Improved Speech Synthesis using Generative Adversarial Networks

Dineshraj Gunasekaran[1 [0000−0003−3479−960X]], Gautham Venkatraj[1 [0000−0002−3553−1608]], Eoin Brophy[2 [0000−0002−6486−5746]], and Tomas Ward[1 [0000−0002−6173−6607]]

[1] Insight SFI Research Centre for Data Analytics, Dublin City University, Dublin
[2] Infant Research Centre, Cork University Maternity Hospital, Ireland
{dineshraj.gunasekaran2,gautham.venkatraj2,eoin.brophy7}@mail.dcu.ie
tomas.ward@dcu.ie

Abstract. Artificially generated audio and speech have been a major area of machine learning research in recent years. Generative Adversarial Networks (GANs) have been at the forefront of the research progress in this domain. WaveGAN and SpecGAN are some of the current state of the art methods for speech synthesis, that utilise generative modelling to produce synthetic speech signals. Of particular interest, SpecGAN uses spectrograms of speech as training data for the GAN. With current available models like WaveGAN, the structure and linguistic information of the audio is not captured and leads to synthesized audio with high noise. In this paper, we propose a method we call Mel-spectrogram GAN (MS-GAN) that instead uses the Mel-Spectrogram of the audio signal as an aid in approximating the human auditory system response more closely than narrow frequency bands. It uses a reconstructed spectrogram with a mel-scale. We demonstrate, using this approach a 23% higher inception score than WaveGAN. Further, we establish an improved version of Conditional MSGAN that quickly learns the data distribution of each of the classes to produce better quality speech and further increases the score by 32%. These results suggest that our Conditional MSGAN architecture is a promising approach for improved speech synthesis using GANs.

Keywords: speech synthesis · GAN · neural network · deep learning · mel-spectrogram

1 Introduction

The primary motive of this study is to improve audio synthesis by training Generative Adversarial Networks (GANs) more effectively and efficiently. Raw time series speech signals like number narrations and word utterances are used as input in this investigation, and this allows us to capture the phonetics and intonations of the speech directly. Natural speech data that is more relevant in real-world scenarios is used in this analysis to produce speech that has a natural sounding quality [1]. In the genres of speech synthesis, TTS (text to speech) is

one of the key areas where many effective algorithms have been established and so now the focus is shifting towards voice conversion and style transfer [2], both of which can be addressed through GAN approaches.

One of the ways to obtain more natural sounding audio is by modelling the probability distribution in a parametric and non-parametric way in the training stage and relating the synthetic speech to these distributions. The existing WaveGAN [3] is used as a benchmark and improved results have been obtained using our proposed Mel-Spec GAN (MSGAN) and Conditional Mel-Spec GAN (CMSGAN) as is further detailed in this report.

In the scenario involving audio files, the conversion of the time series to spectrograms provides the input to these models. Mel-Spectrogram is a type of spectrogram that is constructed by converting the frequencies into a Mel-scale. This reduces the loss of information when translating an audio file into an image or pictorial representation. Humans are better at detecting differences between the different sounds that are in lower frequencies rather than in the higher decibels. This Mel-scale was constructed to represent the equal distances in pitch that will sound equally distinct to the listeners [4]. Using this Mel-Spectrogram, Mel-Spec GAN architecture is implemented in this study to perform speech synthesis that has been seen as the efficient implementations in the domains of computer vision. Also, an improved version of this proposed architecture is implemented by using conditional labels that better the quality of synthetic speech.

2 Related Work

Van den Oord et al., 2016 [5] proposed a novel generative model capable of successfully generating raw audio in the form of speech and music named WaveNet. WaveNet's experiment on text to speech synthesis was able to synthesize human-like speech by mapping linguistics and contours. To deal with long-range temporal dependencies for audio production, a dilated casual convolution model was developed with an advantage of storing the input resolution throughout the network. Saito et al., 2017 [6] have performed Statistical Parametric Speech Synthesis by using GAN. This method successfully alleviates the over smoothening effect, that is not being able to capture the nuances in the audio wave, to produce the optimal results than other conventional models, like the Markov model and the Gaussian mixture model [7]. Engel et al., 2019 [8] have demonstrated that GANs are better at efficiently producing audio with faster magnitude than their autoregressive counterparts. Several key findings were observed in this paper, for instance, more coherent waveforms are produced when generating log-magnitude spectrograms and the phases directly with the GANs. We have harnessed the power of the GAN architecture in our research to synthesize the audio signals and evaluated the same using methodology proposed by Engel et al., 2019 [8].

