DEEPEMOCLUSTER: A SEMI-SUPERVISED FRAMEWORK FOR LATENT CLUSTER REPRESENTATION OF SPEECH EMOTIONS

Wei-Cheng Lin, Kusha Sridhar, Carlos Busso

Multimodal Signal Processing (MSP) lab, Department of Electrical and Computer Engineering
The University of Texas at Dallas, Richardson TX 75080, USA

wei-cheng.lin@utdallas.edu, kusha.sridhar@utdallas.edu, busso@utdallas.edu

ABSTRACT

Semi-supervised learning (SSL) is an appealing approach to resolve generalization problem for *speech emotion recognition* (SER) systems. By utilizing large amounts of unlabeled data, SSL is able to gain extra information about the prior distribution of the data. Typically, it can lead to better and robust recognition performance. Existing SSL approaches for SER include variations of encoder-decoder model structures such as *autoencoder* (AE) and *variational autoecoders* (VAEs), where it is difficult to interpret the learning mechanism behind the latent space. In this study, we introduce a new SSL framework, which we refer to as the DeepEmoCluster framework, for attribute-based SER tasks. The DeepEmoCluster framework is an end-to-end model with mel-spectrogram inputs, which combines a self-supervised pseudo labeling classification network with a supervised emotional attribute regressor. The approach encourages the model to learn latent representations by maximizing the emotional separation of K-means clusters. Our experimental results based on the MSP-Podcast corpus indicate that the DeepEmoCluster framework achieves competitive prediction performances in fully supervised scheme, outperforming baseline methods in most of the conditions. The approach can be further improved by incorporating extra unlabeled set. Moreover, our experimental results explicitly show that the latent clusters have emotional dependencies, enriching the geometric interpretation of the clusters.

Index Terms— Semi-supervised learning (SSL), speech emotion recognition (SER), unsupervised clusters

1. INTRODUCTION

Recognizing human's emotional states is crucial to modern *human computer interaction* (HCI) systems [1]. *Speech emotion recognition* (SER) also plays an important role in various fields such as education [2] and healthcare [3]. Although recent advances in deep learning approaches have led to high performance in areas such as image classification [4] and *automatic speech recognition* (ASR) [5], SER is still a challenging problem with often poor generalization across different conditions (e.g., environment, recording settings or speakers) [6, 7]. One of the major factors leading to low performance is the availability of the amount of training data. Comparing to image or speech recognition tasks, the sizes of speech emotional corpora are relatively small (e.g., IEMOCAP corpus for SER [8] $\approx$ 12 hrs versus the LibriSpeech corpus for ASR [9] $\approx$ 1,000 hrs). To alleviate this limitation, studies have investigated domain adversarial techniques that utilize information from multiple corpora. Studies have explored either adopting additional domain classifier with reverse gradient layers [10] or a critic network [11] to obtain a more generalized intermediate representation, reducing mismatches from different domains. Other studies have also explored data augmentation methods to improve robustness of SER models. More specifically, *generative adversarial networks* (GANs) have been applied to artificially synthesize emotional samples to reduce the sparsity in the distribution of the train set [12, 13]. However, these adversarial training approaches may suffer from unstable convergence issues or be limited by the size of the datasets. Another appealing approach is *semi-supervised learning* (SSL). The key concept of SSL is to leverage large amounts of unlabeled data to improve the robustness and performance of the supervised task [14]. Unlike labeled data, the collection of unlabeled resources is often easy and inexpensive. Therefore, it is easy to obtain a large dataset of unlabeled data that augments the often reduced labeled set. SSL methods learn additional structure of the input distribution by using this unlabeled data [15]. One of the approaches for utilizing unlabeled data in SSL is to make use of the *cluster assumption*, which says that if two data samples in the input space belong to the same cluster, they are likely to belong to the same class or region in the target space. DeepCluster [16] follows this assumption, achieving great success in extracting discriminative features from completely unsupervised representation learning. It utilizes self-supervised training technique to learn cluster-based pseudo class labels, which achieves competitive performance compared to supervised learning models.

