Deep architecture enhancing robustness to noise, adversarial attacks, and cross-corpus setting for speech emotion recognition

Siddique Latif, Rajib Rana, Sara Khalifa, Raja Jurdak, Björn W. Schuller

Angaben zur Veröffentlichung / Publication details:

Latif, Siddique, Rajib Rana, Sara Khalifa, Raja Jurdak, and Björn W. Schuller. 2020. "Deep architecture enhancing robustness to noise, adversarial attacks, and cross-corpus setting for speech emotion recognition." In *Interspeech 2020, Shanghai, China, 25-29 October 2020*, edited by Helen Meng, Bo Xu, and Thomas Zheng, 2327–31. ISCA. https://doi.org/10.21437/interspeech.2020-3190.

INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Deep Architecture Enhancing Robustness to Noise, Adversarial Attacks, and Cross-corpus Setting for Speech Emotion Recognition

Siddique Latif[1,2], *Rajib Rana*[1], *Sara Khalifa*[2,3], *Raja Jurdak*[4], *Björn W. Schuller*[5,6]

[1]University of Southern Queensland, Australia
[2]Distributed Sensing Systems Group, Data61, CSIRO Australia
[3]University of New South Wales, Australia
[4]Queensland University of Technology, Australia
[5]GLAM – Group on Language, Audio, & Music, Imperial College London, UK
[6]Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
siddique.latif@usq.edu.au

Abstract

Speech emotion recognition systems (SER) can achieve high accuracy when the training and test data are identically distributed, but this assumption is frequently violated in practice and the performance of SER systems plummet against unforeseen data shifts. The design of robust models for accurate SER is challenging, which limits its use in practical applications. In this paper we propose a deeper neural network architecture wherein we fuse Dense Convolutional Network (DenseNet), Long short-term memory (LSTM) and Highway Network to learn powerful discriminative features which are robust to noise. We also propose data augmentation with our network architecture to further improve the robustness. We comprehensively evaluate the architecture coupled with data augmentation against (1) noise, (2) adversarial attacks and (3) cross-corpus settings. Our evaluations on the widely used IEMOCAP and MSP-IMPROV datasets show promising results when compared with existing studies and state-of-the-art models.

Index Terms: speech emotion, mixup, data augmentation, convolutional neural networks, DenseNet, highway network.

1. Introduction

Despite the significant progress in Speech Emotion Recognition (SER) through Deep Neural Networks (DNNs), SER systems still perform poorly in noisy environments [1, 2], and when the imperceptible adversarial perturbation is added to test examples [3]. The performance of state-of-the-art SER also degrades in the cross-corpus setting when an acoustic mismatch between training and testing exists [4]. This shows that SER systems lack robustness and generalisation which makes them susceptible to unknown test data shifts. Researchers have developed various methods to improve the performance of SER in noisy environment [2, 5] and cross-corpus setting [6], however, significant performance improvement is still required.

Robustness in DNNs can be enhanced by utilising very deep architectures. Many layers in a deep architecture allow learning very complex patterns, therefore anomalies through the noise and other conditions imposed by cross-corpus can be isolated relatively easily. Studies on speech have shown that utilisation of very deep architectures has led to robustness against noisy situations by learning complex representations [7, 8]. However, deep networks are difficult to train as it is subject to falling into local extrema, also takes longer time and powerful computational resources, e.g. GPU. Very recently, deep networks like Dense Convolutional Network (DenseNet) [9] and highway networks [10] have gained popularity, as they allow for easy training of very deep feedforward networks with considerably fewer parameters. DenseNets are densely connected Convolutional Networks and Highway networks closely follow the structure of the Long Short-Term Memory (LSTM) through gating mechanisms. Several studies in vision and Automatic Speech Recognition (ASR) have shown the benefit of using DenseNet and highway networks, however, their performances for SER need to be evaluated. SER introduces the temporal dimension differentiating it from images. It is also complicated than ASR, as ASR mainly deals with verbal messages, but SER works on vocal expressions that mesh verbal messages coded arbitrarily and categorically with nonverbal affect signalling system coded in an iconic and continuous fashion [11].

