A prototypical network approach for evaluating generated emotional speech

Alice Baird, Silvan Mertes, Manuel Milling, Lukas Stappen, Thomas Wiest, Elisabeth André, Björn W. Schuller

Angaben zur Veröffentlichung / Publication details:

Baird, Alice, Silvan Mertes, Manuel Milling, Lukas Stappen, Thomas Wiest, Elisabeth André, and Björn W. Schuller. 2021. "A prototypical network approach for evaluating generated emotional speech." In Interspeech 2021, Brno, Czechia, 30 August - 3 September 2021, edited by Hynek Heřmanský, Honza Černocký, Lukáš Burget, Lori Lamel, Odette Scharenborg, and Petr Motlicek, 3161–65. Baixas: ISCA. https://doi.org/10.21437/interspeech.2021-1123.

INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

A Prototypical Network Approach for Evaluating Generated Emotional Speech

Alice Baird [1], *Silvan Mertes* [2], *Manuel Milling* [1], *Lukas Stappen*[1], *Thomas Wiest* [1], *Elisabeth André* [2], *Björn W. Schuller* [1 3]

[1] Chair of Embedded Intelligence for Health Care & Wellbeing, University of Augsburg, Germany
[2] Chair of Human-Centered Artificial Intelligence, University of Augsburg, Germany
[3] GLAM – Group on Language, Audio, & Music, Imperial College London, UK

alicebaird@ieee.org

Abstract

The collection of emotional speech data is a time-consuming and costly endeavour. Generative networks can be applied to augment the limited audio data artificially. However, it is challenging to evaluate generated audio for its similarity to source data, as current quantitative metrics are not necessarily suited to the audio domain. We explore the use of a prototypical network to evaluate four classes of generated emotional audio with this in mind. We first extract spectrogram images from WAVEGAN generated audio and other audio augmentation approaches, comparing similarity to the class prototype and diversity within the embedding space. Furthermore, we augment the source training set with each augmentation type and perform a classification to explore the generated audio plausibility. Results suggest that quality and diversity can be quantitatively observed with this approach. In the chosen context, we see that WAVEGAN generated data is recognisable as a source data class (F_1-score 43.6 %), and the samples add similar diversity as unseen source data. This result leads to more plausible data for augmentation of the source training set – achieving up to 63.9 % F_1 which is a 3.5 % improvement over the source data baseline.

Index Terms: generative adversarial networks, prototypical networks, audio generation, speech emotion recognition.

1. Introduction

There are many advantages to generating new audio data computationally, mainly the scarcity of actual data, particularly in the speech emotion domain [1]. The time-dependent nature of audio makes sourcing and annotating such data an extremely time-consuming process [2, 3]. Generative models such as Generative Adversarial Networks [4] (GANs) can be used as an augmentation method. However, one challenge for generative models remains a consensus on appropriate strategies for evaluating the generated data in comparison to the original source [5].

With this in mind, to evaluate the generated output of a GAN, researchers typically focus on aspects including, *similarity*, *diversity*, and *plausibility* [5, 6, 7]. Currently, for generated audio, time-consuming human listening studies are a common qualitative approach to observe these aspects. There are also various metrics, e. g. , the inception score [8], that quantitatively evaluates the quality and diversity of generated data [5]. However, to the best of the authors' knowledge, there is no agreement on which approach is most valid, particularly in the audio domain, and a pitfall of the inception score for the domain of emotional speech is primarily its need for large amounts of data to obtain a robust score [8]. Furthermore, in emotion-based spectrogram images, the classes' subjective nature may be less intuitive to metrics such as the inception score, which are based on more objective class-categories, from e. g. , CIFAR-10 [9], or MNIST [10]. The Fréchet Inception Distance [11] and Kernel Inception Distance [12] scores are extensions of the inception score and introduce a comparison to the source data. However, these approaches remain partly based on objective image classes. Further to this, for audio in general, quality is often an aspect of interest. There are metrics such as the perceptual evaluation of speech quality [13], and methods including Contrastive learning for perceptual audio similarity [14], which evaluate quality by similarity to a reference point. However, these focus on different audio domains and do not observe diversity and plausibility as is needed for generated audio.

Data augmentation is another quantitative approach for evaluating the plausibility of generated audio [15, 1]. In this case, generated samples are added to the training set of a classification paradigm, and where an increase in test accuracy is observed, the samples are deemed to be of value. However, this approach alone is not completely interpretable and limits any conclusions that can be made concerning the generation method.