Inspired by the DeepCluster framework, this study proposes a novel formulation for SER, where we modify the original unsupervised structure into a semi-supervised framework. The approach derives emotional speech clusters, namely - DeepEmoCluster. Specifically, we implement an additional supervised emotional attribute-based regressor (i.e., arousal, dominance and valence) to jointly train with the unsupervised cluster classifier, encouraging the model to learn emotionally discriminative contents under a *maximum latent clusters separation* constraint. The proposed model is an *end-to-end* (E2E) *convolutional neural network* (CNN) architecture (i.e., VGG-16 [17]), where the input feature is a 128D-mel spectrogram. We evaluate our proposed framework using the MSP-Podcast corpus [18], using the *concordance correlation coefficient* (CCC) as the metric to evaluate model performance. The results show that DeepEmoCluster framework achieves competitive prediction performances under fully supervised scheme for arousal, valence and dominance, outperforming baseline models most of the time. We find that the DeepEmoCluster framework can further improve the prediction performance while using the semi-supervised setting, where reinforcing the information gains from the unlabeled data is helpful for the supervised task. As part of the evaluation, we explore the optimal number of clusters needed in the DeepEmoCluster framework. The results indicate that the number of clusters should be fine-tuned depending on the size of the unlabeled set and the emotional attribute. Finally, our latent cluster analysis demonstrates that the DeepEmoCluster approach creates latent clusters that depend on the emotional content, enriching the geometric interpretation of the clusters.

This study was funded by the National Science Foundation (NSF) under grant CNS-2016719 and CAREER IIS-1453781. Github link-https://github.com/winston-lin-wei-cheng/DeepEmoClusters

The key contribution of this paper is the new SSL DeepEmoCluster framework for the SER task, which is a semi-supervised variant of the DeepCluster framework [16]. By leveraging unlabeled data and jointly training with supervised networks, the DeepEmoCluster approach improves attribute-based SER model performance with explicit geometric interpretations in the latent representation.

2. RELATED WORK

One of the recent popular trends in SER is to build E2E learning systems. E2E models do not require predefined handcrafted features and can directly extract emotionally relevant features from either time domain waveforms [19], or frequency domain raw spectrograms [20]. These approaches often rely on *deep neural networks* (DNNs). Satt *et al.*[20] demonstrated that an E2E SER model with raw spectrogram inputs could be easily combined with a noisy reduction solution (e.g., harmonic filtering), enabling SER systems to resist low SNR environments. Li *et al.*[21] proposed a multitask model with self-attention mechanism based on spectrogram features. Their method achieved state-of-the-art performance on the IEMOCAP dataset [8]. Trigeorgis *et al.*[19] built their E2E model from raw waveforms using a CNN-BLSTM structure. The key component was the 1D convolutions operating on the discrete-time domain waveforms, which could be considered as a feature refinement system to remove background noise. Their results showed that features derived from the E2E model outperforms handcrafted acoustic features (e.g., eGeMAPS [22]). In this study, we also employ an E2E model with mel-spectrogram inputs to better represent emotional cues.

Various studies have explored SSL and latent representation learning approaches to utilize unlabeled data for SER tasks. Deng *et al.*[23] presented the *semi-supervised autoencoder* (SSAE), which combined an unsupervised *deep denoising autoencoder* (DDAE) with a supervised learning objective. They introduced an extra class label for the unlabeled data, forcing models to learn a bottleneck latent representation by incorporating prior information from unlabelled data. Following a similar concept, Latif *et al.*[24] proposed a *variational autoencoders* (VAEs) framework for deriving a latent representation from speech signals. In contrast to DDAE, VAEs learns a probability distribution representing the inputs in a latent space instead of compressing them in the bottleneck layer. Parthasarathy and Busso [25, 26] adopted the *ladder networks* framework (Γ-model in Rasmus *et al.*[27]) to improve attribute-based SER performance. The Γ-model imposes consistency regularization on skip connections between noisy encoder and decoder, aiming to obtain invariant intermediate representations toward noise perturbations. These approaches rely on the encoder-decoder structure, where the bottleneck latent representation was not trained to explicitly involve strong geometric properties by pulling apart data with different emotional content. Therefore, it is difficult to understand the mechanism behind the latent space or the bottleneck layer. This study aims to create a meaningful latent representation using a SSL approach based on the DeepCluster framework [16], which imposes strong geometric constraints (i.e., maximum emotional separation between clusters) without requiring decoder networks.

