MP3 Diffusion:
Audio Generation in the MDCT Domain

Dan Jacobellis
University of Texas at Austin
danjacobellis@utexas.edu

Matthew Qin
University of Texas at Austin
matthwlqin@utexas.edu

Abstract

Current approaches that adapt diffusion models for audio generation operate in the domain of log Mel-scale magnitude spectrograms (LMMS), a type of lossy time-frequency representation. Operating in this domain provides some advantages, but also introduces distortion and requires a separate learned vocoder model for conversion back to the time domain. In this work, we use low-rank adaptation (LoRA) to fine-tune an existing diffusion-based foundation model to generate audio using an alternative time-frequency representation based on the modified discrete cosine transform (MDCT) which is inspired by lossy audio compression standards such as MPEG Layer III. We propose a simple companding transform based on a generalized Gaussian model of sub-band audio statistics which provides (1) less round-trip distortion, (2) less design complexity, and (3) less computational complexity compared to LMMS+Vocoder, even when combined with 8-bit quantization for storage. Finally, we introduce MDCT-1k, a variant of the Google music captions dataset, based on this transform. Our code is available online:
https://github.com/danjacobellis/audio_diffusion

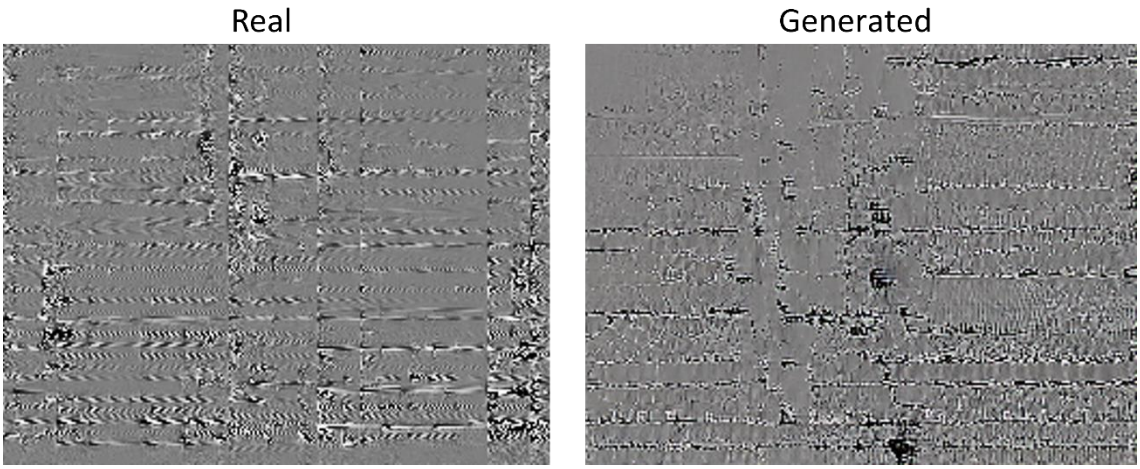

Figure 1. Proposed transform applied to a real audio clip from the MusicCaps dataset (left), and sample generated from the proposed method (right). The transform consists of companding the MDCT coefficients of an audio signal and is resilient to quantization while being invertible without the need for a phase vocoder.

1 Introduction

The rapid development of generative models has opened up new possibilities for audio synthesis, including generating high-fidelity music from textual descriptions. Recently, diffusion-based generative models have shown promising results in audio generation by operating on log Mel-scale magnitude spectrograms (LMMS), a popular but lossy time-frequency representation. However, these approaches introduce distortion and require an additional learned vocoder model for conversion back to the time domain.

In this work, we propose a novel approach to diffusion-based audio generation using an alternative time-frequency representation based on the modified discrete cosine transform (MDCT), which is inspired by lossy audio compression standards such as MPEG Layer III. We present a simple companding transform, leveraging a generalized Gaussian model of sub-band audio statistics to achieve less round-trip distortion, lower design complexity, and reduced computational complexity compared to LMMS+Vocoder-based methods. Furthermore, our the companded MDCT representation is fairly robust to 8-bit quantization for storage. To facilitate training and evaluation, we introduce MDCT-1k, a variant of the Google Music Captions dataset based on our proposed transformation.

