International Conference on Electrical, Computer &
Telecommunication Engineering (ICECTE) 26-28 December, 2019

Audio Future Block Prediction with Conditional Generative Adversarial Network

Md. Rahat-uz-Zaman*, Shadmaan Hye, Mahmudul Hasan
Department of Computer Science and Engineering
Khulna University of Engineering & Technology, Bangladesh
Email: *rahatzamancse@gmail.com, praptishadmaan@gmail.com, mahmudul@cse.kuet.ac.bd

***Abstract*—Signal processing is a vast subfield of electrical and computer science where audio signal processing has secured a remarkable position to restore corrupted or missing audio blocks. However, generating possible future audio block from the previous audio block is still a new idea that can help to reduce both audio noise and partially missing an audio segment. In this paper, a generative adversarial network (GAN) along with a pipeline is proposed for the prediction of possible audio after an input audio sequence. The proposed model uses short-time Fourier transformation of audio to make it an image. The image is then fed to a conditional GAN to predict the output image. After that, Inverse short-time Fourier transform is then applied to that predicted image, generating the predicted audio sequence. For a small audio sequence prediction, the proposed methodology is quite fast, robust and has achieved a loss of 0.43. So it is may work well if deployed on a voice call and broadcasting applications.**

***Index Terms*—Audio, Short Time Fourier Transform, Conditional Generative Adversarial Network**

I. Introduction

Overcoming the network instability is one of the major issues in online voice call. Due to network instability, the call can be interrupted for a moment. If this interruption happens several times, the user experience of the calling service is considered as low. For a seamless audio calling experience, it is imperative for the service providers to maintain network stability, prevent audio interruption, or take any step to make it less hampering.

Rather than increasing the network bandwidth or quality of the medium, many low-cost methodologies can be used to interpolate the audio segment in the places where the packet has been lost. Alongside the interpolation technique, in some applications, audio from the previous segment is played where the packet is lost [1].

A better approach to audio generation due to audio time frame loss or corruption is to use modern deep learning techniques. These techniques always predict one-time frame ahead of the current time frame. So, if the audio time frame is not found within a specific small time, the generated audio will be played. This paper introduces a similar methodology where Generative Adversarial Network [2] is used as the deep learning model.

Current works of future block prediction, specifically for audio, are discussed in the next section. In the rest of the paper, the methodology along with experiments and results are provided. Then a discussion of the GAN used for future time frame prediction is provided. Also, it is demonstrates how audio time frame can be converted to images with short time Fourier transform (STFT) and fed to the convolutional neural network.

II. Literature Review

By far, there is minimal research work similar to the proposal served in this paper. So it is difficult to give any comparison study with state-of-the-art achievements for this work. Elowsson and Friberg [3] has generated music audio with ensemble learning and achieved the loss R2 of 0.82. Feature extraction of the audio is done with sectional spectral flux, and then the generated features are trained with ensembled multilayer perceptron. All of these works are not needed for the proposed architecture because of automatic feature extraction on deep learning.

This type of next sequence prediction is also extended in videos. Reda et al. [4] introduced such analysis on videos. They used past frames and past optical flow as a feature and introduced spatially-displaced convolution (SDC) module for frame prediction. Their loss measurement is done in SSIM and was 0.904 on high-definition YouTube-8M videos [5]. This paper does not require past optical flow because it is irrelevant for unnatural images such as spectrograms.

Compared to other approaches, this paper turns the extracted audio features into images and then predict the next possible image. Later the image is converted to the audio again. If the images can be thought of as a frame, then this methodology is most similar to video next frame prediction.

III. Audio Future Block Prediction

The whole architecture of the proposed system is illustrated in Fig. 1. In the final application, only the generator part of the adversarial network has been used with trained weights. The audio is first transformed with short time Fourier transformation. Then the matrix of magnitudes of complex numbers is used as the input of the generator network. The output of the generator is the magnitude of the predicted audio images spectrogram. The experimental result shows that for a minimal time, the phase can be neglected without affecting that much to the final predicted audio. The final predicted audio can be found by inverse short time Fourier transformation of the predicted spectrum.

