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

Analysis of Deep Learning Architectures for Cross-corpus Speech Emotion Recognition

Jack Parry, Dimitri Palaz, Georgia Clarke, Pauline Lecomte, Rebecca Mead, Michael Berger, Gregor Hofer

Speech Graphics Ltd, Edinburgh, United Kingdom
{jack.parry,dpalaz,g.clarke,lecomte,mead,berger,hofer}@speech-graphics.com

Abstract

Speech Emotion Recognition (SER) is an important and challenging task for human-computer interaction. In the literature deep learning architectures have been shown to yield state-of-the-art performance on this task when the model is trained and evaluated on the same corpus. However, prior work has indicated that such systems often yield poor performance on unseen data. To improve the generalisation capabilities of emotion recognition systems one possible approach is cross-corpus training, which consists of training the model on an aggregation of different corpora. In this paper we present an analysis of the generalisation capability of deep learning models using cross-corpus training with six different speech emotion corpora. We evaluate the models on an unseen corpus and analyse the learned representations using the t-SNE algorithm, showing that architectures based on recurrent neural networks are prone to overfit the corpora present in the training set, while architectures based on convolutional neural networks (CNNs) show better generalisation capabilities. These findings indicate that (1) cross-corpus training is a promising approach for improving generalisation and (2) CNNs should be the architecture of choice for this approach.

Index Terms: speech emotion recognition, neural networks

1. Introduction

Speech Emotion Recognition (SER) has seen a growing number of applications in recent years. An important application is human-computer interaction, typically in the context of conversational agents. Users of agents such as Siri or Google Assistant will attest that these systems lack relatability and fail to elicit empathy from the user. One way to improve the relatability of such systems is to give them the capacity to detect emotion from speech, allowing the system to respond in a more appropriate manner.

Deep learning architectures, such as Convolutional Neural Networks (CNNs) [1] and highway networks [2], have been shown to yield state-of-the-art performance on this task. However, being able to use these models "in the wild" (i.e. on unseen data with varying characteristics) is still an open question because these models seem to generalise poorly [3]. Recently, several approaches have been investigated to improve generalisation. One promising approach is cross-corpus training, which consists of aggregating several corpora to create the training set. This approach is appealing because (1) the diversity and varying contextual factors contained in the training set should help the models learn a robust representation of emotion and thus improve performance, and (2) it allows models to be trained on more data, which should improve generalisation, as has been shown for several pattern recognition tasks, such as image recognition [4] and speech recognition [5]. The main drawback of this approach is that these models still tend to overfit the corpora in the training set and also display poor generalisation capabilities to out-of-domain data [6].

In this paper, we present an analysis of cross-corpus training for speech emotion recognition. We select three common deep learning models based on CNNs and Long Short-Term Memory (LSTM) [7]. We first evaluate the performance of these models trained on a single corpus on the in-domain test set (i.e. on the same corpus) for comparison with the literature, before testing on out-of-domain corpora (i.e. corpora not part of the training set). We then evaluate the performance of the models trained on cross-corpus data. A comparison with single-corpus training shows that the cross-corpus approach improves generalisation on out-of-domain corpora for all model architectures. Then, in order to discern which of the architectures displays the best generalisation capabilities, we present two studies. In the first study, we use an unseen corpus as out-of-domain test set and we show that the LSTM model yields inferior performance compared with CNN models consisting of several convolution and max-pooling layers. In the second study, we apply the t-SNE [8] visualisation technique on the learned representations of each model and show that the LSTM model seems to cluster the data according to corpus rather than emotion.

The main contribution of this paper is to show that deep learning architectures composed of several convolution and max-pooling layers improve the generalisation capabilities of the model, alleviating the issue of corpus overfitting for cross-corpus training.

The remainder of the paper is organised as follows: a literature review is first presented in Section 2; the methodology is then presented in Section 3, including the models and the experimental setup; Section 4 presents the results of the studies, and Section 5 concludes the paper.

