Unsupervised Regularization-Based Adaptive Training for Speech Recognition

Fenglin Ding, Wu Guo, Bin Gu, Zhenhua Ling, Jun Du

National Engineering Laboratory for Speech and Language Information Processing
University of Science and Technology of China, Hefei, China
{flding,bin2801}@mail.ustc.edu.cn, {guowu,zhling,jundu}@ustc.edu.cn

Abstract

In this paper, we propose two novel regularization-based speaker adaptive training approaches for connectionist temporal classification (CTC) based speech recognition. The first method is center loss (CL) regularization, which is used to penalize the distances between the embeddings of different speakers and the only center. The second method is speaker variance loss (SVL) regularization in which we directly minimize the speaker interclass variance during model training. Both methods achieve the purpose of training an adaptive model on the fly by adding regularization terms to the training loss function. Our experiment on the AISHELL-1 Mandarin recognition task shows that both methods are effective at adapting the CTC model without requiring any specific fine-tuning or additional complexity, achieving character error rate improvements of up to 8.1% and 8.6% over the speaker independent (SI) model, respectively.
Index Terms: speaker adaptive training, regularization, speech recognition, connectionist temporal classification

1. Introduction

Mismatches between training and testing conditions are a common problem in modern pattern recognition systems. It is particularly critical in perceptual sequence learning tasks such as automatic speech recognition (ASR) and speech emotion recognition (SER). For example, the performance of deep neural network (DNN) based ASR [1, 2] systems experience mismatches between training and testing conditions, which are caused by the different characteristics of acoustic variability such as speakers, channels and environmental noises. Adaptation techniques to transform a model to match the testing condition or augment the inputs to match a model have been investigated. In ASR, speaker adaptation (SA) techniques are used to minimize the mismatch between the training and testing conditions due to the speaker variability.

Speaker adaptation techniques for DNN based ASR can be categorized into two broad approaches: feature space and model space adaptation. In feature space adaptation, the traditional technique is to transform the acoustic features to a normalized space and then the adapted features are used to train the acoustic model. The maximum likelihood linear regression (MLLR) and its feature-space variant (fMLLR) [3, 4] are two of the most widely used methods. For a deep neural network (DNN) based acoustic model, another effective method is to provide the network with auxiliary features that characterize speaker information to perform adaptation such as the i-vector [5, 6, 7, 8] and speaker code [9, 10]. In model space adaptation, speaker dependent (SD) parameters are estimated from a trained speaker independent (SI) model using additional adaptation data. The DNN Adaptation techniques can also be categorized into two broad approaches: regularized adaptation and subspace or subset adaptation. For model adaption, a straightforward idea is to retrain all the SI model parameters. To avoid overfitting, regularization approaches such as L2 regularization using a weight decay [11], the Kullback-Leibler divergence (KLD) [12] and adversarial multitask learning (MTL) [13] have been proposed. There are also many approaches that have been proposed in which small subsets of the network parameters are adapted [14, 15, 16].Linear transformations, which augment the SI network with certain speaker-specific linear layer(s), including linear input network (LIN) [17, 18], linear hidden network (LHN) [19] and linear output network (LOH) [18], were investigated. Furthermore, parameterized hidden activation functions have also been widely explored [20, 21, 22] and have achieved good performances.

Recently, researchers began training adaptive models on the fly instead of estimating the adaptive parameters from a well-trained SI model [23, 24]. In such approaches, SD auxiliary networks are adopted to improve adaptive training and are jointly optimized with the main network. These methods greatly simplify model adaptation by using only one-pass training and not requiring additional adaptation data.

Although the methods using SD auxiliary networks [23, 24] make adaptive training easier, they usually add an extra burden to the acoustic model. On the other hand, the regularization-based adaptation techniques in [11, 12, 13] do not require additional processing. Inspired by the work mentioned above, we integrate the regularization approaches into adaptive training and propose two novel regularization-based speaker adaptive training methods. The first method is center loss (CL) [25] regularization, where the center loss is used to penalize the distances between the embeddings of different speakers and the only center of all speaker classes. For the second method, we propose a novel regular loss function called the speaker variance loss (SVL). We directly minimize the speaker interclass variance during model training by using SVL regularization.