Shen et al., 2017 [9] proposed an entirely neural approach for speech synthesis using recurrent sequence-to-sequence Tacotron model. Even though features such as linguistics, phoneme and log fundamental frequency are primary components of speech, this approach incorporates the Mel-spectrogram, which had

a considerable advantage in reducing the size of WaveNet architecture. Wang et al., 2018 [10] proposed an architecture, that uses a similar method proposed by Shen et al., [9], with an extra layer called the "Style Token" layer for measuring the similarities between the embeddings, created by reference encoder. Style control includes token identification of each attribute, such as pitch, speaking rate and emotion, and changing its magnitude to animate the speech. On inheriting the above notion in our proposed research, the network was able to generate speech in any human-like voice when we train it using Mel-Spectrograms.

Jia et al., 2019 [11] proposed a neural network based system for multi-speaker speech synthesis and TTS methodology. The synthesizer generated not only high-quality speech from the learned speaker but also speakers never seen before. As the diversity in the training set increases, the synthesizer was able to learn the variation between each speaker and produce realistic audio. Donahue et al., 2018 [12] applied WaveGAN to generate raw audio through unsupervised training. WaveGAN is capable of synthesizing the second slice of audio waveforms with global coherence, ideal for generation of sound effects. The analysis finds that, without labels, WaveGAN learns to produce intelligible words when trained on a small-vocabulary speech dataset, and can also synthesize audio from other domains such as drums, bird vocalizations, and piano. We took inspiration from this literature and made a network much efficient than these models.

3 Generative Adversarial Networks

GAN are deep learning models that comprise of two neural networks competing against one another to generate realistic synthetic samples from its respective data distribution $P_{\text{data}}(x)$ [13, 14]. Typical GAN architecture contains a generator and a discriminator.

The generator $G(z)$ neural network accepts a random variable z (Gaussian noise) from the prior distribution and maps it to the pseudo-data distribution through the hidden layers to generate a complex distribution $P_g(z)$. The ultimate goal of the generator is that the generated distribution $P_g(x)$ and the actual data distribution $P_{\text{data}}(x)$ should be as similar as possible [15]. Therefore, the target of the generator is to balance $G*$, as shown in equation (1).

$$G^* = \arg\min_G \text{Div}\left(P_g, P_{\text{data}}\right) \tag{1}$$

To calculate the difference between the two distribution, the original Generative Adversarial Network maneuver a binary classifier [16] model called Discriminator $D(z)$. During training, the output of the discriminator should be 1, if the input is a real sample x. Otherwise, the output is 0. Goodfellow et al. [14] used binary cross-entropy function to define the discriminator, which is popularly used for problems with binary classification. The sample to a discriminator can come either from the actual distribution P_{data} or from the model predicted distribution P_g. Therefore, the complete object function for discriminator is obtained in the following equation (2).

$$\mathbf{V}(\mathbf{G},\mathbf{D}) = E_{x\ P_{\text{data}}}[\log D(x)] + E_{x\ P_g}[\log(1 - D(x))] \tag{2}$$

By combining the equations, (1) and (2), we get min-max optimization function (3) of the Generative Adversarial Network. In this min-max game, the generator tries to delude the discriminator. The generator attempts to maximize the output of the discriminator when a fake sample is introduced. Instead, the discriminator tries to minimize the loss by differentiating between true and false samples. Specifically, the discriminator, maximizes V (G, D) while the generator aims to minimize V (G, D), thus establishing the min-max relationship [17].

$$\min_G \max_D V(G,D) = \min_G \max_D E_{x\ P_{\text{data}}}[\log D(x)] + E_{z\ P_z}[\log(1 - D(G(z)))] \tag{3}$$

When the generator is training, the discriminator's parameters are fixed. The predicted data from the generator is mapped as a fake sample and given as an input to the discriminator. The error is determined by the output of the $D(G(z))$ discriminator that classifies between the positive sample x from the real data set and the negative sample generated from the generator $G(z)$. Finally, the calculated error is used to modify the generator parameters using backpropagation.