In our experiments, we demonstrate fine-tune a diffusion-based foundation model using low-rank adaptation (LoRA) on transformed audio clips from the MusicCaps dataset. While the round-trip distortion introduced by the proposed transformation (including quantization) is low, we found that the quality of generated samples to be poor. This limitation prompts the exploration of potential future work to improve the quality of synthesized audio. Possible directions include training the full denoising U-Net from scratch to better adapt to the MDCT representation or modifying the Variational Autoencoder (VAE) in a latent diffusion model to operate completely in the time domain. By focusing on these areas, we aim to further refine our approach and ultimately generate high-quality audio samples that closely resemble natural audio signals.

2 Related Work

Representations that are good for lossy coding are also good representations for generative modeling and transfer learning. For example, in [1], speech synthesis is performed by operating on the codes produced by a standardized neural audio codec. In this section, we provide background on time-frequency representation of audio and explain the historical trajectory of how they have been applied to learning and generative modeling tasks. We also compare several recent diffusion-based generative models for audio and describe the type of datasets available for training them.

2.1 Time-frequency transforms and lossy audio coding

In analyzing audio signals, especially music and speech signals which have complex and non-stationary frequency content, operating in the joint time-frequency domain is usually preferable. Some standard, invertible time-frequency representations include:

1. Short-Time Fourier Transform (STFT)
2. Discrete Wavelet Transform (DWT)
3. Block Discrete Cosine Transform (DCT)
4. Modified Discrete Cosine Transform (MDCT)

Each of these methods are capable of mapping N time domain audio samples to N time-frequency cells in $O(N \log N)$ time and each is perfectly invertible assuming infinite precision arithmetic.

Some notable differences among these methods are (1) the time-frequency tiling they induce, (2) their ability to compact natural signals into a sparse representations, and (3) the severity of blocking artifacts introduced by quantization or thresholding. For example, the STFT results in a uniform tiling of the time-frequency plane, while the DWT provides better frequency resolution at lower frequencies and better time resolution at higher frequencies. The DCT possess the ability to compact natural signals into sparse representations with efficiency approaching the optimal Karhunen–Loève transform (KLT).

However, it is well known that block DCT transforms (such as those used in the JPEG codec) introduce blocking artifacts. The Modified Discrete Cosine Transform (MDCT) is an extension of the DCT that allows for partial overlap between blocks without introducing redundancy. This results in a more efficient representation and improved reconstruction of the original signal. MPEG Layer III, also known as MP3, uses the MDCT as the basis for its encoding. Though most commonly applied to audio, it has also been shown in [6] that the MDCT provides similar benefits for image compression by reducing JPEG-like blocking artifacts.

2.2 Log Mel-scale magnitude spectrogram

After the initial success of convolutional neural networks (CNNs) for vision tasks in the 2010s, there was a rush to adapt these models to audio. The log Mel-scale magnitude spectrogram (LMMS) emerged as a the popular representation for training CNNs since it provides some of the benefits of a DWT but in a convenient real-valued matrix form to which 2D convolution operations can be applied.

The LMMS was not designed to be an invertible transform since the phase is discarded, low frequencies are overly redundant, and high frequencies have insufficient time resolution. However, for the speech recognition and classification tasks of the time, an inverse transform was not necessary. However, when adapting the LMMS to generative modeling, a vocoder is required for conversion back to the time-domain audio signal, increasing the computational complexity as well as the design complexity, while addition

additional distortion. This is one reason why early generative models for audio such as WaveNet [10] operated purely in the time domain.

2.3 Diffusion based generative models for audio

Current diffusion-based models such as [2], [3], and [9] operate nearly identically to their image-based counterpart by applying denoising steps to the LMMS representation of audio and using a vocoder to recover time domain samples. The vocoder can either be based on a conventional architecture, such as a Griffin-Lim vocoder in the case of [9], or a pretrained GAN-based learned vocoder as in [2] and [3]. Surprisingly, even though [9] only fine-tunes an image-based model, the quality is fairly good. However, all three approaches share similar quality issues, even when [2] and [3] train the diffusion model from scratch. The combination of LMMS and vocoder, originally developed for speech processing, appears to be the fundamental limitation of these models. As it does not appear to adapt well to music or other types of audio signals.

