Audio Bit Depth Super-Resolution with Neural Networks

Thomas Liu
University of British Columbia
tyangliu@gmail.com

Taylor Lundy
University of British Columbia
tlundy@cs.ubc.ca

William Qi
University of British Columbia
wiqi@cs.ubc.ca

Abstract

Audio bit depth super-resolution is a problem that has yet to be examined through the lens of deep learning, with few effective methods in use today. In this paper, we propose a WaveNet-based architecture to perform upscaling of low-resolution 8-bit audio input to generate high-fidelity 16-bit output, eliminating noise and artifacts in the process. We also explore several different methods to improve the computational tractability of this problem, providing a thorough analysis of the advantages and drawbacks presented by each.

1. Introduction

Within the domain of audio signal processing, there exists a number of upsampling and super-resolution problems that have yet to be examined through the lens of deep learning. One major reason for this relative lack of progress compared to the visual domain has been due to the inherent difficulty in working with high-dimensional audio data. To lessen the computational burden stemming from this "curse of high dimensionality", many previous learning-based methods for signal processing have instead focused on the use of time-frequency audio representations (such as spectrograms) as input [15, 1, 16].

In recent years, results from WaveNet-based [12] methods have shown that it is indeed possible to operate directly on raw audio input under reasonable memory constraints. With improved access to temporal information in data, these generative [8, 7] and discriminative [9] methods have been able to exploit valuable time-dependent features such as signal phase, expanding the scope of audio tasks that machine learning techniques can solve [17].

One task that could benefit from the availability of temporal information is audio bit-depth super-resolution. Bit-depth, as used in this paper, refers to the number of bits used to represent each sample within an audio signal. Importantly, this variable controls the range of volume an audio signal is capable of representing (dynamic range), playing a significant role in determining perceived audio quality. As the accurate representation of a larger dynamic range requires the use of more bits at every sample, it is often the case that a trade-off has to be made between the size and fidelity of an audio file. Commonly performed compression operations include downscaling from 24-bit studio quality input to 16-bit CD-quality output and from 16-bit to 8-bit audio for use in vintage computer or arcade systems.

As many compressed audio encoding formats downsample bit-depth in a lossy format [2], the recovery of high fidelity audio from degraded input is an ill-defined and hardly trivial problem. In order to mitigate the negative effects of bit-depth reduction, a number of techniques have been previously proposed. One method, oversampling, sacrifices audio bit-rate and signal noise in exchange for higher simulated audio bit-depth. Another method, dithering, introduces noise at the downsampling step to increase the perceived dynamic range of the low-resolution (LR) signal. To date, we are not aware of any methods that have taken advantage of deep networks to upscale audio bit-depth by learning relationships between LR input and HR source audio.

Even with the application of these noise-mitigation techniques, there are often still significant quantitative and qualitative differences between source and compressed audio signals. We hypothesize that by taking advantage of the temporal dependencies inherent in the raw audio format, we can further decrease the discrepancy in signal quality by super-resolving a LR signal's bit-depth. This mapping from 8-bit input to 16-bit output can be learned in a supervised manner using a modified version of the WaveNet architecture that we term BDSR.

The success of this approach is reliant on the existence of semantic structure within the audio signal, which may differ between audio sources. To explore this, we evaluated our model on two different commonly-occurring categories of audio, music and speech.

2. Related Work

Audio generation has been explored on different but related tasks, such as text-to-speech, audio denoising, and bandwidth extension. The WaveNet architecture [12] is the audio domain equivalent of the PixelCNN [13]. Wavenet has demonstrated high-quality audio generation, using a dilated CNN structure to retain a large receptive field, while still remaining computationally tractable on raw audio waveform containing 16,000 samples per second. The original WaveNet paper focuses on text-to-speech synthesis, while derivative work has experimented using the architecture on tasks such as music generation [3] and audio denoising [9].

The most relevant derivative work to our task is the speech denoising work done by Rethage et al. Similarly to bit-depth super resolution, speech denoising necessitates a discriminative model with one-to-one correspondence between input and output samples, unlike the autoregressive nature of the original Wavenet architecture. Additionally, due to the discriminative nature of the problem, it is possible to loosen the causality constraint of Wavenet, and condition on input samples from future timesteps during prediction. Unlike with autoregressive generation where future samples simply do not exist during test time, with our problem of discriminative bit-depth super resolution, samples from both past and future may provide valuable contextual information.