3. RESOURCES

This study uses the MSP-Podcast corpus [18], which consists of emotionally rich spontaneous speech recordings collected from publicly available podcasts under creative commons license. The content of the recordings cover various topics such as interviews, sports, academic talks, entertainment and politics. We process these recordings through a speaker diarization tool to segment the podcasts into smaller speaking turns of length between 2.75 and 11 seconds in duration. We employ a number of pre-processing steps to select speaking turns with high *signal to noise ratio* (SNR) and single speaker content without overlapped speech. We remove turns with background noise, music and telephone quality speech. To balance the emotional content of the corpus and have emotionally rich sentences, we run speech segments through emotional retrieval algorithms based on the strategy suggested in Mariooryad *et al.*[28].

This study uses version 1.6 of the corpus which consists of 50,362 sentences (83h 29m). We use *amazon mechanical turk* (AMT) to annotate the sentences in the corpus using a variation of the crowdsourcing protocol discussed in Burmania *et al.*[29]. The sentences are annotated for their primary and secondary emotional content (categorical classes), as well as the emotional attributes arousal (calm versus active), valence (negative versus positive) and dominance (weak versus strong). This study uses emotional attributes, formulating the SER task as a regression problem. Each annotator assessed the emotional content with a seven point likert-type scale using *self-assessment manikins* (SAMs). Each sentence in the corpus has five or more annotations and the ground-truth labels are obtained by averaging the scores across subjects. The test set has 10,124 samples from 50 speakers, the development set has 5,958 samples from 40 speakers, and the train set has 34,280 samples from the rest of the speakers. This partition aims to keep speaker independent sets. There are around 500, 000 additional speech segments that have not been retrieved or annotated. These speaking turns form our unlabeled data in the corpus.

4. PROPOSED DEEPEMOCLUSTER APPROACH

Figure 1 shows the proposed DeepEmoCluster approach, which is a SSL framework that creates meaningful emotional dependent clusters as a latent representation.

4.1. Acoustic Features and Pre-processing

We use the toolkit *librosa* [30] to extract the 128D-mel spectrogram as the input acoustic representation to our end-to-end SER model. First, the magnitude spectrogram of the waveform signal is calculated with a window size of 32 ms (512 sample points), with 16 ms overlap between windows. Then, we map the magnitude spectrogram into 128D mel-scale filters. We perform z-normalization on these features, where the mean and standard deviation used in the normalization are estimated over the train set. After obtaining the normalized 128D-mel spectrogram for each sentence, we split them into smaller data chunks (i.e., sub-images of the original spectrogram). This study follows the chunk segmentation approach proposed by Lin and Busso [31] that splits different duration sentences into a fixed number of chunks with fixed duration. We achieve this segmentation by dynamically adjusting the overlap between chunks per sentence. The parameters of this segmentation are the desired chunk length w_c, and the maximum duration of a sentence in the dataset T_{max}. Based on these parameters, Equation 1 defines the number of chunks C per sentence. Equation 2 provides the target chunk step size Δc_i for a given sentence i with duration T_i. Then, we can split a sentence into fixed C chunks without relying on zero-paddings. Figure 2 shows a visual example of the chunking process.

$$C = \left\lceil \frac{T_{\max}}{w_c} \right\rceil \tag{1}$$

$$\Delta c_i = \frac{T_i - w_c}{C - 1} \tag{2}$$

The input of our model is the split sub-images of the spectrogram, which have the same image size. During training, the same sentence-level emotional label is assigned to all the sub-images obtained from original spectrogram. While more sophisticated meth-

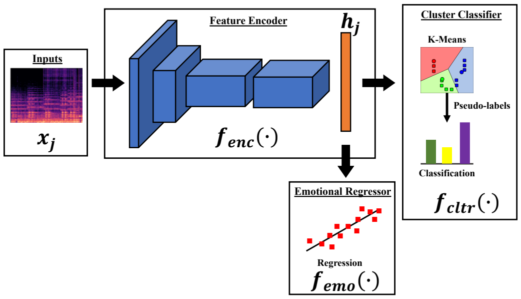


Fig. 1. The semi-supervised DeepEmoCluster framework for SER.

ods can be used to combine the chunk-based decisions [31], in this study we directly average the prediction outputs of the sub-images to derive a final prediction for the sentence.

4.2. The DeepEmoCluster Framework

The proposed semi-supervised DeepEmoCluster framework consists of three networks: 1) the *feature encoder* $f_{enc}(\cdot)$, 2) the *cluster classifier* $f_{cltr}(\cdot)$, and 3) the *emotional regressor* $f_{emo}(\cdot)$ (Fig. 1).

