VOICE2SOUND: Devocalization for Timbral Control

Mason Wang
Stanford University
ycda@stanford.edu

Wanyue Zhai
Stanford University
wzhai702@stanford.edu

Cody Ho
Stanford University
codyho@stanford.edu

Abstract

The human mind can imagine an infinite variety of unique sounds, yet existing sound design methods fail to map these comprehensively into real waveforms. Traditional synthesizers offer detailed control over a limited range of timbres, while modern text-to-audio models, despite their diversity, lack fine-grained control. To bridge this gap, we introduce VOICE2SOUND, a model that transforms vocal imitations into their intended sounds by leveraging the rich expressive control inherent in human vocalization. Our primary approach utilizes the vast generative capabilities of a large, pretrained audio diffusion model to morph the timbre of a vocal imitation while applying a spectral mask to maintain pitch and timing. Consequently, VOICE2SOUND offers a new form of sound synthesis that complements existing methods and expands the explored regions of the timbral space.

1 Introduction

The human mind is capable of imagining nearly any sound. The loudspeaker can generate any of these sounds when given a waveform. The goal of sound design is to unite these two: to map the vast space of unique imaginary sounds, or the *timbral space* (Wessel, 1979), into real waveforms that can be played from a loudspeaker.

Large swaths of the timbral space remain unmapped by existing methods. Synthesizers can generate sound using controllable oscillators and envelopes, but require extensive training and but only parameterize a small subset of the timbral space. In contrast, text-to-audio models (Dhariwal et al., 2020; Agostinelli et al., 2023; Copet et al., 2024; Huang et al., 2023; Liu et al., 2023) can generate an incredibly diverse set of sounds, but lack the fine-grained control of a synthesizer. Text can specify a broad variety of sounds, but very coarsely - for instance, it is difficult to use text to reconstruct or even distinguish between the various notification sounds on your phone.

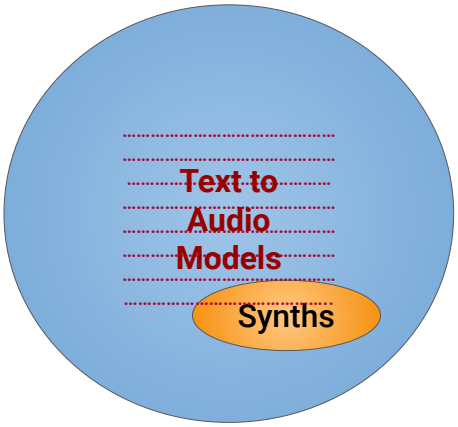


Figure 1: A visualization of the timbral space, or the space of all possible unique sounds. Synths allow a small subset of sounds to be mapped into real waveforms, with a high degree of control. Text-to-Audio models can generate a broad range of sounds, but lack precise control.

Thus, we propose VOICE2SOUND- a model that allows you to imitate a sound with your mouth and then creates the sound you are trying to imitate. Unlike text-to-audio models, human vocalization provides a rich form of low-level control, allowing us to express dynamic envelopes, transients, resonances, tremolos, vibratos, and so on. Unlike synthesizers, VOICE2SOUND is simple and natural, and can specify a wide variety of complex sounds. Thus, VOICE2SOUND provides a unique form of controlled sound design that can complement existing methods, and map complementary portions of the timbral space.

2 Related Works

Query by Vocal Imitation. Our task closely relates to query by vocal imitation (QBV), where users provide an audio imitation example as a query to find a desired sound in the database. Due to its nature as a search problem, previous attempts

have focused on retrieval-based methods. Kim and Pardo (2019) leveraged audio feature extraction methods from existing models to measure the similarity between CNN features. They also incorporated user feedback, allowing additional positive or negative imitations to update the QBV measures. Zhang et al. (2020) focused on the deployment and user evaluation of QBV systems. They developed the *Vroom!* system, replacing CNN layers in TL-IMINET with a combination of CNN and GRU units. Their results, compared to the baseline text search engine *TextSearch*, demonstrated that complementary QBV can enhance both search performance and user experience. While these studies provide valuable insights into data and metrics, they differ substantially from our task because they focus on retrieval instead of generation.