The essential idea of both proposed methods is to adapt the speaker variability by encouraging speaker interclass compactness, which measures the degree of mismatches. Both methods achieve the purpose of training an adaptive model on the fly by adding regularization terms to the training loss function. More importantly, they hardly add any complexity to the model: the CL only increases the number of parameters of one vector while the SVL does not increase any number of parameters. Considering that there is limited work on speaker adaptive training for the connectionist temporal classification (CTC) [26] model, we applied the proposed methods to CTC-based ASR in this paper. The experiments are conducted on the public Chinese dataset AISHELL-1 [27]. The experimental results show that, both methods are effective at speaker adaptation without requiring any specific fine-tuning or additional complexity, achieving up to 8.1% and 8.6% character error rate improvements over the speaker independent (SI) model, respectively.

The rest of this paper is organized as follows. Section 2

gives a brief description of the related work. We introduce the adaptive training approaches we proposed in Section 3. Section 4 shows our experimental setup and other details, including the experimental results. Finally, the discussion and conclusion are presented in Section 5.

2. Relation to prior work

The center loss (CL) [25] was first proposed to learn discriminative features for Face Recognition (FR) tasks. CL encourages intraclass compactness by penalizing the distances between the features of samples and their centers. To avoid the deeply learned features and centers degrading to zeros, researchers adopted the joint supervision of the softmax loss and CL to train neural networks. Via joint supervision, the CL pulls the features in the same class closer to their class center, and the softmax cross-entropy loss separates the features from different categories. It has performed remarkably on various benchmark datasets for face recognition. Since the CL enjoys the same requirement as the softmax loss and needs no complex recombination of the training samples, it can be easily extended to other tasks. Variants of these methods have also been successfully adopted in Speaker Recognition (SR) tasks [28, 29], automatic speech recognition (ASR) [30] tasks and speech emotion recognition (SER) [31] tasks.

Intuitively, the center loss function pulls the deep features of the same class to their corresponding centers. The purpose of speaker adaptive training in ASR is to normalize the speaker variability between the training and testing conditions. In other words, we want the deep features to contain as little speaker information as possible. The center loss must be tailored for the adaptive training; therefore, we use the center loss in this work to penalize the distances between different embeddings of speakers and the only center of all speaker classes. By minimizing such a center loss, different speaker categories approach the same center to achieve the purpose of normalizing speaker variability.

The purpose of adaptive training is to reduce the interspeaker variability of speech. In addition to the proposed center loss function, we further propose a novel loss function called the speaker variance loss (SVL), which directly minimizes the speaker interclass variance. Similar to the center loss, we use the SVL as a regulation term and adopt the joint supervision of the conventional CTC loss and the proposed SVL in model training. In this way, speaker interclass variance can be normalized as much as possible on the premise of ensuring the accurate classification of acoustic features.

Different from previous regularization-based speaker adaptation approaches [11, 12, 13], our proposed methods train the adaptive model on the fly by adding regularization terms to the training loss function. Therefore, we do not need any additional adaptation data to fine-tune the model parameters. This simplifies adaptive training while introducing little additional model complexity.

3. Proposed methods

An illustration of the proposed regularization-based speaker adaptation approach is shown in figure 1. We directly add the regular loss while training the acoustic model instead of using it to prevent overfitting when estimating speaker dependent (SD) parameters. The regularization encourages the speaker interclass compactness while the CTC loss encourages the separability of features. Consequently, the joint supervision of these two loss function minimizes the speaker variability while keeping the features of different classes separable. More details are discussed as follows.