The first network of the DeepEmoCluster approach is the feature extractor, which extracts discriminative information from the chunks. This block is implemented with the VGG-16 architecture [17] with batch normalization. The CNN-based network extracts the chunk-level feature representations h_j from the input mel-spectrograms x_j. The second network is the cluster classifier, which is an unsupervised (or self-supervised) network that classifies pseudo-class labels given by the K-means clustering results based on the latent feature space. The number of classes depends on the number of K-means clusters and these clustering pseudo labels are re-assigned after every training epoch. The third network is the emotional regressor, which is a supervised regression network that predicts the target emotional attribute (i.e., arousal, dominance and valence). The addition of this supervised network brings emotional dependencies into the latent space.

The semi-supervised training scheme of the proposed DeepEmoCluster framework is divided into two stages within a single training epoch. First, data points from the unlabeled set are considered for the cluster classifier (i.e., unsupervised path). During this step, we freeze the $f_{emo}(\cdot)$ model and the gradients only backpropagate through the $f_{enc}(\cdot)$ and $f_{cltr}(\cdot)$ networks. Second, the data-label pairs from the labeled set are jointly used to backpropagate the gradients to the entire model. During this training stage, we consider the network as a multitask learning model with the loss function given in Equation 3, where CCC is the *concordance correlation coefficient* (CCC) loss for the regression task, and CE is the *cross entropy* (CE) loss computed for the self-supervised cluster classifier. The parameter λ controls the importance of the unsupervised task.

$$Loss = (1 - CCC) + \lambda \times CE \tag{3}$$

5. EXPERIMENTAL RESULTS

5.1. Experimental Settings

We split the 128D mel-spectrogram feature of every sentence into $C = 11$ sub-images by setting the desired chunk window size w_c to 1 sec (Eq. 1). Note that the maximum sentence duration of the MSP-Podcast dataset is 11 secs [18]. The detailed architecture of the feature encoder $f_{enc}(\cdot)$ model is shown in Table 1, which follows the VGG-16 structure [17]. The CNN-blocks in Table 1 consist of a 2D-CNN with BatchNorm and ReLU activation function.

The cluster classifier and emotional regressor models have the same structure, but they are constructed using fully connected layers (ReLU activation) and a task-dependent output layer (i.e., softmax output for the cluster classifier and linear combination output for the emotional regressor). The weighting factor λ of the loss function (Eq. 3) is set to 1 and the cluster number of K-means is a fine-tuned parameter depending on the size of the dataset. We discuss its value in Section 5.3. We train our model with a batch size of 64, with Adam optimizer for the emotional regressor (lr=0.0005), and *stochastic gradient descent* (SGD) optimizer for the feature encoder and cluster classifier (lr=0.001). We save the best models with an early stopping criterion based on the validation loss. All models are implemented with PyTorch.

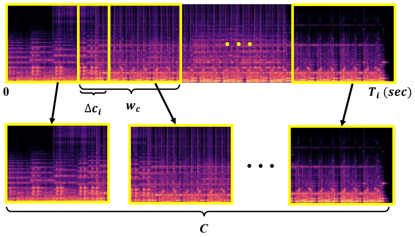


Fig. 2. A visualization example of the chunk segmentation procedure [31]. Sentences with different durations are split into C chunks with fixed duration (w_c) by adjusting the chunk step size (Δc_i).

Table 1. The VGG-16 feature encoder model used in this study.

Layer	Channels/Nodes	Kernel	Stride	Activation
Input	1	N/A	N/A	N/A
CNN-block (×2)	32	(3, 3)	1	ReLU
CNN-block (×2)	64	(3, 3)	2	ReLU
CNN-block (×3)	128	(3, 3)	2	ReLU
Flatten	N/A	N/A	N/A	N/A
Linear	256	N/A	N/A	ReLU
Dropout	$p = 0.5$	N/A	N/A	N/A