Text-to-Audio Models. At a high level, we can view our task of voice-to-audio synthesis as doing two things: we want to both remove the vocal characteristics from the input and to 'fill-in' the remaining components of the intended 'real' sound. The latter process requires a model with prior knowledge of a broad distribution of real sounds so that it can 'recognize' the sound that we are trying to imitate and generate a realistic version of it. Recent years have shown an explosion in text-to-audio models text-to-audio models (Dhariwal et al., 2020; Agostinelli et al., 2023; Copet et al., 2024; Huang et al., 2023; Liu et al., 2023). These models have been trained on an extremely large number of text-audio pairs, and thus are capable of generating a wide variety of realistic sounds. We would like to distill the generative abilities from one of these large models to enhance our synthesis procedure.

Thus, our approach builds off of Riffusion (Forsgren and Martiros, 2022), a text-to-audio latent diffusion model based on Stable Diffusion (Rombach et al., 2022). Riffusion consists of two components: First, it has a variational autoencoder (VAE), trained to compress and decompress mel-spectrograms into low-dimensional latent representations (Kingma et al., 2019). Second, it has a diffusion model trained to iteratively denoise Gaussian noise signals into latent vectors representing real-world sounds. The diffusion model is conditioned on a text prompt, which is used to guide the denoising process into producing a desired sound. For instance, if the text prompt is "Guitar", the diffusion model will generate a latent vector that, when fed to the VAE's decoder, generates a guitar sound.

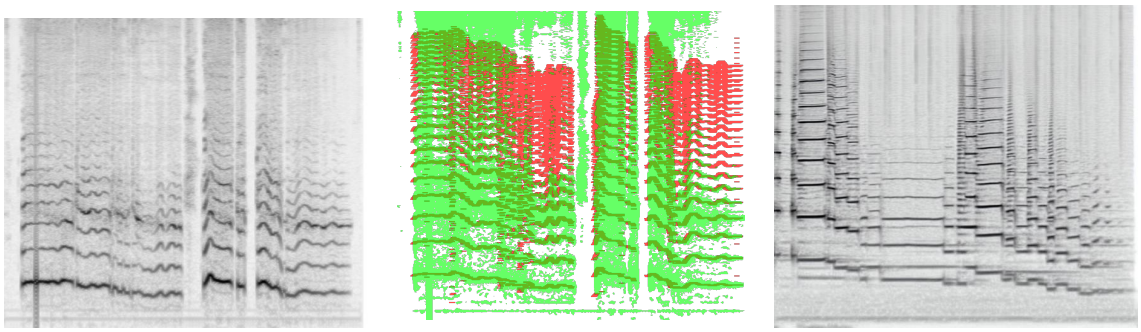
Harmonic Modeling. VOICE2SOUND should ideally allow for control over the vocal input's pitch, timing, and dynamics while adapting its timbre to fit that of the targeted 'real' sound. Sines plus noise models are one way of disentangling a sound's timbre from its non-timbral qualities (Serra, 2013). These models decompose sound into sinusoids (which represent the sound's harmonic components) and filtered noise (which represent the sound's transients, and other textural qualities like buzziness or grittiness). A huge variety of sounds can be modeled as a sum of sinusoids and noise, allowing for such models to be used in interesting tasks in music processing, for instance, transferring the timbre from one instrument to another (Engel et al., 2020). It should also be noted that for monophonic sounds (like the voice), the frequencies of the sinusoids occur as multiples of the fundamental frequency and are called 'harmonics'. The relative amplitudes of these harmonics largely determine the sound's timbre - for instance, the clarinet only has odd harmonics (Wolfe, 2024).

3 Methods

We present several approaches to this problem with varying degrees of success. For all methods, we concern ourselves with the task of transforming the mel-spectrogram of the vocal input into the mel-spectrogram of a realistic sound. The mel-spectrogram can then be inverted into a playable waveform using the Griffin-Lim algorithm (Griffin and Lim, 1984).

3.1 Devocalize and Diffuse

Our primary approach seeds Riffusion's diffusion model with the vocal input, and performs the diffusion process in a manner that preserves the vocal input's pitch, timing, and dynamics, while allowing the timbre of the vocal input to morph. To do this, we use traditional signal processing tools to generate a mask for the vocal input's spectrogram. Our approach applies two kinds of conditioning to the diffusion model. First, the mask ensures that the diffusion occurs only in regions of the spectrogram relating to timbre, while preserving everything else. Second, the timbre of the vocal input should have some influence over the generation (e.g., to control buzziness), but should not be retained completely. Thus, we condition the generation weakly on the



(a) Vocal Input (b) Harmonic and Noise Masks (c) Devocalized Output

Figure 2: Mel spectrograms for the vocal input, the harmonic and noise masks, and the generated output using the "Devocalize and Diffuse" method. The noise mask is shown in green, and the harmonic mask is shown in red.