Network structures based on RNNs, traditionally a natural fit for sequential data, have also been explored. Notably, SampleRNN [6] proposes a hierarchical GRU model to generate audio with quality on-par with WaveNet while using a simpler network and providing faster inference. The hierarchical structure attempts to address the empirically-shown poor receptive field of RNNs, which would otherwise be problematic over the several tens of thousands of samples across a few seconds of audio.

A problem to note is that, due to intractability of evaluating softmax distributions over 16-bits (65536 probabilities per sample), both the original WaveNet and SampleRNN produce 8-bit audio as output, lowering the output to only 256 probabilities per sample. Since most modern audio is encoded in 16-bits, and there exists an interesting use-case of exploring bit-depth super-resolution on traditionally 8-bit tracks (such as video game music), we seek a method to tractability produce 16-bit output. A potential solution proposed in PixelCNN++ [10], for the similar task of per-pixel image generation, is to model the 256-way, or in this case 65536-way, categorical distribution using a discretized logistic mixture. This technique is adopted for audio generation in the parallelized WaveNet optimizations [13] to generate high-fidelity 16-bit voice clips, currently deployed in Google Assistant.

3. Bit-Depth Super Resolution

3.1. Setup

Given a low-resolution signal x with bit depth b_1, the goal of our model is to reconstruct a high-resolution version y of x with bit depth $b_2 > b_1$. For example, if x is a sample of music compressed to 8-bit depth with audible hisses and decreased dynamic range; y would be a high-resolution 16-bit reconstruction of the original audio track.

To recover the under-defined signal, we attempt to learn a model $p(y|x)$ of the higher-resolution signal y, conditioned on its lower-resolution input x. As our model is fully convolutional, it is capable of upscaling audio samples of any length.

3.2. Architecture

The architecture of our model is structurally similar to the modified WaveNet used by Rethage et al. [9] for audio denoising, with both models built from blocks containing stacked layers of dilated convolutions. These dilated convolutions serve to increase the receptive field by "skipping" over input samples with some frequency. By beginning with dilation of 1 in the first layer (equivalent to a standard convolution) and doubling the dilation factor at each layer, a model's receptive field can be increased exponentially using only a linearly growing number of layers. This serves a role similar to pooling layers in CNNs, but preserves input dimensionality, which is important as the output length of the audio must match with the input. We detail the modifications made to WaveNet to adapt this problem to bit-depth super resolution.

3.2.1 Discriminative Prediction

Unlike the autoregressive nature of the original WaveNet, which learns to predict the next sample given all previous samples, the bit-depth super resolution problem entails a discriminative model, where the model is trained to output samples with one output sample per input sample. In this case, prediction is conditioned purely on input samples and further leads to the changes described below including non-causality and larger convolutional kernels.

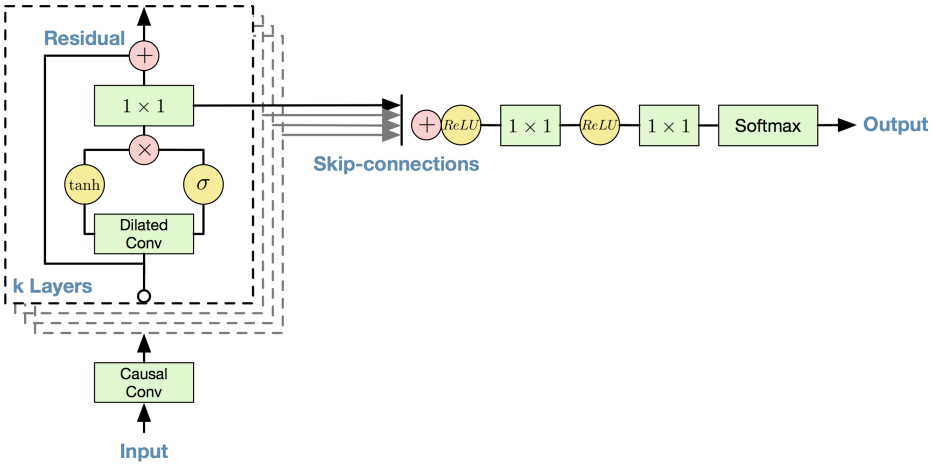

Figure 1. High level overview of the WaveNet architecture.[12]