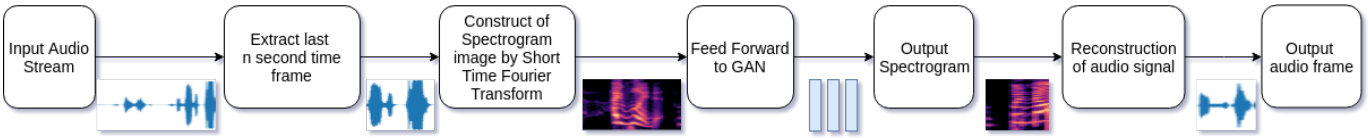

Fig. 1. Block Diagram of proposed complete system.

Conditional GAN is preferred to Deconvolutional neural network [6] (DeconvNet) because generator of GAN is faster than DeconvNet. The workflow of Conditional GAN with proper input and output is briefly visualized in Fig. 2. GANs are generative models that take in mapping from irregular clamor vector z to yield picture y, G: z $\rightarrow$ y. Whereas, Conditional GANs take in mapping from watched image x and irregular commotion vector z to y, G: $\{x, z\} \rightarrow y$ [7]. The generator G is prepared to create images that can not be recognized as fake by an adversarial network discriminator, D which is prepared to identify the generator's "fakes".

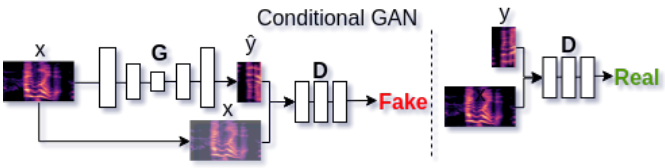

Fig. 2. Workflow of Conditional GAN.

In the context of this paper, the output of the model is the magnitude spectrum of short time Fourier transformed audio time frame. The phase of the Fourier transform is taken to be 1 for all the time and frequency points in the spectrogram.

IV. GAN Architecture

GAN is a composition of two systems, a generator model, and a discriminator model, as stated in section III. These two systems can be neural networks, extending from convolutional neural systems [8] to Long short term memory RNNs [9]. The main prosperity of GAN has come to the field of content generation. There are numerous image generation researches done with the help of GAN. It is evident to use GAN for this audio generation task. Moreover, GAN with Deconvolutional networks [6] has promising results on several image generation tasks that are faster than other equivalent models. Hence in this paper, Deep Convolutional GAN (DCGAN) [10] was decided to use for audio generation. The generator and the discriminator network architectures are demonstrated in Fig. 3.