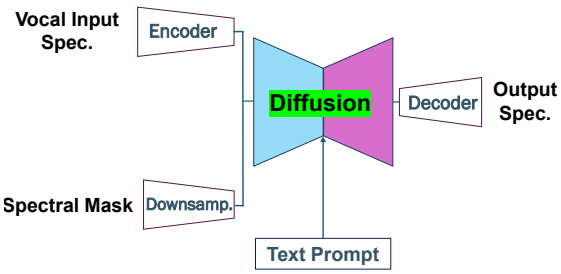


Figure 3: The Devocalize and Diffuse method of generating realistic audio from vocal imitations. The Vocal Input and Outputs are both mel-spectrograms. A text prompt can be added to jointly condition the generation on the vocal input and text.

vocal input's timbre by seeding the diffusion process with the vocal input's spectrogram. Figure 3 shows the process in action.

Mask Generation. The mask determines where the diffusion model is allowed to modify the vocal input's spectrogram, and where it should keep it the same. Since our goal is to modify the vocal input's timbre but keep its pitch, timing, and dynamics, the mask should forbid the diffusion model from generating in regions that would result in new transients or new fundamental frequencies.

Inspired by sines plus noise modeling, the mask contains two components - one that corresponds to the harmonics of the vocal input's fundamental frequency (which would be modeled by sines), and one that corresponds to the vocal input's noisy qualities (which would be modeled by filtered noise).

To generate the **harmonic mask**, we perform pitch detection on the vocal input using the Robust Epoch And Pitch EstimatoR (REAPER) (Talkin, 2014). This provides us with a time-series of the vocal input's fundamental frequency $\mathbf{f_0}$. We then compute its harmonics ($k\mathbf{f_0}$ for $k = 1, 2, ...$). We highlight the harmonics on a mel-spectrogram and allow the diffusion model to generate in these regions.

To generate the **noise mask**, we threshold the energy of the vocal input's spectrogram to highlight bins louder than the volume of the loudest spectrogram bin minus 60 dB. We highlight these regions on the mel-spectrogram, and allow for the diffusion model to generate in these regions. Note that using the noise mask alone would prevent the diffusion model from adding in harmonics that were not originally present in the vocal input. Figure 2 shows the difference between the two masks.

Diffusion. We provide the diffusion model with our vocal input as a starting point, and run 50 denoising steps on it. Riffusion allows for negative text prompts, which allow you to discourage certain outputs. We provide Riffusion with the negative prompt of "Voice" to encourage vocal ablation. The text input is used to encourage certain outputs, e.g. "Trumpet", so that both text and vocal control can be used to control the generation. Our qualitative results show that this text prompt can meaningfully guide our generation. The denoising process yields a latent vector, which can be decoded into an output mel-spectrogram.

3.2 VAE Inspired Methods

The VAE of a text-to-audio diffusion model is trained to reconstruct the mel-spectrogram it is given after compressing it into a lower-dimensional latent representation. By training the VAE to reconstruct a large number of audio examples, the goal is to procure a model that can encode mean-

ingful variations in high-dimensional audio data, by reducing it to a low-dimensional latent space.

We hypothesize that there is a mapping between the distribution of vocally imitated recordings and the distribution of real recordings within the latent space of Riffusion's VAE. We attempt to learn this mapping by encoding all of our training-set reference and imitation recordings, and trying two approaches:

U-Net. U-Net (Ronneberger et al., 2015) is originally developed for biomedical image segmentation, and uses a U-shaped structure with symmetric encoder and decoder paths that allow for precise segmentation. We train an adaptation of the U-Net model for regression on the task of reconstructing the latent vector of a 'real' mel-spectrogram from the latent vector of its corresponding vocal imitation. We optimize with respect to the mean-squared error between our predicted latent space and the latent space of the real recording on the training set.

AdaIn. AdaIn (Huang and Belongie, 2017) is an adaptive instance normalizing layer for style transfer that adjusts the mean and variance of the content input to match the measures for the targeted style. In our case, we measure the mean and variance of the latent encodings of our reference recordings, as well as the mean and variance of the latent encodings of our imitation recordings. To transform the vocal imitation, we apply a linear transformation 1 to get the aligned encodings.

$$AdaIN(x) = \sigma_Y\Big(\frac{x - \mu_X}{\sigma_X}\Big) + \mu_Y \qquad (1)$$

In the equation, μ_X, μ_Y represent the mean of the latent vectors of the vocally imitated recordings and 'real' reference recordings, respectively, while σ_X, σ_Y represent their variances. We apply the transformation to x, which is the latent representation of a vocal imitation.

