A Differentiable Simulator for Human Vocal Learning

Eric M. Chen
MIT CSAIL
echen@csail.mit.edu

Vincent Sitzmann
MIT CSAIL
sitzmann@csail.mit.edu

Abstract

Speech is a complex sensorimotor process that requires the coordination of hundreds of muscles, yet is something we humans can do almost automatically. Although computational models of how the vocal tract maps to speech have been studied for almost a century, the inverse process—reconstructing the dynamic vocal tract shape from its radiated speech—remains a challenge. Prior works which attempt to model the inverse process with acoustic simulation rely on sampling, which is inefficient due to the high dimensionality of the search space. To overcome this problem, we propose to model speech production with a differentiable simulator, which directly couples the vocal tract geometry and its acoustic output with analytic gradients, thereby efficiently modeling both the forward and inverse processes. Experiments show that our simulator can reconstruct human-like speech, enabling potential applications in studying infant language acquisition, psycholinguistics, and sensorimotor control.

1 Can you hear the shape of a vocal tract?

When you learn to say a new word or phrase, the process can be reminiscent of gradient descent. Try saying the tongue-twister "she sells sea shells by the sea shore." Or maybe "sea shells she shells by the she shore?" Either way, as you wallow through, you quickly realize how small changes in how you place your tongue or constrict your throat lead to dramatically different sounds. What do you do next? Perhaps you experiment. You shift your tongue front or back, adjust how tightly you stretch your mouth, and just as with gradient descent, you repeat the process—iteratively tuning your speech until you get closer and closer to the correct phrase: *she sells sea shells by the sea shore.*

This iterative process can be formulated as an optimization problem: the objective is to match your produced utterance with the target utterance, while the learnable parameters are your tongue position, mouth opening, velum position, etc. But is this optimization problem, inferring the motion of the vocal tract from only the speech it produces, even tractable in the first place? Put simply, "can you hear the shape of a vocal tract?"[1]

Mathematically, the question is ill-posed. Different geometries can produce identical resonant frequencies [Panchapagesan and Alwan, 2011]. A well-known example occurs in American English. The /r/ sound, as in but**er** and doll**ar**, can be pronounced in two ways that are acoustically indistinguishable but anatomically distinct: either with a bunched tongue or a retroflex [Arai, 2014]. Yet in practice, humans routinely appear to do something close to "hearing the shape of a vocal tract". Infants learn to speak by hearing alone—without seeing other people's vocal muscles. Longitudinal studies have shown that blind children, who cannot read lips, learn language at a pace similar to sighted children [Landau and Gleitman, 1985]. While we do not attempt to model the cognitive

[1]This phrasing is a play on the classic mathematics question "Can one hear the shape of a drum?" [Kac, 1966]. Although different shapes can produce identical resonances [Gordon and Webb, 1996], it is unclear how often these ambiguities arise in human speech.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Data on the Brain & Mind.

processes for speech acquisition in this paper, what we can take away from human experience is that the vocal learning process, though ambiguous, is still learnable. Framing this sensorimotor faculty as an optimization process, we introduce a *differentiable simulator* to map dynamic vocal tract shapes to speech, and perform the inverse: reconstructing the shape of the vocal tract from its radiated sound.

By modeling the fluid dynamics of the vocal tract, the simulator provides a physically-grounded representation for speech modeling, and potentially enables myriad applications in cognitive science. For example, it could offer a computational framework to test hypotheses on infant speech acquisition [Kuhl and Meltzoff, 1996, Beguš et al., 2023], study the neuroscience of speech comprehension [Gwilliams et al., 2024], and provide insights into theory of mind [Caren et al., 2024].

