INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

A Preliminary Study on Augmenting Speech Emotion Recognition using a Diffusion Model

* *Mohammad Ibrahim Malik*[1], *[*]*Siddique Latif*[2], *Raja Jurdak*[2], *and Björn W. Schuller*[3,4]

[1]Emulation AI
[2]Trusted Networks Lab, Queensland University of Technology (QUT), Australia [3]GLAM – Group on Language, Audio, & Music, Imperial College London, UK
[4]Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
siddique.latif@qut.edu.au

Abstract

In this paper, we propose to utilise diffusion models for data augmentation in speech emotion recognition (SER). In particular, we present an effective approach to utilise improved denoising diffusion probabilistic models (IDDPM) to generate synthetic emotional data. We condition the IDDPM with the textual embedding from bidirectional encoder representations from transformers (BERT) to generate high-quality synthetic emotional samples in different speakers' voices[1]. We implement a series of experiments and show that better quality synthetic data helps improve SER performance. We compare results with generative adversarial networks (GANs) and show that the proposed model generates better-quality synthetic samples that can considerably improve the performance of SER when augmented with synthetic data.

Index Terms: speech emotion recognition, synthetic speech, generative models, data augmentation.

1. Introduction

Speech emotion recognition (SER) aims at enabling machines to perform emotion detection using deep neural networks (DNNs) models [1]. SER has a wide range of applications in customer centres, healthcare, education, media, and forensics, to name a few. Various studies have explored different deep learning (DL) models including deep belief networks (DBN) [2], convolutional neural networks (CNN) [3], and long short-term memory (LSTM) networks [4] to improve the performance of SER systems. However, the SER performance is hindered by the unavailability of larger labelled datasets. Developing high-quality emotional datasets can be a time-consuming and costly process.

Data augmentation is considered an effective method to generate synthetic samples to tackle the data scarcity problem in SER. Various studies (e.g., [5, 6, 7]) in SER have shown the effectiveness of audio data augmentation techniques including SpecAugment [8], speed perturbation [9], and noise addition [10]. However, speech variations like speed perturbations do not change the semantic content and they may have an effect on emotional expressions. Another approach is to utilise generative models including a conditional generative adversarial network (GAN) [11], Balancing GAN [12], StarGAN [13], and CycleGAN [14] to generate emotional features to augment the SER system (e. g., [15, 16]). Studies have found that synthetic data by generative models can help improve the performance of SER systems. However, vanilla GANs face convergence issues due to smaller emotional corpora and are unable to produce high-quality emotional synthetic features [17, 18]. To address these issues, we propose to use diffusion models to generate synthetic data to augment the SER system. In contrast to GANs, diffusion models provide better training stability and produce high-fidelity results for audio and graphics [19, 20]. To the best of our knowledge, this paper is the first to explore diffusion models for SER.

The key contribution of this paper is the use of improved denoising diffusion probabilistic models (IDDPM) to generate synthetic data to augment the training of the SER system. In order to generate high-quality synthetic samples, we condition the IDDPM with text embedding from bidirectional encoder representations from transformers (BERT) [21]. We present a comprehensive analysis by evaluating the SER system in (i) within the corpus and (ii) cross-corpus settings on 4 publicly available datasets. We empirically show that synthetic data generated by the proposed framework considerably improves the SER results compared to recent studies.

2. Related Work

Various data augmentation techniques are used to improve SER performance. Speed perturbation [9] is a popular technique that is widely used in SER to generate augmented data. For example, studies [22, 5] use speed perturbation as a data augmentation method and evaluate it in SER. Based on the results, they show that data augmentation helps improve SER performance. SpecAugment [8] is another augmentation technique that was proposed for automatic speech recognition (ASR). Studies [7, 6] explore the SpecAugment technique in the SER domain and show that the data SpecAugment improves the generalisation and performance of the systems. Recently, the mixup [23] data augmentation technique is also being explored in SER. Mixup generates the synthetic sample as a linear combination of the original samples. Latif et al. [24] use mixup to augment the SER system in order to achieve robustness. Based on the results, they found that augmentation helps improve generalisation by generating diverse training data. Other studies [5, 25] also use data augmentation to improve the performance of SER by increasing the training data. However, these studies do not use generative models to generate synthetic data.