We compare three baseline models in this study: *CNN-regressor*, *CNN-AE* and *CNN-VAE*. The *CNN-regressor* is a regular CNN-based emotional regression model. More specifically, the model contains the feature encoder and emotional regressor networks, which can only be applied to fully supervised learning. Instead of using the cluster classifier (i.e., pseudo labeling), the *CNN-AE* and *CNN-VAE* are attached with an additional decoder network to reconstruct the inputs as an unsupervised task. By combining a supervised task with the autoencoder, these two baselines can also be extended to semi-supervised frameworks providing appropriate baselines for our proposed DeepEmoCluster framework. The decoder network mirrors the layers of the encoder (i.e., the feature encoder in Table 1), where the transposed convolutional layer is used to reconstruct the original input feature maps. The major difference between *CNN-AE* and *CNN-VAE* is that the *CNN-AE* baseline directly reconstructs the outputs from the bottleneck, whereas, the *CNN-VAE* baseline reconstructs the input from the reparametrized latent vectors. We evaluate the model performance using CCC. We randomly split the original test set into four subsets with the same size and run five trails for all models, reporting the average based on these results (i.e., 4 subsets × 5 trails = 20 results per model per attribute). We implement this strategy to conduct statistical analysis using a two-tailed T-test over the 20 results. We assert statistical significance when p-value <0.05.

5.2. Evaluation with Fully Supervised Learning

In this section, we focus on comparing fully supervised performance of the proposed DeepEmoCluster framework with the three baseline models. All the models are trained with the labeled set without considering any unlabeled data. We fix the number of clusters to 10 (i.e., 10-clusters) for all emotional attributes in the DeepEmoCluster

Table 2. CCC performance for the DeepEmoCluster approach and three baselines for the supervised condition. Results tagged with $*$, † and ‡ indicate that the CCC values are statistic significant better than the results for *CNN-regressor*, *CNN-AE* and *CNN-VAE*, respectively.

	CNN-regressor	*CNN-AE*	*CNN-VAE*	*DeepEmoCluster* (10-clusters)
Aro.	0.6177	0.6338	0.5586	$\mathbf{0.6502}^{*\dagger\ddagger}$
Dom.	0.4928	0.5111	0.4800	$\mathbf{0.5426}^{*\dagger\ddagger}$
Val.	0.1696	0.1354	**0.1826**	$0.1510^{\dagger}$

Table 3. CCC performances of the semi-supervised DeepEmoCluster framework as we increase the number of clusters. Results tagged with $*$ indicates that the values are statistic significant better than the results of the model trained with only 10-clusters.

DeepEmoCluster (10-clusters)	*fully supervised*	*15K unlabeled*	*40K unlabeled*
Aro.	0.6502	0.6504	$\mathbf{0.6611}^{*}$
Dom.	**0.5426**	0.5400	0.5400
Val.	0.1510	$\mathbf{0.1714}^{*}$	0.1572

framework. Table 2 shows the CCC performances of all models for the three emotional attributes. DeepEmoCluster achieves the best prediction scores for arousal and dominance, which significantly outperforms all the baseline models. Although the valence result of the DeepEmoCluster does not reach the highest performance, it still obtains competitive results. These results indicate that discretizing latent representation by encouraging maximum separations between data points can be beneficial to a continuous regression task, since it reduces the complexity of the learned latent space which is constrained under discrete clusters.

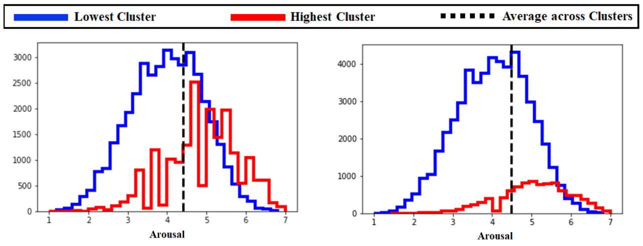
5.3. Evaluation of Unlabeled Set Size and Number of Clusters

Since the proposed DeepEmoCluster framework is a semi-supervised framework, we can utilize unlabeled data to further improve the model performance. Table 3 reports the CCC performances, as we add more unlabeled data (0, 15K and 40K unlabeled segments) under a semi-supervised training scheme. We use 10 clusters for the DeepEmoCluster approach. Table 3 shows that training with additional unlabeled data leads to improved model prediction results, especially for the valence attribute. This result validates the effectiveness of our SSL approach to better represent features by utilizing unlabeled data information.

We notice that while we increase the size of the unlabeled set, the performance does not keep significantly increasing. Therefore, we hypothesize that the clusters' number depends on the size of the unlabeled set. To validate our hypothesis about the relation between the clusters' number and the size of the unlabeled set, we fix the unlabeled data to 40K segments. Then, we evaluate the approach with different number of clusters. Table 4 shows the increasing trends in the prediction performance as we increase the number of clusters in the DeepEmoCluster framework for dominance and valence. However, the prediction for arousal achieves the best result with only 10-clusters. This result suggests that the number of clusters is a fine-tuning parameter that depends on the size of unlabeled set and the emotional attribute.