2 Related Work

Articulatory Synthesis and Inversion. While articulatory synthesis has been studied for almost a century [Chiba and Kajiyama, 1941, Fant, 1960], the inverse process, reconstructing the dynamic vocal tract shape from its radiated speech, remains elusive. Take VocalTractLab [Birkholz, 2013] for example, a modern speech simulator that simulates the acoustics of the vocal tract. Because VocalTractLab is not differentiable, prior works that attempt to model vocal learning with VocalTractLab rely on gradient-free optimization. These approaches search the articulatory space via sampling, but are slow because of the high dimensionality of the parameter space. These models have only been able to produce short syllable clusters [Krug et al., 2023, Xu et al., 2024]. On the other hand, a differentiable simulator directly couples the vocal tract geometry and acoustic output with analytic gradients, accelerating the process.

Neural Surrogate Models. To overcome the difficulties of creating physical simulations, recent works have introduced various neural surrogate models: neural networks trained to map articulatory parameters directly to speech. However, their quality fundamentally depends on paired training data, which is sparse. Krug et al. [2025] train a surrogate model to map anatomical controls to synthetic sounds generated from VocalTractLab. And as shown in the experiments, it does not generalize as well to real speech as our simulator. Surrogates have also been trained on other physiological signals, such as vocal tract MRI [Nguyen et al., 2024] and electromagnetic articulography (EMA) data [Cho et al., 2024]. However, these data are costly to collect, as they require specialized machinery. On the other hand, our method does not require paired data between articulatory parameters and speech.

3 Background

We base our differentiable speech simulator on the source-filter models of Maeda [1982], Sondhi and Schroeter [1987] and Birkholz [2013]. For a comprehensive introduction to the physics of speech, we refer the interested reader to Flanagan [1972].

The Source-Filter Model. Speech is formed as air pushed from the lungs travels through the vocal tract, forming resonances and turbulence, which is eventually radiated from the mouth and nostrils. While a high-fidelity simulation of these fluid dynamics could be modeled with the Navier-Stokes equations, speech is well approximated with the simpler source-filter model. The model divides speech production into two stages. First, a sound source $u_0(t)$—such as vocal fold vibration or broad-spectrum noise—is generated at the glottis. Second, the geometry of the vocal tract acts as an acoustic filter h_{tract}. The radiated pressure, $p_{\text{out}}(t)$, which we know as speech, is the convolution of the source and filter: $p_{\text{out}}(t) = h_{\text{tract}} * u_0(t)$. The peak resonances of the transfer function h_{tract} are called *formants*, and the frequencies of these formants are critical to how humans disambiguate vowels. When constrictions are formed in the vocal tract to produce consonants, additional noise sources $u_i(t)$ from turbulence are added to the system. The radiated pressure becomes the superposition of multiple sources and filters:

$$p_{\text{out}}(t) = h_{\text{tract}} * u_0(t) + \sum_{i=1}^{n} h_{\text{fric}_i} * u_i(t). \tag{1}$$

The Transmission Line Model. For frequencies up to 4kHz, the 3D geometry of the vocal tract is well approximated by a cylindrical tube. The most significant factor affecting h_{tract} is the tube's

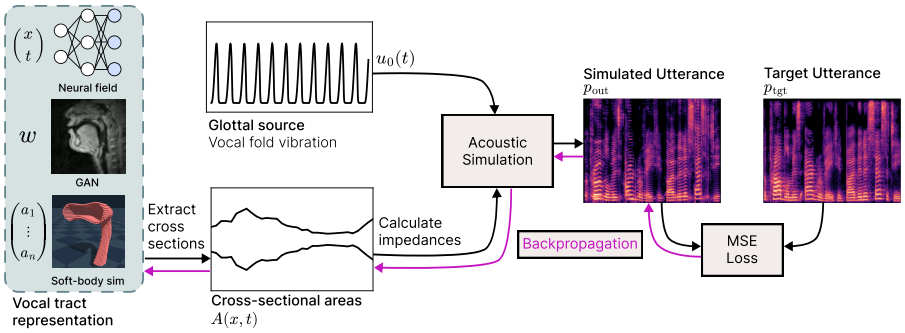

Figure 1: **End-to-End Vocal Simulation and Animation.** Because our acoustic simulator is fully differentiable, we can couple a variety of vocal tract representations to their acoustics, allowing us to animate both the muscle movements and speech. This enables a user to visualize how articulatory gestures map exactly to sound, with guarantees of physical plausibility.

