STARGAN FOR EMOTIONAL SPEECH CONVERSION: VALIDATED BY DATA AUGMENTATION OF END-TO-END EMOTION RECOGNITION

Georgios Rizos[1], *Alice Baird*[2], *Max Elliott*[1], *Björn Schuller*[1,2]

[1]GLAM–Group on Language, Audio and Music, Imperial College London, UK
[2]ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
georgios.rizos12@imperial.ac.uk

ABSTRACT

In this paper, we propose an adversarial network implementation for speech emotion conversion as a data augmentation method, validated by a multi-class speech affect recognition task. In our setting, we do not assume the availability of parallel data, and we additionally make it a priority to exploit as much as possible the available training data by adopting a cycle-consistent, class-conditional generative adversarial network with an auxiliary domain classifier. Our generated samples are valuable for data augmentation, achieving a corresponding 2 % and 6 % absolute increase in Micro- and Macro-F1 compared to the baseline in a 3-class classification paradigm using a deep, end-to-end network. We finally perform a human perception evaluation of the samples, through which we conclude that our samples are indicative of their target emotion, albeit showing a tendency for confusion in cases where the emotional attribute of valence and arousal are inconsistent.

***Index Terms*—** adversarial networks, data augmentation, end-to-end affective computing, emotional speech synthesis

1. INTRODUCTION

An individual's emotional state is among many human attributes which are transferred via the speech signal [1], and the ability to correctly recover the carried emotion is typically an unconscious task for most humans [2]. Despite this, machine detection of speech emotion is an unresolved domain with many challenges. The availability of good quality annotated data is one major limitation, as the process can have high monetary and time costs, requiring thorough planning [3]. Recent developments in deep, end-to-end models that do not depend on domain knowledge for feature design are promising [4, 5], but such models are even more dependent on dataset sizes for learning features tailored to the case. Training a generative model on a well-annotated training set in order to augment the original set has shown to offer improvement for prediction results [6], suggesting a promising way of addressing this challenge. Data augmentation via generative adversarial networks (GANs) [7] may offer a solution for this, and has been an expanding topic across domains of research such as vision [8] and affective computing [9].

1.1. Related Work

Generation of emotional speech is a challenging research area that has been approached in multiple ways: dedicated models for each emotion [10], speaker [11] or each emotion-to-emotion transformation [12] have been used, or sometimes [10], [13], [14], the authors make *the very strong assumption of the availability of a parallel dataset*, i. e., one where each source sample is paired with a ground truth target sample. *We do not make this assumption as it requires multiple times the number of annotated samples and we also aim to train a unified model with all available data, a resource that is already severely budgeted.*

The *cycle-consistency loss* [15] was devised in the context of GANs to overcome the need for parallel datasets. One attempts to minimise a measure of distance between a real sample and a sample generated by converting the same real sample to a target class and then converting it back to the original. A CycleGAN has been used for augmenting an affective computing dataset before, with positive results [9]. The StarGAN [16] was an improvement upon the CycleGAN [15] in that it requires only one generator/discriminator pair that is parameterised by class (instead of a dedicated pair for each class-to-class transformation as in a CycleGan [16]) and additionally a domain classifier that verifies the class correctness of both real and converted samples. The StarGAN framework has been used for generating images evocative of emotion [17] and also for speaker voice conversion [18]. The StarGAN we use is similar to the one in [18], *albeit adapted to the particularities of emotional speech conversion.*

1.2. Contributions

Many humans have the ability to manipulate their emotional expression, whilst keeping the lexical structure unchanged [19]. In this study, we utilise a StarGAN that transforms real emotive speech samples into different target emotions. This way, instead of applying computational perturbations on the signal (e. g., as proposed in [20] or the addition of jitter in [5]), we utilise the available emotion modelling information in the training dataset to bootstrap our original samples into retaining as much as possible speech content and mannerisms, but alter the carried emotion. We take great care in evaluating the quality of the generated samples and we perform a two-fold evaluation: a) we use the generated samples for data augmentation in a multi-class affective computing task and b) we perform a human evaluation of a set of generated samples to ascertain whether the StarGAN is indeed learning to imbue emotion into the samples instead of learning to trick the classifiers (e. g., due to mode collapse). *For the purpose of reproducibility, we will provide an implementation of the proposed method in the project's GitHub page upon acceptance, to preserve originality.*