4 Methodology

4.1 Benefits of Short-time Fourier Transform

A signal is a variation in a certain quantity over time. For audio, the quantity that varies is air pressure [18]. Typically, the amplitude is measured as a function of the pressure shift around the microphone or receiver unit that initially picked up the audio. The amplitude is a function of its transition over a duration (usually time). A waveform is a visual representation of an audio signal, typically depicted as a representation of the time series, where the value of the y-axis is the amplitude of the waveform. However, it is barely a two-dimensional representation of this dynamic and vibrant audio signal. The waveform itself does not deliver proper class information i.e it is difficult to distinguish between the digits. Therefore, there are GAN models that use a waveform (both as an image and 1-D sequence) to generate data.

We are only observing the resulting amplitudes of the measurements of signal taken over time. Multiple single-frequency sound waves make up an audio signal. Therefore, we used Fourier Transform(FT), which is another mathematical representation of the signal processing to extract useful information. The FT is a numerical method, that decomposes a signal into its frequencies and magnitude. The original audio signal is broken down into a series of sine and cosine waves adding up to the original signal [20]. In other words, it converts the time domain signal into a frequency domain signal. The resulting frequency-domain signal is known as a spectrum.

Audio signals such as music and speech, are referred to as non-periodic signals, because their frequency varies over time. To represent these signals as a spectrum, several small Fourier transforms are calculated on multiple windowed fragments of the signal. This is called as the short-time Fourier Transform (STFT). STFT provides the time-localized frequency information for situations in which frequency components of a signal vary over time [19]. In contrast, the standard Fourier transform provides the frequency information, averaged over the entire signal time interval [21].

4.2 Reasons for Using Mel-Spectrogram

A spectrogram is a visual illustration of a signal's frequency spectrum that changes over time. The STFT is applied over each fragment of the audio signal to obtain a power spectra of the signal. The power spectrum of a time series explains how power is distributed in frequency components (energy per unit time) that compose the signal [22]. The power spectrum is stacked on top of each other to become spectrogram of an audio signal.

However, humans can only perceive a small and concentrated range of frequencies and amplitudes. The calculated spectrogram will not discern between human-perceivable frequencies. Therefore the y-axis (frequency f) is converted to a log scale and the color dimension to decibels [23] (log scale of amplitude). This technique is called Mel-Scale, which approximates the human auditory system's response more closely than narrow frequency bands. Stevens, Volkmann and Newman proposed Mel-Scale as a unit of pitch such that equal distances in pitch sounded equally distant to the listener [24]. Therefore a spectrogram where the frequencies are converted to the Mel-scale in Y-axis is called a Mel-Spectrogram.

In our approach, the audio wave files of speech command digit dataset are dynamically represented as images of Mel-Spectrograms and synthesized in GAN models. Finally, the original audio is then retrieved from an STFT sequence, by taking an inverse transform of each frame, which is overlapped and added iteratively. This algorithm of reconstruction of an audio signal from spectrogram by solving phase recovery is known as Griffin Lim [25]. The restoration of the audio signal will regain its phase with the increasing number of iterations. In our application, we have used 60 repetitions to bring audio back from Mel-Spectrogram.

4.3 Mel-Spec GAN (MSGAN)

In the field of Computer Vision, Generative Adversarial Networks have seen tremendous success in the past years. We can now produce incredibly realistic images which are indistinguishable from the actual ones, showing how far GAN technology has indeed progressed. Since audio signals can be represented as Mel-Spectrogram, it can be incorporated into GANs to synthesize new Mel-Spectrogram that in turn can be converted back to audio signals.

The core of the proposed Mel-Spec GAN model takes inspiration from Deep Convolutional GAN (DCGAN) [26]. Unlike a grayscale image, the audio signals are represented in the form of 3-Dimensional vector frequency bins with single-channel input (128,128,1) and a steady sample rate of 22050. Subsequently, each frequency bin is normalized with mean and standard deviation and rescaled to [-1, 1].

We incorporated the convolutional layers while designing the models of generator and discriminator. The discriminator model takes as input one 128×128x1 Mel-spectrogram and outputs a binary prediction of whether the audio is real or fake. The hidden layers are customized with downsampling blocks consisting of 2×2 stride 2-Dimensional convolutional layers equipped with a 5x5 filter. Inspired from the architecture of WaveGAN [3], the generator model is refashioned by installing 2D transposed convolutional layers and achieved upsampling without using max-pooling and nearest neighbours.