cross-sectional area, which we call $A(x)$. The transmission line model provides a method for calculating h_{tract} efficiently by transforming the area function into an analogous RLC circuit. First, $A(x)$ is discretized into sections $A_1, A_2, \ldots, A_n$, and is used to compute acoustic impedances for each respective section. As done in Sondhi and Schroeter [1987], Birkholz and Jackel [2004], the impedances are then used to form a chain of 2D matrices $K_1, K_2, \ldots, K_n$, which map the input acoustics to the output acoustics in each section. The resulting matrix defining the entire input/output behavior of the vocal tract is the matrix product $K_{\text{tract}} = K_n \ldots K_2 K_1$. Finally, given a boundary condition defined by the impedance at the lips, Z_{lips}, the frequency response for h_{tract} is

$$H_{\text{tract}}(\omega) = \frac{P_{\text{out}}}{U_0} = \frac{Z_{\text{lips}}}{K_{\text{tract}_{0,0}} - K_{\text{tract}_{1,0}} Z_{\text{lips}}}. \tag{2}$$

This model establishes the foundation for our method. However, differentiability is a problem when modeling consonants, since transmission line models represent consonants with discontinuities.

4 Method

Given the transmission line model, our objective is to find a time-varying area function $A(x,t)$ that best reconstructs a target utterance $p_{\text{tgt}}(t)$. Our loss function is the mean-squared error loss between the log Mel spectrograms of the utterances.

$$\mathcal{L}(p_{\text{out}}, p_{\text{tgt}}) = \| \log |\text{mel}(p_{\text{out}})| - \log |\text{mel}(p_{\text{tgt}})| \|_2. \tag{3}$$

With the Adam optimizer [Kingma and Ba, 2014], we minimize the loss by backpropagating through a differentiable version of the transmission line model.

Glottal and Noise Sources. From $p_{\text{tgt}}(t)$, we first extract the glottal pulse frequency using the PYIN algorithm [Mauch and Dixon, 2014], and parameterize the pulse with the Liljencrants-Fant model Fant et al. [1985]. Unvoiced speech is represented with white noise. We call this signal $u_0(t)$. To make noise sources caused by frication differentiable, we choose to add a noise source at every section in the vocal tract model, and smoothly gate its volume with a softplus function. As the cross section of the vocal tract becomes smaller, the Reynolds number becomes larger, eventually exceeding the gate's threshold.

GPU Parallelization. To extend the transmission line model to represent time-varying geometries, we model the area function as quasi-static within short time windows. The geometry is assumed to be constant over windows of 40 ms, after which the h_i filters for the next frame, staggered by 10 ms, are computed. For each frame, we synthesize the output pressure in the frequency domain, and then map the signal to the time domain using a Hann window and overlap-add synthesis.

Unlike time domain synthesis, overlap-add synthesis can be easily parallelized on GPUs, as each window's computation is independent of the others. The backward pass is also efficient, as it is numerically stable and does not require adjoint optimization. At a sampling rate of 44.1 kHz, it takes only 51 ms to synthesize a 1 second sequence, 20x faster than real-time.

Table 1: **LibriTTS-R Reconstruction.** We reconstruct 100 random LibriTTS-R [Koizumi et al., 2023] audio samples using TensorTract2 [Krug et al., 2025], a surrogate model, and our model. Our model is able to better reconstruct the speech samples, leading to a lower word error rate.

	Ground Truth	**TensorTract 2 (Audio to Motor)**	**Ours**
English WER	0.027	0.119	**0.073**

Escaping Local Minima. One of the most fundamental issues with gradient-based optimization is local minima, and speech is no exception. We have found that if we fit a list of variables representing the discrete tube sections, gradient descent fails to fit even static geometries of vowels like /a/ and /e/. However, work in differentiable rendering has shown that overparameterizing the representation, with for example a neural network, can actually escape local minima and thus learn high-frequency signals [Tancik et al., 2020]. Inspired by this line of work, we choose to parameterize the continuous area function $A(x, t)$ with a *neural field*; see Xie et al. [2021] for a survey. By representing the area function as a continuous field, it significantly smooths the loss landscape compared to directly optimizing each tube segment individually.