2. EMOTIONAL SPEECH STARGAN

StarGAN [17] is an adversarial network model for non-parallel data, that allows for multi-class to multi-class domain conversion. The generator G, given an input sample $\mathbf{x}$ and a target domain label c, outputs a translated version of $\mathbf{x}$, conditioned on c. The discriminator D, given a real or fake sample $\mathbf{x}$ with label c, outputs the probability of $\mathbf{x}$ being real. The input of the classifier is either a real or fake sample $\mathbf{x}$ and its output is the softmax probability of the sample being of the correct class. The loss functions for G, D and C are:

$$\mathcal{L}_G = \mathcal{L}_{adv}^{G} + \lambda_{cls}\mathcal{L}_{cls}^{G} + \lambda_{cyc}\mathcal{L}_{cyc} + \lambda_{id}\mathcal{L}_{id}, \tag{1a}$$

$$\mathcal{L}_D = -\mathcal{L}_{adv}^{D}, \tag{1b}$$

$$\mathcal{L}_C = \mathcal{L}_{cls}^{C}, \tag{1c}$$

where $\mathcal{L}_{adv}^{\{G,D\}}$, $\mathcal{L}_{cls}^{\{G,C\}}$, $\mathcal{L}_{cyc}$ and $\mathcal{L}_{id}$ are the *adversarial loss*, *auxiliary classifier loss for a real or fake input sample*, *cycle-consistency loss* and *identity loss*, respectively, and λ_{cls}, λ_{cyc} and λ_{id} are regularisation parameters for the losses. The superscripts denote the relevant component. The losses are defined as follows:

$$\begin{aligned}\mathcal{L}_{adv}^{D} = &-\mathbb{E}_{\mathbf{x}\in X_{real}}[\log(D(\mathbf{x}, c_x)] \\ &- \mathbb{E}_{\mathbf{x}\in X_{real}, c\in C}[\log(1 - D(G(\mathbf{x}, c), c)],\end{aligned} \tag{2a}$$

$$\mathcal{L}_{adv}^{G} = -\mathbb{E}_{\mathbf{x}\in X_{real}, c\in C}[\log(D(G(\mathbf{x}, c), c)], \tag{2b}$$

$$\mathcal{L}_{cls}^{C} = -\mathbb{E}_{\mathbf{x}\in X_{real}, c\in C}[\log(C(\mathbf{x})], \tag{2c}$$

$$\mathcal{L}_{cls}^{G} = -\mathbb{E}_{\mathbf{x}\in X_{real}, c\in C}[\log(C(G(\mathbf{x}, c))], \tag{2d}$$

$$\mathcal{L}_{cyc} = \mathbb{E}_{\mathbf{x}\in X_{real}, c'\in C, c'\neq c\in C}[||\mathbf{x} - G(G(\mathbf{x}, c), c')||_1], \tag{2e}$$

$$\mathcal{L}_{id} = \mathbb{E}_{\mathbf{x}\in X_{real}, c\in C}[||\mathbf{x} - G(\mathbf{x}, c)||_1], \tag{2f}$$

where X_{real} is the set of real data in the training partition, and $||\cdot||_1$ represents the L1 vector norm. L_{cls} is the auxiliary classification loss for the set of either real or fake samples. L_{cyc} is the cycle-consistency loss and L_{id} is the identity loss, i. e., a measure of distance between the source sample and itself after being converted to the source class. The classifier, discriminator and generator are updated every 1-1-3 epochs, respectively. We follow the model architecture as given in [18] and use Wasserstein gradient penalties [21].