The MusicCaps dataset, introduced in [5], is a valuable resource for training diffusion-based generative models for audio. MusicCaps consists of 5.5k music-text pairs, with rich text descriptions provided by human experts. These descriptions encompass various aspects of the audio, such as the style, instruments, and mood, making it a suitable dataset for training models that generate high-fidelity music from text inputs.

Low-Rank Adaptation (LoRA) [7] is an efficient approach for fine-tuning foundation models, such as the approach used in [9]. Instead of retraining all model parameters, LoRA freezes the pre-trained model weights and introduces trainable rank decomposition matrices into each layer. This significantly reduces the number of trainable parameters, leading to substantial savings in memory and computational resources. As a result, fine-tuning models with LoRA can be performed on a single consumer GPU in just a few hours.

3 Proposed Approach

We propose the following approach to diffusion-based audio generation:

1. Create a time-frequency representation $X[n,k]$ of each audio sample $x[n]$ in the training set using the MDCT.
2. Use the procedure outlined in [4] to estimate variance and shape parameters $\hat{\sigma}^2$ and $\hat{\gamma}$ of a generalized gaussian distribution (GGD) for each frequency subband across many samples from the dataset.
3. Using the estimated GGD parameters, perform companding according to the GGD cumulative distribution function $F_{\mu,\sigma,\gamma}(x)$, thus approximately mapping the distribution of $X[n,k]$ to Uniform[0,1]. Perform uniform 8-bit quantization of this result for storage. We shall refer to the overall transformation of the audio samples $\mathcal{T}\{x[n]\}$ as the output of this step, i.e,

$$Y[n,k] = \mathcal{T}\{x[n]\} = \text{round}\left(256 \cdot F_{0,\hat{\sigma},\hat{\gamma}}(\text{MDCT}\{x[n]\})\right)$$

4. Adapt a pretrained, diffusion-based foundation model by using LoRA [7] and training on transformed audio clips $\mathcal{T}\{x[n]\}$ paired with text captions.
5. Perform inference of the fine-tuned model, then apply the inverse transformation $\mathcal{T}^{-1}\{Y[n,k]\}$ to synthesize time domain audio waveforms. The inverse transformation is defined as

$$\hat{x}[n] = \mathcal{T}^{-1}\{Y[n,k]\} = \text{IMDCT}\left\{\frac{1}{256} \cdot F^{-1}_{0,\hat{\sigma},\hat{\gamma}}(Y[n,k])\right\}$$

4 Experiments

4.1 Sub-band Audio Statistics

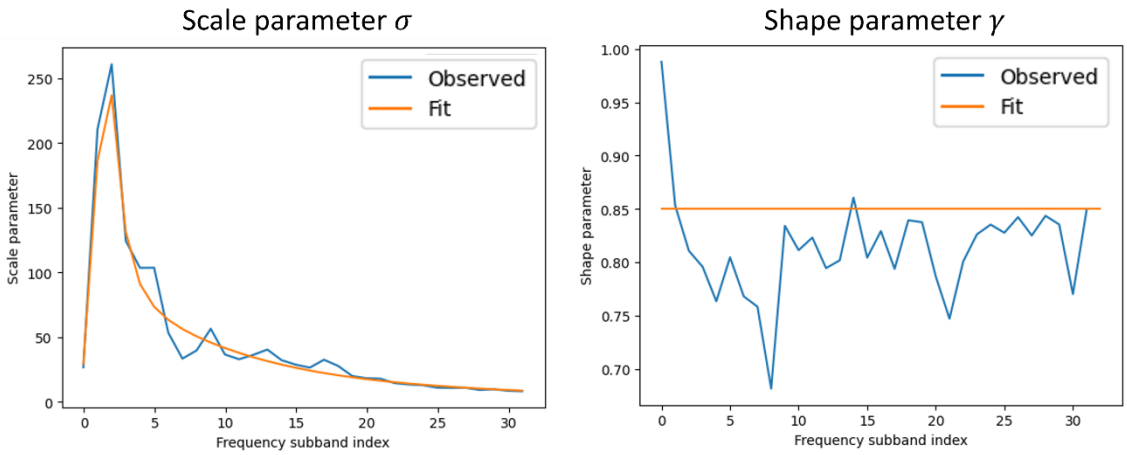

Figure 2. Estimated generalized Gaussian scale (left) and shape (right) parameters of audio from the MusicCaps dataset. The lowest subband index corresponds to 0 Hz and the highest subband index corresponds to 12kHz. The orange curves show the values used for the transformation $\mathcal{T}$ and its inverse $\mathcal{T}^{-1}$.