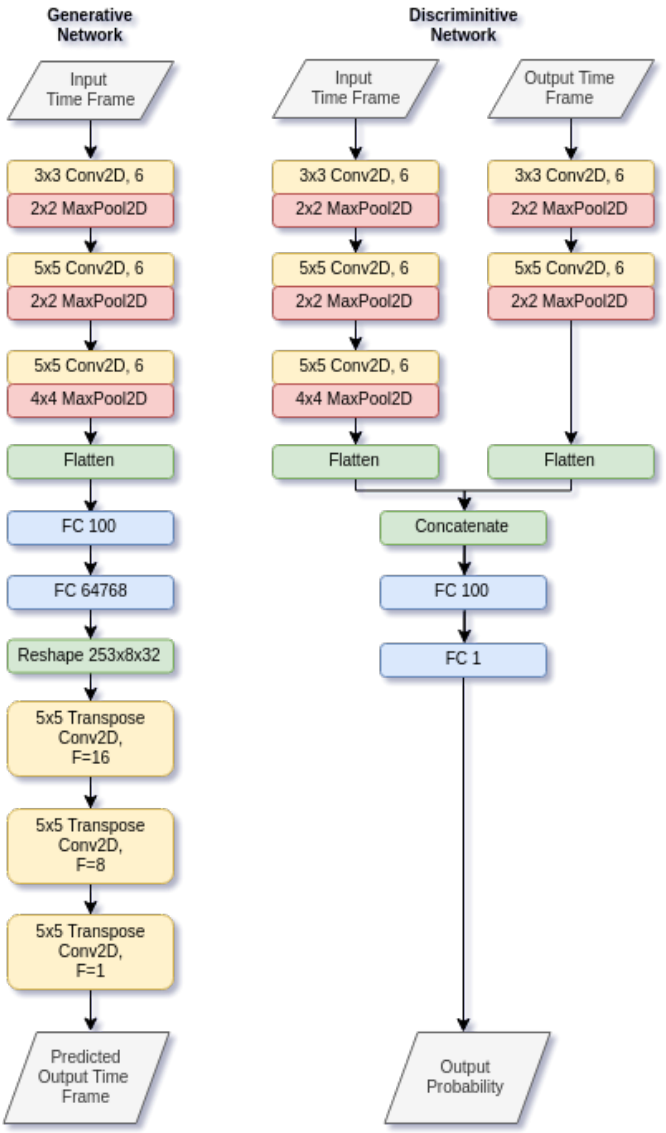

Fig. 3. Proposed architecture of GAN.

In the proposed model in this paper, Stochastic Adam optimizer [11], Binary Cross Entropy loss function [12] and Leaky Rectified Linear Unit [13] activation function have been used. During the training process in each iteration, the audio spectrogram is fed to the generator for getting the output time frame. Then the predicted output time frame and the input time frame are passed through the discriminator neural network, which yields the probability of the two audio being from different distributions. The discriminator is trained on the input time frame and the predicted output time frame by the generator with a label of 1. Here, a performance increase is noticed when using a range from 0.8 to 1 instead of using only 1 as a label. The discriminator is also trained with the input time frame and the actual time frame with the label of 0 to 0.2. Finally, the generator network is trained on input time frames with labels always equal to 1.

V. Audio Sampling and Short Time Fourier Transformation

For the training process, proper inputs and targets need to be generated from the existing ones. First, all the audio files of the dataset are sampled to fixed length audio. This fixed length audio is then split into two audio files. The first file of the pair of audio files is stated as an input time frame and the second file is stated as an output time frame. This notation will be used for the rest of the paper. The input and output time frame may not be equal in length.

Spectrogram is a time series visual representation of the spectrum of frequencies of a signal. To represent spectrograms as images in this paper, a two-dimensional graph is used. The x-axis represents time and the y-axis represents frequency. Another third dimension is used to indicate the amplitude of a particular frequency at a particular time which is represented by the intensity or color of each point in the image. On Fig. 4, the spectrum of an audio of a lady on phone call (first 5 seconds) from freesound.org [14] is shown. The corresponding log scaled spectrogram is also shown with a sampling rate of 22.05 kHz. The spectrogram is found with short time Fourier transformation of the wavelet with (1) where the audio is $x[n]$ and the window is $w[n]$.

$$STFT\{x[n]\}(m,\omega) \equiv X(m,\omega) = \sum_{n=-\infty}^{\infty} x[n]w[n-m]e^{-j\omega n} \tag{1}$$

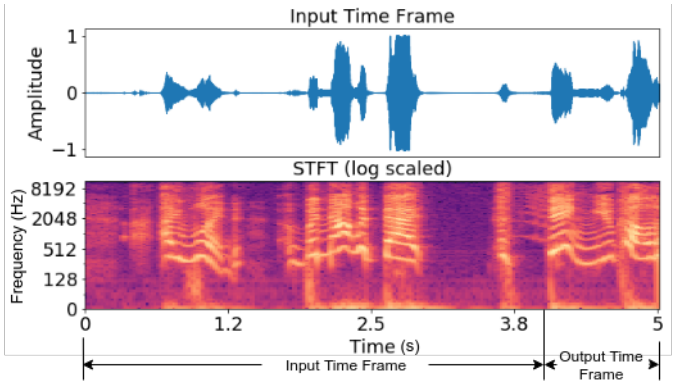

Fig. 4. Audio signal amplitude and log scale spectrogram by STFT.

