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 introduce a new audio processing technique that increases the sampling rate of signals such as speech or music using deep convolutional neural networks.

使用深度卷积神经网络提高了语音或音乐等信号的采样率。

Our model is trained on pairs of low and high-quality audio examples; at test-time, it predicts missing samples within a low-resolution signal in an interpolation process similar to image super-resolution.

针对成对的低质量和高质量音频例子进行训练； 在使用时，以类似于图像超分辨率的插值方法来预测低分辨率信号中丢失的样本。

Our method is simple and does not involve specialized audio processing techniques;

方法很简单，不涉及专门的音频处理技术。

in our experiments, it outperforms baselines on standard speech and music benchmarks at upscaling ratios of 2×, 4×, and 6×.

在2倍，4倍和6倍的放大比例上, 超过了参考基准。

The method has practical applications in telephony, compression, and text-tospeech generation; it demonstrates the effectiveness of convolutional architectures on an audio generation task.

该方法在电话，压缩和文本语音转换中具有实际应用。 它展示了卷积架构在音频生成任务上的有效性。

# INTRODUCTION

The generative modeling of audio signals is a fundamental problem at the intersection of signal processing and machine learning; recent learning-based algorithms have enabled advances in speech recognition (Hinton et al., 2012), audio synthesis (van den Oord et al., 2016; Mehri et al., 2016), music recommendation systems (Coviello et al., 2012; Wang & Wang, 2014; Liang et al., 2015), and in many other areas (Acevedo et al., 2009). Audio processing also raises basic research questions pertaining to time series and generative modeling (Haykin & Chen, 2005; Bilmes, 2004).

One of the most significant recent advances in machine learning-based audio processing has been the ability to directly model *raw signals* in the time domain using neural networks (van den Oord et al., 2016; Mehri et al., 2016). Although this affords us the maximum modeling flexibility, it is also computationally expensive, requiring us to handle *>* 10*,*000 audio samples at every second.

音频信号的生成建模是信号处理和机器学习相交的一个基本问题。最近基于学习的算法在语音识别（Hinton等，2012），音频合成（van den Oord等，2016; Mehri等，2016），音乐推荐系统（Coviello等，2012）方面取得了进展； Wang＆Wang，2014； Liang等，2015），以及其他许多领域（Acevedo等，2009）。音频处理也提出了有关时间序列和生成模型的基础研究问题（Haykin＆Chen，2005； Bilmes，2004）。

In this paper, we explore new lightweight modeling algorithms for audio. In particular, we focus on a specific audio generation problem called *bandwidth extension*, in which the task is to reconstruct high-quality audio from a low-quality, down-sampled input containing only a small fraction (1550%) of the original samples. We introduce a new neural network-based technique for this problem that is inspired image super-resolution algorithms (Dong et al., 2016), which use machine learning techniques to interpolate a low-resolution image into a higher-resolution one. Learning-based methods often perform better in this context than general-purpose interpolation schemes such as splines because they leverage sophisticated domain-specific models of the appearance of natural signals.

在基于机器学习的音频处理中，最近最重要的进展之一是能够使用神经网络在时域中直接对原始信号进行建模（van den Oord等人，2016; Mehri等人，2016）。尽管这为我们提供了最大的建模灵活性，但它在计算上也很昂贵，要求我们每秒处理超过10,000个音频样本。

在本文中，我们探索了音频的新型轻量级建模算法。特别是，我们专注于特定的音频生成问题，即带宽扩展，其中的任务是从仅包含原始样本的一小部分（1550％）的低质量，下采样的输入中重建高质量的音频。我们针对此问题引入了一种基于神经网络的新技术，即启发式图像超分辨率算法（Dong等人，2016），该算法使用机器学习技术将低分辨率图像插值为高分辨率图像。在这种情况下，基于学习的方法通常比样条之类的通用插值方案表现更好，因为它们利用了自然信号外观的复杂领域特定模型。

As in image super-resolution, our model is trained on pairs of low and high-quality samples; at testtime, it predicts the missing samples of a low-resolution input signal. Unlike recent neural networks for generating raw audio, our model is fully feedforward and can be run in real-time. In addition to having multiple practical applications, our method also suggests new ways to improve existing generative models of audio.

## CONTRIBUTIONS

From a practical perspective, our technique has applications in telephony, compression, text-tospeech generation, forensic analysis, and in other domains. It outperforms baselines at 2×, 4×, and 6× upscaling ratios, while also being significantly simpler than previous methods. Whereas most existing audio enhancement methods make substantial use of signal processing theory, our