Generative models aim to generate new data points with some variations by learning the true data distribution of the training set. GANs are popular generative models due to their ability to learn and generate data distributions. Different studies have explored GANs in SER. For instance, Sahu et al. [17] explored a vanilla GAN and a conditional GAN to generate a synthetic emotional feature vector from a low-dimensional (2-d) feature space. They use a support vector machine (SVM) as a classifier and computed the results on both real and synthetic data. They found that the vanilla GAN faces convergence issues

[*]These authors contributed equally to this work
[1]synthetic samples url: https://www.emulationai.com/diffusion-model-ser.

10.21437/Interspeech.2023-1080

due to the smaller emotional corpus, however, a conditional GAN could generate better synthetic features that help improve the SER performance. Recently, the authors in [18] attempted to augment their GAN with mixup [23] augmentation and achieved better SER performance. In contrast to these studies, we use the diffusion model to generate high-quality synthetic samples to augment the SER system.

3. Proposed Approach

We use improved denoising diffusion probabilistic models (IDDPM) to generate emotional data. Figure 1 shows the proposed model that takes spectrogram conditioned on text embedding to generate synthetic emotional spectrograms. The details of the proposed framework are presented next.

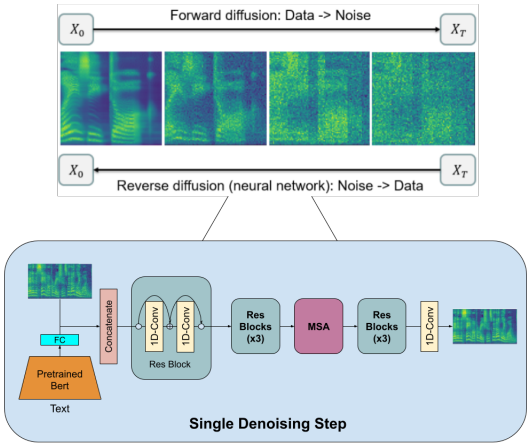

Figure 1: *Illustration of the forward and reverse diffusion process. In the forward phase, we add Gaussian noise on each timestep until the sample becomes an isotropic Gaussian distribution. In the reverse phase, we estimate the noise for each timestep using a neural network and denoise the corrupted sample.*

3.1. Diffusion for Emotional Data Synthesis

Diffusion models are state-of-the-art generative models inspired by non-equilibrium thermodynamics. They are fundamentally different from other popular generative models including GANs and variational auto-encoders (VAEs) [26]. They define the diffusion process as a Markov chain by slowly adding random noise to the input data for a total of T times and learn to reverse this process by reconstructing the desired data samples from the noise. Various diffusion model architectures have been proposed, however, we utilise the improved denoising diffusion probabilistic models (IDDPM) [27], which is an extended version of denoising diffusion probabilistic models (DDPM) [28]. The main motivation for the utilisation of IDDPM is its improved log-likelihoods and that it requires fewer timesteps to generate high-fidelity outputs.

Given an emotional data point x_0 sampled from a real data distribution, we add Gaussian noise for T timesteps using a forward noising process q that provides latent for each timestep. The sample $\mathbf{x_T}$ becomes isotropic Gaussian noise for a large $T \rightarrow \infty$. This is highlighted in Figure 1 which shows that an initially clean Mel-spectrogram (x_0) is completely transformed into Gaussian noise after adding noise for each timestep t. The noise is according to a schedule of β_t and the forward noising process can be defined as:

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}), \tag{1}$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$. We illustrate a single denoising step in Figure 1 that shows the approximation of noise at each timestep done by a neural network p_θ :

$$p_\theta(\mathbf{x_{t-1}}|\mathbf{x_t}) := \mathcal{N}(\mathbf{x_{t-1}}; \mu_\theta(\mathbf{x_t}, t); \Sigma_\theta(\mathbf{x_t}, t)), \tag{2}$$

where $\mu_\theta(\mathbf{x_t}, t)$ and $\Sigma_\theta(\mathbf{x_t}, t)$ are the mean and variance parameters respectively for the noisy sample at timestep t.

