AUDIO SUPER-RESOLUTION USING NEURAL NETS

Volodymyr Kuleshov, S. Zayd Enam, and Stefano Ermon
Department of Computer Science,
Stanford University
{kuleshov,ermon}@cs.stanford.edu
zayd@stanford.edu

ABSTRACT

We propose a neural network-based technique for enhancing the quality of audio signals such as speech or music by transforming inputs encoded at low sampling rates into higher-quality signals with an increased resolution in the time domain. This amounts to generating the missing samples within the low-resolution signal in a process akin to image super-resolution. On standard speech and music datasets, this approach outperforms baselines at $2\times$, $4\times$, and $6\times$ upscaling ratios. The method has practical applications in telephony, compression, and text-to-speech generation; it can also be used to improve the scalability of recently-proposed generative models of audio.

1 INTRODUCTION

Modeling audio is an important problem at the intersection of signal processing and representation learning. Recently, machine learning techniques have enabled advances in audio generation (van den Oord et al., 2016; Mehri et al., 2016), speech recognition (Zhang et al., 2017), and classification (Aytar et al., 2016).

Most of these recent works model *raw audio* signals over time; although this affords us the maximum modeling flexibility, it is also computationally expensive, requiring us to handle $> 10,000$ audio samples at every second. Our work takes a step towards alleviating this difficulty by proposing a technique for reconstructing high-quality audio from input containing only a small fraction (15-50%) of the original signal's information. Our technique has applications in telephony, compression, and text-to-speech generation and suggests new architectures for generative models of audio.

2 SETUP AND BACKGROUND

Audio signal processing. We represent an audio signal as a function $s(t) : [0, T] \to \mathbb{R}$, where T is the duration of the signal (in seconds) and $s(t)$ is the amplitude at t. Taking a digital measurement of s requires us to discretize the continuous function $s(t)$ into a vector $x(t) : \{\frac{1}{R}, \frac{2}{R}, ..., \frac{RT}{R}\} \to \mathbb{R}$. We refer to R as the *sampling rate* of x (in Hz). Sampling rates may range from 4 KHz (low-quality telephone speech) to 44 Khz (high-fidelity music).

In this work, we interpret R as the resolution of x; our goal is to increase the resolution of audio samples by predicting x from a fraction of its samples taken at $\{\frac{1}{R}, \frac{2}{R}, ..., \frac{RT}{R}\}$. Note that by basic signal processing theory, this is equivalent to predicting the higher frequencies of x.

Bandwidth extension. Audio upsampling has been studied in the audio processing community under the name *bandwidth extension* (Ekstrand, 2002; Larsen & Aarts, 2005). Several learning-based approaches have been proposed, including Gaussian mixture models (Cheng et al., 1994; Park & Kim, 2000) and neural networks (Li et al., 2015). These methods typically involve hand-crafted features and use relatively simple models (e.g., neural networks with at most 2-3 densely connected layers) that are often part of a larger, more complex systems. In comparison, our method is conceptually simple (operating directly on the raw audio signal), scalable (our neural networks are fully convolutional and fully feed-forward), more accurate, and is also among the few to have been tested on non-speech audio.

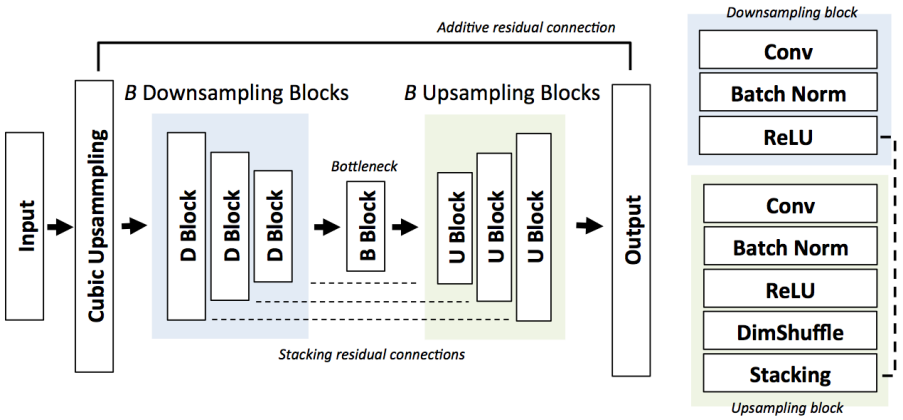

Figure 1: Deep residual network used for audio super-resolution. We extract features via B residual blocks; upscaling is done via stacked SubPixel layers.

3 METHOD

Given a low resolution signal $x = \{x_{1/R_1}, ...x_{R_1T_1/R_1}\}$ sampled at a rate R_1, our goal is to reconstruct a high-resolution version $y = \{y_{1/R_2}, ...y_{R_2T_2/R_2}\}$ of x that has a sampling rate $R_2 > R_1$. For example, x may be a voice signal transmitted via a standard telephone connection at 4 KHz; y may be a high-resolution 16 KHz reconstruction of the orignal. We use $r = R_2/R_1$ to denote the *upsampling ratio* of the two signals, which in our work equals $r = 2, 4, 6$. We thus expect that $y_{rt/R_2} \approx x_{t/R_1}$ for $t = 1, 2, ..., T_1R_1$. We compute $y = f_\theta(x)$ via a function f_θ parametrized by a neural network with parameters θ. The neural network is fully convolutional and can be run on inputs of an arbitrary length. We determine θ by training the neural network on a large dataset of examples x_i, y_i.