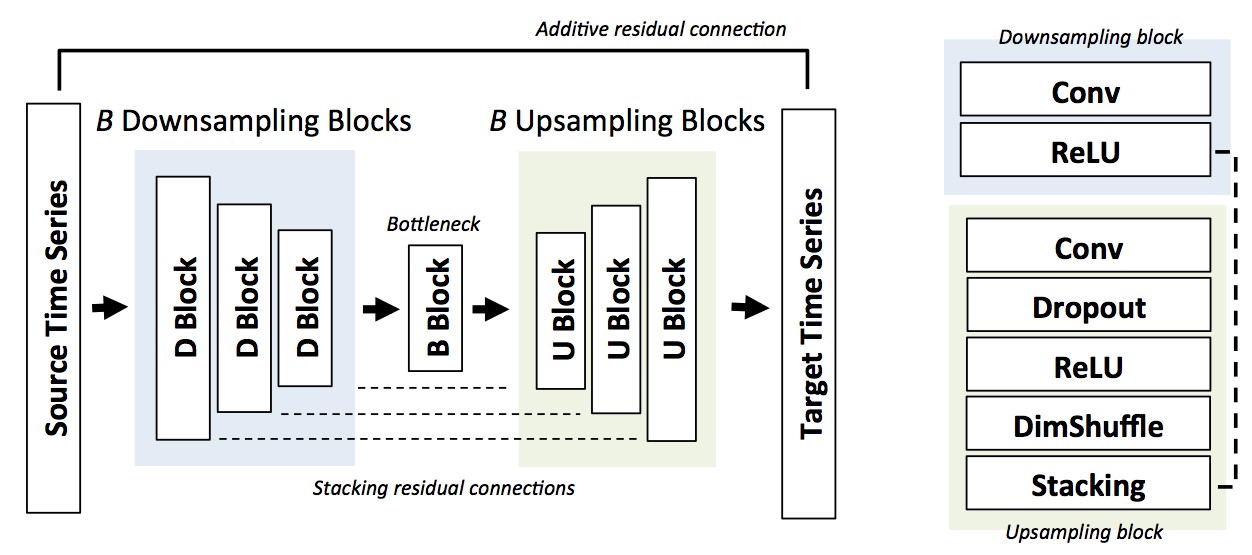


Figure 1: Deep residual network used for audio super-resolution. We extract features via *B* residual blocks; upscaling is done via stacked SubPixel layers.

approach is conceptually very simple and requires no specialized knowledge to implement. Our neural networks are simply trained to map one audio time series into another. Our approach is also among the first to use convolutional architectures for bandwidth extension; as a result, it scales better with dataset size and computational resources relative to current alternatives.

从实际的角度来看，我们的技术可以应用于电话，压缩，文本语音转换，法医分析以及其他领域。它在2倍，4倍和6倍的放大比例下优于基准，同时也比以前的方法简单得多。尽管大多数现有的音频增强方法都充分利用了信号处理理论，但是我们的方法在概念上非常简单，不需要专门知识即可实施。我们的神经网络经过简单培训，可以将一个音频时间序列映射到另一个音频时间序列。我们的方法也是最早使用卷积体系结构进行带宽扩展的方法之一。结果，相对于当前的替代方案，它在数据集大小和计算资源上的扩展性更好。

From a generative modeling perspective, our work demonstrates that purely feedforward architectures operating in a non-discretized output space can achieve good performance on an important audio generation task. This hints at the possibility of designing improved generative models for audio that combine both feedforward and recurrent components.

从生成建模的角度来看，我们的工作表明，在非离散输出空间中运行的纯前馈体系结构可以在重要的音频生成任务上实现良好的性能。这暗示了设计结合前馈和递归分量的改进的音频生成模型的可能性。

# SETUP AND BACKGROUND

Audio processing. We represent an audio signal as a function *s*(*t*) : [0*,T*] → 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. 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).

音频处理。 我们将音频信号表示为s（t）的函数：[0，T]→R，其中T是信号的持续时间（以秒为单位），而s（t）是在t处的振幅。 对s进行数字测量需要我们将连续函数s（t）离散为一个向量。 我们将R称为x的采样率（以Hz为单位）。 采样率的范围可以从4 KHz（低质量电话语音）到44 Khz（高保真音乐）。

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. Note that by basic signal processing theory, this is equivalent to predicting the higher frequencies of *x*.

在这项工作中，我们将R解释为x的分辨率。 我们的目标是通过根据在采集的样本中的一小部分来预测x，从而提高音频样本的分辨率。 注意，根据基本信号处理理论，这等效于预测x的较高频率。

Bandwidth extension. Audio upsampling has been studied in the audio processing community under the name *bandwidth extension* (Ekstrand, 2002; Larsen & Aarts, 2005). Several learningbased 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 handcrafted 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.

带宽扩展。 在音频处理社区中，以带宽扩展的名称研究了音频上采样（Ekstrand，2002； Larsen＆Aarts，2005）。 已经提出了几种基于学习的方法，包括高斯混合模型（Cheng等，1994； Park＆Kim，2000）和神经网络（Li等，2015）。 这些方法通常涉及手工特征并使用相对简单的模型（例如，具有最多2-3个紧密连接的层的神经网络），通常是更大，更复杂的系统的一部分。 相比之下，我们的方法从概念上讲是简单的（直接对原始音频信号进行操作），可扩展的（我们的神经网络是完全卷积和完全前馈的），更准确，并且是少数未经语音测试的方法 音频。

# METHOD

## SETUP

Given a low resolution signal *x* = {*x*1*/R*1*,...xR*1*T*1*/R*1} sampled at a rate *R*1, our goal is to reconstruct a high-resolution version *y* = {*y*1*/R*2*,...yR*2*T*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