The Rectified Linear Unit(ReLU) activation function $f(x) = x^+ = \max(0, x)$ is implemented in the generator network [29], and the weights are initialized with a slightly positive initial bias using a truncated normal distribution with zero mean and 0.02 standard deviation to avoid dead neurons [27]. The LeakyReLU activation function $f(x) = \max(0, x) + a \cdot \min(0, x)$ with a slope(a) of 0.2 is fashioned in every layer of discriminator except the dense output layer [28]. We regulated the vectors and accelerated the learning process with batch normalization in all layers except the input and output layer. Ultimately, an Adam version of stochastic gradient descent optimizer with a learning rate of 0.0002 and a momentum of 0.5 is adapted in both the models.

We used sigmoid activation function $f(x) = 1/\left(1 + e^{-x}\right)$ with binary cross-entropy loss for the output layer of the discriminator and a tanh activation function $f(x) = \left(e^x - e^{-x}\right)/\left(e^x + e^{-x}\right)$ for the generator to generate a 128x128x1 Mel-spectrogram in the range of -1 to 1.

The Mel-Spec GAN is trained with a pre-processed input audio wave files and used the generator to synthesize new plausible spectrogram, which is reconstructed to audio waveform, using Griffin-Lim algorithm.

4.4 Conditional-Mel-Spec GAN (CMSGAN)

Although Mel-SpecGAN model can fabricate new random plausible audio from a given domain, there is no other way to monitor the types of audios that are generated than trying to find out the strong correlation between the generator's latent space input and the audio produced.

Therefore, we took the Digit labels (0-9) information into consideration to forge Conditional-Mel-SpecGAN (CMSGAN). These labels are transmuted into one-hot encoded vectors y of dimension (data size x 10). By feeding the class label vector y into the generator and the discriminator, the original Mel-SpecGAN is expanded to a conditional model [30]. In this way, the improved Mel-SpecGAN can quickly learn the data distribution of each class independently and generate the samples in accordance with the given condition label y [31]. The loss function of modified GAN is depicted in formula (4).

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x \mid y)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z \mid y)))] \tag{4}$$

To encode the class labels into the discriminator model [32], we embedded input class layer to a fully connected dense layer (128x128) with a linear activation that scales the embedding to the size of the spectrogram (128x128). Furthermore, the embedded input is reshaped into single activation map (128x128x1) before concatenating it with the model as an additional feature map.

In the generator model, we concatenated the latent vector and the input class vector before embedding to the fully connected layer with the activation of 16384 vectors to match the activations of the unconditional generator model. The new ten feature-map is appended as one more channel to existing (Nx100), resulting in (Nx110) feature-maps and up-sampled as in the previous model.

We then trained the GAN model with latent vector and class label as an input, and spawn a prediction of whether the input was genuine or counterfeit. We optimized the GAN by smoothening the real and fake tags of the discriminator with the intention that the loss does not converge to 0. Both GAN models are evaluated by using the Inception score as a stopping criterion.

5 Experimental Protocol

5.1 SC09 Dataset and The Effects of Background Noise

Our research focuses on the Dataset of Spoken Commands [33]. Google brain created this dataset through several speakers recording single words under unregulated recording environments. We analyzed a subset of the spoken command "zero" through "nine" and referred to this subset as the Speech Commands Digits dataset (SC09) [12]. Each recording is one second in duration with different alignment in time. Although this dataset is deliberately similar to the famous MNIST written digit dataset, we note that SC09 examples (128x128) are far higher in dimensions than MNIST examples (28x28). These ten words contain several phonemes, and two are multiple syllables. The training set includes 1850 utterances of every digit, resulting in 5.3 hours of speech [12].

We calculated the amount of background noise in the underlying dataset, using a speech quality measure called Signal to Noise Ratio (SNR) [34]. However, to compare the waveform directly in the time domain, the synchronization of the original and distorted signal was necessary. Since we converted time domain signal to the spectral domain, we can compute SNR using speech parts, typically between 20 and 30 MS long [35]. This method is known as Segmented Signal to Noise Ratio (SegSNR) [35]. This approach is more reliable than their predecessor and less sensitive to signal alignments. Certain digits like 'Nine', 'Six' and 'Five' have negative SegSNR and are highly prone to errors in synthesis. Therefore, the ambiguity of alignments, speakers, and recording environments makes this a challenging modeling dataset [12].