We therefore propose an evaluation framework based on a prototypical network trained on the source data to tackle the problems above. The Prototypical Network was initially proposed for image-based few-shot learning by [16], and assumes that a prototypical representation exists for each class and calculates the Euclidean distance of unseen data points to a trained class prototype (cf. Section 3). This method is based on class prototypes and their embeddings and requires fewer data. With this in mind, when applied to synthetic emotional speech, in combination with data visualisation, it may allow for a more human-interpretable and fine-grained evaluation of similarity, diversity and plausibility.

Our approach is based on the calculation of class prototypes from the generated and source audio. First, we train a WAVEGAN [4] on the GEMEP corpus of nonsense emotional speech and generate several audio samples for four emotional classes. We then train a prototypical network to learn a task-specific embedding space, in which *prototypes* for each class can be built. These prototypes are then used to evaluate data generated by a WAVEGAN , as well as data generated by other augmentation approaches, i. e. , time-shifting, additive noise, and spectrogram masking [17]. In order to do so, we (i) measure the Euclidean distance of the generated samples to the learnt prototypes, thus evaluating *similarity*, (ii) we use the trained prototypical network to extract embeddings from the generated data and analyse their *diversity*, and (iii) retrain the prototypical network as a classifier while incorporating the augmented data into the training set to evaluate *plausibility*.

http://dx.doi.org/10.21437/Interspeech.2021-1123

Table 1: *Speaker-independent folds, for the four emotional classes utilised from a sub-set of the* GEMEP *corpus.*

	Fold-1	Fold-2	Fold-3	$\sum$
Speakers (M:F)	6 (3:3)	2 (1:1)	2 (1:1)	10
Pleasure	60	18	12	90
Anger	60	18	12	90
Elation	48	18	24	90
Sadness	48	18	24	90
$\sum$	216	72	72	360

2. The GEMEP Corpus

For our experiments, we utilise a sub-set of the Geneva Multimodal Emotion Portrayals (GEMEP) corpus [18]. The GEMEP corpus was utilised in the 2013 Computational Paralinguistics Challenge (COMPARE) [19], and includes ten native French actors (five female) speaking nonsense utterances to avoid cultural, and lexical bias. The sub-set includes four emotional classes, Hot-Anger (referred to as Anger), Elation, Sadness, and Pleasure. The emotions are selected to cover the four quadrants of Russel's circumplex for affect [20] (e. g. , Elation = High Arousal, High Valence, and Sadness = Low Arousal, Low Valence), allowing for a controlled emotional setting, with perceived diversity in the classes. Furthermore, given the small sub-set that we utilise (16 m:32 s), this fits with the realistic use-case of data augmentation, as many well-gathered corpora in the emotional speech domain are a smaller size.

When processing the raw audio, we first convert to 16 kHz, 16 bit, mono, WAV format, and split the data into three (speaker-independent) folds, cf. Table 1. The partitioning chosen is applied for all experiments, and considers as best possible a balance between classes and speaker demographics. As a note, Fold-1 is utilised as the core training set for data plausibility experiments and WAVEGAN training, and for this reason is larger than Fold-2 and Fold-3, which are used as validation and test sets.

2.1. Audio generation

We utilise WAVEGAN to generate new audio data, as first proposed in [4]. We have chosen WAVEGAN , as it shows promise for a range of audio generation tasks, in the domain of emotional speech [21], and music [4], as well as being successfully adapted for the task of data augmentation [22]. Typically, in a GAN paradigm, a generator produces new samples and competes against the discriminator, attempting to classify the instances as fake or real. WAVEGAN is a GAN developed explicitly for audio and is based mainly on *Deep Convolutional GANs* (DC-GANs) [23]. In the GEMEP corpus, samples are of varied length. As WAVEGAN requires fixed-length data, we randomly select 1-second chunks from the samples during training. The WAVEGAN we apply was trained using the default parameters described in [4] for 100 000 training steps. For our experiments, we generate samples until the quantity is equal to the classes within the source training data (total of 526 1-second samples)[1].

From a qualitative evaluation of the generated audio, the samples seem to have similar attributes as the source speakers. Of note, as is typical for GAN generated audio, there is a shimmer type artefact in the high-frequency range, which is also visible in the extracted spectrograms, cf. Figure 1. In future work, inclusion of a processing step (denoising, or low-pass filtering) to remove such artefacts may be of value for comparison.

[1]To listen to a selection of the generated samples for each of the four classes visit `shorturl.at/mwDZ1`.