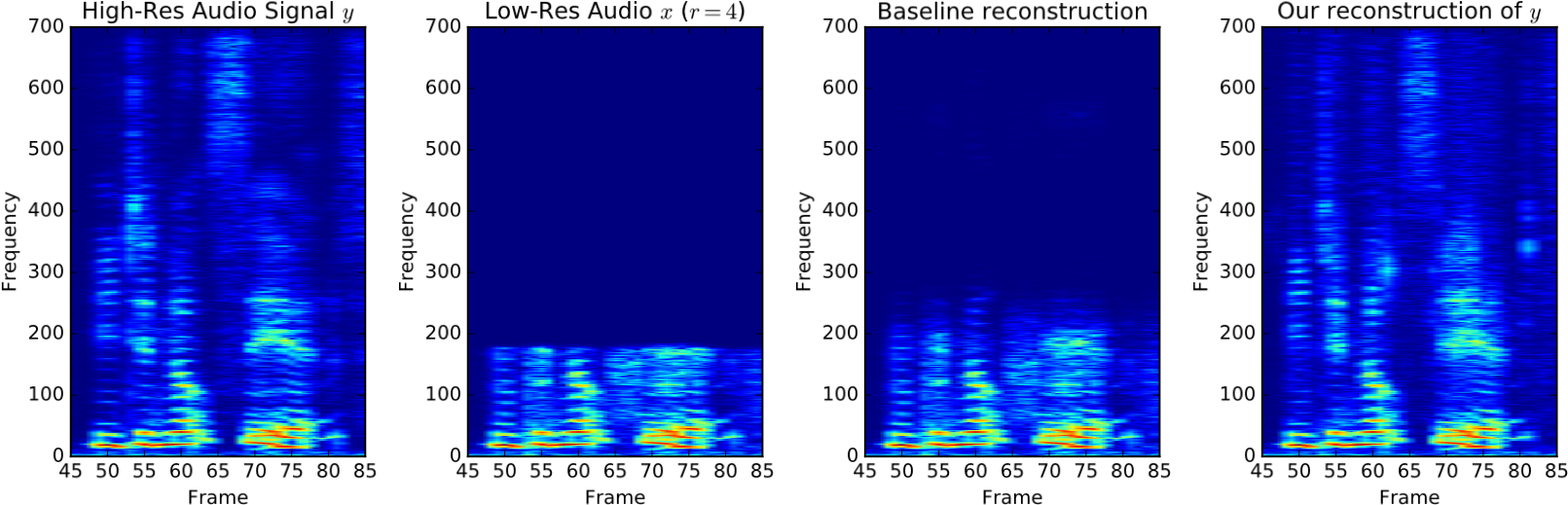


Figure 2: Audio super-resolution visualized using spectrograms. A high-quality speech signal (leftmost) is subsampled at *r* = 4, resulting in the loss of high frequencies (2nd from left). We recover the missing signal using a trained neural network (rightmost), greatly outperforming the cubic baseline (second from right).

the *upsampling ratio* of the two signals, which in our work equals *r* = 2*,*4*,*6. We thus expect that *yrt/R*2 ≈ *xt/R*1 for *t* = 1*,*2*,...,T*1*R*1.

给定以速率R1采样的低分辨率信号x = {x1 / R1，... xR1T1 / R1}，我们的目标是重建y的高分辨率版本y = {y1 / R2，... yR2T2 / R2} x的采样率R2> R1。 例如，x可以是通过标准电话连接以4 KHz传输的语音信号； y可能是原始的高分辨率16 KHz重建。 我们使用r = R2 / R1表示两个信号的上采样率，在我们的工作中等于r = 2,4,6。 因此，我们期望对于t = 1,2，...，T1R1，yrt / R2≈xt / R1。

To recover the under-defined signal, we learn a model *p*(*y*|*x*) of the higher-resolution *y*, conditioned on its low-resolution instantiation *x*. We assume that the relationship between the time series *x,y* follows the equation  where is Gaussian noise and *fθ* is a model parametrized by *θ*. Our framework also extends to more complex noise models which the user may provide as a prior or that may be themselves parametrized by the model (similarly to how one parametrizes the normal distribution in a variational autoencoder).

为了恢复未定义的信号，我们学习高分辨率y的模型p（y | x），条件是其低分辨率实例化x。 我们假设时间序列x，y之间的关系遵循方程，其中高斯噪声和fθ是由θ参数化的模型。 我们的框架还扩展到更复杂的噪声模型，用户可以事先提供这些模型，也可以由模型自己对其进行参数化（类似于一个参数化变量自动编码器中的正态分布的方式）。

The above formulation naturally leads to a mean squared error (MSE) objective

### (1)

for determining the parameters *θ* based on a dataset of source/target time series pairs. Since our model is fully convolutional, we may take the *xi,yi* to be small patches sampled from the full time series.

上述公式自然会导致均方误差（MSE）目标

（1）

用于基于源/目标时间序列对的数据集确定参数θ。 由于我们的模型是完全卷积的，因此我们可以将xi，yi看作是从整个时间序列中采样的小块。

## MODEL ARCHITECTURE

We parametrize the function *f* with a deep convolutional neural network with residual connections; our neural network architecture is based on ideas from Shi et al. (2016), Dong et al. (2016), and Isola et al. (2016), and is shown in Figure 1. We highlight its main features below.