If we know the exact reverse distribution $q(x_{t-1}|x_t)$, we can sample $x_t \sim \mathcal{N}(0, I)$ and generate x_o by running the process in reverse. As explained by Ho et al. [28], a denoising neural network can be trained to predict $\mathbf{x_{t-1}}$ from $\mathbf{x_t}$ at timestep t using the following:

$$\mathbf{x}_{t-1} = \mathcal{N}(\mathbf{x}_{t-1}; \frac{1}{\sqrt{\alpha_t}}\Big(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\Big), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)). \tag{3}$$

Essentially, the network learns to predict the mean and variance parameters of noise for $t-1$ at each t. This noise is then subtracted from the $\mathbf{x_t}$ and we obtain $\mathbf{x_{t-1}}$. Ho et al. [28] kept β_t fixed as constants and set $\boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t) = \sigma_t^2\mathbf{I}$, and σ_t is either set to β_t or $\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t} \cdot \beta_t$. We use IDDPM [27] that learns $\boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)$ and select a cosine-based noise schedule instead of a fixed β_t. This helps reduce the time for sampling by utilising a strided sampling schedule. The sampling is updated after every $[T/S]$ step, which reduces the sampling time from T to S. The new sampling schedule for generation is $\{\tau_1, \ldots, \tau_S\}$, where $\tau_1 < \tau_2 < \cdots < \tau_S \in [1, T]$ and $S < T$. We generate the synthetic samples by training the model for both forward and reverse processes. The training process and model configuration are explained next.

3.2. Model Configuration and Training

In our IDDPM model, we select a commonly used modified version of the UNet [29] for the denoising process. In the modified UNet, a self-attention layer between the bottleneck and CNN layers is used. For our experiments, we observed better results in terms of audio synthesis by using 1-dimensional CNN layers instead of the more commonly used 2-dimensional CNN for image generation. This also helps reduce the memory footprint of the model and speed-up training. To ensure that each synthetic sample had a consistent speaker emotion and does not contain nonsensical gibberish as speech, we conditioned our denoising network on a representation of target samples. We passed the corresponding text with emotion and speaker information of each target sample to a pre-trained bidirectional encoder representations from transformers (BERT) [21] model to get the representation and extrapolated it simply by using fully connected layers with a linear sigmoid unit (SiLU) activation followed by a self-attention layer. This representation was concatenated with the input at each timestep t. This modified input is passed through a block of 8 1-dimensional CNN layers with 1536 filters, each. Each layer has a SiLU activation and we provide residual connections between consecutive layers which we call res-blocks. A self-attention layer is provided afterwards and we add 3 more res-blocks. A final 1-d CNN layer is added to bring the number of filters back to 80.

For our training procedure, we use a cosine-based noise schedule and choose 4000 diffusion steps to learn the variance

along with the mean of noise for each timestep t. We trained the model on an Nvidia Rtx 3090 GPU with a batch size of 64 for about 120,000 steps. During inference, we control the emotion and speaker's voice in the output samples using the condition vector.

4. Experimental Setting

4.1. Datasets Details

We selected four publicly available popular emotional datasets for SER evaluations. The details of these datasets are presented below.

IEMOCAP: The interactive emotional dyadic motion capture (IEMOCAP) [30] is a multimodal dataset containing English dyadic conversations. The dataset spans over 10 sessions and two speakers for each session. The annotation is performed by 3-4 assessors in 10 emotions. For consistency with previous studies [31, 18, 17], we use four emotions (angry, sad, happy, and neutral) for our experiments. The total samples for these selected emotions are 5531.

MSP-IMPROV: For cross-corpus evaluation, we select the "acted corpus of dyadic interactions to study emotion perception" (MSP-IMPROV) [32]. Similar to IEMOCAP, this corpus also contains recordings of English dyadic conversations. It consists of six sessions, with utterances from two speakers per session. In total, there are 7,798 utterances with four emotions: neutral, sad, angry, and happy. All samples from the corpus are used for our experiments.

CREMA-D: The crowd-sourced emotional multimodal actors dataset (CREMA-D) [33] is a data set of 7,442 clips from 91 actors. We select this corpus for cross-corpus evaluations across datasets having different distribution and recording scenarios. The clips in this corpus are from 48 male and 43 female actors between the ages of 20 and 74 coming from a variety of races and different ethnicities. In our experiments, we only use four emotions for our experiments.

RAVDESS: The Ryerson audio-visual database of emotional speech and song (RAVDESS) [34] is another popular multimodal database. This corpus is gender balanced consisting of 24 professional actors, vocalising lexically-matched statements in a neutral North American accent. Similar to CREMA-D, we select this data for cross-corpus evaluations. We select four emotions from this data similar to other datasets used in this paper.