Supervised Fine Tuning. Supervised fine-tuning involves leveraging a pre-trained model and subsequently training it on a labeled dataset specific to the target task. We utilize Riffusion's pre-trained VAE as the initial model. We then refine this model through supervised learning on a dataset consisting of paired mel-spectrograms and their corresponding vocal imitations by minimizing the mean-squared error between the predicted mel-spectrogram and the ground-truth mel-spectrogram for each pair in the training set. The goal of this fine-tuning is to enhance the model's capability to generate more accurate and contextually relevant audio outputs, by honing the learned representations in the latent space for both the mel-spectrograms and their vocal imitations. Through this targeted training, the model adapts better to the specific nuances of the task, potentially benefiting from the priors learned during the initial VAE training phase.

4 Experiments

4.1 Dataset

The dataset we are using for evaluation and/or training is the Voice Imitation Dataset (Kim et al., 2018), a dataset of over 11,000 recordings of high-quality vocal imitations. The dataset contains 302 unique reference sounds, which are the real sounds that the imitations in the dataset are attempting to replicate. The 302 reference sounds include musical instrument and synthesizer sounds, sounds from animals and pets, and recordings of everyday sounds. The vocal imitations are crowdsourced using Amazon Mechanical Turk, meaning that random users online were paid to record vocal imitations of real sounds. For quality control purposes, researchers in the Interactive Audio Lab rated each vocal imitation. For each reference sound, there are many vocal imitations performed by several imitators. We performed substantial data preprocessing, filtering out all imitations with a score below 30, resampling the audio to 22050 Hz, removing trailing and leading silence from all audio recordings, and discarding recordings shorter than three seconds. Our resulting filtered dataset contains 2825 imitations of 233 unique sounds.

We evaluate all of our methods on our test-split of the Vocal Imitation Set. For each method, we attempt to reconstruct the reference recording given the vocal imitation.

4.2 Quantitative Results

In Table 1, we show the mel-spectral MSE between the output of each of our methods and the reference recording on the test split of the Vocal Imitation Set. Our "Devocalize and Diffuse" method appears to be outperforming the others, although evaluations using spectral-MSE are difficult for several reasons: the vocal imitations and their corresponding references are not exactly time or pitch-aligned, and a vocal imitation recording may have multiple valid

	Mel-Spectral MSE
Devocalize and Diffuse	129.9
U-Net	45320
AdaIn	123500
Supervised Fine Tuning	13570

Table 1: Comparison between our methods and their respective baseline. The numbers shown are the mean-squared error between the mel-spectrograms of the model's output, and the reference recording they are trying to reproduce. Numbers are averaged across the test set.

'real' sounds.

4.3 Qualitative Results

Qualitative examples of our primary method are available at masonlwang.com/devocalization. We observe that the results generally preserve the pitch and rhythm of the input recording without preserving any of the vocal characteristics. Our primary method also tends to produce high-quality, realistic sounds. The quality of our results is often correlated with the quality of the vocal imitation. For instance, the vocal "Chainsaw" imitation is already very high qualty, but our model devocalizes it further, removing any evidence of the sound being vocally imitated. It follows the pitch of the vocal imitation very well, and even the dynamics - as the vocal imitation gets quieter, the chainsaw seems to get further away in our model's output. It also adds in the rapid, stuttery motor noise of a real chainsaw, which is difficult or impossible to imitate using the mouth.

Embellishment. While our model consistently follows the pitch track of the vocal imitation, sometimes it adds in extra notes, which is the case for the "Trumpet" example. We hypothesize this to be the result of two things: First, the frequency bins in the mel-spectrogram are not small enough to specify exact notes. Thus, the mel-spectral harmonic mask does not constrain the output to an exact sequence of notes, but rather a small range of notes in each time interval. Second, the training distribution likely contains lots of trumpet solos, who tend to embellish their performances by adding in lots of extra notes. Thus, the diffusion model tends to generate these embellishments. Since the mel-spectral input representation and the training dataset are both parts of the pretrained Riffusion model. Embellishment can also be a desired behavior, since it enhances the user's creativity, and can produce passages thare are too fast to imitate with the human voice.