Table 4. The CCC performances of semi-supervised DeepEmoCluster framework training with additional 40K unlabeled set by different settings of clusters number. Results tagged with $*$ means statistic significant greater than *10-clusters* setting.

DeepEmoCluster (40K unlabeled)	*10-clusters*	*20-clusters*	*30-clusters*
Aro.	**0.6611**	0.6491	0.6416
Dom.	0.5400	0.5459	$\mathbf{0.5490}^{*}$
Val.	0.1572	$\mathbf{0.1756}^{*}$	$\mathbf{0.1752}^{*}$



(a) Unsupervised Clusters (b) Supervised Clusters

Fig. 3. Emotional distributions of the clusters with the highest and lowest average level of arousal. The distributions are farther apart with the addition of the supervised SER task.

5.4. Cluster Analysis

The DeepEmoCluster framework possesses high geometric interpretation, which enables us to explicitly show the mechanism behind the learned emotional latent space contributing to performance improvements. One approach to validate whether the emotional content plays a role in the latent clusters is to visually compare the emotional distribution for each cluster. We can sketch histograms of emotional labels corresponding to clusters to evaluate the separations between different latent clusters according to the ground truth annotations. Figure 3 is an example graph which shows the arousal distribution for two different clusters obtained with supervised and unsupervised models. The unsupervised model is trained by removing the emotional regressor in Figure 1. The goal is to evaluate the role of the regressor network in the clusters obtained by the DeepEmoCluster framework. To simplify the figure, we only present the clusters with the highest (red) and lowest (blue) separation. For the unsupervised implementation, Figure 3(a) shows only a small difference in the emotional distributions for the clusters with the highest and lowest emotional means. However, the supervised version of the DeepEmoCluster framework presents a larger separation between clusters (Fig. 3(b)). These results show the implicit emotional dependencies induced in the clusters by adding the supervised regression task.

6. CONCLUSIONS

We presented the DeepEmoCluster framework, a new semi-supervised approach that learns better latent representations for SER from labeled and unlabeled sets. The DeepEmoCluster approach is able to construct a latent feature space based on clusters that depend on the emotional content. We achieve this goal by adding a supervised emotional regression task. The maximum latent cluster constraint during the joint training procedure enables our approach to achieve the best prediction scores for arousal and dominance, and competitive performance for valence under a fully supervised training scheme. By applying a semi-supervised training strategy, the DeepEmoCluster approach can further improve the model performance, reaching the best overall recognition accuracy for emotional attributes. A future work of this study is to strengthen the connections between the latent clusters and the target emotions. We can adopt additional information theoretic-based loss function such as mutual information, directly associating the cluster classifier and emotional regressor. Another promising future direction is to build a multimodal DeepEmoCluster based on video, audio and language, forming meaningful behavioral emotional clusters.

7. REFERENCES

[1] N. Fragopanagos and J.G. Taylor, “Emotion recognition in human-computer interaction,” *Neural Network*, vol. 18, no. 4, pp. 389–405, May 2005.

[2] C. Milne and T. Otieno, “Understanding engagement: Science demonstrations and emotional energy,” *Science Education*, vol. 91, no. 4, pp. 523–553, June 2007.

[3] M. Sidorov and W. Minker, “Emotion recognition and depression diagnosis by acoustic and visual features: A multimodal approach,” in *International Workshop on Audio/Visual Emotion Challenge (AVEC 2014)*, Orlando, FL, USA, November 2014, pp. 81–86.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in *Advances in Neural Information Processing Systems (NIPS 2012)*, Lake Tahoe, CA, USA, Dec. 2012, vol. 25, pp. 1097–1105.

[5] H. Sak, A. Senior, and F. Beaufays, “Long short-term memory recurrent neural network architectures for large scale acoustic modeling,” in *Interspeech 2014*, Singapore, September 2014, pp. 338–342.

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

[7] C. Busso, M. Bulut, and S.S. Narayanan, “Toward effective automatic recognition systems of emotion in speech,” in *Social emotions in nature and artifact: emotions in human and human-computer interaction*, J. Gratch and S. Marsella, Eds., pp. 110–127. Oxford University Press, New York, NY, USA, November 2013.