Robustness can also be enhanced using data augmentation techniques, which improve the generalisation and help DNNs to provide robustness against unseen real-time situations [12, 13]. Recently "mixup" [13] shows great promise for data augmentation by augmenting a synthetic sample as a linear combination of the original sample. It can make the training data more diverse and the regularisation effect more powerful, so as to further improve the generalisation ability of the network [14]. Despite its potential, the performance of mixup is not validated for SER.

Besides noise, several studies have shown that SER systems are also susceptible to adversarial attacks and their performance can significantly drop due to the attack [3, 15]. Adversarial attacks are developed by malicious adversaries to craft adversarial examples by the addition of unperceived perturbation to elicit wrong responses from machine learning (ML) models. Methods to achieve robustness against the adversarial attacks in SER systems have not been evaluated.

This paper makes several contributions.

1. We propose a deep SER model built on DenseNet and highway networks for robust SER.
2. To further improve robustness, we propose mixup as a data augmentation method.
3. We comprehensively evaluate the robustness of the proposed model in three distinct settings (a) noisy conditions, (b) adversarial attacks, and (c) cross-corpus scenarios.

2. Related Work

In recent years, several studies in the image domain [16] and in Automatic Speech Recognition (ASR) have used deep archi-

http://dx.doi.org/10.21437/Interspeech.2020-3190

tectures to achieve robustness. In [8], the authors used a very deep convolutional residual network (VDCRN) for noise-robust speech recognition. They empirically showed VDCRN is more robust to noise and able to significantly reduce the word error rate (WER). Studies in vision [17, 16] showed DenseNet is more robust compared to the other state-of-the-art models including ResNet [18] due to its efficient feature reuse ability. Similarly, studies [19, 20] have found that DenseNet can help achieve robustness and generalisability in ASR. However, none of these studies evaluates DenseNet for robust SER. In this work, we modify DenseNet architecture and use it as a feature extractor in our proposed model.

Although we could not find any study using very deep architectures for SER, several studies have considered DNNs to achieve robustness for SER. Huang et al. [21] used a convolutional neural network (CNN) – long-short term memory (LSTM) CNN-LSTM model for robust SER. They found that CNN demonstrates a certain degree of noise robustness. However, this study does not utilise very deep architectures. In [2], the authors utilised deep residual networks as an enhancement architecture to remove noise from speech while preserving enough information for an SER system. This study was focused on speech enhancement instead of robust representation learning. Some studies [1, 22] also explored different noise removal methods for SER in noisy environments.

Data augmentation is a well-known practice to enlarging the size of the training set in many machine-learning applications. Recently, it has been shown that mixup data augmentation can enhance the classifiers' robustness for unseen test data [13]. It also improves generalisation performance and model robustness against adversarial examples [23]. The regularisation effect of mixup has been evaluated for ASR [24], however, no study has evaluated mixup in SER. Here, we use mixup to improve generalisation and robustness of proposed model against the noisy environment, adversarial attacks, and cross-corpus setting.

3. Proposed Model

Our proposed model is a hybrid architecture, where we use a DenseBlock for temporal feature extraction, LSTM for context aggregation and fully connected layers in highway configuration for discriminative feature learning. A schematic diagram of the proposed model is shown in Figure 1.

3.1. Temporal Feature Capturing using DenseNet

The first element of our model is DenseNet. DenseNet enables learning temporal features by introducing direct connections from each layer to all subsequent layers. Consequently, the l^{th} layer (x_l) receives the feature maps of all preceding layers as input:

$$x_l = H_{l,G}([x_o, x_1, ..., x_{l-1}]) \quad (1)$$

Here, $H_{l,G}(.)$ represents to a composite function including batch normalisation (BN) [25], a rectified linear unit (ReLU) [26] and a convolution (Conv) layer. G is the growth rate that represents the number of output feature maps. Cascading multiple layers of composite functions and feature map concatenations forms a so-called DenseBlock (L, G), which has L layers and a growth rate of G. The concatenations (Equation 1) in the DenseBlock causes the input size to be increased as the number of layers increases in DenseBlock. For downsampling, a transition layer is used after each DenseBlock. The transition layer consists of a batch normalisation layer, a 1×1 convolutional layer and a 2×2 average pooling layer.