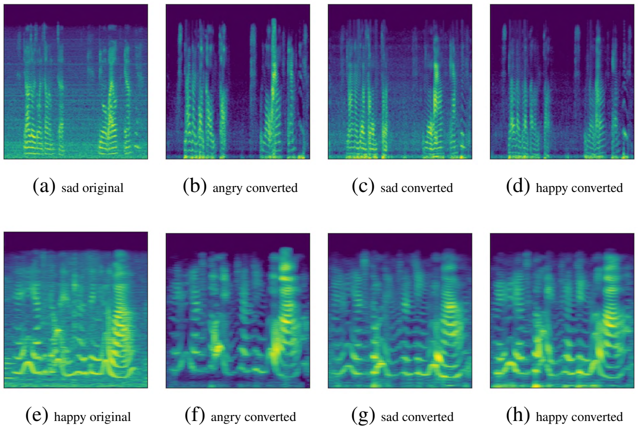
2.1. Speech Sample Conversion

The harmonic frequency content contains emotional indices [22, 1]; thus, we opted for our StarGAN to learn how to transform the spectral envelope, approximated by a set of 36 cepstral coefficients, which are then min-max normalised. In order to perform a source-target conversion, we use the WORLD [23] vocoder decomposition and synthesis functions. For each source sample, we extract three feature groups: a) aperiodicity parameters, b) the 36 cepstral coefficients, and c) the fundamental frequency (f0) contour. Before synthesising, we transform the cepstral coefficients using our trained StarGAN and also the logarithm Gaussian normalised transformation (LGNT) [24] for the f0 contour, as in [18] for voice conversion, which requires the calculation of first order statistics of log(f0) for each speaker in the dataset. In our paradigm for emotion conversion, there are *both* different speakers *and* emotions, so we do not want to average across speakers in order to calculate the statistics per emotion. If, instead, we calculate them per emotion per speaker, then we rely on speaker identity information for conversion, which makes the process inapplicable to new speakers in testing.

We propose the relative LGNT of the f0 contour for converting an utterance from emotion e_1 to emotion e_2:

$$\log(\mathrm{f0}_{new}) = \log(\mathrm{f0}) - \mu(\mathrm{f0})\ \frac{\sigma(\mathrm{f0}) + \Delta\sigma_{e_1,e_2}}{\sigma(\mathrm{f0})} + \mu(\mathrm{f0}) + \Delta\mu_{e_1,e_2}, \tag{3}$$

where μ(f0) and σ(f0) are the within-utterance mean and variance of $\log(f0)$. $\Delta\mu_{e_1,e_2}$ is the average difference in the mean $\log(f0)$ between emotions e_1 and e_2 and is calculated by finding the average difference in the mean $\log(f0)$ value between utterances of emotion e_1 and emotion e_2 for each speaker independently, then averaging. $\Delta\sigma_{e_1,e_2}$ is the average change in the variance of $\log(f0)$ between utterances of emotion e_1 and emotion e_2.



(a) sad original (b) angry converted (c) sad converted (d) happy converted

(e) happy original (f) angry converted (g) sad converted (h) happy converted

Fig. 1: Spectrogram representation of two source speech utterances from the IEMOCAP database and the conversions to all three target emotions, including the original. Male voice (`Ses03M_impro06_M008`) is shown in the top row and female voice (`Ses03M_impro07_F000`) in the lower.

3. THE IEMOCAP DATASET

We use the widely applied Interactive Emotional dyadic MOtion CAPture (IEMOCAP) database, which contains approximately 12 hours of audio-visual recordings of acted emotion in five sessions; a different pair of experienced actors interacting within each [25]. The recordings are segmented in utterances and annotated by at least two different annotators as being examples of five emotions; including neutral. We utilise utterances both from scripted and improvised recordings and retain only the ones for which a majority vote exists such that a ground truth label can be defined. The sampling rate of all recordings is equal to 16 kHz.

4. QUANTITATIVE EVALUATION

We evaluate quantitatively the quality of the samples generated by StarGAN, by using them in two multi-class emotion classification experiments using a variant of a state-of-the-art, deep, end-to-end model [4]. A three-class emotion (angry-sad-happy) StarGAN is trained with manual stopping when the domain classifier loss reaches a plateau, with sessions 1-3 used for training. Afterwards, each training sample is converted into all 3 emotions, including the original. The spectrograms of such generated samples[1] in comparison to the original are depicted in Figure 1.