3.2.2 Non-Causal Dilations

One of the modifications to WaveNet incorporated in BDSR from Rethage's [9] speech denoising model is the removal of causality. As causal convolutions force predictions at each time step to rely only on inputs from previous time steps, we found this to be unnecessary for the task of bit-depth upscaling. Unlike the task of novel audio generation for which WaveNet was originally designed for, the entirety of the input audio signal is available from the beginning. This is implemented in the model by shifting the receptive field forward in time by half of its size, changing the source of input from past samples only to a 50/50 mix of samples from the past and future.

3.2.3 Smoothing Convolution Kernel

Since our model is no longer autoregressive and prediction is not conditioned on previous output, it is possible for discontinuities to be introduced in the output, resulting in the perception of audio artifacts. To prevent this from occurring, we change the 1x1 convolutions on the skip connection outputs to 3x1 convolutions [9] which contributes smoothness to the output. This change in kernel size is the only change from the high level overview of the WaveNet model that presented in Figure 2.

3.2.4 Non-Linear Activation

The remaining components of our model's architecture are similar to the original implementation of WaveNet. We use the same non-linear activation function first introduced by PixelCNN [13], which outperforms the usual rectified linear activation (ReLU) function. The activation function used in BDSR is defined as:

$$z = tanh(W_{f,k} * x) \cdot \sigma(W_{g,k} * x)$$

where x denotes the input and W is a set of learned parameters; k refers to the index of the current layer, f and g denote either filter or gate. We also make use of parameterized skip connections, allowing for training of deeper models by propagating representations learned at lower layers directly to the output [9].

3.3. High Resolution Generation and Loss

Within many domains of audio, 16-bit depth is often the minimum required to accurately reproduce sound free of artifacts and defects. However, with a wide range of possible values at each time step, working with 16-bit audio can often be prohibitively expensive. The original WaveNet model addressed this issue through the use of a pre-processing step with μ-law companding to quantize input to 8 bits. In contrast, BDSR operates directly on lower-quality 8-bit audio input and attempts to learn a mapping to HR 16-bit audio.

Producing 16-bit audio directly from 8-bit or lower audio input has been computationally intractable in the past, due to this necessitating a 65536-way categorical distribution as output. Containing the probability for each of the 2^{16} possible amplitudes, 16,000 of these distributions would need to be generated per second of audio, one for each sample. This places an enormous computational burden on models operating on 16-bit audio, affecting both training and inference. We explored two approaches to practically predicting 16-bit audio output.

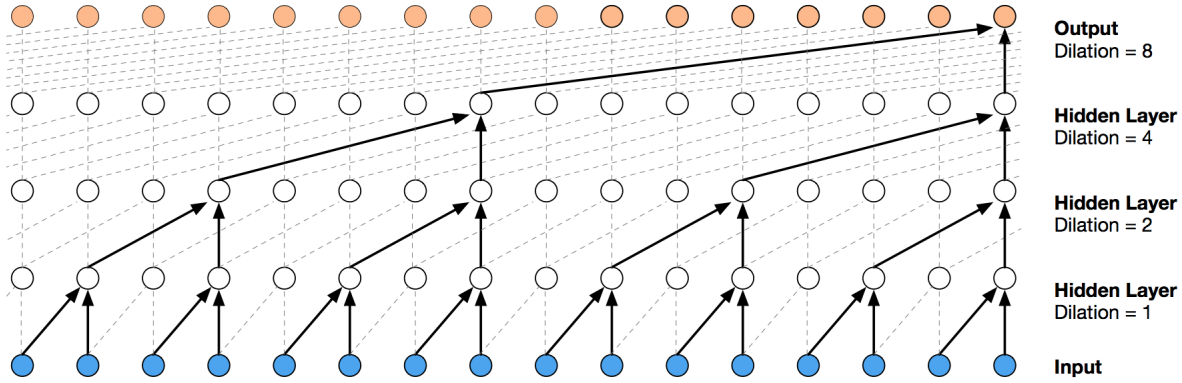

Figure 2. An example of dilated causal convolutional layers with exponentially increasing step sizes.[12]