2. Related Work

Automatic speech emotion recognition has been an active area of research for decades [16, 17]. Considerable effort was put into designing a relevant set of features, which was used with simple classifiers, like linear classifiers or kNN [18]. This led to "standard" sets of features, like the *GeMAPS* feature set [19]. Recently, the deep learning approach, which is based on complex classifiers that can learn features from raw data, has been shown to drastically improve performance on several pattern recognition tasks [4, 5]. Deep learning models have been shown to yield state-of-the-art performance on the SER task. For instance, learning features from the raw speech using CNNs has been proposed in [20]. A CNN-LSTM model taking spectrograms as input was also proposed in [1]. CNNs with an atten-

http://dx.doi.org/10.21437/Interspeech.2019-2753

Table 1: *Corpus information*

Corpus	**Language**	**# utterances**	**Duration (hh:mm)**	**# speakers**	**Label index** (corresponding emotion in Table 4)
EMOVO [9]	Italian	588	00:31	6	1,6,7,11,16,19,23
Emo-DB [10]	German	535	00:25	10	1,6,7,11,15,20,23
IEMOCAP [11]	English	7529	09:32	10	1,6,7,8,11,14,15,19,23
EPST [12]	English	2409	01:00	13	2,3,4,5,6,9,10,11,12,13,15,18,20,22,23
RAVDESS [13]	English	2542	02:47	24	1,6,7,11,15,19,21,23
SAVEE [14]	English	476	00:30	4	1,6,7,11,15,19,23
TESS [15]	English	2800	01:36	2	1,6,7,11,15,17,23

tion mechanism were investigated in [21] and [2] proposed an architecture composed of convolutional highway networks and LSTMs.

SER models trained on a single corpus tend to overfit, leading to poor performance on out-of-domain data, as presented in [3]. To address this issue, several techniques have been proposed: (1) the data augmentation approach, which consists of generating additional training samples by duplicating and often modifying the original training set, using techniques such as vocal tract length perturbation [22] or variation of tempo, loudness and background noise [23]; (2) multi-task learning, in which the models are trained on additional tasks, such as gender or domain identification [24, 25]; (3) the transfer learning approach, in which the models are first trained on a given domain and then adapted to the task at hand [26, 27]; and (4) cross-modal transfer, in which an image-based emotion recognition model is used to improve SER [28].

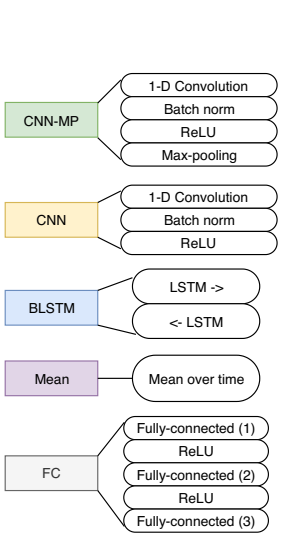
Finally, cross-corpus SER has been studied to improve generalisation. It was shown in [3] that training models on aggregated corpora improves the performance, but leads to overfitting the training set. This approach has been investigated in conjunction with multi-task learning using Extreme Learning Machine (ELM) [25]. In [29], a data augmentation technique based on mixing up samples using an LSTM model was proposed for valence-arousal prediction, showing only a limited gain in addressing the overfit issue. Feature normalisation strategies are presented in [30], where low level descriptors undergo a cascade of normalisation, including speaker-level and feature vector-level, and are used as input to ELM models. This approach seems to improve generalisation using low-level features, however it is unclear if neural network models using high-level features such as filterbank could benefit from these techniques.

3. Methodology

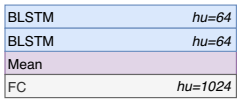
3.1. Models

The models used for the experiments and their respective hyperparameters are described in Figures 1b, 1c and 1d as a sequence of modules, which are described in Figure 1a.

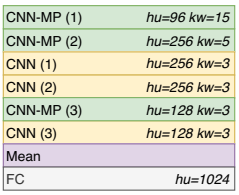
These architectures were chosen due to their prevalence in the literature and proven effectiveness at the task at hand. All three models are capable of taking arbitrarily long sequences as input. The process of classifying an utterance is similar for each model: the initial layers are designed to extract important local features of the input; the mean layer combines this information to produce a dense, global representation of the sequence; the final module uses this dense representation to classify the sequence.



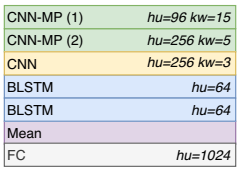
(a) *Network modules*



(b) *LSTM model*



(c) *CNN model*



(d) *CNN-LSTM model*

Figure 1: *Model architectures.* kw *denotes the kernel width,* dw *denotes the stride and* hu *the number of hidden units. All max-pooling layers have* $kw = 3$ *and* $dw = 2$*.*

3.2. Cross-corpus training

The corpora used in this paper are summarised in Table 1, showing relevant information for each corpus. More details can be found in their respective papers. Each corpus is split between training (80%), validation (10%) and test sets (10%). For all corpora, the speakers in the validation and test sets do not appear in the training set.

