corpora. The CE loss function incorporating CWW is as follows:

$$\mathcal{L}_{\text{CWW}}(F) = -\frac{1}{D}\sum_{d=1}^{D}\frac{w_d}{M^{(d)}}\sum_{i=1}^{M^{(d)}}\alpha_i^{(d)}y_i^{(d)}\cdot\log F(x_i^{(d)}) \tag{2}$$

where $w_d$ is the corpus-wise weight for the $d$-th training corpus. The corpus-wise weights are initialized as 1 and updated at the end of every training epoch using the unweighted accuracies (UAs) $\{\mathrm{U}_d\}_{d=1}^{D}$, which are the averages of the accuracies for individual emotional classes. The CWW is designed to give higher weights to the corpora that are difficult to classify with the current model. One such example to assign corpus-wise weights can be

$$w_d = \frac{(1-U_d)^{\lambda_W}}{\frac{1}{D}\sum_{d=1}^{D}(1-U_d)^{\lambda_W}} \tag{3}$$

in which $\lambda_W$ controls the degree of compensation for the difference in corpus-wise recognition accuracy. The procedure to train an SER model incorporating CWW is described in Algorithm 1.

**Algorithm 1**

```
Input: $D$ emotional speech corpora, randomly initialized neural network $F$
Output: Learned neural network $F$
Initialize $\{w_d\}_{d=1}^{D}$ as $\{1\}_{d=1}^{D}$
while stopping criterion is not met do
    Randomly sample minibatches for each corpus
    for $t = 1$ to #minibatches on the smallest corpus do
        Compute $\mathcal{L}$ in Eq. 2 with $t$-th minibatch
        Update $F$ for $\nabla_F\mathcal{L}$
    end for
    Update $\{w_d\}_{d=1}^{D}$ using Eq. 3
end while
```

### 2.2. Pesudo-emotion labels for unlabeled speech corpora

Emotional speech databases are rather small compared with the corpora for automatic speech recognition (ASR). Although the emotional labels are not available for those corpora, the large sizes of the corpora make them cover many factors of speech including various lexical contents, speaking styles, and speakers. One way to utilize unlabeled speech corpora is to use them with autoencoders to learn features that keep all the essential information to reconstruct the speech [12, 13, 14, 15] to facilitate generalization ability. An alternative way to exploit large speech databases without emotional labels is to assign pseudo-emotion labels (PELs) to the data and use them in training. One approach to assign PELs is to assign "neutral" labels for all the data in the ASR corpus, as many ASR corpora consist of sentences without much emotional expression, which we denote as $\text{PEL}_{\text{NL}}$. The other approach is to use all-one vectors instead of one-hot vectors as pseudo-label vectors, admitting that we do not have any information on the emotional class for the unlabeled ASR corpus, which is denoted as $\text{PEL}_{\text{AL}}$. Using one of these PEL vectors $\bar{y}$, the additional loss term can be used on top of the CE-based loss, which is given by

$$\mathcal{L}_{\text{PEL}}(F) = -\frac{\lambda_L}{M}\sum_{i=1}^{M}\bar{y}\cdot\log F(x_i^{Unlabeled}) \tag{4}$$

where $x_i^{Unlabeled}$ is the $i$-th input feature in the minibatch of size $M$ from the unlabeled corpus, and $\lambda_L$ is a weight for the PEL loss.

## 3. Experiments

### 3.1. Databases

In our experiments, we employed 4 different emotional speech corpora in English. Three corpora out of the CREMA-D (CRE) [21], IEMOCAP (IEM) [22], MSP-IMPROV (IMP) [23], and MSP-Podcast (POD) [24] databases were utilized for training, while the remaining corpus was used for testing. Within each corpus, we considered 4 categorical emotions including neutral, happy, sad, and angry. The specifications on the corpora are summarized in Table 1. CRE is an audiovisual corpus with 91 professional actors acting a target emotion for a pre-defined list of 12 sentences [21]. IEM is an audiovisual dyadic conversational corpus that contains conversations from 5 sessions. In each session, one actor and actress converse about a pre-defined topic [22]. To balance the class distribution in IEM, the excitement class is merged into happy. IMP is a multimodal emotional corpus spoken by 12 actors performing dyadic interactions in 6 sessions similar to IEM. IMP also collected natural speech by recording the colloquial discussions while the actors were not acting [23]. POD is collected from podcast recordings and contains various linguistic contents [24]. We used the released version 1.8 which consists of 28602, 4772, and 12787 samples for the train, validation, and test sets, respectively. In POD, the number of labeled speakers is 1285 but also contains samples without speaker labels. For the speech corpus without emotion labels used with pseudo-emotion labels, we utilized the Librispeech dataset [25], which contains 1000 hours of English speech read from audio-books. As in [12], we used a subset of 100 hours, which contains 28539 samples.