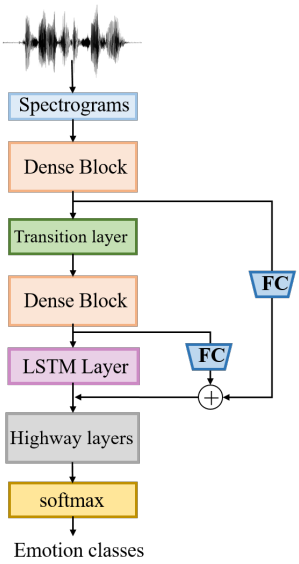

Figure 1: *Schematic diagram of the proposed model.*

In the DenseNet architecture, there is a global average pooling layer that we replace with a reshape layer to adapt the dimension of feature maps to LSTM layer for context learning.

3.2. Emotional-context Modelling using LSTM

Emotions are context-dependent [27] and contexts are embedded in temporal dimension of the data [28]. LSTM [29] have gated architectures, which enables memory, helps to model temporal relationships. We, therefore, use LSTM after the DenseBlock for context modelling.

3.3. Discriminative Feature Extraction using Highway Network

Finally, we use the highway network [10] to extract high-level discriminative features. We also use two skip connections from each DenseBlock and the sum of these features are concatenated with the LSTM features and given to the highway network. Skip connections introduce the shortcut connection that shuttles different levels of abstractions and also improves the gradient flow in the network [30]. The output y of a highway block is given by:

$$y = H(x, W_H).T(x, W_T) + x.C(x, W_C), \quad (2)$$

where H (parameters by W_H) is a nonlinear transformation on its associated input x, and C and T represent carry gates and transform gates, respectively. The layer indices and biases have been excluded in Equation 2 for simplification.

Gates in a highway network control information flow and speed up convergence enabling effective training of DNNs across several layers without degradation [10]. This helps the highway network to learn high-level discriminative representation [31] that help to achieve better classification performance. Therefore, we use the highway network layers prior to softmax layer for classification.

4. Experimental Setup

4.1. Dataset

We evaluate our model on two popular datasets: Interactive Emotional Dyadic Motion Capture (IEMOCAP) [32] and MSP-IMPROV [33]. The detail of these datasets is given below. **IEMOCAP:** This corpus contains five sessions, where each session has utterances from two speakers (one male and one female). Overall, there are 10 unique speakers. We consider four emotions including angry, happy, neutral and sad. To be consistent with previous studies [34], we merge excitement with happiness and consider it as one class: happy. **MSP-IMPROV:** The MSP-IMPROV dataset contains six sessions, where each session comprises of utterances from two speakers, one male, and one female. There are four emotion categories in MSP-IMPROV: angry, neutral, sad, and happy. All were used in the experiments. **DEMAND** We choose the Diverse Environments Multichannel Acoustic Noise Database (DEMAND) dataset [35] as a source of our noise signal. This data contains the recording of various real-world noises in a variety of settings. We select noise recordings of 16 kHz sampling rates.

4.2. Data Pre-processing

We use spectrogram as our starting point. We compute them using Short-Time Fourier Transform (STFT) with an overlapping Hamming window of size 25 ms with a 10 ms shift. We select the height of spectrogram equal to 128. We apply a context window of 256 frames to reach a fixed width of segments of spectrograms following the procedure used in [36]. Each segment is assigned the emotion labelling of the corresponding utterance. We train all the models using short segments, however, utterance level prediction is calculated by averaging the posterior probabilities of the respective sub-segments.

4.3. Data Augmentation

We use the "mixup" data augmentation technique, which has not been used in SER. It creates training samples using following equations:

$$\tilde{x} = \lambda x_i + (1 - \lambda) x_j \quad (3)$$

$$\tilde{y} = \lambda y_i + (1 - \lambda) y_j. \quad (4)$$

Here, (x_i, y_i) and (x_j, y_j) are two randomly selected examples from training data, and $\lambda \in [0, 1]$. Mixup can be employed on raw speech as well as on features [13]. We use mixup on spectrograms.