我们用带有残差连接的深度卷积神经网络对函数f进行参数化； 我们的神经网络架构基于Shi等人的想法。 （2016），Dong等。 （2016），以及Isola等。 （2016），如图1所示。我们在下面重点介绍其主要功能。

Bottleneck architecture. Our model contains *B* successive downsampling and upsampling *blocks*: each performs a convolution, batch normalization, and applies a ReLU non-linearity. Downsampling block *b* = 1*,*2*,...,B* contains max(26+*b,*512) convolutional filters of length min(27−*b* +1*,*9) and a stride of 2. Upsampling block *b* has max(27+(*B*−*b*+1)*,*512) filters of length min(27−(*B*−*b*+1) + 1*,*9).

瓶颈架构。 我们的模型包含B个连续的下采样和上采样块：每个块执行卷积，批量归一化并应用ReLU非线性。 下采样块b = 1,2，...，B包含长度为min（27−b +1,9），跨度为2的max（26 + b，512）个卷积滤波器。上采样块b具有max（27+ （b-b + 1），512）个长度为min（27-（B-b + 1）+ 1,9）的滤波器。

Thus, at a downsampling step, we halve the spatial dimension and double the filter size; during upsampling, this is reversed. This bottleneck architecture is inspired by auto-encoders, and is known to encourage the model to learn a hierarchy of features. For example, on an audio task, bottom layers may extract wavelet-style features, while higher ones may correspond to phonemes Aytar et al. (2016). Note that the model is fully convolutional, and may run on input sequences of arbitrary length.

因此，在下采样步骤中，我们将空间尺寸减半并将滤波器尺寸加倍； 在上采样期间，这是相反的。 这种瓶颈架构受自动编码器的启发，并且众所周知可以鼓励模型学习功能的层次结构。 例如，在音频任务上，底层可以提取小波样式的特征，而较高的可能对应于音素Aytar等。 （2016）。 请注意，该模型是完全卷积的，并且可以在任意长度的输入序列上运行。

Skip connections. When the source series *x* is similar to the target *y*, downsampling features will be also be useful for upsampling (Isola et al., 2016). We thus add additional skip connections which stack the tensor of *b*-th downsampling features with the (*B*−*b*+1)-th tensor of upsampling features. We also add an additive residual connection from the input to the final output: the model thus only needs to learn *y* − *x*, which in practice speeds up training.

跳过连接。 当源序列x与目标y相似时，下采样特征也将对上采样有用（Isola等人，2016）。 因此，我们添加了附加的跳过连接，这些连接将第b个下采样特征的张量与第（Bb + 1）个上采样特征的张量堆叠在一起。 我们还添加了从输入到最终输出的附加残差连接：因此，该模型仅需要学习y-x，从而在实践中加快了训练速度。

Subpixel shuffling layer. In order to increase the time dimension during upscaling, we have implemented a one-dimensional version of the Subpixel layer of Shi et al. (2016), which has been shown to be less prone to produce artifacts (Odena et al., 2016).

亚像素改组层。 为了在放大过程中增加时间维度，我们实现了Shi等人的Subpixel层的一维版本。 （2016年），已被证明不太容易产生伪像（Odena等人，2016年）。

An upscaling block’s convolution maps an input tensor of dimension *F* ×*d* into one of size *F/*2×*d*. The subpixel layer reshuffles this *F/*2×*d* tensor into another one of size *F/*4×2*d* (while preserving the tensor entries intact); these are concatenated with *F/*4 features from the downsampling stage, for a final output of size *F/*2 × 2*d*. Thus, we have halved the number of filters and doubled the spatial dimension.

放大块的卷积将尺寸为F×d的输入张量映射为尺寸为F / 2×d的张量。 子像素层将此F / 2×d张量改组为另一个大小为F / 4×2d的张量（同时保持张量条目完整）； 这些与降采样阶段的F / 4功能串联在一起，最终输出为F / 2×2d。 因此，我们将过滤器的数量减少了一半，空间尺寸增加了一倍。

# EXPERIMENTS

Datasets. We use the VCTK dataset (Yamagishi) — which contains 44 hours of data from 108 different speakers — and the Piano dataset of Mehri et al. (2016) (10 hours of Beethoven sonatas). We generate low-resolution audio signal from the 16 KHz originals by applying an order 8 Chebyshev type I low-pass filter before subsampling the signal by the desired scaling ratio.

We evaluate our method in three regimes. The SINGLESPEAKER task trains the model on the first

223 recordings of VCTK Speaker 1 (about 30 mins) and tests on the last 8 recordings. The MULTISPEAKER task assesses our ability to generalize to new speakers. We train on the first 99 VCTK speakers and test on the 8 remaining ones; our recordings feature different voices and accents (Scottish, Indian, etc.) Lastly, the PIANO task extends audio-super resolution to non-vocal data; we use the standard 88%-6%-6% data split.