VI. Experiments and Results

A. Dataset

The primary dataset used for this research is audioset speech dataset [15] provided by Google and librispeech [16] contributed by Vassil Panayotov. From the audioset dataset, 1,010,480 videos are labeled as human speech. These speech videos are further divided into 5,735 balanced train, 999,421 unbalanced train, and 5,324 evaluation videos. This paper is only concerned with the audio of the whole balanced train set and audio of only 10% handpicked video of the unbalanced train set. Also, the model is evaluated with the provided evaluation set.

The Librispeech dataset is well structured into several folders and provides CSV file as metadata. The original librispeech contains 1000 hours of English reading sampled at 16kHz taking about 87 gigabytes of storage space. Several preprocessed and sampled smaller datasets can be found. Our chosen dataset from these sampled librispeech datasets is prepared for machine learning purposes and is divided into train and test set. The train set contains 360 hours, and the test set includes 5.6 hours of clean English speech of both males and females.

B. Implementation details

The complete implementation is done with python programming language, PyTorch [17] and Keras [18] deep learning library. Librosa [19] library is used for audio processing. Firstly, the input audio files are preprocessed, and spectrogram images of each audio file are saved on disk. These images have huge widths (time) and fixed height (frequency) and took a considerably large amount of disk space (300gb). The model is trained on a custom-built PC (Intel Core i7 @ 3.5 GHz, RAM: 32 GB with GPU NVIDIA GeForce GTX 1080 Ti).

C. Result Analysis

The training is done with 10 fold cross-validation on the training set for approximately 8 hours and evaluated on the evaluation set. The graph of iteration vs. loss of both the generator and the discriminator during the training process is shown in Fig. 5. It can be pointed out that the loss is not converged even after 800 epochs. So, the loss would have decreased more if it was trained for a more extended period. The average loss on the training set of 10 folds was 0.782 (output of discriminator). After evaluating on the test set, the loss was found to be 0.753. Due to the proximity of losses between train and test set, we can conclude that the model is not overfitted on the dataset.

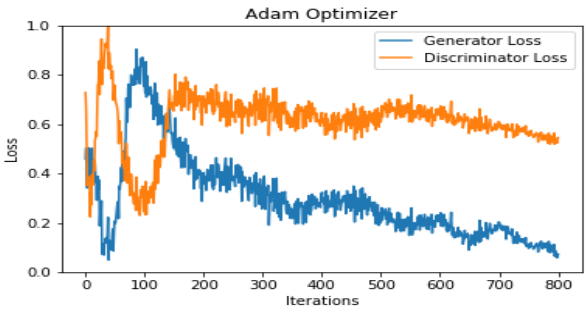

Fig. 5. Number of iterations vs loss of generator and discriminator model.

After training the model, the corresponding actual output and predicted output of the sample input are given in Fig. 6. The similarity of output time frame and output of discriminator proves that the model is indeed working as expected.

Similar models like Fig. 3 are trained on the different selected length of time frame shown in Table I. The neural networks are similar in overall shape but are tuned for different