### 3.2. Experimental design and cross-validation settings

We evaluated the performance of the out-of-corpus SER using three training corpora for the proposed CWW and other ways to strengthen generalization including unsupervised representation learning [12], soft label [16], label smoothing with uniform and unigram distributions [17], focal loss [18], and confidence penalty [20]. We conducted a leave-one-corpus-out

Table 1: *Numbers of speakers and numbers of utterances for 4 emotions in the emotional corpora used in the experiments.*

| Corpus | Speakers | Neutral | Happy | Sad | Angry |
|---|---|---|---|---|---|
| CRE [21] | 91 | 1,087 | 1,271 | 1,270 | 1,271 |
| IEM [22] | 10 | 1,708 | 1,636 | 1,084 | 1,103 |
| IMP [23] | 12 | 3,477 | 2,644 | 885 | 792 |
| POD [24] | 1285+ | 26,009 | 14,285 | 2,649 | 3,218 |

cross-validation (CV) that utilizes three corpora for training and validation and one corpus for testing in turn. POD is already divided into training and validation sets, and CRE, IEM, and IMP were split as 58 and 33 speakers, 4 and 1 sessions, and 5 and 1 sessions for training and validation, respectively. The test set of POD was not used for training. In the evaluation phase, the whole dataset was used when the testing corpus was CRE, IEM, or IMP, while the test set of POD was utilized when evaluating performance for POD.

As a benchmark, we also provided the performances for the within-single-corpus SER in which the models were trained and tested in the same corpus. For within-single-corpus SER, CV was conducted with different settings for each corpus except for POD, for which the training, validation, and test sets are already well-defined and of which the sizes are large. For CRE, 10-fold leave-ten-speaker-out CV was performed. We randomly selected 10, 9, and 72 speakers for the testing, validation, and training, respectively. For IEM, 10-fold leave-one-speaker-out CV was conducted, i.e, 8 speakers in 4 sessions were used for the training, utterances from one speaker in the last session were used for the validation, and the last speaker's speech was used for testing. For IMP, 12-fold leave-one-speaker-out CV was conducted in a similar way to IEM.

### 3.3. Acoustic features and model configuration

As the input $x$, we used the IS10 [26] feature set consisting of 1582 acoustic features. We conducted z-normalization on these features with the training set. The emotion classifier $F$ comprises five fully connected (FC) layers with 1024, 1024, 512, 512, and 4 units, respectively, where the activation function for the last layer was softmax. The activation functions for all other layers were ReLU, and the dropout rate was 0.5. The hyper-parameters $\lambda_W$ and $\lambda_L$ were tested over $[1, 2, 3, 4, 5]$ and $[1, 0.1, 0.01, 0.001, 0.0001]$, respectively, and 4 and 0.001 produced the best results. For the unsupervised representation learning [12], we constructed the reconstruction module with two FC layers, 1024 and 1582, and attached the module on the second layer of $F$. The hyper-parameters for the compared methods were also tuned to result in the best performance.

We used Pytorch [27] to train the models and used the Adam optimizer [28] with a learning rate of 0.0002. The mini-batch size $M^{(d)}$ and $M$ were set to be 32. The average of the UAs for the validation sets from individual training corpora was used as a stopping criterion. An early stopping strategy with a patience of 5 was employed for the averaged UA for training corpora. In addition, we used a learning rate scheduler, which reduced the learning rate by multiplying by 0.1 after 2 patience. For each fold of CV, we experimented with 5 random seed initializations and reported the averaged results.