数据集。 我们使用VCTK数据集（Yamagishi）（其中包含来自108个不同说话者的44小时数据）和Mehri等人的Piano数据集。 （2016）（贝多芬奏鸣曲10小时）。 在通过所需的缩放比例对信号进行二次采样之前，我们通过应用8级Chebyshev I型低通滤波器，从16 KHz原始信号生成低分辨率音频信号。

我们在三种情况下评估我们的方法。 SINGLESPEAKER任务在VCTK Speaker 1的前223个记录上训练模型（约30分钟），并在最后8个记录上进行测试。 MULTISPEAKER任务评估了我们推广新演讲者的能力。 我们训练了前99位VCTK扬声器，并测试了其余8位扬声器； 我们的录音具有不同的声音和口音（苏格兰，印度等）。最后，PIANO任务将音频超分辨率扩展到非语音数据； 我们使用标准的88％-6％-6％数据拆分。

Methods. We compare our method relative to two baselines: a cubic B-spline — which corresponds to the bicubic upsampling baseline used in image super-resolution — and the recent neural network-based technique of Li et al. (2015),

The latter approach takes as input the short-time Fourier transform (STFT) of the input and predicts directly the phase and the magnitudes of the high frequency components using a dense neural network with three hidden layers of size 2048 and ReLU nonlinearities. Li et al. (2015) have shown that this method is preferred over Gaussian Mixture Models in 84% of cases in a user study. This model requires that the scaling ratio be a power of 2, hence it is not applicable when *r* = 6.

方法。 我们将我们的方法相对于两个基线进行了比较：三次B样条曲线（对应于图像超分辨率中使用的三次三次上采样基线）以及Li等人最近基于神经网络的技术。 （2015），

后一种方法将输入的短时傅立叶变换（STFT）作为输入，并使用具有2048个大小的三个隐藏层和ReLU非线性的密集神经网络直接预测高频分量的相位和幅度。 Li等。 （2015年）表明，在用户研究中84％的案例中，该方法优于高斯混合模型。 此模型要求缩放比例为2的幂，因此在r = 6时不适用。

We instantiate our model with *B* = 4 blocks and train it for 400 epochs on patches of length 6000 (in the high-resolution space) using the ADAM optimizer with a learning rate of 10−4. To ensure source/target series are of the same length, the source input is pre-processed with cubic upscaling. We do not compare against previously-proposed matrix factorization techniques (Bansal et al., 2005; Liang et al., 2013), as they are typically trained on *<* 10 input examples (Sun & Mazumder, 2013) (due to the cost of jointly factorizing a large number of matrices), and do not scale to the size of our datasets.

我们使用B = 4个块实例化模型，并使用学习速率为10−4的ADAM优化器在长度为6000的小块上（在高分辨率空间中）训练400个纪元。 为了确保源/目标序列的长度相同，请使用三次放大对源输入进行预处理。 我们没有将其与之前提出的矩阵分解技术进行比较（Bansal等，2005； Liang等，2013），因为它们通常在少于10个输入示例上进行训练（Sun＆Mazumder，2013）（由于成本高昂）。 联合分解大量矩阵），并且无法缩放到我们数据集的大小。

Metrics Given a reference signal *y* and an approximation *x*, the Signal to Noise Ratio (SNR) is defined as

度量给定参考信号y和近似值x，信噪比（SNR）定义为

## SNR*.* (2) LSD*,* (3)

The SNR is a standard metric used in the signal processing literature. The Log-spectral distance (LSD) (Gray & Markel, 1976) measures the reconstruction quality of individual frequencies as follows:

SNR是信号处理文献中使用的标准度量。 对数谱距离（LSD）（Gray＆Markel，1976）如下测量各个频率的重构质量：

where *X* and *X*ˆ 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 *`* and *k* index frames and frequencies, respectively; in our experiments, we used frames of length 2048.

其中X和Xˆ分别是y和x的对数谱功率大小。 这些定义为X = log | S | 2，其中S是信号的短时傅立叶变换（STFT）。 我们分别使用`和k索引帧和频率； 在我们的实验中，我们使用了长度为2048的帧。

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | SingleSpeaker | | | MultiSpeaker | | |  | Piano |  |
| Ratio | Obj. | Spline | DNN | Ours | Spline | DNN | Ours | Spline | DNN | Ours |
| *r* = 2 | SNR | 20.3 | 20.1 | 21.1 | 19.7 | 19.9 | 20.7 | 29.4 | 29.3 | 30.1 |
|  | LSD | 4.5 | 3,7 | 3.2 | 4.4 | 3.6 | 3.1 | 3.5 | 3.4 | 3.4 |
| *r* = 4 | SNR | 14.8 | 15.9 | 17.1 | 13.0 | 14.9 | 16.1 | 22.2 | 23.0 | 23.5 |
|  | LSD | 8.2 | 4.9 | 3.6 | 8.0 | 5.8 | 3.5 | 5.8 | 5.2 | 3.6 |
| *r* = 6 | SNR | 10.4 | n/a | 14.4 | 9.1 | n/a | 10.0 | 15.4 | n/a | 16.1 |
|  | LSD | 10.3 | n/a | 3.4 | 10.1 | n/a | 3.7 | 7.3 | n/a | 4.4 |

Table 2: Accuracy evaluation of audio-super resolution methods (in dB) on each of the three superresolution tasks at upscaling ratios *r* = 2*,*4*,*6.

MultiSpeaker Sample

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 | Average |
| Ours | 69 | 75 | 64 | 37 | 61.3 |
| DNN | 51 | 55 | 66 | 53 | 56.3 |
| Spline | 31 | 25 | 38 | 47 | 35.3 |

Table 1: MUSHRA user study scores. We show scores for each sample, averaged individual users. Average across all samples is also displayed

Evaluation The results of our experiments are summarized in Table 2. Our objective metrics show an improvement of 1-5 dB over the baselines, with the strongest improvements at higher upscaling factors. Although, the spline baseline achieves a high SNR, its signal often lacks higher frequencies; the LSD metric is better at identifying this problem. Our technique also improves over the DNN baseline; our convolutional architecture appears to use our modeling capacity more efficiently than a dense neural network, and we expect such architectures will soon be more widely used in audio generation tasks.