As for end-to-end model training, we normalise all utterances to zero mean and unit standard deviation, calculating these statistics from sessions 1-3. The normalised raw waveform is then input

[1] The converted samples can be found in the project page: https://github.com/glam-imperial/EmotionalConversionStarGAN

	Train	Val	Test	$\sum$
Happy	387	65	143	595
Sad	696	143	245	1,084
Angry	606	327	170	1,103
$\sum$	1,689	535	558	4,040

Table 1: Speaker independent Train, (Val)idation, and Test partitions used for the baseline method in our emotion classification with data augmentation experiment. Created from the IEMOCAP dataset.

Train on	**%Micro-F1**	**%Macro-F1**
Sessions 1-3	(79.5) 61.6	(56.8) 50.4
Converted	(53.2) 60.0	(45.4) 51.7
Sessions 1-3 + Converted	(65.1) **63.7**	(55.9) **56.4**

Table 2: Results of the end-to-end emotion classification with data augmentation experiment on the IEMOCAP dataset. Validation and testing were performed on sessions 4 and 5, respectively. We report both validation (in parentheses) and testing set measure values.

into a temporal pattern extraction component, which is then followed by a sequence modelling component. The former consists of three stacked one-dimensional convolutional neural network (CNN) layers and the latter of two stacked bidirectional, long short-term memory recurrent neural network (LSTM-RNN) layers. The number of filters and widths of the convolutional layers are 64-128-256 and 8-6-6, respectively, and each one is followed by a max pooling layer that undersamples at a rate and stride of 10-8-8. The hidden units of both the RNN layers are equal to 256. The hidden state sequences produced by the two directional RNNs are merged by summation. In the study performed in [4], the authors utilised an attention layer for merging the hidden state sequence, but we achieved slightly better results by applying global max pooling. Finally we pass the result through a dense layer with 3 outputs to produce the logits. We use the softmax cross-entropy loss for training. For evaluation we opt for two means of averaging the F1 score, i. e., the harmonic mean of precision and recall: a) *Micro-F1*, by calculating the harmonic mean of the overall precision and recall scores and b) *Macro-F1*, by performing an unweighted average of class-specific precision and recall scores. We evaluate test set performance with the best model based on the validation set with respect to the Macro-F1 score, in order to place equal emphasis on each of the classes, regardless of respective number of samples and additionally report Micro-F1. We set the batch size equal to 10. All results are averaged across 10 trials.

4.1. StarGAN for Data Augmentation

We assess whether the generated samples can be used for the purpose of increasing the predictive performance in emotion classification via data augmentation. For the purpose of this experiment, as a baseline, we treat the IEMOCAP sessions 1-3 as the training set, and sessions 4 and 5 as validation and testing, respectively, conforming to the study performed in [4]. In Table 1, we summarise the class sizes for the partitions. We compare with the case of training only on the StarGAN generated data based on the sessions 1-3 as a training set and one final time using both the original and the generated training sets. We denote the three aforementioned methods as *Sessions 1-3*, *Converted*, and *Sessions 1-3+Converted*, respectively. We want the model to learn from a comparatively similar amount of data, so the corresponding number of epochs for which we train each method is 40, 14, and 10. The results are summarised in Table 2.

Test on	**%Micro-F1**	**%Macro-F1**
Sessions 1-3	(75.4) 67.5	(60.7) 58.9
Converted	(75.4) 48.9	(60.4) 43.1

Table 3: Results of the end-to-end emotion classification of the generated samples experiment on the IEMOCAP dataset. Training and validation were performed on sessions 5 and 4, respectively. We report both validation (in parentheses) and testing set measure values. The converted samples are correctly classified to a degree.

4.2. Machine Classification of Generated Samples

We now look at this evaluation from an opposite standpoint and measure the testing predictive performance of the end-to-end classifier on the converted samples. Here, sessions 5 and 4 are used as the training and validation sets, respectively, and the generated samples from the StarGAN model trained using sessions 1-3 are used for testing. We also test on the original session 1-3 data. We train for 100 epochs in order to adhere to the amount of samples seen by the model in the previous subsection. The results are summarised in Table 3.