3.3.1 Logistic Mixture Approximation

The first approach was to formulate the upscaling problem by approximating the 65536-way categorical distribution representing the range of amplitudes with a lower-dimensionality mixture distribution. Our early experiments with this involved the use of a discretized logistic mixture likelihood loss (introduced by PixelCNN++ [10]). This would model the full categorical distribution of 16-bit audio with only a mixture of 10 logistic distributions (μ, π, s per distribution, totaling to 30 output channels as opposed to 65536). In previous work, this type of loss has been shown to be effective at modeling both 3-channel visual output as well as 16-bit audio output, the latter of which was used in Google's parallelized implementation of WaveNet [7]. In practice, despite significant memory savings, training for clean output was difficult to achieve, and we ultimately shifted focus to our next approach which demonstrated better results.

3.3.2 Delta Prediction

An observation leading to an alternative method of 16-bit prediction was that in most reasonable downsampling procedures of 16-bit audio sequence $\boldsymbol{h}$ to 8-bit sequence $\boldsymbol{l}$, the loss of information of the amplitude at timestep t is constrained such that $\boldsymbol{h}_t = \boldsymbol{l}_t \times 256 \pm d$, where $d \leq \frac{2^{16}}{2^8}$. This is intuitively true and verified empirically when we aligned our 8-bit and 16-bit audio and examined the range of amplitude differences. If we model the problem as learning the inverse mapping of the downsampling procedure under this constraint, we can simplify the task of our prediction to predict only the delta between the low-resolution 8-bit sequence and the 16-bit high-resolution sequence. Our objective then is to predict as accurately as possible one of 512 amplitude deltas between -256 and 255.

3.3.3 Real-Valued Prediction

As an extension of delta prediction, we observed better performance by modeling the output directly as a real value between -256 and 255 rather than a discrete 512-way categorical distribution. This method was shown to be effective for denoising in [9]. In addition to slightly reducing the model size, using real-valued prediction enforced smoothness of the output distribution, as neighboring values of amplitude imply similarity in audio quality. We empirically verified that volume of background noise was closely correlated with the mean squared error of deltas between prediction and source audio.

4. Results and Evaluation

The goal of our project was to demonstrate the effectiveness of deep neural networks for audio bit-depth super-resolution. We hoped to show that our proposed model is capable of outperforming existing approaches that naively map signals from the low-resolution to high-resolution domain using a constant scale factor, which does not recover any additional audio detail.

4.1. Setup

4.1.1 Datasets

We evaluated all of our models on two datasets: the VCTK dataset [14], which contains 44 hours of data from 108 different speakers and the Mehri Piano dataset [6], which contains 10 hours of Beethoven sonatas. In order to generate the low-resolution audio signal from the 16-bit originals, we applied dithering before degrading the signal by the desired downscaling ratio.

Following the structure of experiments conducted in [4], we evaluated several variants of the BDSR model in two

regimes. In the single speaker task, the model was trained on the first 223 recordings of VCTK Speaker 1 (30 mins of audio) and tested on the last 8 recordings. In the piano task, we evaluated the ability of BDSR to generalize across different types of audio, using a standard training/validation/test split of 88%-6%-6%.

An additional multi-speaker task intended to assess BDSR's ability to generalize across speakers by training on samples from the first 99 VCTK speakers and testing on the 8 remaining speakers was unfortunately not evaluated due to time constraints. This would however make a great starting point for future exploration.

4.1.2 Metrics

In order to provide an unbiased quantitative comparison of BDSR's ability to recover audio detail through bit-depth super-resolution, the peak signal-to-noise ratio (PSNR) metric was chosen and computed between generated samples y and source audio x.

$$PSNR(x, y) = 10log\frac{MAX_v^2}{||x - y||_2^2}$$

While PSNR is a widely reported metric in the assessment of super-resolution performance within both image and audio domains [11, 4], it has been shown to not be entirely accurate in the assessment of perceptual quality [5]. As a result, we also manually evaluated the perceptual quality of generated samples against baseline to develop a better understanding of mappings learned by various models.