2.2. Data augmentation

To compare any results obtained with WAVEGAN generated data, we also utilise several low-resource audio augmentation approaches, namely, time-shifting and additive noise. As a more state-of-the-art approach, we also apply spectrogram warping with time and frequency masking utilising the SPECAUGMENT (SPECAUG) method [17]. We choose these types of augmentation for the audio to give a broad range of representations to compare to. As can be seen in Figure 1, the time-shift representation is most similar to the source; and subjectively, the additive noise or SPECAUG approaches are the most dissimilar. We duplicate the total number of samples from the training set of the GEMEP corpus for each of these augmentation approaches. The audio signal is moved by a maximum of 0.5 seconds from the end of the signal for time-shifting the audio samples, selecting the value for time-shift randomly for each sample. For additive noise, white noise is injected at a level of $1\,\%$ of the amplitude from the source. The spectrogram augmentation approach (SPECAUG), unlike the other approaches, is applied to the spectrogram image itself. This approach masks segments on the frequency and time axis, as well as *warping* the time axis. For more detailed information on this approach, cf. [17].

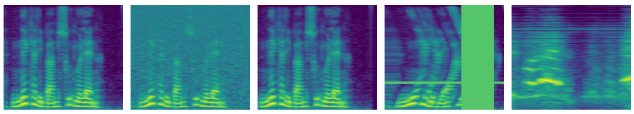

Figure 1: *Spectrogram representation (with a range of 0–8 kHz) for source and augmented audio samples for the anger class.*

2.3. Spectrogram extraction

As an input source for the prototypical network, we extract spectrogram images (cf. Figure 1) from the raw audio. The spectrograms are extracted with a pixel dimensionality of 256×256 (a dimension within a common range as applied to other speech emotion studies [24]). We limit the spectrograms to a maximum frequency of 8 kHz to reduce the presence of high-frequency non-speech related activity. The colour-map for the spectrograms is *viridis* as this has shown promise in other spectrogram-based speech emotion recognition tasks [25].

3. The Prototypical network

For our experiments[2], we adapt the prototypical network first presented in [16]. Prototypical networks have shown to achieve robust performance for few-shot learning classification of images [16], text [26], and audio [27]. To the best of the authors' knowledge, a prototypical network has not yet been utilised for evaluating generated audio samples. In the following section, we describe our network; however, we would refer the interested reader to [16], for further details of specific terminology not explicitly defined herein.

A prototypical network is searching for a prototypical (i. e. , typical) representation of a class k (cf. Figure 2 for an example of the class-prototypes within the embedding space) from data points provided as a *support* set S_k – a set of spectrogram images $\mathbf{x}_i$ and labels $\mathbf{y}_i$ for each class k, function as an anchor for the class-prototypes $\mathbf{c}_k$ – and the distance of these is compared to a

[2]GitHub repository: `https://github.com/EIHW/prototypical-network-audio-evaluation`.

query set – as with S_k, excluding $\mathbf{y}_i$. Essentially, a prototypical network learns an embedding function f_ϕ, which maps spectrograms to a N-dimensional embedding space. The prototype for k is calculated from the average of the support set embedding as

$$\mathbf{c}_k = \frac{1}{|S_k|} \sum_{(\mathbf{x}_i, y_i) \in S_k} f_\phi(\mathbf{x}_i). \quad (1)$$

Further, the euclidean distance d between a class-prototype and the embedding of a *query* is input to a softmax function, allowing the network to additionally serve as a classifier.

3.1. Model architecture

The embedding function (f_ϕ), which learns data representations based on the four classes, applies a Convolutional Neural Network (CNN) architecture. For our model we implement four convolutional blocks, consisting of a convolutional layer with batch normalisation and ReLU activation as well as a max-pooling layer. The first three blocks have a 3×3 filter, 64 channel output, and the last 3×3 block has a reduced channel output of 32. For the first two convolutional blocks, a 2×2 max-pooling layer is applied, and for the third layer, the max-pooling is increased to 4×4. As a measure to avoid overfitting the network, we apply a 40 % dropout before the last convolutional block. The same model is used for embedding both support and query sets. The model is trained with the Adam optimiser applying an initial learning rate of 10^{-3}, which is halved every epoch of 100 episodes. An episode can be referred to as a mini-batch. We have four classes for training, 20 spectrograms per class in the support set and 20 per class in the query set are presented to the classifier. From preliminary experiments, we choose to stop training after 10 epochs. The episodic sampling approach mitigates the class imbalance in the data as samples are evaluated in randomly selected class pairings. Given this, we report the F_1-score (F_1) as an evaluation metric. To evaluate the trained models, a smaller episode size of 50 is applied, with 5 samples in both the query and support sets. We perform the evaluation of each model five times and report the result with the highest F_1.

