# Can Deep Generative Audio be Emotional? Towards an Approach for Personalised Emotional Audio Generation

Alice Baird[1], Shahin Amiriparian[1], Björn Schuller[1,2]

*Abstract*—**The ability for sound to evoke states of emotion is well known across fields of research, with clinical and holistic practitioners utilising audio to create listener experiences which target specific needs. Neural network-based generative models have in recent years shown promise for generating high-fidelity based on a raw audio input. With this in mind, this study utilises the WaveNet generative model to explore the ability of such networks to retain the emotionality of raw audio speech inputs. We train various models on 2-classes (happy and sad) of an emotional speech corpus containing 68 native Italian speakers. When classifying the combined original and generated audio, hand-crafted feature sets achieve at best 75.5 % unweighted average recall, a 2 percent point improvement over the original only audio features. Additionally, from a two-tailed test on the predictions, we find that the audio features from the original speech concatenated with the generated audio features provides significantly different test result compared to the baseline. Both findings indicating promise for emotion-based audio generation.**

## I. INTRODUCTION

Deep generative networks (including Generative Adversarial Networks (GANs) [1]) have found an abundance of use cases within the field of machine learning in recent years. Particularly in the computer audition community as well as for vision, applications include domain adaptation [2] and data manipulation [3], amongst others [4], [5]. With generative methods, e. g., for speech enhancement [6], [7] showing substantial improvements over previous methods, including multi-task learning of long short-term memory recurrent neural networks (LSTM-RNN) [8].

Higher fidelity audio comes with higher computational costs, making generative networks not yet fully applicable for realistic real-time audio-based applications. Although, through the utilisation of pre-trained networks, real-time processing does show promise for tasks including audio denoising [9]. However, given the often lower dimensionality of data sources, more effective reinforced real-time frameworks have been applied to vision tasks, e. g., image correction [10].

Given that synthetic audio has the ability to immerse a listener in an emotional environment, and transmit an emotional state [11], there is much room for research in the realm of generative networks towards personalised variations of emotional soundscapes. Implemented in real-time and with specific penalisation ability, such an audio environment could be implemented in daily life scenarios, e. g., in the work place to improve general quality of life [12].

Emotion is a subtle aspect of audio transmission, which may not be captured via deep generative approaches. Generative networks are however, able to reach near human replication in the field of speech synthesis [13], and the perception of various approaches, including the state-of-the-art has been evaluated [14]. As well as this, conversion of emotional speech states utilising the WaveNet Vocoder framework has recently shown promise [15], and approaches for deriving representations of emotional speech features found deep convolutional generative adversarial networks (DC-GANs) to be of most benefit for feature generation, as compared to convolutional neural networks (CNNs) architectures [5], [16]. However, to the best of the authors' knowledge, the advantages of data augmentation utilising emotional data have not yet been explored in the audio domain, although it is a topic that has shown to be successful for emotion-based visual data [17].

As an initial step in exploring this topic, we utilise the large and highly emotionally diverse DEMoS corpus of Italian emotional speech [18]. Applying pitch-based augmentation, 2-classes (happy and sad) from the corpus are then used as training data for several speaker independently partitioned WaveNet models [19]. Post-audio generation from the WaveNet models, we extract both state-of-the-art and conventional feature sets including, the deep representations of the DEEP SPECTRUM toolkit [20], as well as hand-crafted features from the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [21]. We choose these feature sets due to their known strength for similar emotion recognition [22] and classification [23] tasks. From both the generated and original data, a series of classification experiments were performed to ascertain if the generated audio is able to improve results, assuming that this implies the inclusion of subtle emotion-related features in the generated audio.

This paper is organised as follows. In the following section (Section II), the corpus used in our experiments is presented, including data processing, partitioning and augmentation. We then describe our experimental settings for both the generative model Section III, and the following classification paradigm. Followed by a discussion of results, and concluding remarks in Section IV and in Section VI, respectively.

[1] Alice Baird, Shahin Amiriparian, and Björn Schuller are with the ZD.B. Chair of Embedded Intelligence for Health Care & Wellbeing, Univeristy of Augsburg, Germany. {alice.baird, shahin.amiriparian, bjoern.schuller}@informatik.uni-augsburg.de

[2] Björn Schuller is also faculty at GLAM – Group on Language, Audio & Music, Imperial College London, U. K.