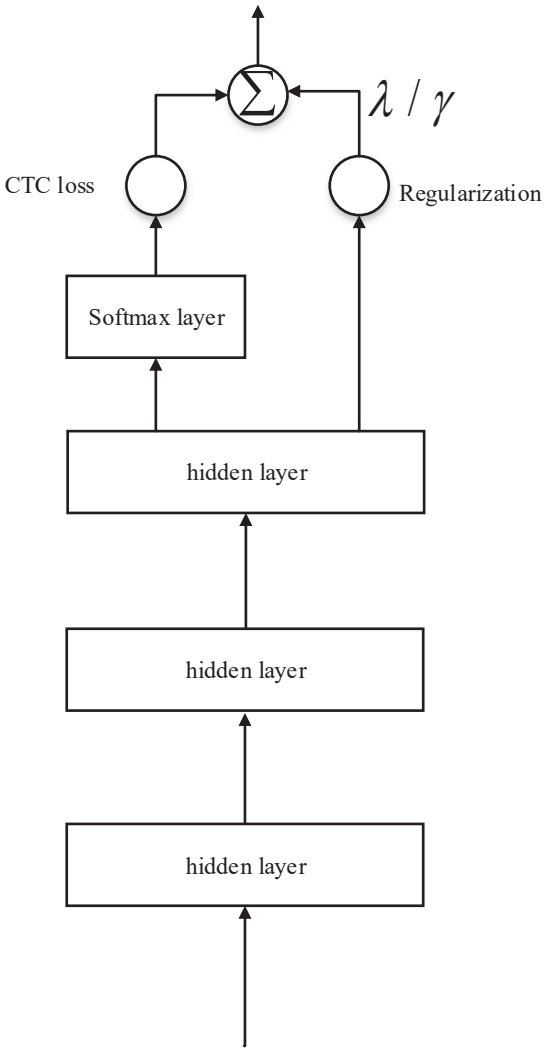

Figure 1: *Illustration of the proposed unsupervised regularization-based speaker adaptive training approach. The scalar λ or γ is used for balancing the CTC loss and the regular loss.*

3.1. Center Loss for adaptive training

Assuming that the training set contains a total of k speakers, we define the center loss function for adaptive training as follows:

$$\mathcal{L}_{\mathrm{c}} = \sum_{i=1}^{k} ||\mathbf{S}_i - \mathbf{C}||_2^2 \tag{1}$$

where $\mathbf{S}_i$ denotes the deep features of speaker i, $\mathbf{C}$ denotes the center of all speaker classes.

When we minimize the center loss, different speaker categories approach the same center, which is beneficial for the final sequence classification. As mentioned in [22], the center should be updated as the deep features change. In other words, we need to use the entire training set in each iteration, which is

inefficient and even impractical. Therefore, we make the necessary modification. Instead of updating the center with respect to the entire training set, we perform the update based on the mini-batch. Under this modification, the k in eq. (1) is redefined as the number of speaker classes within a mini-batch.

Then, the representation of speaker i, $\mathbf{S}_i$, can be easily calculated as follows:

$$\mathbf{S}_i = \frac{1}{\sum_t \mathbf{1}[s_t = i]} \sum_t \mathbf{1}[s_t = i]\mathbf{h}_t \tag{2}$$

where s_t denotes the speaker label of the t^{th} sample in the mini-batch, and $\mathbf{1}[.]$ is the indicator function that evaluates to 1 when its argument holds. $\mathbf{h}_t$ denotes the hidden activation of the second last layer.

We adopt the joint supervision of the classification loss and center loss in order to minimize the speaker variability as much as possible while retaining accurate sequence classification. The formulation is given as follows:

$$\mathcal{L}{=}\mathcal{L}_{\rm ctc} + \lambda\mathcal{L}_c \tag{3}$$

where the scalar λ is used for balancing the two loss functions. $\mathcal{L}_{\rm ctc}$ is the CTC loss function given by the following:

$$\mathcal{L}_{ctc} = -\sum_{(\mathbf{x},\mathbf{z})} \ln(\mathrm{p}(\mathbf{z}|\mathbf{x})) \tag{4}$$

where $(\mathbf{x}, \mathbf{z})$ are the training data pairs.