Areas of Improvement. Our model sometimes produces a generation that fits the vocal input, but in a seemingly unintended way. As shown on our website, our model takes a "ringtone" imitation, and turns it into an alien-like laser blaster. The output is still devocalized and realistic, and follows the pitch, rhythm, and timbre of the input, but might not be what the imitator intended to produce. Decoding user intent is difficult since the best 'realistic' pair to a vocal imitation is often requires subjective human judgement.

Other Methods The generated recordings for the other methods are less satisfying. U-Net is not always able to capture the latent representation of the input recordings and thus generates noise that was present in the original recording. It also suffers from time alignment issues due to the unrestricted spectrogram generation. AdaIn, on the other hand, was able to preserve the pitch and time in the original recording, but are not stylistically aligned with all recordings. Given the wide range of sound categories, it is expected that the style imposed on the input recording is a mix of features from the target recordings. Therefore, further exploration of style constraints can be done to improve the performance. Supervised fine tuning sufferred from similar issues – it failed to meaningfully capture the latent representations of any sounds other than what were directly included in the dataset, likely because there were a relatively small number of samples, many of which were similar and/or had the same subject. It is difficult to evaluate whether a larger dataset would improve performance on this task.

5 Conclusion

In this paper, we introduced VOICE2SOUND, a novel model designed to transform vocal imitations into realistic sound outputs, addressing the limitations of existing sound design methods. By leveraging the expressive capabilities of human vocalization, VOICE2SOUND provides a unique approach to controlled sound synthesis, enabling fine-grained manipulation of timbre, dynamics, and other sonic characteristics. Our methods, including the "Devocalize and Diffuse" approach and VAE-inspired techniques, demonstrate the potential to

map the vast timbral space more comprehensively than traditional synthesizers and text-to-audio models.

Our experiments highlight the strengths and challenges of our approach. While quantitative evaluations using spectral-MSE provide some insights, they also reveal the need for more robust qualitative metrics to fully capture the effectiveness of our model. Future work will focus on improving the alignment of pitch and timing in generated outputs, refining evaluation methods, and enhancing the overall quality of sound synthesis.

References

Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. 2023. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. 2024. Simple and controllable music generation. *Advances in Neural Information Processing Systems*, 36.

Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. 2020. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341*.

Jesse Engel, Lamtharn Hantrakul, Chenjie Gu, and Adam Roberts. 2020. Ddsp: Differentiable digital signal processing. *arXiv preprint arXiv:2001.04643*.

Seth Forsgren and Hayk Martiros. 2022. Riffusion - Stable diffusion for real-time music generation.

Daniel Griffin and Jae Lim. 1984. Signal estimation from modified short-time fourier transform. *IEEE Transactions on acoustics, speech, and signal processing*, 32(2):236–243.

Qingqing Huang, Daniel S Park, Tao Wang, Timo I Denk, Andy Ly, Nanxin Chen, Zhengdong Zhang, Zhishuai Zhang, Jiahui Yu, Christian Frank, et al. 2023. Noise2music: Text-conditioned music generation with diffusion models. *arXiv preprint arXiv:2302.03917*.

Xun Huang and Serge Belongie. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510.

Bongjun Kim, Madhav Ghei, Bryan Pardo, and Zhiyao Duan. 2018. Vocal imitation set: a dataset of vocally imitated sound events using the audioset ontology. In *DCASE*, pages 148–152.

Bongjun Kim and Bryan Pardo. 2019. Improving content-based audio retrieval by vocal imitation feedback. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4100–4104. IEEE.

Diederik P Kingma, Max Welling, et al. 2019. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392.

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. 2023. Audioldm: Text-to-audio generation with latent diffusion models. *arXiv preprint arXiv:2301.12503*.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer.

Xavier Serra. 2013. Musical sound modeling with sinusoids plus noise. In *Musical signal processing*, pages 91–122. Routledge.

David Talkin. 2014. Reaper: Robust epoch and pitch estimator. Accessed: 2024-05-31.

David L Wessel. 1979. Timbre space as a musical control structure. *Computer music journal*, pages 45–52.

Joe Wolfe. 2024. Clarinet acoustics. Accessed: 2024-05-31.

Yichi Zhang, Junbo Hu, Yiting Zhang, Bryan Pardo, and Zhiyao Duan. 2020. Vroom! a search engine for sounds by vocal imitation queries. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*, pages 23–32.

6 Contributions

- Cody conducted experiments around supervised fine tuning of the Riffusion VAE.
- Mason came up with the idea for this project, did most of the data preprocessing and preparation, and was responsible for the Devocalize and Diffuse experiments.
- Wanyue experimented with with U-Net and AdaIn.

All members contributed equally. We had no external collaborators.