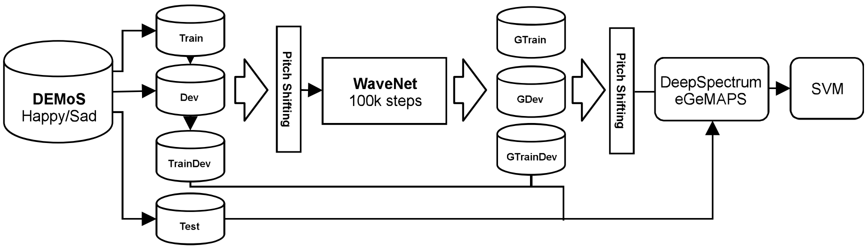

Fig. 1: An overview of the implementation utilised in this study for emotional audio generation with WaveNet. Features were extracted using both the OPENSMILE and DEEP SPECTRUM toolkits, resulting in eight feature sets, (Baseline DEMoS , (G)enerated DEMoS , Augmented DEMoS , and Generated Augmented DEMoS ). A Support Vector Machines was utilised for classification experiments, with optimisation of the complexity only for the 2-class classification.

TABLE I: Speaker independent partitions, Train, (Dev)elopment, Test created from the DEMoS emotional speech database. Including the distribution of the 2-classes (Happy and Sad).

| | Train | Dev. | Test | $\sum$ |
|---|---|---|---|---|
| Speakers | 24 | 22 | 22 | 68 |
| Gender M:F | 15:9 | 15:7 | 15:7 | 45:23 |
| Happy | 447 | 434 | 514 | 1395 |
| Sad | 493 | 486 | 551 | 1530 |
| $\sum$ | 940 | 920 | 1 065 | 2 925 |

## II. The Corpus

For this study, we utilise the Database of Elicited Mood in Speech (DEMoS ) [18]. DEMoS is a corpus of induced Italian emotional speech, including the 'big 6' emotions, plus neutral, and additionally guilt. This dataset is comprised of 9 365 emotional instances and 332 neutral samples produced by 68 native speakers (23 females, 45 males).

For this first step study, we chose to use only the emotional samples of happiness and sadness, as these fall in opposite positions when observing the the valence and arousal emotional circumflex [24], and will possibly allow for a more significant difference between the generated classes. In this way, the subset of DEMoS that we utilised for the study has in total 2 925 instances, with a duration of 2 h:47 m:41 s.

### A. Data Pre-Processing

As a first step, the DEMoS the data was normalised across all speakers. The data then remained at the provided format of monophonic WAV 44.1k Hz.

As a means of avoiding any speaker dependency during training, partitioning of the data was made prior with consideration to gender. There is a gender bias in the dataset (45:23, male:female), and future work could be to consider gender-independent models, however, with consideration to gender we balance each (train, development, and test) equally, with the addition of balancing the instances for the 2-classes of interest happy and sad (cf. Table I for the partitioning applied)[1].

### B. Data Augmentation

In general, it is known that deep networks require a large amount of data to achieve usable results. With this in mind, although for an emotional speech corpus, the DEMoS database is reasonable in size, when training the WaveNet system we applied pitch shifting to augment the input data.

Pitch shifting has shown to be a strong choice over others such as time and noise augmentation for similar tasks, including environmental sound classification [25]. Augmentation was only applied to the Train and Development partitions, keeping the Test set entirely unchanged. Utilising the LibROSA toolkit [26], we choose to raise or lower the pitch of the audio samples, and keep the duration of the samples unchanged. Each utterance was pitch shifted by a factor 10; 5 lower: {0.75, 0.80, 0.85, 0.90, 0.95} and 5 higher {1.05, 1.10, 1.15, 1.20, 1.25} increments, which are audibly observed to have made minimal change to the original data. Resulting in an DEMoS augmented data set (not including unchanged Test set) of 19 h:01 m:22 s.

## III. Experimental Settings

As a first step to explore the potential of emotional audio generation utilising generative networks, we utilise a TensorFlow implementation of the WaveNet generative framework for modelling raw audio [19][2]. We choose WaveNet as this is a standard framework in the field of audio generation (cf. Figure 1 overview of the experimental setting).

WaveNet is an audio implementation of PixelCNN [27], and is a generative network for modelling features of raw audio, represented as 8-bit audio files, with 256 possible values. During the training process, the model predicts audio signal values (with a temporal resolution of at least 16k Hz) at each step comparing to the true value, using cross-entropy as a loss function. Hence, the WaveNet model implements a 256 class classification [28]. As a means of decreasing the computational expenses, WaveNet applies the method of stacked dilated casual convolutions, reducing the receptive field, and minimising the loss in the resolution [29].