4.3. Quantitative Evaluation Results & Discussion

In the data augmentation experiment, we observe that we achieve absolute improvements of 2 % and 6 % in Micro- and Macro-F1, respectively. In the presence of class imbalance, we consider Macro-F1 to be the more meaningful performance measure, as it places equal weight for each class in the averaging of F1 scores. However, Micro-F1 is also important, and with this kind of augmentation we observe *a larger portion of correctly classified samples overall and strong indications that this behaviour holds with respect to each class individually.* By training the model only on the generated samples, we note that even though the Micro-F1 score is lower than the baseline, it is still not insignificant. Furthermore, we achieve higher Macro-F1, which may at least partially be justified by the fact that the training set is now perfectly balanced.

In the reverse experiment, we achieve much higher than random or majority-vote performance when testing on the converted samples based on sessions 1-3, which is another quantitative indication that these samples carry emotion information, although the performance is lower than when testing on the original data from 1-3. On its own, this might be indicative that the generated data are sampled from a different latent probability distribution than the original. However, their great usefulness in data augmentation, even when solely used, shows us that there is valuable emotion information within this set.

5. QUALITATIVE EVALUATION

Given that emotion is typically perceived on a human-specific, subjective basis, we perform a human evaluation to observe how effectively the converted audio samples convey their intended emotions. 27 fluent English-speaking participants were asked to evaluate 70 audio samples; 30 source samples, 10 generated by a two-class emotion model, and 30 from a three-class emotion model, trained on all the IEMOCAP data.

5.1. Human Evaluation Experimental Design

For each sample, the participants were given a pair of emotions (i. e., angry and sad) and asked to rate it between two emotions on a Likert scale from -2 to 2, the extremes corresponding to source class and target class. For example, if the angry-sad pair is perceived as definitely angry or somewhat sad, these correspond to ratings of -2 and 1. Overall, the ordering of the emotion pair and sample order was randomised to remove any bias. Firstly, the subject is provided with a source sample, e. g., of the angry class and a paired emotion is given as well. There were 5 questions for the pair angry-sad and 5 questions for angry-happy, and then accordingly for utterances of the sad and happy classes. Next, the participants were asked the same questions but using the converted versions of each sample. For example, if a listener rated an angry source sample on the scale between angry-happy, then the next sample provided would be angry converted to happy. This process was applied for both the two-class emotion model and the three-class emotion model samples. We attempted to assess the degree to which samples with altered perceived emotions will be rated further towards the target end of the scale.

5.2. Human Evaluation Results & Discussion

The results for this evaluation are shown in Figure 2. For each emotion pair, the source emotion is -2 on the y-axis, and the target emotion is 2. For example, the original samples score towards the negative end of the scale as no conversion has been applied there.

The two-class emotion model (angry and sad) shows a tendency towards successful conversions. We note that the scores skew towards the positive end of the scale, showing that, on average, the listeners were able to correctly identify the intended emotion. We observe similar performance for angry-sad and sad-angry, albeit slightly better for the former.

The three-class emotion model (angry, sad, and happy) actually performs better than the two-class emotion model for the two common transformations between them, i. e., angry-sad, and sad-angry. It scores 0.12 points higher on angry-sad and 0.33 higher on sad-angry, indicating that the participants generally found the three-class emotion model samples represented the intended target emotions better than those of the two-class model. Our hypothesis for this behaviour is that it has to do with the overall increase in training data for the three-class emotion model: while it had the same amount of training samples of the sad and angry classes, *the StarGAN was also trained on the happy samples*, which could be the reason for the higher emotion fidelity performance. When converting between the remaining emotion pairs, the three-class emotion model has mixed results. It struggles when converting sad-happy as seen by the score of -1.27, meaning that most participants thought the converted samples were still sad. On the other hand, the score for converting from happy to sad is 0.77, the highest of all emotion pairs. For both angry-happy and happy-angry, the converted samples were generally perceived as the same emotion as the source clip. The results for happy-angry are particularly interesting, as even the source happy samples were perceived by the participants as less happy, and towards angry, scoring an average of -0.56 on the scale. Both of these emotions are on the high arousal half of the two-dimensional emotion circumplex, and this type of misclassification between them has been previously observed in a computational comparison based on acoustic features [22]. *The above results indicate a certain difficulty on the part of the model for converting to high arousal samples, if the valence content needs to be changed as well.*