5.2 Pre-Processing and Experimental Setup

We parallelized the pre-processing of the audio signals, while converting them into Mel-Spectrogram images. The functions from Librosa library were utilized to load the WAV formatted audio files. In addition, the audio utility package was constructed using core Librosa functions to compute Mel-Spectrogram and its Inverse. To calculate STFT, we sampled the audio signals with windows of size n_fft=2048, hops of size hop_length=512 to transform from the time domain to the frequency domain. We then took the entire frequency spectrum, and separated it into n_mels=128 evenly spaced frequencies. For each window, we decomposed the magnitude of the signal into frequencies and scaled the corresponding frequencies into a log scale. The uniform dimension (128x128) was maintained for the spectrogram by attributing the difference in value with -80 decibels. Finally, we normalized the Mel-frequencies with mean and standard deviation and scaled to the range of [-1,1]. The Mel-Spectrogram and its corresponding mean and standard deviation are saved as NPZ file so that the model can further synthesize it.

We trained our networks, using batches of size 32 on an NVIDIA TESLA P100 GPU in Google Colab Pro. During our quantitative assessment of SC09, our Mel-SpecGAN networks converged by their evaluation criteria (Inception score) within 5 hours of training (around 80K epochs) and produced speech-like audio after 30k epochs. Our Conditional Mel-SpecGAN networks converged more quickly, within 3 hours (about 50k epochs) and produced better results with much higher Inception score.

6 Evaluation Methodology

6.1 Inception Score

Tim Salimans et al [36]. proposed the Inception Score, which is an empirical metric for measuring the quality, and the semantic discriminability of the image generated by the GAN models [12]. The performance of the Generative Adversarial Networks is monitored by a pre-trained deep learning image classification model, which is incorporated to classify the generated images and uses the conditional probability as a base to calculate the Inception Score [37]. The Inception Score has a minimum value of 1.0 and a maximum value of the number of classes provided by the classification model.

To measure the Inception Score, we trained an audio classifier model on Mel-Spectrogram features of Speech Command Digit Dataset. The pre-processed and normalized Mel-Frequencies were used as an input to the network and the one-hot encoded labels, as the output class vectors. We built our classifier network with four layers of 2D convolutional and pooling, followed by two layers of dense activations, projecting the result to a softmax layer with ten classes [12]. The network was compiled with categorical cross-entropy along with Ada-Delta optimizer. We ran upto 50 epochs with early stopping on the minimum negative log-likelihood of the testing dataset and achieved the accuracy of 99.82% and

saved the model for evaluating the GAN. To calculate the Inception Score, we first used our pre-trained deep learning Mel-Spectrogram classifier model to estimate the conditional probability of generated audio spectrograms ($p(y|x)$). After that, the marginal probability was calculated as the average of the conditional probabilities for the spectrograms in the group ($p(y)$).

$$\text{KL divergence } = p(y \mid x)^{\star}(\log(p(y \mid x)) - \log(p(y))) \tag{5}$$

We combined these metrics and calculated the Kullback–Leibler divergence (KL divergence) for each spectrogram as the product of conditional probability with the log of the same minus the log of the marginal likelihood [36] [38] as exhibited in the formula(5). Finally, the final Inception score is computed by taking the exponent of the summation of the KL divergence and averaged over all classes.

6.2 Quality Human Evaluation

To support the algorithmic evaluation of the model, we incorporated another evaluation metric called Mean Opinion Score (MOS). International Telecommunication Union (ITU) defines the Mean Opinion Score (MOS) as a numerical ranking of the human-judged overall performance of the system quality (voice or video) [39]. MOS is calculated on the scale of 1 (lowest perceived quality) to 5 (highest perceived quality), which is the arithmetic mean of individual values of human-scored parameters.

We measured the ability of human annotators by creating a survey on Amazon Mechanical Turk to rank the generated speech. We identified our best Mel-Spec GAN (MSGAN) and Conditional Mel-Spec GAN (CMSGAN) models by their core evaluation metrics and produced random samples. We created 20 batches of each digit (labelled by Classifier Model) for CMSGAN model and 200 random digits samples for MSGAN model. We customized a form-based layout using Crowd-HTML to develop the UI for the survey and linked with MTurk services for the human-evaluation. We asked the 100 annotators to assign subjective values of 1-5 for sound quality and reported the score in Table 1.