[1] The Speaker IDs for each partition are as follows: (training) 01 − 17, 21, 22, 29, 31, 36 − 38, (development) 18 − 20, 23 − 28, 30, 32 − 34, 39 − 41, 43 − 48 and (test) 42, 49, 50 − 69.

[2] https://github.com/ibab/tensorflow-wavenet

### A. Model Training

As the input for the WaveNet model, we supply 6 training sets (3 for each class). We then train three models separately on the augmented Training, Development, and Training plus Development partitions. The WaveNet model was iterated for 100 000 steps and a silence threshold $s$ of 0 was set. $s$ acts as filter, ignoring samples of silence. Given the subtle nature of emotion-based speech features $s$ was set to zero to avoid loss of information. To reach 100 000 steps, ca. 22 hours was needed on an Nvidia GTX TITAN X with 12 GB of VRAM for each training set model.

### B. Audio Generation

After training each WaveNet model for 100 000 steps, we generate new speech audio samples. Based on an approximation of the mean duration from each partition of the original DEMoS data (Train = 2.6 s , Dev. = 2.6 s, Train+Dev. = 2.6 s ), we generated samples of 2.6 s for the Train, Development and the combined Train and Development models, with total instances of 2 123. Due to limited computational processing time, we only reach ca. 75 % of the original DEMoS quantity [3]. For generation, the hyperparameters remain the same, with the temperature threshold $t$ of 1.0 applied – lowering $t$ causes the model to focus on higher probability predictions. Following this, we also apply data augmentation in the same manner as described in Section II-B to each of these generated partitions. Spectrogram plots of the generated audio in comparison to the original audio data can be seen in Figure 2. From a qualitative analysis of the generated data, attributes of the original speech, i. e. accent and intonation are audible, despite the presence of noise that has occurred in excess during generation.

### C. Feature Extraction and Fusion

For both datasets, the generated and the original, we extract conventional hand-crafted features, and state-of-the-art deep representations. Resulting in 8 features sets (4 hand crafted, and 4 *deep*), from 4 variations of the data: (1) Original DEMoS, (2) Augmented DEMoS, (3) Generated DEMoS, and (4) Augmented Generated DEMoS.

Given the success of the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [21], we utilise this as a conventional handcraft approach. From each instance, the eGeMAPS acoustic features are extracted with the OPENSMILE toolkit [30]. Using the default parameter settings from OPENSMILE for the low-level descriptors (LLDs) of each feature set, the higher level suprasegmental features were extracted over the entire audio instance.

Additionally, we extract a 4 096 dimensional feature set of deep data-representations with the DEEP SPECTRUM toolkit [31][4]. DEEP SPECTRUM has shown success for other emotion-based speech tasks [23]. For this study, we extract mel-spectrograms with a *viridis* colour map, using the default DEEP SPECTRUM settings and the VGG16 pre-trained imagenet model [32], with no window size or overlap applied.

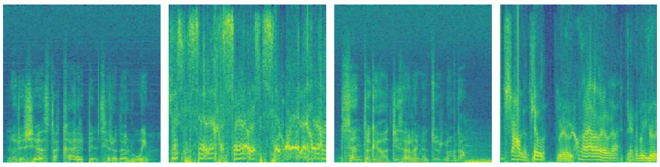
(a) NP_f_47_tri01b (b) tri_2.6_308 (c) NP_f_47_gio03c (d) gio_2.6_10

Fig. 2: Spectrogram representation of speech files. 'Sad' original audio (a) and 'Sad' generated audio (b), as well as 'Happy' original audio (c) and 'Happy' generated audio (d). Files names indicated in caption. High frequency noise can be seen in excess for the generated audio, however, vocal features such as formants are also seen to be replicated in the lower frequency range.