[8] C. Busso and S.S. Narayanan, “Scripted dialogs versus improvisation: Lessons learned about emotional elicitation techniques from the IEMOCAP database,” in *Interspeech 2008 - Eurospeech*, Brisbane, Australia, Sep. 2008, pp. 1670–1673.

[9] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An ASR corpus based on public domain audio books,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2015)*, South Brisbane, QLD, Australia, April 2015, pp. 5206–5210.

[10] 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, December 2018.

[11] J. Gideon, M. McInnis, and E. Mower Provost, “Improving cross-corpus speech emotion recognition with adversarial discriminative domain generalization (ADDoG),” *IEEE Transactions on Affective Computing*, 2020.

[12] F. Bao, M. Neumann, and N.T. Vu, “CycleGAN-based emotion style transfer as data augmentation for speech emotion recognition,” in *Interspeech 2019*, Graz, Austria, September 2019, pp. 2828–2832.

[13] A. Chatziagapi *et al.*, “Data augmentation using gans for speech emotion recognition,” in *Interspeech 2019*, Graz, Austria, September 2019, pp. 171–175.

[14] O. Chapelle, B. Schlkopf, and A. Zien, *Semi-Supervised Learning*, MIT Press, Cambridge, Mass., USA, Sep. 2006.

[15] V. Verma *et al.*, “Interpolation consistency training for semi-supervised learning,” in *International Joint Conference on Artificial Intelligence (IJCAI 2019)*, Macao, China, August 2019, pp. 3635–3641.

[16] M. Caron *et al.*, “Deep clustering for unsupervised learning of visual features,” in *European Conference on Computer Vision (ECCV 2018)*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., vol. 11217 of *Lecture Notes in Computer Science*, pp. 139–156. Springer Berlin Heidelberg, Munich, Germany, September 2018.

[17] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *International Conference on Learning Representations (ICLR 2015)*, San Juan, Puerto Rico, May 2015, pp. 1–10.

[18] 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.

[19] G. Trigeorgis *et al.*, “Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, Shanghai, China, March 2016, pp. 5200–5204.

[20] A. Satt, S. Rozenberg, and R. Hoory, “Efficient emotion recognition from speech using deep learning on spectrograms,” in *Interspeech 2017*, Stockholm, Sweden, August 2017, pp. 1089–1093.

[21] Y. Li, T. Zhao, and T. Kawahara, “Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning,” in *Interspeech 2019*, Graz, Austria, September 2019, pp. 2803–2807.

[22] F. Eyben *et al.*, “The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing,” *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, April-June 2016.

[23] J. Deng, X. Xu, Z. Zhang, S. Frühholz, and B. Schuller, “Semisupervised autoencoders for speech emotion recognition,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 31–43, January 2018.

[24] S. Latif, R. Rana, J. Qadir, and J. Epps, “Variational autoencoders for learning latent representations of speech emotion,” in *Interspeech 2018*, Hyderabad, India, September 2018, pp. 3107–3111.

[25] S. Parthasarathy and C. Busso, “Semi-supervised speech emotion recognition with ladder networks,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2697–2709, September 2020.

[26] S. Parthasarathy and C. Busso, “Ladder networks for emotion recognition: Using unsupervised auxiliary tasks to improve predictions of emotional attributes,” in *Interspeech 2018*, Hyderabad, India, September 2018, pp. 3698–3702.

[27] A. Rasmusi *et al.*, “Semi-supervised learning with ladder networks,” in *Advances in neural information processing systems (NIPS 2015)*, Montreal, Canada, December 2015, pp. 3546–3554.

[28] S. Mariooryad, R. Lotfian, and C. Busso, “Building a naturalistic emotional speech corpus by retrieving expressive behaviors from existing speech corpora,” in *Interspeech 2014*, Singapore, September 2014, pp. 238–242.

[29] A. Burmania, S. Parthasarathy, and C. Busso, “Increasing the reliability of crowdsourcing evaluations using online quality assessment,” *IEEE Transactions on Affective Computing*, vol. 7, no. 4, pp. 374–388, October-December 2016.

[30] B. McFee *et al.*, “librosa: Audio and music signal analysis in python,” in *Python in Science Conference (SciPy 2015)*, Austin, TX, USA, July 2015, pp. 18–25.

[31] W.-C. Lin and C. Busso, “An efficient temporal modeling approach for speech emotion recognition by mapping varied duration sentences into fixed number of chunks,” in *Interspeech 2020*, Shanghai, China, October 2020, pp. 2322–2326.