3.2. Speaker Variance Loss for adaptive training

According to the intuition behind the center loss, we propose a novel speaker variance loss (SVL) for speaker adaptation, which directly minimizes the speaker interclass variance. The formula is given as follows:

$$\mathcal{L}_{\rm sv} = ||\mathrm{var}(S_1, ..., S_i, ...S_k)||_2^2 \tag{5}$$

where $\mathrm{var}(S_1, ..., S_i, ...S_k)$ denotes the interclass variance of the k speakers. We then take into account the modification to update the mini-batch. The interclass variance can be calculated as follows:

$$\sigma_s^2 = \frac{1}{k}\sum_i (S_i - \mu_s)^2 \tag{6}$$

where μ_s is the mean of the k speaker classes. It is given by the following:

$$\mu_s = \frac{1}{k}\sum_i S_i \tag{7}$$

Then, the joint supervision of the CTC loss and SVL can be given as follows:

$$\mathcal{L}{=}\mathcal{L}_{\rm ctc} + \gamma\mathcal{L}_{sv} \tag{8}$$

where γ is the balance factor.

In fact, the SVL effectively characterizes the interclass variations of speakers. Compared with the modified center loss, the SVL targets more directly on the learning objective of the speaker interclass compactness, which is very beneficial for reducing speaker variability. More importantly, the SVL does not introduce any learnable parameters. In this way, it may avoid the local optimization caused by the random initialization of parameters to a certain extent.

Note that both proposed loss functions only take the hidden activation and do not need a complex recombination of the training samples. Therefore, their application for neural networks is more flexible. For example, the hidden activation used to calculate the speaker representation can be taken from any certain layer. In addition, each layer can add the regular loss direction to achieve layer-wise speaker adaptation. These will be discussed in detail in the experimental part.

4. Experiments

4.1. Dataset

We evaluate the proposed methods on an open-source Mandarin speech corpus AISHELL-1 [23]. All the speech files are sampled at 16K Hz with 16 bits. AISHELL-1 has 7,176 utterances from 20 speakers for evaluation (10 hours). We use 120,098 utterances from 340 speakers (150 hours) as the training set and 14,326 utterances from 40 speakers (20 hours) as the development set. The speakers of the training, development and test sets do not overlap.

4.2. Model setup

The PyTorch toolkit [32] is used in our model training process. All the model parameters are randomly initialized and updated by Adam [33]. The acoustic feature is 108-dimensional filter-bank features (36 filter-bank features, delta coefficients, and delta-delta coefficients) with mean and variance normalization. According to the statistical information of the transcripts, there are 4294 Chinese characters in the training set. Along with the added blanks, 4295 modeling units are used in the grapheme-based CTC system. The trigram language model is used in the decoding procedure.

The network is trained to minimize the CTC loss function with an initial learning rate of 0.0001. The development set is used for learning rate scheduling and early stopping. We start to halve the learning rate when the relative improvement falls below 0.004, and the training ends if the relative improvement is lower than 0.0005, which is usually approximately 13 epochs.

4.3. Network architecture

The acoustic modeling adopts a combination of CNNs and LSTM based RNNs for good performance as well as high efficiency. For this baseline, the bottom two layers are 2D convolution layers with 64 and 256 output channels. Each convolution layer is followed by a max-pooling layer with a stride of 2 in the time dimension to finally down sample an utterance to a quarter of its original length. After the CNN layers, there are three LSTM layers, each of which is a bidirectional LSTM layer with 512 units. We also use a dropout rate of 0.3 for the LSTM layers to avoid overfitting.

4.4. Results

We first investigate the sensitiveness of the balance factor of the standard CL regularization and SVL regularization in which only the last LSTM layer is adapted.