4. Experimental Settings

When augmenting a training set, considering the quality and diversity of the new data to the already known data is a necessary factor [28]. To this end, to explore the use of prototypical networks as an evaluation method for these aspects – similarity, diversity as well as plausibility – of generated data, we perform three core experiments which are described as follows:

Generated data similarity: As a first-step to observe the similarity of the generated samples, we train two prototypical networks on the source data, one which uses Fold-1, and the other using a concatenation of Folds 2 and 3 (cf. Table 1). We then test these models with data from Fold-1, for each of the data augmentation types. Samples are classified based on the distance between support class prototypes and query samples, and therefore samples with higher distance (lower similarity) to the support class-prototypes will be miss-classified.

Pairwise-embedding space diversity: To investigate diversity of the generated data, we look at distances between samples in a trained model's embedding space. A representation of a sample in the embedding space is a data point. We assume that two similar samples lead to similar representations and, therefore, to a small distance of points in the embedding space. Counter to this, two samples from a diverse corpus are expected to lead to separation in the embedding space. We measure the distance between generated and source data points with the target of generating diverse data that is distinct from the source data. For this purpose, we build point pairs such that each generated data point is matched with its closest source data point according to the Euclidean distance. Finally, we calculate the mean Euclidean distance between points in a pair. As a reference point, we compare generated data to source data from a different fold and the source data from the same fold.

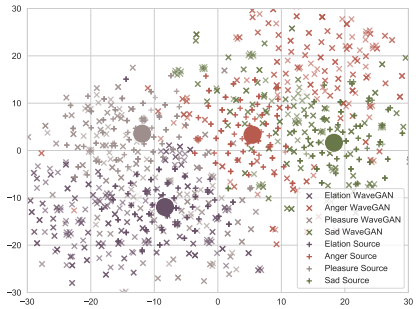

Figure 2: *t-distributed stochastic neighbour embedding (t-SNE) representation of the four classes of interest from the* WAVEGAN *generated and Source Fold-1 data within the prototypical embedding space of the F2+F3 trained model, where the prototype is based on the source data support set.*

Generated data plausibility: As mentioned, a common use-case for generated emotional speech is to improve classification accuracy through data augmentation[1]. Therefore, we augment the training set (Fold-1) of the source data with the four data augmentation approaches for this experiment. As well as showing the plausibility of the data for the specific task, will validate and compare to the previous experiments, analysing if similarity and diversity in the prototypical embeddings space impact the results when using the generated data to augment.

5. Discussion of Results

The results for each of the three experiments described in Section 3, are given in, Table 2 for *generated data similarity*, Figure 3 for *pairwise-embedding space diversity*, and Table 3 for *generated data plausibility*. For ease of discussion, we will discuss the results from these experiments individually.

Generated data similarity: The prototypical networks trained for evaluating prototype similarity are reporting accuracies above chance level (25 %). These results suggest that the network can differentiate between the four classes to a reasonably high degree. For example, in Figure 2, we see the class-prototypes from the Source-F2+3 experiments as a t-SNE representation, appear to have definition within the embedding space, with source data clusters very close to the class-prototype. However, we would like to note that the model's overall performance was not the focus of our experiments, so although a higher F_1 may have been obtainable through a hyperparameter search, we consider the results obtained to be robust for our comparison. Specifically for evaluating the augmentation types, we see the time-shifting approach has a consistently strong test accuracy, which would confirm that this is the most similar to the source data. For all other augmentation types between the models, the results are less clear, and when evaluating with unseen data (Fold-2+3 model), Noise, SPECAUG and WAVEGAN fall within a similar range with WAVEGAN showing to be the lowest performing. Furthermore, the SPECAUG approach appears to perform reasonably well, which could be due to retaining aspects

Table 2: *Reporting F_1 (%) obtained for the **generated data similarity** experiment. Training models on (F)old-1, and Fold-2+3 source data, and evaluating with all data combinations – Source, Additive (Noise),* SPECAUG *, Time-shift, and* WAVEGAN .