7 Results and Discussion

We implemented the MSGAN and CMSGAN models with Inception Score as evaluation and stopping criteria. The Mel-Spec GAN achieved the Inception Score of 5.76 while improving the same, GAN with label conditions gave a substantial score of 7.64. We compared the scores with other similar implementations of Adversarial Audio Synthesis Networks, like WaveGAN and SpecGAN [12], as shown in the Table 1. To endorse the assessment, we validated the models using a crowd-sourced experiment. While the MSGAN averaged a MOS of 3.01, CMSGAN achieved a higher MOS of 3.74. By leveraging the Mel-Spectrograms and label embedding, the speech quality of the proposed model improved by 63.6%.

Table 1. Comparing Inception Scores of the models

S.No	*Network*	*Inception Score*	*MOS*
1	Real (train)	9.18	
2	Real (test)	8.01	4.02
3	WaveGAN (Phase shuffle n=2) [12]	4.67	
4	SpecGAN [12]	6.03	
5	Mel-Spec GAN (proposed)	**5.76**	3.01
6	Conditional Mel-Spec GAN (improved)	**7.64**	3.74

8 Conclusion

In this paper, we synthesized the speech signals using various GAN models and introduced a novel methodology for generating audio waveform. We used frequency domain representation of the audio signal and converted wave files to Mel-Spectrogram image files, using the short-time Fourier transform. By incorporating ideas from image synthesizing DCGAN, we customized the Mel-Spec GAN architecture by leveraging the Mel-Spectrogram of the audio signal. The generated Mel-Spectrograms are converted back to the actual waveform, using fast Griffin Lim algorithm by solving conversion loss. We then improved the proposed architecture by embedding label as an input, so that the audio tags condition the output. Not only did the Inception Score of the model improve from 5.7 to 7.6, but also, the model converged quickly to achieve extraordinary results.

In its current form, Mel-Spec GANs can be used for real-time speech synthesis. In our future work, we plan to extend the potential of GANs to operate on variable and longer length audios and multiple accents. By providing a template for speech synthesis and Mel-spectrogram generation models to serve on speech signals, we hope that this research will catalyze future audio-synthesis experiments of GANs.

9 Acknowledgements

This work is supported in part by by Science Foundation Ireland (Grant Nos. SFI/12/RC/2289_P2 and 17/RC-PhD/3482). We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp used for this research.

References

1. Saito, Y., Takamichi, S., Saruwatari, H.: Statistical Parametric Speech Synthesis Incorporating Generative Adversarial Networks. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 26, 84-96 (2018).

2. Pasini, M.: MelGAN-VC: Voice Conversion and Audio Style Transfer on arbitrarily long samples using Spectrograms, https://arxiv.org/abs/1910.03713.
3. Donahue, C., McAuley, J., Puckette, M.: Synthesizing Audio with GANs. ICLR, https://openreview.net/forum?id=r1RwYIJPM.
4. Biswas, S., Solanki, S.: Speaker recognition: an enhanced approach to identify singer voice using neural network. International Journal of Speech Technology. (2020).
5. Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., Kavukcuoglu, K.: WaveNet: A Generative Model for Raw Audio, https://arxiv.org/abs/1609.03499.
6. Saito, Y., Takamichi, S., Saruwatari, H.: Statistical Parametric Speech Synthesis Incorporating Generative Adversarial Networks. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 26, 84-96 (2018).
7. Awad, M., Khanna, R.: Hidden Markov Model. Efficient Learning Machines. 81-104 (2015).
8. Engel, J., Agrawal, K., Chen, S., Gulrajani, I., Donahue, C., Roberts, A.: GANSynth: Adversarial Neural Audio Synthesis, https://arxiv.org/abs/1902.08710.
9. Shen, J. Pang, R. Weiss, R.J. Schuster,M. Jaitly,N .Yang,Z. Zhang,Y. Wang,Y. Skerrv-Ryan, R. Saurous,R.A. Agiomvrgiannakis,Y. Wu,Y.:Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions.IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 4779-4783 (2018).
10. Wang, Y., Stanton, D., Zhang, Y., Skerry-Ryan, R., Battenberg, E., Shor, J., Xiao, Y., Ren, F., Jia, Y., Saurous, R.: Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis, https://arxiv.org/abs/1803.09017.
11. Jia, Y., Zhang, Y., Weiss, R., Wang, Q., Shen, J., Ren, F., Chen, Z., Nguyen, P., Pang, R., Moreno, I., Wu, Y.: Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis, https://arxiv.org/abs/1806.04558.
12. Donahue, C., McAuley, J., Puckette, M.: Adversarial Audio Synthesis, https://arxiv.org/abs/1802.04208.
13. Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., Bharath, A.: Generative Adversarial Networks: An Overview. IEEE Signal Processing Magazine. 35, 53-65 (2018).
14. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative Adversarial Networks, https://arxiv.org/abs/1406.2661.
15. Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S., Lew, M.: Deep learning for visual understanding: A review. Neurocomputing. 187, 27-48 (2016).
16. Deng, L.: The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation, and Machine Learning. Technometrics. 48, 147-148 (2006).
17. Lan, L., You, L., Zhang, Z., Fan, Z., Zhao, W., Zeng, N., Chen, Y., Zhou, X.: Generative Adversarial Networks and Its Applications in Biomedical Informatics. Frontiers in Public Health. 8, (2020).
18. Allen,D.: Chapter 1. Sounds and Signals. In Think DSP: Digital Signal Processing in Python, First edition., pp 1–11, Sebastopol, CA: O'Reilly Media, Inc (2016).
19. Kehtarnavaz,N.: Digital Signal Processing System Design. Amsterdam. Academic Press, (2008).
20. van den Bogaert, B.: When Frequencies Change in Time; Towards the Wavelet Transform. Data Handling in Science and Technology. 33-55 (2000).