Table 1 shows the character error rate (CER) of the adaptive model with CL regularization and SVL regularization under different hyper parameters λ and γ. As shown in the table, at first, as the balance factor increases, the CER gradually decreases; and then when the balance factor continues to increase, the CER gradually increases. It is speculated that when the interclass variance is excessively penalized, some features become indistinguishable, which is not conducive to sequence classification. Finally, with only the third layer punished, the CL and SVL adaptive models achieve CER reductions of 5.6% and 5.8% reduction over SI model, respectively.

Table 1: *The CERs (%) of the CL and SVL adaptive models under different balance factors.*

CL		SVL	
λ	CER(%)	γ	CER(%)
0	9.96	0	9.96
0.005	9.88	1	9.91
0.01	9.69	10	9.64
0.1	**9.40**	25	**9.38**
1	9.68	30	9.43
5	9.92	50	9.80

Table 2: *The CERs (%) of the CL and SVL models with different adapted LSTM layers.*

Adapted LSTM layer	CL	SVL
SI	9.96	9.96
1	9.55	9.55
2	9.46	9.49
3	9.40	9.38
2,3	9.31	9.30
1,2,3	**9.15**	**9.10**

In the following experiments, we investigate how many and which hidden layers should be used in adaptive training. The combinations of different adaptation layers are investigated. The results are summarized in Table 2. Note that each layer corresponds to a separate center for multilayer adaptation in the CL adaptive model. For each adaptive model, the balance factor has been adjusted to achieve the best performance. It can be seen that the highest layer is most important for adaptation. If only one LSTM layer is adapted, the higher layer (3rd layer) can achieve better performance than the lower layer (1st layer). Furthermore, the CER steadily decreases as the number of adapted layers increases. When all three LSTM layers are used for adaptive training, the proposed CL and SVL adaptive models achieve CER reductions of 8.1% and 8.6% over the SI model, respectively.

5. Conclusions

In this work, we propose two novel regularization-based speaker adaptation approaches, center loss (CL) regularization and speaker variance loss (SVL) regularization. The idea of the proposed methods is to reduce the speaker variability by encouraging speaker interclass compactness, which measures the degree of mismatches. Different from previous work, both methods train an adaptive model on the fly by adding regularization terms to the training loss function. Moreover, they hardly add any complexity to the acoustic model: the CL only increases the number of parameters of one vector while the SVL does not increase any number of parameters. The experimental results show that both methods are effective at adapting the CTC model, achieving CER improvements of up to 8.1% and 8.6% over the SI model, respectively.

In the following work, we will investigate how to achieve more effective speaker representation for calculating the regular loss. Attention mechanisms and other schemes may be introduced to enhance the hidden activation used to extract speaker embeddings.

6. Acknowledgements

This work was partially funded by the National Key Research and Development Program of China (Grant No. 2016YFB100 1303) and the National Natural Science Foundation of China (Grant No. U1836219).

7. References

[1] G. Hinton, L. Deng, D. Yu *et al.*, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal processing magazine*, vol. 29, 2012.

[2] D. Yu and J. Li, "Recent progresses in deep learning based acoustic models," *IEEE/CAA Journal of Automatica Sinica*, vol. 4, no. 3, pp. 396–409, 2017.

[3] A.-r. Mohamed, T. N. Sainath, G. E. Dahl *et al.*, "Deep belief networks using discriminative features for phone recognition." in *ICASSP*, 2011, pp. 5060–5063.

[4] F. Seide, G. Li, X. Chen, and D. Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*. IEEE, 2011, pp. 24–29.

[5] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*. IEEE, 2013, pp. 55–59.

[6] Y. Miao, H. Zhang, and F. Metze, "Towards speaker adaptive training of deep neural network acoustic models," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[7] A. Senior and I. Lopez-Moreno, "Improving dnn speaker independence with i-vector inputs," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 225–229.