4.2. Results

We measured the PSNR after bit-depth super resolution on both the Mehri Piano and VCTK speech datasets, as detailed in Table 1. A higher PSNR value is correlated with lower mean squared error and therefore should correspond to an output that is closer to the source audio. In our evaluation, we compare the PSNR from the output of our BDSR variants to the baseline of naively upscaled 8-bit audio.

As seen in Table 1, predicting a categorical distribution using the delta BDSR model did not lead to baseline-surpassing results. We postulate that the unimpressive performance of this model can be attributed to discontinuity in the output, resulting in deviation from the true value. Although the results indicate that the model is predicting deltas relatively accurately, human perceptual tests have shown that the model often fails to reduce noise, only

we observed that the discontinuity led to more structured noise but not necessarily reduction. That is, in contrast to the static background noise of the 8-bit input audio, the categorically super-resolved 16-bit audio exhibited dynamic noise that correlated with the loudness of the track.

Dataset	Naive Upscaling	BDSR (12280)	BDSR (3080)	BDSR (Real)
Mehri	58.94	57.28	58.40	60.25
VCTK	58.91	56.53	57.02	60.12

Table 1. Comparison of bit-depth super-resolution performance (measured in PSNR) on Mehri piano and VCTK datasets. Methods evaluated include the naive upscaling baseline, delta BDSR with 12280 receptive field, delta BDSR with 3080 receptive field, and BDSR with real-value output.

Increasing the receptive field from 3080 to 12280 did not lead to any meaningful improvement in super-resolution quality. In fact, equivalent models with larger receptive fields actually performed worse in PSNR metrics. We speculate that due to the local and independent nature of artifacts introduced during the downsampling process, they may not share many relationships with long-term audio structure. Models with extremely large receptive fields could potentially be less effective, as they are exposed to a greater amount of noise and require more time to train. This discrepancy is especially pronounced for the Mehri dataset, as a receptive field of 12280 covers almost a second of audio, an eternity in the highly dynamic domain of piano music. In contrast, speech from the VCTK

In our experiment with real-valued output, we observe better PSNR than baseline, and with a double blind listening test between the 8-bit audio and super-resolved 16-bit audio, there is a subtle but consistent reduction in shrillness of background noise. Due to time and credit constraints, we terminated training early. However observing that the validation loss as well as PSNR score were still decreasing approximately linearly at the time of training termination, we are confident that further training (approximately one week of training in total) will further improve the results to a less subtle level to the human ear.

4.3. Early Experiments

In early experiments from our presentation, we encountered deceptively high PSNR scores due to not having filtered out quiet sections of audio while producing training and validation clips from audio, typically from beginning of tracks. These sections are evidently easier to predict correctly and led to overfitting that did not generalize well when we attempted to diversify the validation clips. We have since corrected this by training on middle sections of tracks which we identified to frequently contain high activity.

With adequately diversified and dense audio clips in the dataset, the original model from proposal which predicted full audio waveform as a categorical distribution over amplitude steps did not perform well within reasonable training time. This is in part due to the noisiness of the categorical output distribution, and in part due to the model spending significant amount of time learning to simply replicate

the quality of the input let alone improving upon it.

These observations motivated our experimentation into real-valued and delta prediction, which addressed both problems respectively. The inherent continuity of the real number space allowed the model to produce better results in significantly less time than with categorical outputs. Despite being generally less assumptious of output distribution shapes, the disjoint categories proved counterproductive to and unintuitive in modelling the conceptually continuous amplitude space (though discretized in computer representation). Delta prediction reduced the output parameters and removed the redundancy that led to significant portion of training time spent before even matching the quality of identity mapping.

4.4. Zero Shot Audio Super-Resolution

Inspired by the work of Shocher et al. [11] in zero-shot super-resolution for images, we also investigated the prevalence of internal recurrence across bit-depth scale within a single audio sample. To do so, we attempted to upscale the bit-depth of an audio sample in the absence of any additional training data or pre-trained models.