### D. Classification Approach

A support vector machine (SVM) implementation with linear kernel from the open-source machine learning toolkit Scikit-Learn [33] is used for our experiments. During the development phase, we trained a series of SVM models, optimising the complexity parameters ($C \in 10^{-4}$, $10^{-3}$, $10^{-2}$, $10^{-1}$, 1), evaluating their performance on the Development set. For original DEMoS data we re-trained the model with the concatenated Train and Development set, and evaluate the performance on the Test set. For the Generated DEMoS , we utilise the data which has been generated from the combined Training and Development WaveNet, and evaluated on the original DEMoS test. Further, upon creation of the 8 aforementioned features sets, we prepared 5 experiments which were repeated for each feature set type (DEEP SPECTRUM and eGeMAPS), in various combinations of the data, with all tested on the original unseen DEMoS Test partition:

1) Baseline (original data for Training, Development).
2) Generated (generated speech for Training, Development).
3) Baseline + generated (combined baseline and generated in Training, Development).
4) Generated + augmentation of original (combined generated with pitch shifting augmentation of original for Training, Development).
5) Augmented generated + augmented original (combined pitch shifting augmentation of generated speech with augmented original for Training, Development).

## IV. Results and Discussion

When observing the results found in Table II, it can be seen that for both DEEP SPECTRUM and eGeMAPS results, there is an improvement on the classification baseline when applying the generated audio to the Training set. We will discuss experiments in relation to the number indicated in the results table, as previously described in section III-D. To evaluate the significant (or not) difference between predictions, we conduct a two-tailed T-test, rejecting the null-hypothesis at a significance level of $p < 0.05$ and below. For this, we

[3] For the interested reader, a selection of generated data can be found here: https://bit.ly/32KzXTf

[4] https://github.com/DeepSpectrum/DeepSpectrum

checked each Test set prediction result for normality using a Shapiro-Wilktest [34].

From the Deep Spectrum results we see slight improvement when utilising the generative data with the original data. Of most interest is experiment 3 in which the result improves over the original baseline by 0.9 percent points, at the same classification complexity of C = $10^{-2}$. When performing a T-test with test predictions of experiment 3 against experiment 1, we obtain $p = 0.05$, which would suggest a borderline significant difference in this improvement. We also see improved results for experiment 5 although this is not found to be significantly different to the baseline. Better results were found for Test with larger complexity optimisation values; however, this occurred due to overfitting on the Development set and therefore the result is not reported.

Experiment 2 from the Deep Spectrum results, which utilises the generated data only in the Training set, received below chance level, implying that the original data is needed for this scenario. However, when we observe the eGeMAPS result for experiment 2, there is a 5 percent point increase in comparison to Deep Spectrum features. This shows promise, although through significance testing between experiment 2 of Deep Spectrum and eGeMAPS results, there is no significant difference found.

Continuing with the results from eGeMAPS features, the best result is seen in experiment 4, with a $75.5\,\%$ UAR, 1.9 percent higher than the baseline experiment 1. This results would suggest that the known emotionality of the eGeMAPS feature set is more able to capture the emotion from the generated data, as compared to the Deep Spectrum result for this experiment. However, no significant difference is found between these experiments when evaluating with a T-test.

The result of the eGeMAPS experiment 5 are significantly different from the baseline experiment 1 with $p = 0.006$. This does show promise for additional pitch based augmentation on the generated data; however, for experiment 4 which is our highest result, no significant difference over the baseline was found ($p = 0.069$).

## V. Limitations

When considering the limitation of this study, we see that the results do show promise, but there is minimal significance to the improvement. It may be of benefit to consider deeper networks for classification of the generated data rather than the conventional SVM. Specifically, networks which incorporate the time-dependency which are inherent to audio, e. g., RNNs or convolutional RNNs [35]. As well as this, incorporating multiple data sets, and more emotional classes may be fruitful for evaluation, given the tendencies we have seen arise from this first step 2-class setup. Additionally, the pitch shifting may also be altering emotional attributes, therefore we would consider exploring alternative augmentation methods including additive noise and time-shifting.

In this way, our results are also limited by the single WaveNet architecture that we have implemented, and it would be of best interest to evaluate alternative generative networks including DC-GANs [36], allowing for deeper hyperparameter optimisation. Additionally, other deep generative networks implementations which have shown success, e. g., SpecGAN [37] may be useful for this task.

TABLE II: Results for 2-class classification (happy vs sad) across all experimental setups on the DEMoS corpus as described in Section III-D. Utilising a SVM, optimising $C$, and reporting unweighted average recall (UAR) on both Deep Spectrum and eGeMAPS feature set (Dim)ensions, including (O)riginal, (G)enerated and (A)ugmented data. Chance level for this task is $50\,\%$ UAR. * indicates significant difference over the baseline (1).