4.2. Pre-processing and Input Representation

A popular method in SER is to represent speech as Mel-spectrograms. Likewise, we compute the Mel-spectrograms using a short-time Fourier transform of size 1024, 256 hop-size, and a window size of 1024. We select the frequency range of 0-8 kHz and extract 80 Mel frequency bands scaled linearly in the range of $[-1, 1]$. To cater for the varying audio length, we use a segment-based approach for training the model as used in [3]. We use the segment length of three seconds. The larger utterances are segmented and the smaller ones are zero-padded. This results in a Mel-spectrogram of shape (80 x 256) for each segment.

5. Experiments and Evaluations

In this section, we present the results of our diffusion model and our SER.

5.1. Synthetic Data

We generate synthetic data using IDDPM for different emotions. In Figure 2, we plot synthetic Mel-spectrograms for each emotion and compare them with the corresponding ground truth samples. To evaluate the quality of our synthetic audios, we use a pre-trained HiFi-GAN [35], a vocoder that can synthesise high-fidelity audio waves from Mel-spectrograms. We choose a sampling rate of 22.5 kHz as HiFi-GAN is trained on this sampling rate. We calculate the mean absolute difference (MAD)

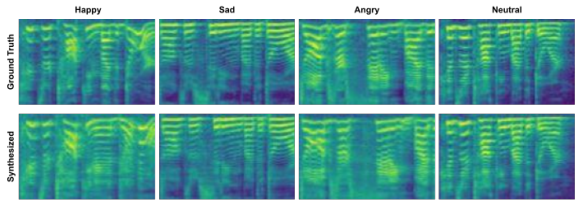

Figure 2: *Comparing ground truth samples of IEMOCAP data with synthetic samples generated using our proposed model.*

between the ground truth samples and our synthesised samples for IEMOCAP and the results are presented in Table 1. This shows that our synthetic data have very small variations from the real data for all emotions. Our synthetic samples for happy and neutral have slightly high MAD, however, we are achiving better score in contrast to the Bao et. al. [31]. These variations in synthetic data help the speech emotion classifier learn from diverse information and improve the SER performance that we highlight in the next experiments.

Table 1: *Mean of absolute difference for different emotions.*

Emotion	Mean of Absolute Difference
Happy	0.0141
Sad	0.0096
Angry	0.0097
Neutral	0.0143
Total	**0.0476**
Bao et. al. [31] ($\lambda^{cls} = 2$)	0.0490

5.2. Speech Emotion Classification

In this section, we perform SER to empirically evaluate the quality of synthetic data. We augment the training data and present the results for within-corpus and cross-corpus settings. We implement a convolutional neural network (CNN) and bidirectional LSTM (CNN-BLSTM) based classifier for SER. We use three 1-D CNN layers with rectified linear layers as activations to learn high-level representations from input Mel-spectrograms. These representations are then passed to BLSTM layers to learn the emotional context. The final output from the BLSTM layer is fed to a fully connected layer that gives an output vector equal to the number of emotions classes. We use batch normalisation to speed up training and add a dropout of 0.1 between the CNN layers and 0.2 between the LSTM layers. We train each model for 100 epochs with a learning rate of 10^{-5} and a batch size of 64. We select all these parameters using the validation set. To have a fair comparison with previous studies [18, 31, 17], we use a leave-one-speaker-out scheme and results are presented by the field's standard measure unweighted average recall.

5.2.1. Within-Corpus

In this experiment, we present SER results using the real, synthetic, and real+synthetic data in Table 2. Results for IEMOCAP