Using just a single LR input sample, an augmented corpus of training data can be generated through downscaling. In our experiment, we achieved this by downscaling the original low resolution signal from 8-bit to both 6-bit and 4-bit signals. We then trained the model to reconstruct the 6 bit signal from the 4 bit signal and the 8 bit signal from the 6 bit signal. The hypothesis behind this approach is that by learning how to upscale bit-depth at lower scales, the trained network will be able to generalize and effectively upscale the input LR sample to the desired output fidelity.

Unfortunately, the zero shot training approach proved to be unsuccessful in practice. Even when training for upwards of 100 epochs on the generated training set, the loss failed to converge. In addition, the samples generated were often degenerate and contained a large amount of audible noise. This suggests a lack of internal recurrence across bit-depth within individual audio samples.

5. Conclusion

In this paper, we presented a successful modification to the WaveNet architecture that is capable of increasing peak signal-to-noise ratio of audio by increasing bit depth. The combination of dilated convolutions and the modified loss function allow the model to remain computationally feasible even with a large receptive field. Our model is able to produce compelling qualitative and quantitative results on both spoken word and music datasets. This model could be used for a variety of applications where memory can be afforded in exchange for an increase in dynamic range and a decrease in signal noise. Promising examples include reconstructing lossless music files and increasing recording quality for speech recognition.

References

[1] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen, et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International Conference on Machine Learning*, pages 173–182, 2016.

[2] B. D'Alessandro and Y. Q. Shi. Mp3 bit rate quality detection through frequency spectrum analysis. In *Proceedings of the 11th ACM Workshop on Multimedia and Security*, pages 57–62. ACM, 2009.

[3] J. Engel, C. Resnick, A. Roberts, S. Dieleman, D. Eck, K. Simonyan, and M. Norouzi. Neural audio synthesis of musical notes with wavenet autoencoders. *arXiv preprint arXiv:1704.01279*, 2017.

[4] V. Kuleshov, S. Z. Enam, and S. Ermon. Audio super resolution using neural networks. *arXiv preprint arXiv:1708.00853*, 2017.

[5] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. *arXiv preprint*, 2016.

[6] S. Mehri, K. Kumar, I. Gulrajani, R. Kumar, S. Jain, J. Sotelo, A. Courville, and Y. Bengio. Samplernn: An unconditional end-to-end neural audio generation model. *arXiv preprint arXiv:1612.07837*, 2016.

[7] A. v. d. Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. v. d. Driessche, E. Lockhart, L. C. Cobo, F. Stimberg, et al. Parallel wavenet: Fast high-fidelity speech synthesis. *arXiv preprint arXiv:1711.10433*, 2017.

[8] T. L. Paine, P. Khorrami, S. Chang, Y. Zhang, P. Ramachandran, M. A. Hasegawa-Johnson, and T. S. Huang. Fast wavenet generation algorithm. *arXiv preprint arXiv:1611.09482*, 2016.

[9] D. Rethage, J. Pons, and X. Serra. A wavenet for speech denoising. *arXiv preprint arXiv:1706.07162*, 2017.

[10] T. Salimans, A. Karpathy, X. Chen, and D. P. Kingma. Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. *arXiv preprint arXiv:1701.05517*, 2017.

[11] A. Shocher, N. Cohen, and M. Irani. " zero-shot" super-resolution using deep internal learning. *arXiv preprint arXiv:1712.06087*, 2017.

[12] A. Van Den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[13] A. van den Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals, A. Graves, et al. Conditional image generation with pixelcnn decoders. In *Advances in Neural Information Processing Systems*, pages 4790–4798, 2016.

[14] C. Veaux, J. Yamagishi, K. MacDonald, et al. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit. 2017.

[15] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller. Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr. In *International Conference on Latent Variable Analysis and Signal Separation*, pages 91–99. Springer, 2015.

[16] W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig. The microsoft 2016 conversational speech recognition system. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 5255–5259. IEEE, 2017.

[17] L.-C. Yang, S.-Y. Chou, and Y.-H. Yang. Midinet: A convolutional generative adversarial network for symbolic-domain music generation. In *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR2017), Suzhou, China*, 2017.