Trained on	Evaluated w/	Test score
Source-F1	Source-F1	95.6
	Noise-F1	61.4
	SPECAUG	77.9
	Time-shift-F1	**87.8**
	WAVEGAN -F1	53.1
Souce-F2+F3	Source-F1	59.3
	Noise-F1	46.0
	SPECAUG	48.3
	Time-shift-F1	**57.5**
	WAVEGAN -F1	43.6

of the source. In this regard, we consider SPECAUG to be the second most similar to the prototype after time-shifting when observing all results. For the WAVEGAN samples, the model can classify the data above chance level, however, results are lower than that of all augmentation approaches. This low performance still shows promise for the WAVEGAN samples, as it shows that it in range of the source class prototypes, but perhaps has higher diversity in the embedding space.

Pairwise-embedding space diversity: In these experiments, we analyse the embedding space from the models of the previous experiment (generated data similarity). The mean Euclidean distance between augmented query set data points and the closest source query set data point in the prototypical (Source-F1, and Source-F2+3) embedding space is calculated. As a reference, we also provide the same measure for the source support samples and the source query samples. Per Section 4, the prototypical network was trained with Fold-1 and Fold-2+3, respectively. We note that any augmentation technique is based on the source data of Fold-1, implying an inherent dependence within the data. Figure 3 shows that the time-shift augmentation has the smallest average pair-distance, which is in line with the assumption that this augmentation technique only slightly modifies the source data. Therefore, the time-shift adds the least diversity to the data. The right plot of Figure 3 implies that the WAVEGAN samples show a very similar average pair-distance as the independent source support samples taken from Fold-2+3. We consider this to show that the WAVEGAN samples add a similar level of diversity to the source query data as additional independent source support data might. Finally, the noise and SPECAUG generated samples show a higher average pair-distance than the other approaches, especially in the case of SPECAUG samples, which suggests that these approaches add more diversity than the previously mentioned. However, as it seems unlikely that any suggested augmentation methods add more variance than independent source data, which would improve augmentation results, the higher average pair-distance might also result from a distortion of the source data. The left side of Figure 3 depicts the pairwise distance where source query and source support data are identical. This again shows a very similar diversity trend for samples augmented with noise, SPECAUG and WAVEGAN approaches.

Generated data plausibility: Finally, to affirm the findings obtained thus far and explore a use-case for such data generation, we augment the source training data with each augmentation type. We train using Fold-1 source data and each of the Fold-1 augmentation types, and evaluate these models with Fold-3 source data. Although improvement with data augmentation is minimal here, the WAVEGAN data has shown the best results of 63.9 % F_1. The augmentation type performing worst was time-shift, which reports results slightly lower than the baseline.

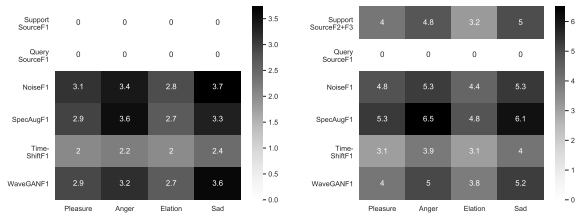

Figure 3: *A heat map representation of the results for the **pairwise-embedding space diversity** experiments. (left) is the Source-F1 trained model, and (right) is the Source-F2+3 trained model. Reporting the mean of each augmentation type's absolute Euclidean distance from the source query samples.*

Table 3: *Results obtained for the **generated data plausibility** experiments, cf. Section 4. Training a prototypical network with source data augmented with Additive (Noise),* SPECAUG *, Time-shifting, and* WAVEGAN *data. Fold-3 is used here for Test evaluation. Reporting F_1 as an evaluation metric.*

Fold-1	Test score
Source Baseline	60.4
Source + Noise	61.8
Source + SPECAUG	61.0
Source + Time-shift	60.2
Source + WAVEGAN	**63.9**

These lower results for time-shift may be explained by higher similarity of the data points to the prototype, as well as their limited diversity. These results establish the plausibility of the emotional WAVEGAN data, and show that there is a relationship in this case between diversity and similarity in the prototype embedding space.

6. Conclusions and Future Work

We proposed the use of a prototypical network to evaluate similarity, diversity, and plausibility of generated emotional speech samples in the present contribution. Despite the complexity of spectrogram images and the limited training data used, we see that all aspects can be observed with this approach and consistently throughout each experiment. The data augmentation results support that the data similarity and diversity in the prototypical embedding space should be neither too close to the prototype nor too far from the prototype. WAVEGAN samples meet this middle point and are therefore found to be plausible data for augmenting the training set.