Table 2: *Unweighted accuracies (%) of speech emotion recognition for the test corpus on top and the average for all four cases. Except for the "Within-single-corpus (CE)," the model was trained with the remaining three corpora. CWW stands for the corpus-wise weights proposed in this study.*

| Method | CRE | IEM | IMP | POD | Avg |
|---|---|---|---|---|---|
| Within-single-corpus (CE) | 66.0 | 60.1 | 49.5 | 46.0 | 55.4 |
| Out-of-corpus (CE) | 51.6 | 50.1 | 38.9 | 31.9 | 43.1 |
| Unsup. learning [12] | 55.2 | 48.9 | 42.8 | 31.3 | 44.6 |
| Soft label [16] | 52.7 | 50.2 | 40.2 | 31.6 | 43.7 |
| Label smoothing [17] | 53.5 | 51.5 | 39.4 | 32.4 | 44.2 |
| Unigram smoothing [17] | 55.0 | 52.6 | 39.0 | 32.7 | 44.8 |
| Focal loss [18] | 51.4 | 49.6 | 40.5 | 32.9 | 43.6 |
| Confidence penalty [20] | 54.3 | 50.7 | 40.5 | 32.5 | 44.5 |
| CE+PEL$_{NL}$ | 51.4 | 49.7 | 39.3 | 37.0 | 44.4 |
| CE+PEL$_{AL}$ | 51.8 | 50.3 | 39.0 | 38.0 | 44.8 |
| CWW | 53.8 | 52.3 | 42.7 | 33.1 | 45.5 |
| CWW+Unsup. learning | **55.8** | 50.4 | **43.3** | 31.8 | 45.3 |
| CWW+Unigram smoothing | 54.6 | **53.7** | 40.4 | 33.8 | 45.6 |
| CWW+PEL$_{AL}$ | 53.5 | 51.2 | 40.7 | **38.7** | **46.0** |

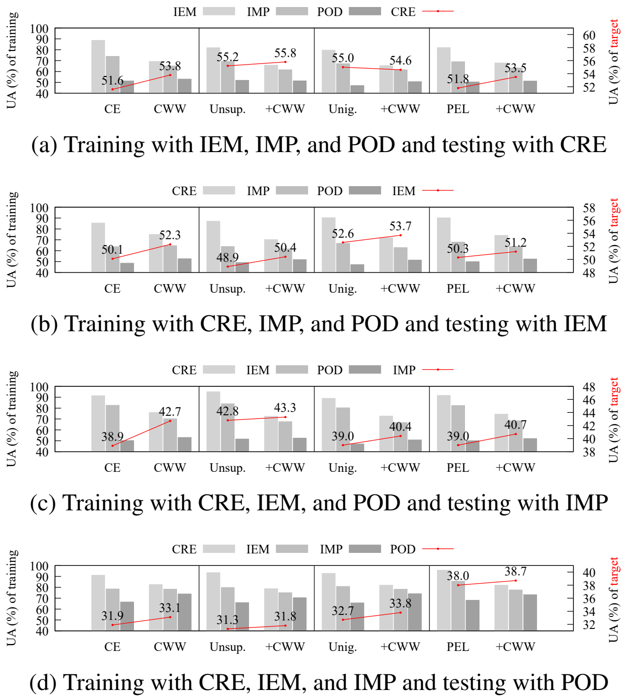
## 4. Results

Table 2 summarizes the UAs for the within-corpus and out-of-corpus SER for the test corpus on the top row. The best performance for each out-of-corpus configuration is marked in bold. The performance for the within-single-corpus SER is also shown in the table, which provides the upper bound of the performance of out-of-corpus SER and implies the complexity of each corpus. The highest UA was observed for CRE, which consisted of limited sentences, and the lowest UA was observed in POD recorded without much restriction [24]. IEM and IMP, which include improvised samples, demonstrated moderate performance.

Compared with a basic out-of-corpus SER system employing CE loss, most of the techniques to enhance generalization resulted in performance improvement except for a few experimental configurations. The introduction of the unlabeled ASR corpus with pseudo-emotion labels improved the accuracy for POD significantly, while it did not enhance the SER performance for other databases much. This may be because the lexical contents in POD that could not be covered by CRE, IEM, and IMP might be included in the Librispeech dataset. PEL$_{AL}$ showed slightly better performance than PEL$_{NL}$ on average.

The adoption of the proposed CWW resulted in an average UA of 45.5%, which is higher than other approaches. Although the average performance was higher with the CWW, the higher performances for individual experimental configurations could be achieved with the unsupervised learning [12], unigram label smoothing [17], and pseudo-emotion labels with all-one vectors when the test set was CRE and IMP, IEM, and POD, respectively. As the proposed CWW can be used with all of these methods, we introduced CWW into the aforementioned three methods. The results are also shown in Table 2. In most of the configurations, the systems with CWW outperformed those without CWW. On average, the SER system with CWW and PEL showed the best UA of 46.0%.