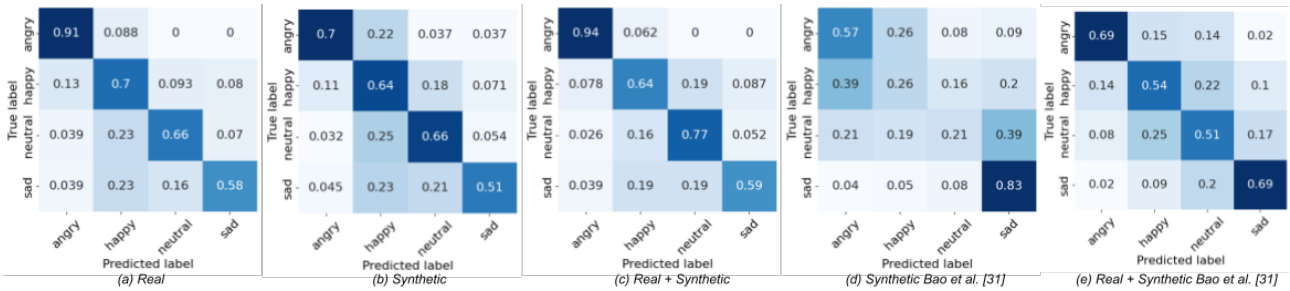

Figure 3: *Confusion Matrix Results for IEMOCAP data.*

data are compared with recent studies [17, 31, 18]. In [17], the authors use conditional GANs to augment the ground truth training data with synthetic samples to improve speech emotion classification. In [31], the authors utilise a CycleGAN-based model for the augmentation of real data. In contrast to these studies, we achieve considerable improvement for synthetic and real+synthetic data. In [18], the authors introduce a framework that utilises mixup augmentation while training a GAN-based network. They were able to improve SER performance by augmenting the training data. Our results are better compared to Latif et al., [18] without augmentation (see Table 2). We achieve further improvements in UAR ($61.38 \pm 2.04\,\%$) when mixup augmentation is applied to the training data similar to [18].

In Figure 3, we present the confusion matrices and compare the results with [31] for synthetic and real+synthetic data cases. We find improved results for both cases. Most importantly, our results for synthetic data are consistent for all the classes compared to the [31], where they achieve high accuracy only on sad emotion and low on happy and neutral classes. This shows that our proposed model is capturing emotions and generating better emotional samples for all the classes.

Table 2: *Results (UAR (%)) on IEMOCAP for corpus setting.*

Studies	Real	Syn	Real+Syn	Improvement
Sahu et al. [17]	59.42	34.09	60.29	0.87
Bao et al. [31]	59.48±0.71	46.59±0.75	60.37±0.70	0.89
Latif et al. [18]	60.51±0.57	45.75±0.81	61.05±0.68	0.54
Ours	58.62±2.11	**57.96±1.54**	**61.22±1.85**	**2.6**

5.2.2. *Cross-corpus Evaluation*

In cross-corpus SER, we utilise the MSP-IMPROV corpus as target data. We perform experiments using real, synthetic, and real+synthetic data. To be consistent with studies [31, 18, 17] compared in this section, we randomly select 30 % of the samples from MSP-IMPROV as the development set for hyper-parameter tuning and the selection and the remaining 70 % as the test set. Results are presented in Table 3. We compare the performance

Table 3: *Comparing results for cross-corpus evaluation.*

Studies	Real	Syn.	Real+Syn.
Sahu et al. [17]	45.14	33.96	45.40
Bao et al. [31]	45.58 ± 0.40	41.58 ± 1.29	46.52±0.43
Latif et al. [18]	46.0±0.57	42.15 ± 1.12	46.60 ±0.45
Our	45.81±0.65	**43.53± 1.20**	**48.22±0.51**
Our (+mixup)	**46.51±0.65**	**44.31± 1.10**	**48.58±0.51**

with different studies [18, 17, 31]. All these studies [17, 31, 18] use GAN-based architectures to generate the synthetic features to augment the training of SER. We are achieving improved results compared to these studies for synthetic and real+synthetic data. However, we are achieving comparable results with [18] for real data. In [18], the authors also utilise mixup augmentation to augment training data. We achieve better results with the utilisation of mixup augmentation in our approach (see Table 3).

Most of the previous studies [18, 31, 17] performed cross-corpus evaluations on IEMOCAP and MSP-IMPROV. Both of these datasets are recorded in similar recording situations and have almost similar distributions. In this work, we extend our experiments to other datasets that have different distributions. We use CREMA-D and RAVDESS for these experiments. We trained our model on IEMOCAP synthetic data and evaluations are performed on 50 % of CREMA-D and RAVDESS. The remaining 50 % samples of these corpora are used as a presentation for model adaptation. Results are presented in Figure 4, which shows that synthetic data contains emotional information that helps the classifier to identify emotions across different datasets even when recorded in different situations.