Reconstruction and Animation. Because the acoustic simulator is fully differentiable, we can also couple it with any other differentiable representation of the vocal tract to reconstruct and animate speech. This method is illustrated in Figure 1. For example, given a soft-body model of the vocal tract controlled by 6 muscle groups (the jaw, cheeks, tongue, lips, velum, and glottis), we can differentiate speech through the soft-body simulation to the muscle actuation commands, thus animating the model. We implemented the model using the material point-method [Hu et al., 2019] to differentiably couple the muscle movements to the transmission line model.

The acoustic simulator can also be used to control data-driven models of the vocal tract, like a GAN trained on MRI volumes. Given a GAN $G : w \mapsto I$, which maps a latent vector w to an image I, we can differentiate through the GAN's latent space to find the MRI volumes which best match the utterance. This provides an interpretable view of how the lip posture, tongue geometry, etc. affect the area of the vocal tract, producing sound. Because this technique does not require paired data between MRI volumes and speech, it is fully self-supervised.

5 Evaluation

We evaluate how well we can fit real-world utterances with our differentiable simulator by reconstructing area functions of 100 random samples from the LibriTTS-R test set [Koizumi et al., 2023]. The reconstructions are then input into the `whisper-large-v3` model [Radford et al., 2022] to evaluate how intelligible the utterances are. While there is no prior work that pursues the exact same goal as ours, we can still compare to TensorTract 2 [Krug et al., 2025], a surrogate model which maps audio to articulatory motor controls. As evidenced by the lower word error rates (WER), we see that when compared to the baseline, our model is able to better reconstruct intelligible speech samples.

We also include many speech reconstruction examples from "in-the-wild" utterances in an anonymous GitHub site at `https://anonymous-vocal-reconstruction.github.io/vocal-samples.github.io/`. We encourage the reader to listen to some examples. Overall, the model succeeds at reconstructing speech in multiple languages, and even challenging singing examples. The area functions learned closely correspond to the human vocal tract geometry.

6 Conclusion

So can you hear the shape of a vocal tract? In this work, we argue yes. While it may never be literally possible to "hear the shape of a vocal tract", we know that humans can learn to speak from hearing alone. Almost automatically, we are able to map the intricacies of our hearing to the physical movements that make up speech. In our work, we have taken the first step in modeling this sensorimotor process as an optimization problem, and we hope that our computational framework can be used as a tool to study language learning in a data-driven way.

Acknowledgments

This project would not have been possible without early discussions with Matthew Caren and Kartik Chandra on their work on vocal imitation; and with Morris Alper on phonetics. We also thank Timothy Langlois and Mark Rau for sharing their expertise on acoustic simulation. EC is supported by the National Science Foundation Graduate Research Fellowship Program.

References

Takayuki Arai. Retroflex and bunched english /r/ with physical models of the human vocal tract. In *Interspeech*, 2014. URL https://api.semanticscholar.org/CorpusID:16264125.

Gašper Beguš, Alan Zhou, Peter Wu, and Gopala K Anumanchipalli. Articulation gan: Unsupervised modeling of articulatory learning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

Peter Birkholz. Modeling consonant-vowel coarticulation for articulatory speech synthesis. *PLOS ONE*, 8(4):1–17, 04 2013. doi: 10.1371/journal.pone.0060603. URL https://doi.org/10.1371/journal.pone.0060603.

Peter Birkholz and Dietmar Jackel. Influence of temporal discretization schemes on formant frequencies and bandwidths in time domain simulations of the vocal tract system. In *Interspeech*, 2004. URL https://api.semanticscholar.org/CorpusID:15404079.