We note that both for the more and the less successful conversions, the magnitudes of the scores compared to the baselines are lower. It is noteworthy that something similar happened in our machine classification of the coverted samples, when compared to the source samples (see subsection 4.2). One explanation for this is that the converted samples suffer from lower audio quality where most of the words cannot be interpreted clearly, and therefore lexical meaning cannot be used as an additional guide, thus leading to less confident ratings and misclassifications. On the other hand, even for the less successful conversions, this might be an indication that the conversion process has at least somewhat altered the emotional tone of the audio in the intended way.

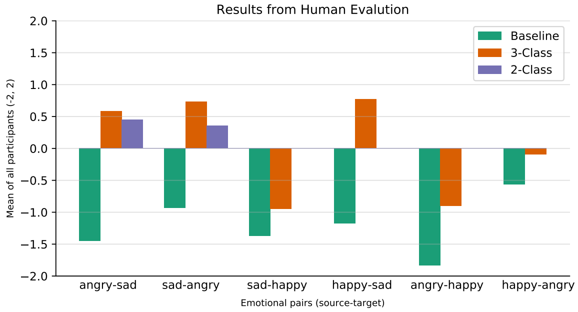


Fig. 2: Mean score of the human perception evaluation experiment. From a 5 point Likert scale (-2 = Definitely the source emotion, and +2 = Definitely the target emotion), the average scores for each emotion pair are depicted. The scores for the baseline (i. e., the unconverted samples), along with the the two-class emotion model, as well as the three-class emotion model are shown side-by-side for each emotion pair. We expect baseline samples to have negative scores, whereas successfully converted samples should have positive scores.

6. CONCLUSION & FUTURE WORK

In this paper we have applied a StarGAN for emotional speech conversion with the goal of improving performance in a deep, end-to-end 3-class classification experiment. We have additionally examined the quality of the samples through a human perception experiment. The insights gathered show that *samples generated by our approach do indeed carry valuable emotion indices that contribute to the achievement of higher predictive performance.* The samples are to a lesser degree successful in conveying the target emotion to humans, as we have identified a tendency for the model to confuse e. g., high arousal emotions as targets.

One possible avenue for future work would be the substitution of the StarGAN components such that we move further away from the need for feature engineering. For example, in [9] the authors propose the usage of a CycleGAN that is applied on spectrograms and the recent developments in applying networks on the raw waveform [4, 5] make this a compelling route for exploration. In this study, we formalised the problem as a 3-class classification task, but seeing as in our human perception experiment we saw indications that the quality of generated samples is correlated with the arousal-valence dimension values, it might hold value to work directly on this two-dimensional circumplex, or even adopt a multi-task framework to use all the available annotation information.

7. ACKNOWLEDGEMENTS

This work is funded by the UK Economic & Social Research Council (UK-ESRC) through the research Grant No. HJ-253479 (ACLEW), the Engineering and Physical Sciences Research Council (EPSRC) Grant No. 2021037, and the Bavarian State Ministry of Education, Science and the Arts in the framework of the Centre Digitisation.Bavaria (ZD.B).

8. REFERENCES

[1] Björn Schuller and Anton Batliner, *Computational paralinguistics: emotion, affect and personality in speech and language processing*, John Wiley & Sons, 2013.

[2] Roddy Cowie, Ellen Douglas-Cowie, Nicolas Tsapatsoulis, George Votsis, Stefanos Kollias, Winfried Fellenz, and John G Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal processing magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[3] Simone Hantke, Zixing Zhang, and Björn W Schuller, "Towards intelligent crowdsourcing for audio data annotation: Integrating active learning in the real world.," in *INTERSPEECH*, 2017, pp. 3951–3955.

[4] Zixing Zhang, Bingwen Wu, and Bjorn Schuller, "Attention-augmented end-to-end multi-task learning for emotion prediction from speech," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing*. May 2019, pp. 6705–6709, IEEE.