We also use the widely used speed perturbation (SP) for data augmentation. We follow [28] to create samples using speed perturbation. For a given utterance, we produce two versions of each utterance by applying the speed effect at the factors of 0.9 and 1.1.

4.4. Model Configuration

In our DenseNet architecture, after the initial convolutional layer, we use two DenseBlocks and each DenseBlock consists of $L = 6$ layers with a growth rate of $G = 24$. After the first DenseBlocks, we place a transition layer for down-sampling the size of feature maps. This consist of a 1×1 convolutional layer followed by a 2×2 average pooling layer with the stride of 2. The transition of temporal features from the DenseBlock to LSTM is made using a reshape layer. The LSTM layer serves to learn contextual information from the temporal features extracted by the DenseBlocks. After the LSTM layer, we insert a dropout layer with a dropout value of 0.5. The sum of features discovered by each DenseBlock, having two skip connections, is concatenated with the contextual features from LSTM and given to the highway network to learn discriminative features. We apply three fully connected layers with 128 hidden units with ReLU activation in the highway block followed by a softmax layer for emotion classification.

We also implement benchmarking models including DenseNet, DenseNet-LSTM, CNN, and CNN-LSTM. In DenseNet, we apply three DenseBlocks and each consists of $L = 6$ layers with a growth rate of $G = 16$. After the DenseBlocks, we place a 3×3 global average pooling layer and a fully connected layer of 1 000 hidden units before a softmax layer.

For DenseNet-LSTM, we use an LSTM layer instead of a global average pooling layer. In all convolutional layers used in the DenseNet based models, batch normalisation [25] is employed before the non-linearities. A rectified linear unit (ReLU) [26] is used as the activation function.

For the CNN-LSTM, we follow the architecture configuration described in [37]. To use CNN for benchmarking, we choose a fully connected layer instead of an LSTM layer in the above CNN-LSTM model.

We train the models using the training set and validation set is used for hyper-parameter selection. For minimisation of the cross-entropy loss function, we choose the Adam optimiser [38]. We start the training with an initial learning rate of 10^{-3}. If the validation accuracy does not improve for five consecutive epochs, we halve the learning rate and stop the process if the validation accuracy does not improve for 20 consecutive epochs. For each model, we repeat the evaluation 10 times and report the mean and standard deviation.

5. Experiments and Results

We apply a leave-one-speaker-out scheme for both datasets and report unweighted average recall (UAR) for both datasets. UAR is a widely used metric used for speech emotion recognition due to the dominance of class imbalanced datasets in this field. In each session, we employ utterances from one speaker for testing and the other speakers' utterances for validation. This configuration is used for all models. Results are computed in noisy, cross-corpus, and adversarial attacks settings.

5.1. Benchmark Results

The comparison of our proposed model with the benchmark models CNN, CNN-LSTM, DenseNet and DenseNet-LSTM are presented in Table 1. All these results are computed without data augmentation. We observe that for both IEMOCAP and MSP-IMPROV datasets, our proposed model performs better than the benchmarking models.

Table 1: *UAR (%) of different models.*

Model	IEMOCAP	MSP-IMPROV
CNN	61.5 ± 2.3	52.6 ± 2.5
CNN-LSTM	62.1 ± 1.8	53.1 ± 2.3
DenseNet	63.2 ± 1.7	54.5 ± 1.9
DenseNet-LSTM	63.5 ± 1.5	55.6 ± 1.6
Proposed	$\mathbf{64.1 \pm 1.3}$	$\mathbf{56.2 \pm 1.5}$
CNN-LSTM [37]	62.0	–
CNN [36]	61.4	55.3

Our proposed model is also performing better compared to standard DenseNet and DenseNet-LSTM, which shows that

Table 2: *UAR (%) of different models on IEMOCAP data in noisy environment setting.*