Model architecture. We give an overview of our architecture in Figure 1. Similar to Dong et al. (2016), we use cubic upsampling to project the input into a high-dimensional space. We pass the result through a series of B feed-forward downsampling blocks. Each block performs a convolution, batch normalization, and applies a ReLU non-linearity. We use a stride of two to reduce the dimensionality of the input, and we increase the number of filters by two at each stage. The image is reconstructed from the learned features via a symmetric series of B upsampling blocks. We add skip connections which stack the tensor of i-th downsampling features with the $(B - i)$-th tensor of upsampling features; this allows us to reuse low-resolution features during upsampling (Isola et al., 2016). We also add an additive residual connection between the cubic upsampling layer and the final output; thus, the model only needs to improve the cubic approximation. Upscaling is performed using a one-dimensional version of the Subpixel dimension shuffling layer of Shi et al. (2016).

We train the model on pairs of high and low-resolution audio patches of length 6000 sampled from a collection of larger signals. Finally, we train the above neural network to minimize the ℓ_2 distance between the high-res patches and their reconstruction.

4 EXPERIMENTS

Setup. We evaluate our method on VCTK — a popular speech dataset which contains 44 hours of data from 109 different speakers — the piano dataset of (Mehri et al., 2016) — containing 32 publicly available Beethoven sonatas (about 10 hours of audio in total) — and MagnaTagATune, which consists of about 200 hours of music from 188 different genres. We split each dataset into a training and a testing set. For VCTK, we used the last 9 speakers for testing; for MagnaTagATune, we used 24,863 random files for training and the remaining 1000 files for testing; for the piano dataset, we use the provided 88%-6%-6% split.

We normalize all files to 16,000 Hz and generate high-resolution patches of length 6000. We instantiate our model with $B = 8$ residual blocks, and train for 400 epochs using the ADAM optimizer with a learning rate of 10^{-4} (with linear decay after 200 epochs).

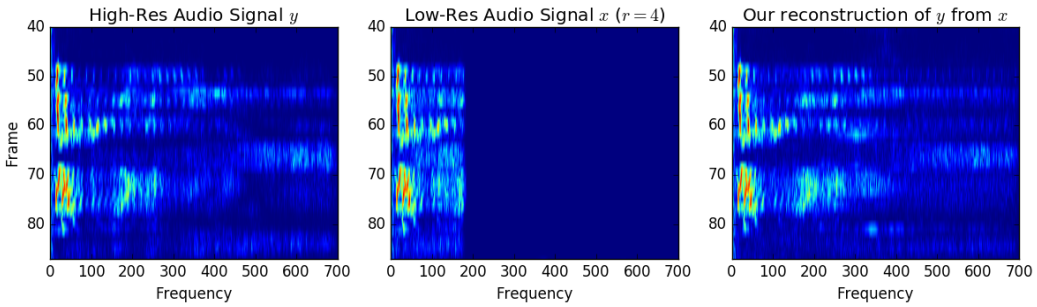

Figure 2: Audio super-resolution explained using spectrograms. A high-quality speech signal (top) is subsampled at $r = 4$, resulting in the loss of high frequencies (middle). We recover the missing signal using a trained neural network (bottom).

Table 1: SNR and LSD on speech and music datasets (in dB)

TASK	RATIO	CUBIC		DNN		AUDIO-SR	
		SNR	LSD	SNR	LSD	SNR	LSD
VCTK Speaker 1	$r = 2$	20.3	4.5	20.1	3.7	21.1	3.2
	$r = 4$	14.8	8.2	15.9	4.9	17.1	3.6
	$r = 6$	10.4	10.3	n/a	n/a	12.3	3.8
VCTK	$r = 2$	19.7	4.4	19.9	3.6	20.7	3.1
	$r = 4$	13.0	8.0	14.9	5.8	16.1	3.5
	$r = 6$	9.1	10.1	n/a	n/a	10.0	3.7
Piano Sonatas	$r = 2$	29.4	3.5	29.3	3.4	30.1	3.4
	$r = 4$	22.2	5.8	23.0	5.2	23.5	3.6
	$r = 6$	15.4	7.3	n/a	n/a	16.1	4.4
MagnaTagATune	$r = 4$	16.1	10.3	n/a	n/a	12.5	3.8

Baselines. We compare our method against cubic B-splines (a standard interpolation technique) and the deep neural network (DNN) based technique of Li et al. (2015). In brief, Li et al. (2015) transform the input into spectral features which are then used to predict high frequencies; 84% of users in a study preferred this method to a standard GMM baseline. The DNN has 3 dense hidden layers of 2048 units with ReLU nonlinearities. Note that, without modification, the features of Li et al. (2015) only apply to scaling ratios $r = 2, 4, 8, ...$