For the cross-corpus approach, all corpora excluding TESS were aggregated together, yielding 11hrs 45min for training, 1hr 30min each for validation and testing. The TESS corpus was reserved as a out-of-domain test set. Each corpus has its own distinct set of emotion labels. In order to avoid discarding data, we elected to train on three classes: negative, positive and neutral. We thus mapped the labels provided with each corpus to this set, described in Table 4.

3.3. Experimental setup

We used Mel filterbank coefficients as input features, computed using the Kaldi toolbox [31]. These features consist of 40 coefficients computed on a 25ms window with a 10ms shift and no

Table 2: *Performance on each corpus test set for models trained only on IEMOCAP. Avg. denotes the average on out-of-domain corpora.*

Model	Unweighted Accuracy [%]							
	IEMOCAP	EMOVO*	Emo-DB*	EPST*	RAVDESS*	SAVEE*	TESS*	Avg.*
LSTM	46.11	33.33	33.33	33.33	33.33	33.33	33.33	33.33
CNN	46.18	33.33	33.33	33.33	33.33	33.33	33.33	33.33
CNN-LSTM	**51.45**	33.33	**41.99**	30.01	33.04	33.33	**36.92**	34.77

**out-of-domain corpora.*

Table 3: *Performance on each corpus test set for cross-corpus training. Avg. denotes the average on in-domain corpora.*

Model	Unweighted Accuracy [%]							
	IEMOCAP	EMOVO	Emo-DB	EPST	RAVDESS	SAVEE	Avg.	TESS*
LSTM	47.45	38.00	59.67	50.68	53.97	60.62	50.69	45.10
CNN	46.86	39.93	58.86	48.78	**65.67**	70.57	**55.11**	48.94
CNN-LSTM	**50.31**	**53.24**	**69.72**	**51.81**	53.08	**72.66**	53.35	**49.48**

**TESS is an out-of-domain corpus.*

Table 4: *Emotion label mapping*

Mapping	Original label
Negative	anger (1), anxiety (2), cold anger (3), contempt (4), despair (5), disgust (6), fear (7), frustration (8), hot anger (9), panic (10), sadness (11), shame (12)
Positive	elation (13), excitement (14), happiness (15), joy (16), pleasant surprise (17), pride (18), surprise (19)
Neutral	boredom (20), calm (21), interest (22), neutral (23)

speed or acceleration coefficients. Features were normalised to zero-mean and unit variance by utterance.

Models were trained using stochastic gradient descent with momentum of 0.9 and a learning rate of 0.01. The loss function was cross entropy. Early stopping based on the validation error rate was used to select the best model. All experiments were implemented with PyTorch [32]. Utterances were padded to a minimum length of 100 frames (1 second of audio). All three models have approximately 2 million total network parameters.

In the studies we use two metrics: the Weighted Accuracy (WA), the overall accuracy across all classes, and the Unweighted Accuracy (UA), the average of the accuracy for each of the classes.

4. Results

4.1. Single corpus training

The first study is focused on single-corpus training in order to validate the selected architecture and serve as comparison for cross-corpus training. We selected the IEMOCAP corpus [11] to this aim, as it has the most data amongst the selected corpora and it is widely used in the literature.

We first present a comparison with the literature. In order to provide a fair comparison with other research, for this single-corpus experiment only, the corpus was cut to include four emotions (anger, happiness, neutral and sadness) and all models were trained on this label set. The performance of the models trained on IEMOCAP is presented in Table 5 using the WA and UA metrics, along with a comparison with the literature. One can see that the performance of the models used in this paper is on par with recently published work on the IEMOCAP corpus. Note that the models' hyper-parameters were not specifically optimised for this task, which probably explains the lower accuracy of the CNN and CNN-LSTM models.

Table 5: *Performance on IEMOCAP testset for models trained on IEMOCAP using the 4 emotions label set.*

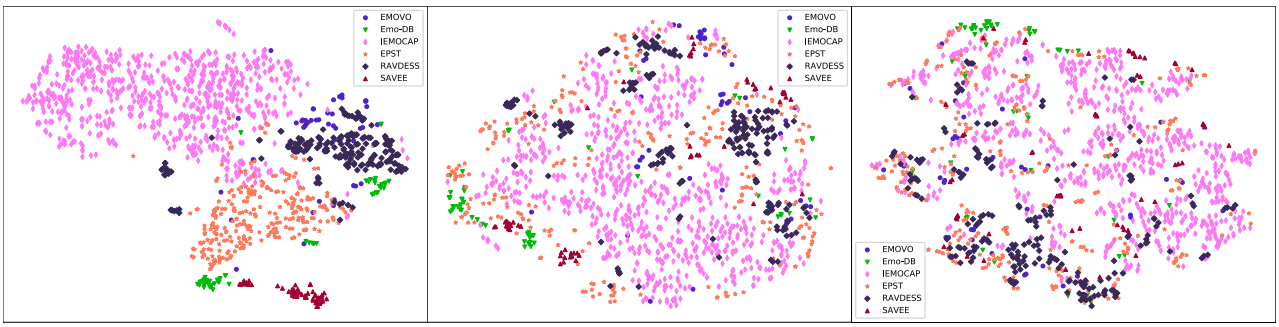
Model	WA [%]	UA [%]
RNN mean pool [33]	56.90	55.30
DNN-ELM [34]	55.00	49.50
LSTM	56.99	53.07
CNN	55.24	49.16
CNN-LSTM	50.17	41.57