4.2 wav2jpeg

To demonstrate the efficacy of this transformation and its resiliency to quantization, we implement and extremely simple lossy audio codec 'wav2jpeg' which consists of $\mathcal{T}$ composed with a standard JPEG encoder. The code and audio example demonstrating the level of distortion for this codec at 21 kbits/s is available at the following URL: huggingface.co/datasets/danjacobellis/MDCT-1k/blob/main/MDCT.ipynb

4.3 Audio generation

We process over 1000 audio samples from the MusicCaps dataset using the proposed transformation. Examples of the transformed data can be explored at the following URL: huggingface.co/datasets/danjacobellis/MDCT-1k. We use the LoRA fine-tuning of a text-to image model following the procedure of [11]. Our justification for this approach is that the quality of audio produced by similar fine tuning using LMMS [9] is not far

from [2] and [3] which train the entire model, including the denoising U-Net from scratch. The fine-tuning procedure ran for 15000 steps (roughly 10 epochs) over the course of 4 hours on a single RTX 2060 GPU. An example demonstrating use of the fine-tuned model for text-to-audio generation along with an audio sample is accessible at: huggingface.co/datasets/danjacobellis/MDCT-1k/blob/main/music_inference.ipynb.

4.4 Discussion

In our experiments, we observed that the proposed transformation based on MDCT and generalized Gaussian companding results in reasonably low round trip distortion, but with significantly increased computational and design efficiency compared to the LMMS+Vocoder. However, we found that the quality of the generated audio samples was not satisfactory, indicating that there are still limitations in our current approach. One possibility is that the fine-tuning process using LoRA may not be sufficient for adapting the model to the new representation effectively. Future work could explore different strategies for adapting the foundation model to the MDCT representation. One option is to train the full denoising U-Net from scratch, allowing the model to learn the features and representations that are specific to the MDCT domain. Another direction is to modify the Variational Autoencoder (VAE) in a latent diffusion model to operate entirely in the time domain.

5 Conclusion

In this work, we explored an alternative approach to diffusion-based audio generation using a time-frequency representation based on the modified discrete cosine transform (MDCT) and a simple companding transform leveraging generalized Gaussian sub-band audio statistics. Our proposed method exhibits reduced round-trip distortion, lower design complexity, and decreased computational complexity compared to LMMS+Vocoder-based methods while maintaining resilience to 8-bit quantization for storage. We also introduced the MDCT-1k dataset, a variant of the Google Music Captions dataset, to facilitate training and evaluation.

Our experiments demonstrated the potential of our approach, but the quality of the generated audio samples still requires improvement. Future work could focus on better adapting the foundation model to the MDCT representation, such as training the full denoising U-Net from scratch or modifying the Variational Autoencoder in a latent diffusion model to operate entirely in the time domain. By addressing these limitations, we hope to further refine our method and ultimately generate high-quality audio samples that closely resemble natural audio signals.

6 References

[1] C. Wang et al. "Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers."

[2] C. Hawthorne et al. "Multi-instrument Music Synthesis with Spectrogram Diffusion."

[3] H. Liu et al. "AudioLDM: Text-to-Audio Generation with Latent Diffusion Models."

[4] K. Sharifi et al. "Estimation of shape parameter for generalized Gaussian distributions in subband decompositions of video."

[5] A. Agostinelli et al. "MusicLM: Generating Music From Text."

[6] R. Muller. "Applying the MDCT to Image Compression."

[7] E. Hu et al. "LoRA: Low-Rank Adaptation of Large Language Models."

[8] S. Park et al. "SeiT: Storage-Efficient Vision Training with Tokens Using 1% of Pixel Storage."

[9] S. Forsgren et al. "Riffusion - Stable diffusion for real-time music generation"

[10] A. van den Oord et al. "WaveNet: A Generative Model for Raw Audio"

[11] P. Cuenca et al. "Using LoRA for Efficient Stable Diffusion Fine-Tuning"