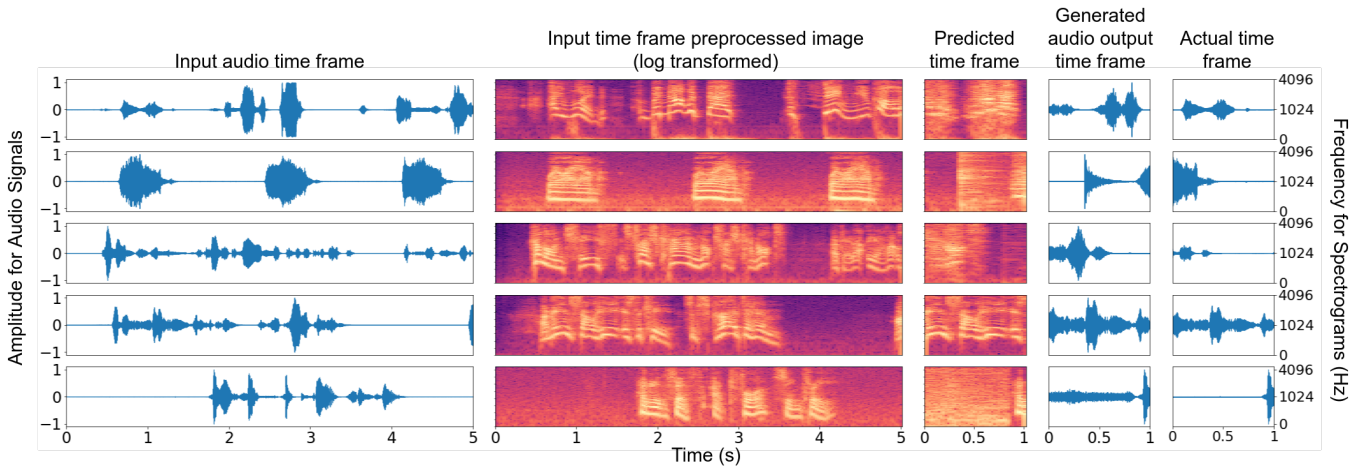

Fig. 6. Audio signal, STFT spectrogram(in log scale), predicted STFT spectrogram(in log scale), predicted future audio block and actual future audio block.

input sizes. It seems on Table I that longer input time frame and shorter output time frame leads to better quality audio prediction. However, the time needed to predict is high for longer inputs.

TABLE I
LOSSES WITH RESPECT TO INPUT AND OUTPUT TIME FRAME LENGTH

Input time frame length and image size	Output time frame length and image size	Loss (400 epochs)	Time for prediction (ms)
5 (1025x216)	2 (1025x87)	0.58	3.10
5 (1025x216)	1 (1025x44)	0.43	2.52
3 (1025x130)	2 (1025x87)	0.71	2.11
3 (1025x130)	1 (1025x44)	0.64	1.23
2 (1025x87)	1 (1025x44)	0.69	1.04
1 (1025x44)	1 (1025x44)	0.75	0.70

VII. Discussion and Conclusion

The results in section VI-C show that the state-of-the-art GAN for image generation can also be used for conditional audio generation. The results are better than a shallow, fully connected neural network fed with audio amplitude sequence because convolution operations can quickly identify spatial features. In section VI-C, it is stated that, by tuning and additional training, the result can be further improved to get a more realistic audio future block. As the time required to predict the future block is very low, this methodology can be used real time for voice calling or broadcasting applications.

References

[1] H. Chen, I. Tsukagoshi, and M. Mehta, "Digital audio decoder having error concealment using a dynamic recovery delay and frame repeating and also having fast audio muting capabilities," Jul. 5 2005, uS Patent 6,915,263.

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[3] A. Elowsson and A. Friberg, "Predicting the perception of performed dynamics in music audio with ensemble learning," *The Journal of the Acoustical Society of America*, vol. 141, no. 3, pp. 2224–2242, 2017.

[4] F. A. Reda, G. Liu, K. J. Shih, R. Kirby, J. Barker, D. Tarjan, A. Tao, and B. Catanzaro, "Sdc-net: Video prediction using spatially-displaced convolution," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 718–733.

[5] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, "Youtube-8m: A large-scale video classification benchmark," *arXiv preprint arXiv:1609.08675*, 2016.

[6] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1520–1528.

[7] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *CVPR*, 2017.

[8] Y. LeCun, Y. Bengio *et al.*, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, vol. 3361, no. 10, p. 1995, 1995.

[9] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

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

[11] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[12] S. Mannor, D. Peleg, and R. Rubinstein, "The cross entropy method for classification," in *Proceedings of the 22nd international conference on Machine learning*. ACM, 2005, pp. 561–568.

[13] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network," *arXiv preprint arXiv:1505.00853*, 2015.

[14] A. et al., "Freesound 2: An improved platform for sharing audio clips," in *Klapuri A, Leider C, editors. ISMIR 2011: Proceedings of the 12th International Society for Music Information Retrieval Conference; 2011 October 24-28; Miami, Florida (USA). Miami: University of Miami; 2011.* International Society for Music Information Retrieval (ISMIR), 2011.

[15] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[16] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[17] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," in *NIPS-W*, 2017.

[18] F. Chollet, "Keras documentation," *keras. io*, 2015.

[19] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, 2015, pp. 18–25.