We then evaluate the generalisation capability of the models trained only on IEMOCAP. We re-train these models to output three emotions according to the mapping in Table 4 and report performance on the out-of-domain corpora. The results are presented in Table 2 on the WA and UA metrics. One can see that the unweighted accuracy for the out-of-domain corpora is very poor, which is in line with previous works [6]. Note that in Table 2, 33.33% UA means that the whole test set is classified as *negative*. This shows that none of these architectures, when trained on a single corpus, generalise at all to out-of-domain data.

4.2. Cross-corpus training

In this study we present the results of cross-corpus training as described in Section 3.2. The performance of the three different models on each of the test sets is presented in Table 3 using the UA metric. On the in-domain corpora, one can see that the cross-corpus training greatly improves the performance for all models. Additionally, the performance on IEMOCAP is similar to single-corpus training (see Table 2).

On the in-domain corpora, The CNN and CNN-LSTM models achieve higher performance than the LSTM model, suggesting that CNN layers are beneficial for the task. On the out-



(a) *LSTM model, corpus-based colouring* (b) *CNN model, corpus-based colouring* (c) *CNN-LSTM model, corpus-based colouring*

(d) *LSTM model, emotion-based colouring* (e) *CNN model, emotion-based colouring* (f) *CNN-LSTM model, emotion-based colouring*

Figure 2: *t-SNE visualisation of trained models.*

of-domain corpus (TESS), the CNN and CNN-LSTM models are very close in performance and both yield higher UA than the LSTM model by around 4% in absolute terms. This suggests that the generalisation capabilities of the CNN-based models are superior compared to the LSTM model.

4.2.1. Visualisation study

In an attempt to understand why the CNN and CNN-LSTM models performed better than the LSTM model on the out-of-domain test set, we present a visualisation of the learned representation of each model. For this we use the t-Distributed Stochastic Neighbor Embedding (t-SNE) [8] technique, allowing us to visualise high-dimensional data in a two dimensional space. We apply the t-SNE to the hidden representation of the last hidden layer of the *FC* modules, after mean-over-time aggregation. Intuitively, at this stage of the network, some clustering based on the emotion labels should be seen. Figure 2 presents the t-SNE plots computed on the aggregated testset. For each model, we present two plots: above, the points are coloured based on the corpus; below, the points are coloured according to the emotion.

On the corpus-coloured plots, the LSTM has formed clusters of data that belong to the same corpus, clearly indicating that the model is overfitting the training corpora. The dense representations taken from the CNN and CNN-LSTM model appear to display more corpus invariance, as points belonging to same corpus are more spread. This analysis shows that CNN and CNN-LSTM models display more generalisation capabilities and could explain why they are better able to generalise to the out-of-domain test data.

This approach can be a useful tool to gain insight into neural networks and potentially identify their weaknesses. For instance, one could easily colour the data points according to speaker, gender or any other variable, to check for overfitting.

4.3. Discussion

The cross-corpus experiments found that the generalisation capabilities of the LSTM model were well below that of the CNN model, as shown by the performance on the out-of-domain test set and the t-SNE visualisation. A possible explanation is the phenomenon of memory loss due to vanishing gradients in the LSTM. This renders the model unable to capture the global features of the utterance, which may be critical to accurate emotion recognition. It appears that the LSTM model is sensitive to variations in the channel conditions and other features specific to the different corpora, perhaps using this information to infer the emotion based on the distribution in the corpus. The combination of convolution and max-pooling layers may allow for unimportant local information to be discarded, leaving features that are more relevant for predicting emotion "in the wild". Therefore, the main limitation of cross-corpus training, which is overfitting on the training corpora, is mitigated to some extent using deep learning architectures composed of several convolution and max-pooling layers.