[8] P. Karanasou, Y. Wang, M. J. Gales, and P. C. Woodland, "Adaptation of deep neural network acoustic models using factorised i-vectors," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[9] O. Abdel-Hamid and H. Jiang, "Fast speaker adaptation of hybrid nn/hmm model for speech recognition based on discriminative learning of speaker code," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7942–7946.

[10] S. Xue, O. Abdel-Hamid, H. Jiang, and L. Dai, "Direct adaptation of hybrid dnn/hmm model for fast speaker adaptation in lvcsr based on speaker code," in *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2014, pp. 6339–6343.

[11] L. Hank, "Speaker adaptation of context dependent deep neural networks," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7947–7951.

[12] D. Yu, K. Yao, H. Su *et al.*, "Kl-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7893–7897.

[13] Z. Meng, J. Li, and Y. Gong, "Adversarial speaker adaptation," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5721–5725.

[14] K. Yao, D. Yu, F. Seide *et al.*, "Adaptation of context-dependent deep neural networks for automatic speech recognition," in *2012 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2012, pp. 366–369.

[15] S. M. Siniscalchi, J. Li, and C.-H. Lee, "Hermitian polynomial for speaker adaptation of connectionist speech recognition systems," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2152–2161, 2013.

[16] L. Samarakoon and K. C. Sim, "Factorized hidden layer adaptation for deep neural network based acoustic modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2241–2250, 2016.

[17] J. Neto, L. Almeida, M. Hochberg, C. Martins, L. Nunes, S. Renals, and T. Robinson, "Speaker-adaptation for hybrid hmm-ann continuous speech recognition system," 1995.

[18] B. Li and K. C. Sim, "Comparison of discriminative input and output transformations for speaker adaptation in the hybrid nn/hmm systems," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[19] R. Gemello, F. Mana, S. Scanzio, P. Laface, and R. De Mori, "Linear hidden transformations for adaptation of hybrid ann/hmm models," *Speech Communication*, vol. 49, no. 10-11, pp. 827–835, 2007.

[20] P. Swietojanski, J. Li, and S. Renals, "Learning hidden unit contributions for unsupervised acoustic model adaptation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 8, pp. 1450–1463, 2016.

[21] P. Swietojanski and S. Renals, "Differentiable pooling for unsupervised speaker adaptation," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4305–4309.

[22] C. Zhang and P. C. Woodland, "Dnn speaker adaptation using parameterised sigmoid and relu hidden activation functions," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5300–5304.

[23] F. Ding, W. Guo, L. Dai, and J. Du, "Attention-based gated scaling adaptative acoustic model for ctc-based speech recognition," *arXiv preprint arXiv:1912.13307*, 2019.

[24] L. Sarı, S. Thomas, and M. A. Hasegawa-Johnson, "Learning speaker aware offsets for speaker adaptation of neural networks," *Proc. Interspeech 2019*, pp. 769–773, 2019.

[25] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *European conference on computer vision*. Springer, 2016, pp. 499–515.

[26] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 369–376.

[27] H. Bu, J. Du, X. Na *et al.*, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," in *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*. IEEE, 2017, pp. 1–5.

[28] S. Yadav and A. Rai, "Learning discriminative features for speaker identification and verification." in *Interspeech*, 2018, pp. 2237–2241.

[29] N. Li, D. Tuo, D. Su, Z. Li, D. Yu, and A. Tencent, "Deep discriminative embeddings for duration robust speaker verification." in *Interspeech*, 2018, pp. 2262–2266.

[30] J. Wang, D. Su, J. Chen, S. Feng, D. Ma, N. Li, and D. Yu, "Learning discriminative features in sequence training without requiring framewise labelled data," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5696–5700.

[31] D. Dai, Z. Wu, R. Li, X. Wu, J. Jia, and H. Meng, "Learning discriminative features from spectrograms using center loss for speech emotion recognition," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7405–7409.

[32] A. Paszke, S. Gross, S. Chintala *et al.*, "Automatic differentiation in pytorch," 2017.

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