Metrics. Signal-to-noise ratio (SNR) is defined as $\text{SNR}(x, y) = 10 \log \frac{||y||_2^2}{||x-y||_2^2}$ for a signal y and its approximation x. The log-spectral distance (LSD; Gray & Markel (1976)) measures the reconstruction quality of individual frequencies as follows: $\text{LSD}(x, y) = \frac{1}{L} \sum_{\ell=1}^{L} \sqrt{\frac{1}{K} \sum_{k=1}^{K} \left(X(\ell, k) - \hat{X}(\ell, k) \right)^2}$, where X and $\hat{X}$ are the log-spectral power magnitudes of y and x, respectively. These are defined as $X = \log |S|^2$, where S is the short-time Fourier transform (STFT) of the signal. We use ℓ and k index frames and frequencies, respectively.

Performance. Our results are summarized in Table 1. On the speech datasets, our method outperforms baselines at all ratios (see Figure 2 for an example). We found it difficult to tell apart the original and the upscaled signals at $2\times$ and sometimes at $4\times$. Relative to the cubic and DNNs baseline, we found the higher frequencies to be more audible (as evidenced by much lower LSD), although our method sometimes introduced slight background noise (resulting in a somewhat smaller improvement in SNR).

We also achieved good reconstruction accuracy on the piano dataset, demonstrating that our method generalizes to non-vocal audio. Our performance on MagnaTagATune was lower (and we were unable to run the DNN baseline); we found that our model was significantly underfitting this large and highly diverse dataset and was limited by our current computational resources. It produced audible improvements in the high-frequency range, but also introduced artifacts that decreased the SNR. We expect improved results with more computational power and a larger model.

REFERENCES

Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 29*, pp. 892–900. Curran Associates, Inc., 2016. URL http://papers.nips.cc/paper/6146-soundnet-learning-sound-representations-from-unlabeled-video.pdf.

J Ballé, V Laparra, and E P Simoncelli. End-to-end optimized image compression. In *Int'l. Conf. on Learning Representations (ICLR2017)*, Toulon, France, April 2017. URL https://arxiv.org/abs/1611.01704. Available at http://arxiv.org/abs/1611.01704.

Yan Ming Cheng, Douglas O'Shaughnessy, and Paul Mermelstein. Statistical recovery of wideband speech from narrowband speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):544–548, 1994.

Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(2):295–307, February 2016. ISSN 0162-8828. doi: 10.1109/TPAMI.2015.2439281. URL http://dx.doi.org/10.1109/TPAMI.2015.2439281.

Per Ekstrand. Bandwidth extension of audio signals by spectral band replication. In *in Proceedings of the 1st IEEE Benelux Workshop on Model Based Processing and Coding of Audio (MPCA02*. Citeseer, 2002.

Augustine Gray and John Markel. Distance measures for speech processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(5):380–391, 1976.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *arxiv*, 2016.

Erik Larsen and Ronald M Aarts. *Audio bandwidth extension: application of psychoacoustics, signal processing and loudspeaker design*. John Wiley & Sons, 2005.

Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew P. Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. *CoRR*, abs/1609.04802, 2016. URL http://arxiv.org/abs/1609.04802.

Kehuang Li, Zhen Huang, Yong Xu, and Chin-Hui Lee. Dnn-based speech bandwidth expansion and its application to adding high-frequency missing features for automatic speech recognition of narrowband speech. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

Soroush Mehri, Kundan Kumar, Ishaan Gulrajani, Rithesh Kumar, Shubham Jain, Jose Sotelo, Aaron Courville, and Yoshua Bengio. Samplernn: An unconditional end-to-end neural audio generation model, 2016. URL http://arxiv.org/abs/1612.07837. cite arxiv:1612.07837.

Kun-Youl Park and Hyung Soon Kim. Narrowband to wideband conversion of speech using gmm based transformation. In *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, volume 3, pp. 1843–1846. IEEE, 2000.

Hannu Pulakka, Ulpu Remes, Kalle Palomäki, Mikko Kurimo, and Paavo Alku. Speech bandwidth extension using gaussian mixture model-based estimation of the highband mel spectrum. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pp. 5100–5103. IEEE, 2011.

Wenzhe Shi, Jose Caballero, Ferenc Huszar, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. pp. 1874–1883, 2016. doi: 10.1109/CVPR.2016.207. URL http://dx.doi.org/10.1109/CVPR.2016.207.

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. *CoRR*, abs/1609.03499, 2016. URL http://arxiv.org/abs/1609.03499.

Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. *ECCV*, 2016.

Ying Zhang, Mohammad Pezeshki, Philemon Brakel, Saizheng Zhang, César Laurent, Yoshua Bengio, and Aaron C. Courville. Towards end-to-end speech recognition with deep convolutional neural networks. *CoRR*, abs/1701.02720, 2017. URL http://arxiv.org/abs/1701.02720.