As we previously mentioned, human listening studies remain common practice for generation evaluation, and so to compare the prototypical network predictions empirically to those obtained by human listeners would be a needed next step. Furthermore, comparing these results to other quantitative evaluation metrics such as the inception score may offer further insight. Of most promise from the current experiments, it appears that we can obtain an understanding of all evaluation criteria, and through various visualisations of the embedding space, begin to have a more interpretable representation of the generated audio.

7. Acknowledgement

This work is funded by the DFG's Reinhart Koselleck project No. 442218748 (AUDI0NOMOUS) as well as the European Union Horizon 2020 research and innovation programme, grant agreement 856879 (PRESENT).

8. References

[1] A. Baird, S. Amiriparian, and B. Schuller, "Can Deep Generative Audio be Emotional? Towards an Approach for Personalised Emotional Audio Generation," in *Proc. International Workshop on Multimedia Signal Processing (MMSP)*. Kuala Lumpur, Malaysia: IEEE, 2019, 5 pages.

[2] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, 2017.

[3] A. Baird and B. Schuller, "Considerations for a more ethical approach to data in ai: on data representation and infrastructure," *Frontiers in Big Data*, vol. 3, p. 25, 2020.

[4] C. Donahue, J. McAuley, and M. Puckette, "Adversarial audio synthesis," *arXiv preprint arXiv:1802.04208*, 2018.

[5] A. Borji, "Pros and cons of gan evaluation measures," *Computer Vision and Image Understanding*, vol. 179, pp. 41–65, 2019.

[6] V. Costa, N. Lourenço, J. Correia, and P. Machado, "Exploring the evolution of gans through quality diversity," in *Proc. of Genetic and Evolutionary Computation Conference*, 2020, pp. 297–305.

[7] K. Shmelkov, C. Schmid, and K. Alahari, "How good is my gan?" in *In Proc. European Conference on Computer Vision (ECCV)*, 2018, pp. 213–229.

[8] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *Advances in neural information processing systems*, 2016, pp. 2234–2242.

[9] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," *University of Toronto*, 2009.

[10] G. Cohen, S. Afshar, J. Tapson, and A. Van Schaik, "Emnist: Extending mnist to handwritten letters," in *Proc. International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2017, pp. 2921–2926.

[11] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *arXiv preprint arXiv:1706.08500*, 2017.

[12] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton, "Demystifying mmd gans," *arXiv preprint arXiv:1801.01401*, 2018.

[13] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. International Conference on Acoustics, Speech, and Signal Processing.*, vol. 2, 2001, pp. 749–752 vol.2.

[14] P. Manocha, Z. Jin, R. Zhang, and A. Finkelstein, "Cdpam: Contrastive learning for perceptual audio similarity," *arXiv preprint arXiv:2102.05109*, 2021.

[15] 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 *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 3502–3506.

[16] J. Snell, K. Swersky, and R. S. Zemel, "Prototypical networks for few-shot learning," *arXiv preprint arXiv:1703.05175*, 2017.

[17] 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," *arXiv preprint arXiv:1904.08779*, 2019.

[18] T. Bänziger and K. R. Scherer, "Introducing the geneva multimodal emotion portrayal (gemep) corpus," *Blueprint for affective computing: A sourcebook*, vol. 2010, pp. 271–94, 2010.

[19] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi *et al.*, "The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proc. INTERSPEECH, Lyon, France*, 2013.

[20] J. Russel, "Core affect and the psychological construction of emotions," *Psychological Review*, vol. 110, pp. 145–172, 2003.

[21] 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." in *Proc. INTERSPEECH*, 2019, pp. 171–175.

[22] S. Mertes, A. Baird, D. Schiller, B. Schuller, and E. André, "An evolutionary-based generative approach for audio data augmentation," in *Proc. International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 2020.

[23] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[24] Y. Zeng, H. Mao, D. Peng, and Z. Yi, "Spectrogram based multitask audio classification," *Multimedia Tools and Applications*, vol. 78, no. 3, pp. 3705–3722, 2019.

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

[26] X. Han, H. Zhu, P. Yu, Z. Wang, Y. Yao, Z. Liu, and M. Sun, "Fewrel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation," *arXiv preprint arXiv:1810.10147*, 2018.

[27] J. Pons, J. Serrà, and X. Serra, "Training neural audio classifiers with few data," in *Proc. in IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP*. IEEE, 2019, pp. 16–20.

[28] X. Wang, K. Wang, and S. Lian, "A survey on face data augmentation," *arXiv preprint arXiv:1904.11685*, 2019.