To illustrate the effect of the CWW, the UA's for the training sets from three corpora and the test set for four systems with and without CWW are shown in Fig. 1. It is noted that the UAs for the training and test sets are shown in different scales on the left and right side of the graphs, because the UA for the test set was much lower in the out-of-corpus SER. With the CWW, we can clearly see that the UAs for different training sets became more



(a) Training with IEM, IMP, and POD and testing with CRE

(b) Training with CRE, IMP, and POD and testing with IEM

(c) Training with CRE, IEM, and POD and testing with IMP

(d) Training with CRE, IEM, and IMP and testing with POD

Figure 1: *Unweighted accuracies (UAs, %) of the training sets from three corpora (bar graphs, UA scale on the left) and the test set (red line graphs, UA scale on the right) for basic system with CE loss (CE), system with unsupervised learning [12] (Unsup.), those with unigram smoothing [15] (Unig.), and pseudo-emotion labels (PEL), with and without the proposed corpus-wise weights (CWW).*

similar, and the performance for the test set was enhanced in most of the cases.

## 5. Analyses

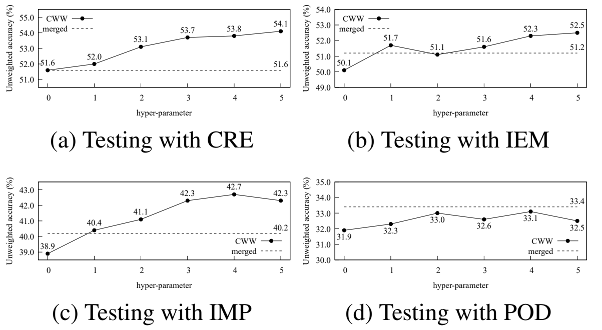
In this section, we analyze the CWW of each out-of-corpus SER and the change of the performance by the hyper-parameter of the CWW.

Table 3 summarizes the CWW computed in the last epoch of training for each training corpus in each of the out-of-corpus SER systems. It is noted that the sum of the CWWs for training corpora for each experiment is $D = 3$. On average, the order of the CWW is CRE<IEM<IMP<POD. CRE, which showed the lowest CWW, is recorded with the most limited sentences among four corpora. POD, which showed the highest CWW, is recorded with completely spontaneous sentences. Both IEM and IMP consist of scripted and spontaneous utterances, but IMP has more various lexical contents.

Fig. 2 shows the UAs depending on the hyper-parameter $\lambda_W$ for each of 4 systems. We can see that the performance was improved with increasing $\lambda_W$ until $\lambda_W$ became 4. We have also provided the performances when we merged all the training corpora into a single corpus. In this case, the size of each training corpus works like CWW in Eq. 2. The average UA was 44.1%, which was higher than that for the original experiment (CE) but lower than that with CWW. This result may imply that performance improvement on unseen corpora with the proposed method was not merely by emphasizing larger databases, but by constructing a model that performed similarly for all training corpora covering different variabilities of the emotional expression.

Table 3: *The corpus-wise weights computed in the last epoch for each training corpus in each of the out-of-corpus speech emotion recognition systems and the averages for the three out-of-corpus cases.*

| testing<br>training | CRE | IEM | IMP | POD | Avg |
|---|---|---|---|---|---|
| CRE | - | 0.17 | 0.17 | 0.35 | 0.23 |
| IEM | 0.37 | - | 0.39 | 0.86 | 0.54 |
| IMP | 0.61 | 0.67 | - | 1.79 | 1.02 |
| POD | 2.02 | 2.16 | 2.44 | - | 2.21 |



(a) Testing with CRE

(b) Testing with IEM

(c) Testing with IMP

(d) Testing with POD

Figure 2: *Unweighted accuracies (%) with CWW depending on different hyper-parameter $\lambda_W$. The dashed line shows the result of the experiment when all training corpora are merged into a single corpus.*

## 6. Conclusions

In this study, for the out-of-corpus speech emotion recognition model trained with multiple corpora, we introduce the corpus-wise weights in the cross-entropy loss function to make the model perform similarly for all training corpora to achieve better generalization. The corpus-wise weights are designed as a function of the unweighted accuracy for the training set from the corresponding corpus to give higher weights for the corpora that are more difficult to classify. We also adopted pseudo-emotion labels for unlabeled speech corpus to further enhance the performance. Experimental results showed that the proposed CWW could enhance the performance of the out-of-corpus SER and could be successfully combined with previously proposed methods to strengthen the generalization capability.