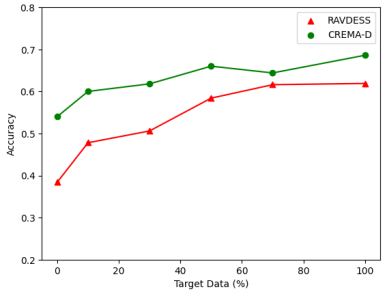

Figure 4: *Cross-Corpus results with varying percentages of target data in the training set.*

6. Conclusions and Future Work

In this work, we have addressed a major challenge of data scarcity in speech emotion recognition (SER) by proposing to use improved denoising diffusion probabilistic models (IDDPM) for synthetic data generation. We conditioned the IDDPM using the textual embedding from Bidirectional Encoder Representations to generate high-quality synthetic data. We used synthetic data to augment SER and the evaluations are performed both in within-corpus and cross-corpus settings using four publicly available datasets. In contrast to the recent studies on GAN-based synthetic data generation, our approach considerably helps improve SER performance with synthetic data utilisation for training data augmentation. In future works, we aim to design an extended version of the proposed framework for addressing the data scarcity issues in cross-lingual SER.

7. References

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

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

[3] P. Yenigalla, A. Kumar, S. Tripathi, C. Singh, S. Kar, and J. Vepa, "Speech emotion recognition using spectrogram & phoneme embedding," *Proc. Interspeech 2018*, pp. 3688–3692, 2018.

[4] M. Wöllmer, A. Metallinou, N. Katsamanis, B. Schuller, and S. Narayanan, "Analyzing the memory of blstm neural networks for enhanced emotion classification in dyadic spoken interactions," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 4157–4160.

[5] 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.

[6] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and B. W. Schuller, "Multitask learning from augmented auxiliary data for improving speech emotion recognition," *IEEE Transactions on Affective Computing*, 2022.

[7] A. Baird, S. Amiriparian, M. Milling, and B. W. Schuller, "Emotion recognition in public speaking scenarios utilising an lstm-rnn approach with attention," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 397–402.

[8] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *Proc. Interspeech 2019*, pp. 2613–2617, 2019.

[9] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Sixteenth annual conference of the international speech communication association*, 2015.

[10] 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," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 854–860.

[11] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[12] G. Mariani, F. Scheidegger, R. Istrate, C. Bekas, and C. Malossi, "Bagan: Data augmentation with balancing gan," in *International Conference on Machine Learning*, 2018.

[13] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8789–8797.

[14] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[15] A. Chatziagapi, G. Paraskevopoulos, D. Sgouropoulos, G. Pantazopoulos, M. Nikandrou, T. Giannakopoulos, A. Katsamanis, A. Potamianos, and S. Narayanan, "Data augmentation using gans for speech emotion recognition," *Proc. Interspeech 2019*, pp. 171–175, 2019.

[16] G. Rizos, A. Baird, M. Elliott, and B. Schuller, "Stargan for emotional speech conversion: Validated by data augmentation of end-to-end emotion recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 3502–3506.

[17] S. Sahu, R. Gupta, and C. Espy-Wilson, "On enhancing speech emotion recognition using generative adversarial networks," *Proc. Interspeech 2018*, pp. 3693–3697, 2018.

[18] S. Latif, M. Asim, R. Rana, S. Khalifa, R. Jurdak, and B. W. Schuller, "Augmenting generative adversarial networks for speech emotion recognition," *Proc. Interspeech 2020*, pp. 521–525, 2020.

[19] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in Neural Information Processing Systems*, vol. 34, pp. 8780–8794, 2021.

[20] R. Huang, M. Lam, J. Wang, D. Su, D. Yu, Y. Ren, and Z. Zhao, "Fastdiff: A fast conditional diffusion model for high-quality speech synthesis," in *IJCAI International Joint Conference on Artificial Intelligence*. IJCAI: International Joint Conferences on Artificial Intelligence Organization, 2022, pp. 4157–4163.

[21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.

[22] 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.

[23] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *International Conference on Learning Representations*.

[24] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and B. W. Schuller, "Deep architecture enhancing robustness to noise, adversarial attacks, and cross-corpus setting for speech emotion recognition," *Proc. Interspeech 2020*, pp. 2327–2331, 2020.

[25] M. Abdelwahab and C. Busso, "Study of dense network approaches for speech emotion recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5084–5088.

[26] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Int. Conf. on Learning Representations*.

[27] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8162–8171.

[28] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[29] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[30] 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.

[31] F. Bao, M. Neumann, and N. T. Vu, "Cyclegan-based emotion style transfer as data augmentation for speech emotion recognition," *Manuscript submitted for publication*, pp. 35–37, 2019.

[32] 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.

[33] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE transactions on affective computing*, vol. 5, no. 4, pp. 377–390, 2014.

[34] 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, 2018.

[35] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 022–17 033, 2020.