| SVM Deep Spectrum | | | | |
|---|---|---|---|---|
| | Dim | C | Dev. | Test |
| (1) Baseline | 4 096 | $10^{-2}$ | 74.0 | 73.5 |
| (2) G | 4 096 | $10^{-2}$ | 59.2 | 49.0 |
| (3) G + O | 4 096 | $10^{-2}$ | 65.3 | **74.4** |
| (4) G + A O | 4 096 | $10^{-3}$ | 80.4 | 73.4 |
| (5) A G + A O | 4 096 | $10^{-3}$ | 85.7 | **74.1** |
| SVM eGeMAPS | | | | |
| | Dim | C | Dev. | Test |
| (1) Baseline | 88 | $10^{-1}$ | 73.7 | 73.6 |
| (2) G | 88 | $10^{-2}$ | 58.6 | 55.8 |
| (3) G + O | 88 | 1 | 76.9 | **74.0** |
| (4) G + A O | 88 | 1 | 79.9 | **75.5** |
| (5) A G + A O | 88 | 1 | 84.5 | **74.1*** |

## VI. Conclusions and Future Work

In this study, we have utilised an emotional speech corpus to take a first step in evaluating the ability for emotional audio to be regenerated via the generative model for raw audio, WaveNet. In this way, we are working towards the use of generative models as a means of generating personalised emotional audio environments, e. g., adapting a stressful audio environment (soundscape) into a more enjoyable space, based on the needs of an individual.

Findings from this have shown promise for emotional generative audio showing an improvement on a binary (happy vs sad) classification paradigm. Deep representations of the audio, and the handcrafted features of eGeMAPS, result in improvements over the dataset baseline. Results suggest that some emotionality is retained in the generated data, in particular we find a slight above chance result for eGeMAPS features from only generated training sets.

Given the promise shown from these results, in future work we would consider expanding our research for generative audio across a variety of emotional audio domains, e. g., music and the soundscape, as a means of exploring immersive use-cases. In this same way, it would be of interest to explore audio generation with larger duration, evaluating the human perception of such emotion-based generation.

## VII. Acknowledgements

This work is funded by the Bavarian State Ministry of Education, Science and the Arts in the framework of the Centre Digitisation.Bavaria (ZD.B).

## References

[1] 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.

[2] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan, "Unsupervised pixel-level domain adaptation with generative adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3722–3731.

[3] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros, "Generative visual manipulation on the natural image manifold," in *European Conference on Computer Vision*. Springer, 2016, pp. 597–613.

[4] Jun Deng, Nicholas Cummins, Maximilian Schmitt, Kun Qian, Fabien Ringeval, and Björn Schuller, "Speech-based diagnosis of autism spectrum condition by generative adversarial network representations," in *Proceedings of the International Conference on Digital Health*. ACM, 2017, pp. 53–57.

[5] Shahin Amiriparian, Michael Freitag, Nicholas Cummins, Maurice Gerczuk, Sergey Pugachevskiy, and Björn Schuller, "A fusion of deep convolutional generative adversarial networks and sequence to sequence autoencoders for acoustic scene classification," in *The European Signal Processing Conference (EUSIPCO)*. IEEE, 2018, pp. 977–981.

[6] Santiago Pascual, Antonio Bonafonte, and Joan Serrà, "Segan: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017.

[7] Daniel Michelsanti and Zheng-Hua Tan, "Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification," *arXiv preprint arXiv:1709.01703*, 2017.

[8] Zhuo Chen, Shinji Watanabe, Hakan Erdogan, and John R Hershey, "Speech enhancement and recognition using multi-task learning of long short-term memory recurrent neural networks," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[9] Dario Rethage, Jordi Pons, and Xavier Serra, "A wavenet for speech denoising," in *The IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5069–5073.

[10] Jie Li, Katherine A Skinner, Ryan M Eustice, and Matthew Johnson-Roberson, "Watergan: unsupervised generative network to enable real-time color correction of monocular underwater images," *IEEE Robotics and Automation Letters*, vol. 3, no. 1, pp. 387–394, 2017.

[11] Emilia Parada-Cabaleiro, Alice Baird, Nicholas Cummins, and Björn W Schuller, "Stimulation of psychological listener experiences by semi-automatically composed electroacoustic environments," in *2017 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2017, pp. 1051–1056.

[12] Irene Van Kamp, Ronny Klaeboe, Hanneke Kruize, Alan Lex Brown, and Peter Lercher, "Soundscapes, human restoration and quality of life," in *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*. Institute of Noise Control Engineering, 2016, vol. 253, pp. 1205–1215.