评估表2总结了我们的实验结果。我们的客观指标表明，与基线相比，改进了1-5 dB，在更高的升频系数下，改进最明显。 尽管样条曲线基线可达到较高的SNR，但其信号通常缺少较高的频率； LSD指标更适合识别此问题。 我们的技术还改善了DNN基线； 我们的卷积架构似乎比密集的神经网络更有效地利用我们的建模能力，并且我们希望这种架构很快会在音频生成任务中得到更广泛的使用。

Next, we confirmed our objective experiments with a study in which human raters were asked to assess the quality of super-resolution using a MUSHRA (MUltiple Stimuli with Hidden Reference and Anchor) test. For each trial an audio sample was upscaled using different techniques[[1]](#footnote-1). We collected four VCTK speaker recordings audio samples from the MULTISPEAKER testing set. For each recording, we collected the original utterance, a downsampled version at *r* = 4, as well as signals super-resolved using Splines, DNNs, and our model (six versions in total). We recruited 10 subjects and used an online survey to ask each of them to rate each sample on a scale of 0 (extremely bad) to 100 (excellent) reconstruction. The results from the experiment are summarized in Table 1. Our method ranked as being the best out of the three upscaling techniques.

接下来，我们通过一项研究证实了我们的客观实验，在该研究中，要求人类评分者使用MUSHRA（具有隐藏参考和锚点的多重刺激）测试来评估超分辨率的质量。 对于每个试验，使用不同的技术来放大音频样本。 我们从MULTISPEAKER测试集中收集了四个VCTK扬声器录音音频样本。 对于每个记录，我们收集原始话语，在r = 4处的降采样版本，以及使用样条线，DNN和模型（共六个版本）进行超分辨的信号。 我们招募了10名受试者，并通过在线调查要求他们每个人以0（极差）至100（极好）重建的等级对每个样本进行评分。 实验结果总结在表1中。在三种放大技术中，我们的方法是最好的。

|  |  |  |
| --- | --- | --- |
|  | LPF (Test) | No LPF (Test) |
|  | SNR LSD | SNR LSD |
| LPF (Train) | 30.1 3.4 | 0.42 4.5 |
| No LPF (Train) | 0.43 4.4 | 33.2 3.3 |

Table 3: Sensitivity of the model to whether low-resolution audio was subject to a lowpass filter (LPF) in dB

Domain adaptation. We tested the sensitivity of our method to out-of-distribution input via an audio super-resolution experiment in which the training set did not use a low-pass filter, while the test set did, and vice-versa. We focused on the PIANO task and *r* = 2. The output from the model was noisier than expected, indicating that generalization is an important practical concern. We suspect this behavior may be common in super-resolution algorithms, but has not been widely documented. A potential solution would be to train on data that has been generated using multiple techniques.

域适应。 我们通过音频超分辨率实验测试了我们的方法对分布外输入的敏感性，在该实验中，训练集没有使用低通滤波器，而测试集却使用了低通滤波器，反之亦然。 我们将重点放在PIANO任务上，并且r =2。模型的输出比预期的要嘈杂，这表明泛化是一个重要的实际问题。 我们怀疑这种行为在超分辨率算法中可能很常见，但尚未得到广泛记录。 潜在的解决方案是训练使用多种技术生成的数据。

In addition, we examined the ability of our model to generalize from speech to music and vice versa. We found that switching domains produced noisy output, again highlighting the specialization of the model.

另外，我们检查了模型从语音到音乐，反之亦然的能力。 我们发现交换域产生了嘈杂的输出，再次突出了模型的专业性。

Architectural analysis. We examined the importance of our various architectural design choices via an ablation analysis on the MULTISPEAKER audio super-resolution task using an upscaling ratio of *r* = 4. The adjacent figure displays the result: the green-ish line display the validation set *`*2 loss of the original model over time; the yellow curve removes the additive residual connection; the green curve further removes the additive skip connection (while preserving the same total number of filters). This shows that symmetric skip connections are crucial for attaining good performance; additive connections add an additional small, but perceptible, improvement.

建筑分析。 我们通过对r = 4的升频比对MULTISPEAKER音频超分辨率任务进行消融分析，研究了各种建筑设计选择的重要性。相邻的图显示了结果：绿线显示了验证集`2损失 随着时间的推移，原始模型的数量； 黄色曲线消除了附加的残余连接； 绿色曲线进一步删除了附加跳过连接（同时保留了相同数量的过滤器）。 这表明对称的跳过连接对于获得良好的性能至关重要。 附加连接增加了其他小的但可察觉的改进。

Model Ablation Analysis

0

1000

2000

3000

4000

5000

0.00

0.01

0.02

0.03

0.04

0.05

Validation L2 loss

No additive or stacking connections

No additive connection

Full model

Step

Figure 3: Model ablation analysis on the MultiSpeaker audio super-resolution task with *r* = 4.