SNR (dB) \ Model	CNN-LSTM	DenseNet	DenseNet-LSTM	Proposed
0	21.0± 2.5	22.3± 1.8	23.1± 1.6	**24.4± 1.5**
10	28.2 ± 2.1	30.0 ± 2.0	31.4 ± 1.8	**32.3 ± 1.1**
20	34.8 ± 1.8	35.5 ± 1.4	36.6 ± 1.6	**37.1 ± 1.6**
speed perturbation				
0	26.5± 2.0	28.5 ± 1.5	30.2 ± 1.3	**31.7 ± 1.4**
10	32.5± 1.9	32.9± 1.3	33.1± 1.4	**34.5± 1.5**
20	38.9 ± 1.4	39.1 ± 1.6	40.2 ± 1.2	**40.8 ± 1.0**
mixup				
0	28.1± 2.4	30.8 ± 1.2	31.1 ± 1.1	**32.5± 1.1**
10	33.6± 1.8	34.7± 1.5	34.5± 1.4	**34.9± 1.6**
20	39.2± 1.4	40.8 ± 1.3	41.6 ± 1.3	**42.5 ± 1.3**
speed perturbation+mixup				
0	30.3± 2.0	33.9 ± 1.4	33.7 ± 1.0	**34.2± 1.2**
10	35.2± 1.2	38.8± 1.2	40.6± 1.2	**40.9± 1.5**
20	40.6± 1.3	42.0 ± 1.1	41.8 ± 1.5	**43.1 ± 1.1**

the use of highway connectivity in our proposed model is helping to achieve better results compared to the variants of CNNs, DenseNet, and recent studies [37, 36].

5.2. Noisy Environment

To evaluate the model in a noisy environment, we select three signal-to-noise ratio (SNR) values [0, 10, 20]. We consider the mismatched condition, where the model is trained on clean data and the test data incorporates noisy samples. We choose five noises including kitchen, park, station, traffic, and cafeteria from the DEMAND dataset. These noises are randomly added to the test set at three different SNR levels. Results on IEMOCAP data are reported in Table 2. It can be noted from Table 2 that the proposed model provides better results compared to all the other models.

We also observe from Table 2 that the data augmentation techniques help to improve robustness. Data augmentation using mixup achieves better results compared to that using speed perturbation, however, the combination of mixup and speed perturbation provides the best results.

5.3. Adversarial Settings

In adversarial settings, we use two adversarial attacks including the Fast Gradient Sign Method (FGSM) [39] and the Basic Iterative Method (BIM) [40] to evaluate the robustness. FGSM creates the adversarial examples by adding a scaled noise in the direction of the gradient of the loss function. Instead of applying adversarial noise in a single step like FGSM, BIM iteratively applies it multiple times. We applied these two attacks with the perturbation factor $\epsilon = 0.08$ on different classifiers and performance is reported in Table 3. We observe that our proposed model performs the best with or without data augmentation. We also observe data augmentation techniques help to improve robustness in the adversarial settings in the same way as in the presence of noise.

5.4. Cross-Corpus Settings

To evaluate the proposed model in a cross-corpus setting, we use IEMOCAP as training data and MSP-IMPROV as the test set. We randomly select 30 % of MSP-IMPROV for parameter selection and 70 % for testing, as used in [36]. We evaluate different models in the cross-corpus setting. The results are given in Figure 2. We observe that the proposed model achieves better performance and data augmentation helps to improve the robustness.

Table 3: *UAR (%) of different models on IEMOCAP data in adversarial setting.*

Attack \ Model	CNN-LSTM	DenseNet	DenseNet-LSTM	Proposed
FSGM	30.1± 1.8	33.5± 1.4	34.5± 1.7	**35.2± 1.3**
BIM	27.2 ± 2.1	29.7 ± 1.8	31.4 ± 1.4	**32.8 ± 1.1**
speed perturbation				
FSGM	34.5± 2.0	38.2 ± 1.4	38.1 ± 1.2	**40.2 ± 1.0**
BIM	30.5± 1.9	33.5± 1.7	33.9± 1.4	**34.6± 1.2**
mixup				
FSGM	36.1± 2.4	39.8 ± 1.7	40.5 ± 1.2	**41.4± 1.4**
BIM	31.3± 1.8	34.4± 1.2	34.0± 1.4	**34.8± 1.2**
speed perturbation+mixup				
FSGM	39.1± 2.4	42.8 ± 1.6	42.5 ± 1.6	**44.0± 1.1**
BIM	32.6± 1.8	35.4± 1.2	36.8± 1.2	**37.4± 1.3**