5. Conclusions

In this paper, we presented an analysis of three deep architectures used for cross-corpus speech emotion recognition. We showed that cross-corpus training improves the generalisation capability of these models. We also showed that LSTM-based models are prone to overfitting on the in-domain corpora and that CNN-based models alleviate this issue. Thus CNNs are the architecture of choice for cross-corpus training, leading towards deployable speech emotion recognition that can be used "in the wild". Our future work will investigate the selection of the training data for the cross-corpus set to further improve generalisation, for instance by considering the length of the utterances as well as corpus-based normalisation strategies.

6. References

[1] A. Satt, S. Rozenberg, and R. Hoory, "Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms," in *Proc. of Interspeech*. ISCA, Aug. 2017, pp. 1089–1093.

[2] J. Kim, G. Englebienne, K. P. Truong, and V. Evers, "Deep Temporal Models using Identity Skip-Connections for Speech Emotion Recognition," in *ACM Multimedia*. ACM Press, 2017, pp. 1006–1013.

[3] B. Schuller, Z. Zhang, F. Weninger, and G. Rigoll, "Selecting training data for cross-corpus speech emotion recognition: Prototypicality vs. generalization," in *Proc. Afeka-AVIOS Speech Processing Conference, Tel Aviv, Israel*. Citeseer, 2011.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. of NIPS*, 2012, pp. 1097–1105.

[5] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *Proc. of ICML*, 2016, pp. 173–182.

[6] B. Schuller, B. Vlasenko, F. Eyben, M. Wollmer, 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, Jul. 2010.

[7] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[8] 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.

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

[10] 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.

[11] 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, Dec. 2008.

[12] M. Liberman and al., "Emotional prosody speech and transcripts ldc2002s28," *Web Download. Philadelphia: Linguistic Data Consortium*, 2002.

[13] 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, p. e0196391, May 2018.

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

[15] 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.

[16] K. R. Scherer, "Vocal affect expression: A review and a model for future research." *Psychological bulletin*, vol. 99, no. 2, p. 143, 1986.

[17] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal processing magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[18] C. M. Lee, S. S. Narayanan *et al.*, "Toward detecting emotions in spoken dialogs," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 2, pp. 293–303, 2005.

[19] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. Andre, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "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, Apr. 2016.

[20] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. W. Schuller, and S. P. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," *Proc. of ICASSP*, pp. 5200–5204, 2016.

[21] M. Neumann and N. T. Vu, "Attentive Convolutional Neural Network based Speech Emotion Recognition: A Study on the Impact of Input Features, Signal Length, and Acted Speech," in *Proc. of Interspeech*, Jun. 2017.

[22] C. Etienne, G. Fidanza, A. Petrovskii, L. Devillers, and B. Schmauch, "Speech Emotion Recognition with Data Augmentation and Layer-wise Learning Rate Adjustment," *arXiv:1802.05630 [cs, eess]*, Feb. 2018, arXiv: 1802.05630.

[23] E. Lakomkin, M. A. Zamani, C. Weber, S. Magg, and S. Wermter, "On the Robustness of Speech Emotion Recognition for Human-Robot Interaction with Deep Neural Networks," *Proc. of IROS*, pp. 854–860, Apr. 2018.

[24] B. Zhang, E. M. Provost, and G. Essi, "Cross-corpus acoustic emotion recognition from singing and speaking: A multi-task learning approach," in *Proc. of ICASSP*. IEEE, 2016, pp. 5805–5809.

[25] 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," in *Proc. of Interpseech*, 2017.

[26] Z. Zhang, F. Weninger, M. Wllmer, and B. Schuller, "Unsupervised learning in cross-corpus acoustic emotion recognition," in *Proc. of ASRU*, Dec. 2011, pp. 523–528.

[27] 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, Sep. 2014.

[28] S. Albanie, A. Nagrani, A. Vedaldi, and A. Zisserman, "Emotion recognition in speech using cross-modal transfer in the wild," in *ACM Multimedia*, 2018.

[29] D. Fedotov, H. Kaya, and A. Karpov, "Context Modeling for Cross-Corpus Dimensional Acoustic Emotion Recognition: Challenges and Mixup," in *Speech and Computer*. Springer International Publishing, 2018, pp. 155–165.

[30] H. Kaya and A. A. Karpov, "Efficient and effective strategies for cross-corpus acoustic emotion recognition," *Neurocomputing*, vol. 275, pp. 1028–1034, Jan. 2018.

[31] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," IEEE Signal Processing Society, Tech. Rep., 2011.

[32] 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.

[33] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *Proc. of ICASSP*. IEEE, 2017, pp. 2227–2231.

[34] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Proc. of Interspeech*, 2014.