Computational performance. Our model is computationally efficient and can be run in real time. On the PIANO task (where all input signals are 12s in length), our method processed a single second of audio in 0.11s on average on a Titan X GPU. Training our models, however, required about 2 days for the MULTISPEAKER task. Unlike sequence-to-sequence architectures our model does not require the complete input sequence in order to begin generating an output sequence.

计算性能。 我们的模型计算效率高，可以实时运行。 在PIANO任务（所有输入信号的长度均为12s）上，我们的方法在Titan X GPU上平均以0.11s处理了一秒钟的音频。 但是，训练我们的模型大约需要2天才能完成MULTISPEAKER任务。 与序列到序列的体系结构不同，我们的模型不需要完整的输入序列即可开始生成输出序列。

## LIMITATIONS

Finally, to explore the limits of our approach, we evaluated our method on the MagnaTagATune dataset, which consists of about 200 hours of music from 188 different genres. This dataset is larger and much more diverse that the ones we considered so far. We found that our model underfit the dataset, with very little reduction in the training error, and no improvement over the spline baseline. Other learning-based baselines fared similarly. However, we expect improved results with a larger model and more computational resources.

最后，为了探索我们方法的局限性，我们在MagnaTagATune数据集上评估了我们的方法，该数据集包含来自188种不同流派的大约200个小时的音乐。 该数据集比到目前为止我们考虑的数据集更大，并且更加多样化。 我们发现我们的模型不适合数据集，训练误差几乎没有减少，并且在样条曲线基线上没有改善。 其他基于学习的基准也有类似的表现。 但是，我们期望使用更大的模型和更多的计算资源来改善结果。

# PREVIOUS WORK AND DISCUSSION

Time series modeling. In the machine learning literature, time series signals have most often been modeled with auto-regressive models, of which variants of recurrent networks are a special case (Gers et al., 2001; Maas et al., 2012; Mehri et al., 2016). Our approach instead generalizes conditional modeling ideas used in computer vision for tasks such as image super-resolution (Dong et al., 2016; Ledig et al., 2016) or colorization (Zhang et al., 2016).

时间序列建模。 在机器学习文献中，时间序列信号最常使用自回归模型建模，其中循环网络的变体是特例（Gers等，2001; Maas等，2012; Mehri等， 2016）。 相反，我们的方法将计算机视觉中用于图像超分辨率（Dong等人，2016; Ledig等人，2016）或着色（Zhang等人，2016）等任务的条件建模思想概括化。

We identify a broad class of conditional time series modeling problems that arise in signal processing, biomedicine, and other fields and that are characterized by a natural alignment among source/target series pairs and differences that are well-represented by local transformations. We propose a general architecture for such problems and show that it works well in different domains.

我们确定了在信号处理，生物医学和其他领域中出现的一系列条件时间序列建模问题，这些问题的特征在于源/目标序列对之间的自然对齐以及以局部转换很好地表示的差异。 我们为此类问题提出了一种通用的体系结构，并表明它在不同的领域都可以很好地工作。

Bandwidth extension. Existing learning-based approaches include Gaussian mixture models (Cheng et al., 1994; Park & Kim, 2000; Pulakka et al., 2011), linear predictive coding (Bradbury, 2000), and neural networks (Li et al., 2015). Our work proposes the first convolutional architecture, which we find to scale better with dataset size and outperform recent, specialized methods. Moreover, while existing techniques involve many hand-crafted features (see e.g., Pulakka et al. (2011)); our approach is fully domain-agnostic.

带宽扩展。 现有的基于学习的方法包括高斯混合模型（Cheng等，1994; Park＆Kim，2000; Pulakka等，2011），线性预测编码（Bradbury，2000）和神经网络（Li等，2015） ）。 我们的工作提出了第一个卷积架构，我们发现它可以随着数据集的大小更好地扩展，并且胜过最近的专用方法。 此外，虽然现有技术涉及许多手工制作的功能（请参阅例如Pulakka等人（2011））; 我们的方法完全与领域无关。

Audio applications. In telephony, commercial efforts are underway to transmit voice at higher rates (typically 16 Khz) in specific handsets; audio-super resolution is a step towards recreating this experience in software. Similar applications could be found in compression, text-to-speech generation, and forensic analysis. More generally, our work demonstrates the effectiveness of feedforward convolutional architectures on an audio generation task.

音频应用。 在电话中，正在进行商业努力以在特定的手机中以更高的速率（通常为16 Khz）传输语音。 音频超分辨率是朝着在软件中重建这种体验迈出的一步。 在压缩，文本到语音生成和法医分析中可以找到类似的应用程序。 更广泛地说，我们的工作证明了前馈卷积体系结构在音频生成任务上的有效性。

# CONCLUSION

Machine learning techniques based on deep neural networks have been successful at solving underdefined problems in signal processing such as image super-resolution, colorization, in-painting, and many others. Learning-based methods often perform better in this context than general-purpose algorithms because they leverage sophisticated domain-specific models of the appearance of natural signals.

基于深度神经网络的机器学习技术已经成功解决了信号处理中定义不足的问题，例如图像超分辨率，着色，绘画等。 在这种情况下，基于学习的方法通常比通用算法具有更好的性能，因为它们利用了自然信号外观的复杂领域特定模型。

In this work, we proposed new techniques that use this insight to upsample audio signals. Our technique extends previous work on image super-resolution to the audio domain; it outperforms previous bandwidth extension approaches on both speech and non-vocal music. Our approach is fast and simple to implement, and has applications in telephony, compression, and text-to-speech generation. It also demonstrates the effectiveness of feedforward architectures on an important audio generation task, suggesting new directions for generative audio modeling.