[13] Aaron van den Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, Koray Kavukcuoglu, George van den Driessche, Edward Lockhart, Luis C Cobo, Florian Stimberg, et al., "Parallel wavenet: Fast high-fidelity speech synthesis," *arXiv preprint arXiv:1711.10433*, 2017.

[14] Alice Baird, Stina Hasse Jørgensen, Emilia Parada-Cabaleiro, Nicholas Cummins, Simone Hantke, and Björn Schuller, "The perception of vocal traits in synthesized voices: Age, gender, and human likeness," *Journal of the Audio Engineering Society*, vol. 66, no. 4, pp. 277–285, 2018.

[15] 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 (ICCE)*. IEEE, 2019, pp. 1–2.

[16] Jonathan Chang and Stefan Scherer, "Learning representations of emotional speech with deep convolutional generative adversarial networks," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 2746–2750.

[17] Xinyue Zhu, Yifan Liu, Zengchang Qin, and Jiahong Li, "Data augmentation in emotion classification using generative adversarial networks," *arXiv preprint arXiv:1711.00648*, 2017.

[18] Emilia Parada-Cabaleiro, Giovanni Costantini, Anton Batliner, Maximilian Schmitt, and Björn W Schuller, "Demos: an italian emotional speech corpus," *Language Resources and Evaluation*, pp. 1–43, 2019.

[19] Aäron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W Senior, and Koray Kavukcuoglu, "Wavenet: A generative model for raw audio.," *SSW*, vol. 125, 2016.

[20] Shahin Amiriparian, Maurice Gerczuk, Sandra Ottl, Nicholas Cummins, Michael Freitag, Sergey Pugachevskiy, Alice Baird, and Björn Schuller, "Snore sound classification using image-based deep spectrum features," in *Proc. INTERSPEECH*, 2017, pp. 3512–3516.

[21] Florian Eyben, Klaus Scherer, Björn Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Devillers, Julien Epps, Petri Laukka, Shrikanth Narayanan, and Khiet Truong, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[22] George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A Nicolaou, Björn Schuller, and Stefanos Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 5200–5204.

[23] Nicholas Cummins, Shahin Amiriparian, Gerhard Hagerer, Anton Batliner, Stefan Steidl, and Björn W Schuller, "An image-based deep spectrum feature representation for the recognition of emotional speech," in *Proceedings of the 25th ACM international conference on Multimedia*. ACM, 2017, pp. 478–484.

[24] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[25] Justin Salamon and Juan Pablo Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, 2017.

[26] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, 2015, pp. 18–25.

[27] Aäron van den Oord, Nal Kalchbrenner, Oriol Vinyals, Lasse Espeholt, Alex Graves, and Koray Kavukcuoglu, "Conditional image generation with pixelcnn decoders," *CoRR*, vol. abs/1606.05328, 2016.

[28] Rachel Manzelli, Vijay Thakkar, Ali Siahkamari, and Brian Kulis, "Conditioning deep generative raw audio models for structured automatic music," *arXiv preprint arXiv:1806.09905*, 2018.

[29] Fisher Yu and Vladlen Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[30] Florian Eyben, Felix Weninger, Florian Gross, et al., "Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor," in *Proc. ACM*, Barcelona, Spain, 2013, pp. 835–838.

[31] Shahin Amiriparian, Maurice Gerczuk, Sandra Ottl, Nicholas Cummins, Michael Freitag, Sergey Pugachevskiy, and Björn Schuller, "Snore Sound Classification Using Image-based Deep Spectrum Features," in *Proc. of INTERSPEECH*, Stockholm, Sweden, 2017, ISCA, 5 pages.

[32] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[33] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[34] Jennifer Peat and Belinda Barton, *Medical statistics: A guide to data analysis and critical appraisal*, John Wiley & Sons, 2008.

[35] Shahin Amiriparian, Alice Baird, Sahib Julka, Alyssa Alcorn, Sandra Ottl, Suncica Petrović, Eloise Ainger, Nicholas Cummins, and Björn Schuller, "Recognition of Echolalic Autistic Child Vocalisations Utilising Convolutional Recurrent Neural Networks," in *Proceedings INTERSPEECH 2018, 19th Annual Conference of the International Speech Communication Association*, Hyderabad, India, September 2018, ISCA, pp. 2334–2338, ISCA.

[36] Alec Radford, Luke Metz, and Soumith Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[37] Chris Donahue, Julian McAuley, and Miller Puckette, "Synthesizing audio with generative adversarial networks," *arXiv preprint arXiv:1802.04208*, 2018.