21. Moorer, J.: A note on the implementation of audio processing by short-term fourier transform. 2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). (2017).
22. Vetterling, W. T., & Press, W. H.: Numerical recipes in Fortran: the art of scientific computing (Vol. 1). Cambridge University Press (1992).
23. Roberts, L.: Understanding the Mel Spectrogram, https://medium.com/analytics-vidhya/understanding-the-mel-spectrogram-fca2afa2ce53.
24. Stevens, S., Volkmann, J., Newman, E.: A Scale for the Measurement of the Psychological Magnitude Pitch. The Journal of the Acoustical Society of America. 8, 185-190 (1937).
25. Griffin, D., Jae Lim: Signal estimation from modified short-time Fourier transform. IEEE Transactions on Acoustics, Speech, and Signal Processing. 32, 236-243 (1984).
26. Radford, A., Metz, L., Chintala, S.: Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks, https://arxiv.org/abs/1511.06434.
27. Goodfellow, I., Bengio, Y. and Courville, A.: Deep Learning.Cambridge, MA: MIT Press, pp.175-250 (2017).
28. Xu, Y., Du, B., Zhang, L.: Can We Generate Good Samples for Hyperspectral Classification? — A Generative Adversarial Network Based Method. IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium. (2018).
29. Ramachandran, P., Zoph, B., Le, Q.: Searching for Activation Functions. IEEE. (2017), https://arxiv.org/abs/1710.05941.
30. Mirza, M., Osindero, S.: Conditional Generative Adversarial Nets, https://arxiv.org/abs/1411.1784.
31. Xu, Y., Du, B., Zhang, L.: Can We Generate Good Samples for Hyperspectral Classification? — A Generative Adversarial Network Based Method. IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium. (2018).
32. Denton, E., Chintala, S., Szlam, A., Fergus, R.: Deep Generative Image Models using a Laplacian Pyramid of Adversarial Networks, https://arxiv.org/abs/1506.05751.
33. 3.Warden, P.: Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition, https://arxiv.org/abs/1804.03209.
34. Prodeus, A., Didkovskyi, V., Didkovska, M., Kotvytskyi, I., Motorniuk, D., Khrapachevskyi, A.: Objective and Subjective Assessment of the Quality and Intelligibility of Noised Speech. 2018 International Scientific-Practical Conference Problems of Infocommunications. Science and Technology (PIC S&T). (2018).
35. Mohamed, S.: Objective Speech Quality Measures, http://www.irisa.fr/armor/lesmembres/Mohamed/Thesis/node94.html.
36. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved Techniques for Training GANs, https://arxiv.org/abs/1606.03498.
37. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the Inception Architecture for Computer Vision. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2016).
38. Kullback, S., Leibler, R.: On Information and Sufficiency. The Annals of Mathematical Statistics. 22, 79-86 (1951).
39. "ITU-T Rec. P.10/G.100 (11/2017) Vocabulary for performance, qualityof service and quality of experience." (2017).