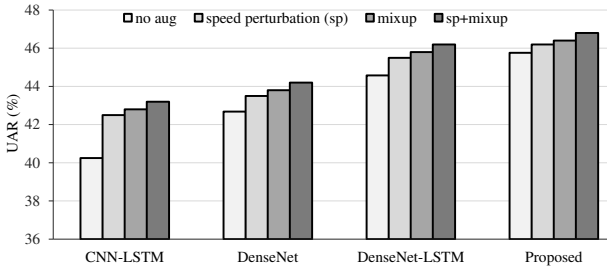

Figure 2: *Comparing different models in cross-corpus SER.*

Table 4: *Comparing cross-corpus results with recent studies.*

Studies	Latif et al. [36]	Sahu et al. [41]	Proposed
UAR (%)	46.41±0.32	40.08	**46.81±0.40**

We also compare our results with previous studies ([36, 41]) in the cross-corpus setting in Table 4. In [36], authors employ a multi-task framework and exploit larger unlabelled data for the auxiliary task to improve the generalisation of the model. In [41], the authors develop a multi-modal technique (audio plus text) for SER based on ASR transcriptions. They demonstrate that the generalisability of ASR models helps to improve the generalisation of emotion classification models. In contrast, we propose a deeper model coupled with data augmentation to achieve improved generalisation. We are achieving better results compared to these studies as reported in Table 4.

6. Conclusions

This paper introduces a new hybrid model to build a robust Speech Emotion Recognition (SER) system. This model exploits a Dense Convolutional Network (DenseNet) for feature extraction, Long Short-Term Memory (LSTM) for contextual learning, and fully connected layers with highway connectivity for discriminative representation learning, and produce robust representation. This paper also proposes data augmentation to further improve the robustness of the architecture. The performance of our proposed technique is evaluated on widely used IEMOCAP and MSP-IMPROV datasets against noise, adversarial attacks, and cross-corpus settings. Results show that our proposed technique is more robust compared to existing methods and other state-of-the-art models. Results also reveal several valuable information, such as mixup is a better augmentation technique for SER compared to the popular speed perturbation. Results also show that DenseNet based models are more robust compared to CNN-LSTM or just CNN. In future work, we aim at further optimising these architectures and augmentation in closer loop.

7. References

[1] M. Pandharipande, R. Chakraborty, A. Panda, B. Das, and S. K. Kopparapu, "Front-end feature compensation for noise robust speech emotion recognition," in *2019 27th European Signal Processing Conference (EUSIPCO)*. IEEE, 2019, pp. 1–5.

[2] A. Triantafyllopoulos, G. Keren, J. Wagner, I. Steiner, and B. Schuller, "Towards robust speech emotion recognition using deep residual networks for speech enhancement," *Proc. Interspeech 2019*, pp. 1691–1695, 2019.

[3] S. Latif, R. Rana, and J. Qadir, "Adversarial machine learning and speech emotion recognition: Utilizing generative adversarial networks for robustness," *arXiv preprint arXiv:1811.11402*, 2018.

[4] S. Latif, J. Qadir, A. Qayyum, M. Usama, and S. Younis, "Speech technology for healthcare: Opportunities, challenges, and state of the art," *IEEE Reviews in Biomedical Engineering*, 2020.

[5] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Qadir, and B. W. Schuller, "Deep representation learning in speech processing: Challenges, recent advances, and future trends," *arXiv preprint arXiv:2001.00378*, 2020.

[6] S. Latif, R. Rana, S. Younis, J. Qadir, and J. Epps, "Transfer learning for improving speech emotion classification accuracy," *Interspeech 2018: Proceedings*, pp. 257–261, 2018.

[7] Y. Qian, M. Bi, T. Tan, and K. Yu, "Very deep convolutional neural networks for noise robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2263–2276, 2016.