在这项工作中，我们提出了利用这种见识对音频信号进行上采样的新技术。 我们的技术将先前有关图像超分辨率的工作扩展到音频领域。 它在语音和非语音音乐方面都优于以前的带宽扩展方法。 我们的方法实现起来快速，简单，并且在电话，压缩和文本到语音生成中都有应用。 它还证明了前馈架构在一项重要的音频生成任务上的有效性，为生成音频建模提供了新的方向。

# REFERENCES

Miguel A Acevedo, Carlos J Corrada-Bravo, Hector Corrada-Bravo, Luis J Villanueva-Rivera, and´ T Mitchell Aide. Automated classification of bird and amphibian calls using machine learning: A comparison of methods. *Ecological Informatics*, 4(4):206–214, 2009.

Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 892–900, 2016. URL [http://papers.nips.cc/paper/](http://papers.nips.cc/paper/6146-soundnet-learning-sound-representations-from-unlabeled-video)

[6146-soundnet-learning-sound-representations-from-unlabeled-video.](http://papers.nips.cc/paper/6146-soundnet-learning-sound-representations-from-unlabeled-video)

Dhananjay Bansal, Bhiksha Raj, and Paris Smaragdis. Bandwidth expansion of narrowband speech using non-negative matrix factorization. In *in Proc. Interspeech*, 2005.

Jeffrey A Bilmes. Graphical models and automatic speech recognition. In *Mathematical foundations of speech and language processing*, pp. 191–245. Springer, 2004.

Jeremy Bradbury. Linear predictive coding. *Mc G. Hill*, 2000.

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.

Emanuele Coviello, Yonatan Vaizman, Antoni B Chan, and Gert RG Lanckriet. Multivariate autoregressive mixture models for music auto-tagging. In *ISMIR*, pp. 547–552, 2012.

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.](http://dx.doi.org/10.1109/TPAMI.2015.2439281)

[1109/TPAMI.2015.2439281.](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.

Felix A Gers, Douglas Eck, and Jurgen Schmidhuber.¨ Applying lstm to time series predictable through time-window approaches. In *International Conference on Artificial Neural Networks*, pp. 669–676. Springer, 2001.

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

Simon Haykin and Zhe Chen. The cocktail party problem. *Neural computation*, 17(9):1875–1902, 2005.

Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.

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.](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.

Dawen Liang, Matthew D. Hoffman, and Daniel P. W. Ellis. Beta process sparse nonnegative matrix factorization for music. In Alceu de Souza Britto Jr., Fabien Gouyon, and Simon Dixon (eds.), *Proceedings of the 14th International Society for Music Information Retrieval Conference, ISMIR 2013, Curitiba, Brazil, November 4-8, 2013*, pp. 375–380, 2013. ISBN 978-0-615-90065-0. URL [http://www.ppgia.pucpr.br/ismir2013/ wp-content/uploads/2013/09/229\_Paper.pdf.](http://www.ppgia.pucpr.br/ismir2013/wp-content/uploads/2013/09/229_Paper.pdf)

Dawen Liang, Minshu Zhan, and Daniel PW Ellis. Content-aware collaborative music recommendation using pre-trained neural networks. In *ISMIR*, pp. 295–301, 2015.

Andrew Maas, Quoc V. Le, Tyler M. ONeil, Oriol Vinyals, Patrick Nguyen, and Andrew Y. Ng. Recurrent neural networks for noise reduction in robust asr. In *INTERSPEECH*, 2012.

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.](http://arxiv.org/abs/1612.07837) cite arxiv:1612.07837.

Augustus Odena, Vincent Dumoulin, and Chris Olah. Deconvolution and checkerboard artifacts. *Distill*, 2016. doi: 10.23915/distill.00003. URL [http://distill.pub/2016/ deconv-checkerboard.](http://distill.pub/2016/deconv-checkerboard)

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 Palomaki, 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.](http://dx.doi.org/10.1109/CVPR.2016.207)

Dennis L. Sun and Rahul Mazumder. Non-negative matrix completion for bandwidth extension: A convex optimization approach. In *IEEE International Workshop on Machine Learning for Signal Processing, MLSP 2013, Southampton, United Kingdom, September 22-25, 2013*, pp. 1–6. IEEE, 2013. doi: 10.1109/MLSP.2013.6661924. URL [http://dx.doi.org/10.1109/MLSP.](http://dx.doi.org/10.1109/MLSP.2013.6661924)

[2013.6661924.](http://dx.doi.org/10.1109/MLSP.2013.6661924)

Aaron 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.](http://arxiv.org/abs/1609.03499)

Xinxi Wang and Ye Wang. Improving content-based and hybrid music recommendation using deep learning. In *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 627–636. ACM, 2014.

Junichi Yamagishi. English multi-speaker corpus for cstr voice cloning toolkit, 2012. *URL http://homepages. inf. ed. ac. uk/jyamagis/page3/page58/page58. html*.

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

1. We have posted a our set of samples to: https://kuleshov.github.io/audio-super-res/. [↑](#footnote-ref-1)