## 7. Acknowledgements

This work was supported in part by the MSIT (Ministry of Sci ence and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2021-2021-0-01835) supervised by the IITP (Institute of Information & Communications Technology Planning & Evaluation) and by Institute of Information & Communications Technology Planning & Eval uation (IITP) grant funded by the Korea government (MSIT) (No.2019-0-01842, Artificial Intelligence Graduate School Program (GIST)).

# 8. References

[1] D. Morrison, R. Wang, and L. C. De Silva, "Ensemble methods for spoken emotion recognition in call-centres," *Speech communication*, vol. 49, no. 2, pp. 98–112, 2007.

[2] T. Kostoulas, I. Mporas, T. Ganchev, and N. Fakotakis, "The effect of emotional speech on a smart-home application," in *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*. Springer, 2008, pp. 305–310.

[3] 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.

[4] R. Milner, M. A. Jalal, R. W. Ng, and T. Hain, "A cross-corpus study on speech emotion recognition," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 304–311.

[5] Y. Ahn, S. J. Lee, and J. W. Shin, "Cross-corpus speech emotion recognition based on few-shot learning and domain adaptation," *IEEE Signal Processing Letters*, 2021.

[6] M. Abdelwahab and C. Busso, "Supervised domain adaptation for emotion recognition from speech," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5058–5062.

[7] J. Gideon, M. McInnis, and E. M. Provost, "Improving cross-corpus speech emotion recognition with adversarial discriminative domain generalization (addog)," *IEEE Transactions on Affective Computing*, 2019.

[8] Y. Xiao, H. Zhao, and T. Li, "Learning class-aligned and generalized domain-invariant representations for speech emotion recognition," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 4, no. 4, pp. 480–489, 2020.

[9] B. Schuller, Z. Zhang, F. Weninger, and G. Rigoll, "Using multiple databases for training in emotion recognition: To unite or to vote?" in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[10] J. Kim, G. Englebienne, K. P. Truong, and V. Evers, "Towards speech emotion recognition" in the wild" using aggregated corpora and deep multi-task learning," *arXiv preprint arXiv:1708.03920*, 2017.

[11] 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. Interspeech 2019*, pp. 1656–1660, 2019.

[12] M. Neumann and N. T. Vu, "Improving speech emotion recognition with unsupervised representation learning on unlabeled speech," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7390–7394.

[13] S. E. Eskimez, Z. Duan, and W. Heinzelman, "Unsupervised learning approach to feature analysis for automatic speech emotion recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5099–5103.

[14] 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.

[15] V. Dissanayake, H. Zhang, M. Billinghurst, and S. Nanayakkara, "Speech emotion recognition'in the wild'using an autoencoder." in *INTERSPEECH*, 2020, pp. 526–530.

[16] S. Steidl, M. Levit, A. Batliner, E. Noth, and H. Niemann, """of all things the measure is man" automatic classification of emotions and inter-labeler consistency [speech-based emotion recognition]," in *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, vol. 1. IEEE, 2005, pp. I–317.

[17] J. Huang, J. Tao, B. Liu, and Z. Lian, "Learning utterance-level representations with label smoothing for speech emotion recognition," *Proc. Interspeech 2020*, pp. 4079–4083, 2020.

[18] Y. Zhong, Y. Hu, H. Huang, and W. Silamu, "A lightweight model based on separable convolution for speech emotion recognition," *Proc. Interspeech 2020*, pp. 3331–3335, 2020.

[19] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[20] G. Pereyra, G. Tucker, J. Chorowski, Ł. Kaiser, and G. Hinton, "Regularizing neural networks by penalizing confident output distributions," *arXiv preprint arXiv:1701.06548*, 2017.

[21] 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.

[22] 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, p. 335, 2008.

[23] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. M. Provost, "Msp-improv: An acted corpus of dyadic interactions to study emotion perception," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 67–80, 2016.

[24] 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.

[25] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[26] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. S. Narayanan, "The interspeech 2010 paralinguistic challenge," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[27] 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.

[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.