[8] T. Tan, Y. Qian, H. Hu, Y. Zhou, W. Ding, and K. Yu, "Adaptive very deep convolutional residual network for noise robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 8, pp. 1393–1405, 2018.

[9] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[10] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Highway networks," *Deep Learning Workshop, ICML'15*, 2015.

[11] P. N. Juslin and K. R. Scherer, "Speech emotion analysis," *Scholarpedia*, vol. 3, no. 10, p. 4240, 2008.

[12] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan, "Augmix: A simple data processing method to improve robustness and uncertainty," *ICLR*, 2019.

[13] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *ICRL*, 2018.

[14] D. Liang, F. Yang, T. Zhang, and P. Yang, "Understanding mixup training methods," *IEEE Access*, vol. 6, pp. 58 774–58 783, 2018.

[15] Y. Gong and C. Poellabauer, "Crafting adversarial examples for speech paralinguistics applications," *DYnamic and Novel Advances in Machine Learning and Intelligent Cyber Security Workshop (DYNAMICS)*, 2018.

[16] M. Guo, Y. Yang, R. Xu, Z. Liu, and D. Lin, "When nas meets robustness: In search of robust architectures against adversarial attacks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 631–640.

[17] L. Zhu, R. Deng, M. Maire, Z. Deng, G. Mori, and P. Tan, "Sparsely aggregated convolutional networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 186–201.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[19] M. Strake, P. Behr, T. Lohrenz, and T. Fingscheidt, "DenseNet BLSTM for acoustic modeling in robust asr," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 6–12.

[20] K. Zhen, M. S. Lee, and M. Kim, "A dual-staged context aggregation method towards efficient end-to-end speech enhancement," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 366–370.

[21] C.-W. Huang and S. S. Narayanan, "Deep convolutional recurrent neural network with attention mechanism for robust speech emotion recognition," in *2017 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2017, pp. 583–588.

[22] Ł. Juszkiewicz, "Improving noise robustness of speech emotion recognition system," in *Intelligent Distributed Computing VII*. Springer, 2014, pp. 223–232.

[23] T. Pang, K. Xu, and J. Zhu, "Mixup inference: Better exploiting mixup to defend adversarial attacks," in *International Conference on Learning Representations*, 2019.

[24] N. A. Tomashenko, Y. Y. Khokhlov, and Y. Estève, "Speaker adaptive training and mixup regularization for neural network acoustic models in automatic speech recognition." in *Interspeech*, 2018, pp. 2414–2418.

[25] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[26] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

[27] S. Latif, R. Rana, J. Qadir, and J. Epps, "Variational autoencoders for learning latent representations of speech emotion: A preliminary study," in *Proc. Interspeech 2018*, 2018, pp. 3107–3111.

[28] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and J. Epps, "Direct modelling of speech emotion from raw speech," *Proc. Interspeech 2019*, pp. 3920–3924, 2019.

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

[30] T. Tong, G. Li, X. Liu, and Q. Gao, "Image super-resolution using dense skip connections," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4799–4807.

[31] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," *Proc. Interspeech 2017*, pp. 4006–4010, 2017.

[32] 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.

[33] 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, 2017.

[34] Z. Aldeneh and E. M. Provost, "Using regional saliency for speech emotion recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 2741–2745.

[35] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings," in *Proceedings of Meetings on Acoustics ICA2013*, vol. 19, no. 1. Acoustical Society of America, 2013, p. 035081.

[36] 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.

[37] A. Satt, S. Rozenberg, and R. Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms." in *INTERSPEECH*, 2017, pp. 1089–1093.

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

[39] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples (2014)," *arXiv preprint arXiv:1412.6572*.

[40] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," *arXiv preprint arXiv:1607.02533*, 2016.

[41] S. Sahu, V. Mitra, N. Seneviratne, and C. Espy-Wilson, "Multimodal learning for speech emotion recognition: An analysis and comparison of asr outputs with ground truth transcription," *Proc. Interspeech 2019*, pp. 3302–3306, 2019.