[5] Georgios Rizos and Björn Schuller, "Modelling sample informativeness for deep affective computing," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2019, pp. 3482–3486.

[6] Alice Baird, Shahin Amiriparian, and Björn Schuller, "Can Deep Generative Audio be Emotional? Towards an Approach for Personalised Emotional Audio Generation," in *Proceedings IEEE 21st International Workshop on Multimedia Signal Processing, MMSP 2019*, Kuala Lumpur, Malaysia, September 2019, IEEE, IEEE, 5 pages, to appear.

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. 2014, pp. 2672–2680, Curran Associates, Inc.

[8] Antreas Antoniou, Amos Storkey, and Harrison Edwards, "Data augmentation generative adversarial networks," *arXiv preprint arXiv:1711.04340*, 2017.

[9] Aggelina Chatziagapi, Georgios Paraskevopoulos, Dimitris Sgouropoulos, Georgios Pantazopoulos, Malvina Nikandrou, Theodoros Giannakopoulos, Athanasios Katsamanis, Alexandros Potamianos, and Shrikanth Narayanan, "Data augmentation using gans for speech emotion recognition," *Proc. Interspeech 2019*, pp. 171–175, 2019.

[10] Heejin Choi, Sangjun Park, Jinuk Park, and Minsoo Hahn, "Emotional speech synthesis for multi-speaker emotional dataset using wavenet vocoder," in *2019 IEEE International Conference on Consumer Electronics*. IEEE, 2019, pp. 1–2.

[11] Jaime Lorenzo-Trueba, Roberto Barra-Chicote, Ruben San-Segundo, Javier Ferreiros, Junichi Yamagishi, and Juan M Montero, "Emotion transplantation through adaptation in hmm-based speech synthesis," *Computer Speech & Language*, vol. 34, no. 1, pp. 292–307, 2015.

[12] Carl Robinson, Nicolas Obin, and Axel Roebel, "Sequence-to-sequence modelling of f0 for speech emotion conversion," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2019, pp. 6830–6834.

[13] Zhaojie Luo, Jinhui Chen, Tetsuya Takiguchi, and Yasuo Ariki, "Emotional voice conversion using neural networks with arbitrary scales f0 based on wavelet transform," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2017, no. 1, pp. 18, 2017.

[14] Huaiping Ming, Dongyan Huang, Lei Xie, Jie Wu, Minghui Dong, and Haizhou Li, "Deep bidirectional lstm modeling of timbre and prosody for emotional voice conversion," in *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association*, 2016, pp. 2453–2457.

[15] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*. Oct. 2017, pp. 2223–2232, IEEE.

[16] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *2018 Conference on Computer Vision and Pattern Recognition*. June 2018, pp. 8789–8797, IEEE.

[17] David Alvarez-Melis and Judith Amores, "The Emotional GAN: Priming adversarial generation of art with emotion," in *2017 NeurIPS Machine Learning for Creativity and Design Workshop*, Dec. 2017.

[18] Hirokazu Kameoka, Takuhiro Kaneko, Kou Tanaka, and Nobukatsu Hojo, "StarGAN-VC: non-parallel many-to-many voice conversion using star generative adversarial networks," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. Dec. 2018, pp. 266–273, IEEE.

[19] Sonja A Kotz and Silke Paulmann, "Emotion, language, and the brain," *Language and Linguistics Compass*, vol. 5, no. 3, pp. 108–125, 2011.

[20] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[21] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville, "Improved training of wasserstein gans," in *Advances in neural information processing systems*, 2017, pp. 5767–5777.

[22] Serdar Yildirim, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Zhigang Deng, Sungbok Lee, Shrikanth Narayanan, and Carlos Busso, "An acoustic study of emotions expressed in speech," in *Eighth International Conference on Spoken Language Processing*, 2004.

[23] Masanori Morise, Fumiya Yokomori, and Kenji Ozawa, "World: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[24] Kun Liu, Jianping Zhang, and Yonghong Yan, "High quality voice conversion through phoneme-based linear mapping functions with straight for mandarin," in *Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007)*. IEEE, 2007, vol. 4, pp. 410–414.

[25] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan, "IEMOCAP: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, Nov. 2008.