Matthew Caren, Kartik Chandra, Joshua Tenenbaum, Jonathan Ragan-Kelley, and Karima Ma. Sketching with your voice: "non-phonorealistic" rendering of sounds via vocal imitation. In *SIGGRAPH Asia 2024 Conference Papers*, SA '24, New York, NY, USA, 2024. Association for Computing Machinery. URL https://doi.org/10.1145/3680528.3687679.

Tsutomu Chiba and Masato Kajiyama. *The vowel: its nature and structure*. Tokyo-Kaiseikan, 1941.

Cheol Jun Cho, Peter Wu, Tejas S. Prabhune, Dhruv Agarwal, and Gopala K. Anumanchipalli. Coding speech through vocal tract kinematics. *IEEE Journal of Selected Topics in Signal Processing*, 18 (8):1427–1440, 2024. doi: 10.1109/JSTSP.2024.3497655.

Gunnar Fant. *Acoustic theory of speech production, with calculations based on X-ray studies of Russian articulations.* Mouton and Co. N.V., The Hague, 1960.

Gunnar Fant, Johan Liljencrants, and Qi-guang Lin. A four-parameter model of glottal flow. *STL-QPSR*, 4(1985):1–13, 1985.

James L. Flanagan. *Speech Analysis Synthesis and Perception*. Communication and Cybernetics. Springer Berlin, Heidelberg, 2 edition, 1972. doi: 10.1007/978-3-662-01562-9.

Carolyn Gordon and David Webb. You can't hear the shape of a drum. *American Scientist*, 84(1): 46–55, 1996. ISSN 00030996. URL http://www.jstor.org/stable/29775597.

Laura Gwilliams, Ilina Bhaya-Grossman, Yizhen Zhang, Terri Scott, Sarah Harper, and Deborah Levy. Computational architecture of speech comprehension in the human brain. *Annual Review of Linguistics*, 2024. URL https://api.semanticscholar.org/CorpusID:273423180.

Yuanming Hu, Luke Anderson, Tzu-Mao Li, Qi Sun, Nathan A. Carr, Jonathan Ragan-Kelley, and Frédo Durand. Difftaichi: Differentiable programming for physical simulation. *ArXiv*, abs/1910.00935, 2019. URL https://api.semanticscholar.org/CorpusID:203626832.

Mark Kac. Can one hear the shape of a drum? *The American Mathematical Monthly*, 73(4):1–23, 1966. URL http://www.jstor.org/stable/2313748.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. URL https://api.semanticscholar.org/CorpusID:6628106.

Yuma Koizumi, Heiga Zen, Shigeki Karita, Yifan Ding, Kohei Yatabe, Nobuyuki Morioka, Michiel Bacchiani, Yu Zhang, Wei Han, and Ankur Bapna. Libritts-r: A restored multi-speaker text-to-speech corpus. *ArXiv*, abs/2305.18802, 2023. URL https://api.semanticscholar.org/CorpusID:258967444.

Paul Konstantin Krug, Peter Birkholz, Branislav Gerazov, Daniel Rudolph van Niekerk, Anqi Xu, and Yi Xu. Artificial vocal learning guided by phoneme recognition and visual information. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1734–1744, 2023. doi: 10.1109/TASLP.2023.3264454.

Paul Konstantin Krug, Christoph Wagner, Peter Birkholz, and Timo Stich. Precisely controllable neural speech synthesis. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10890772.

Patricia K. Kuhl and Andrew N. Meltzoff. Infant vocalizations in response to speech: vocal imitation and developmental change. *The Journal of the Acoustical Society of America*, 100 4 Pt 1:2425–38, 1996. URL https://api.semanticscholar.org/CorpusID:25406740.

Barbara Landau and Lila R. Gleitman. *Language and Experience: Evidence from the Blind Child*. Cognitive Science Series. Harvard University Press, 1985. ISBN 9780674039896.

Shinji Maeda. A digital simulation method of the vocal-tract system. *Speech Communication*, 1 (3):199–229, 1982. ISSN 0167-6393. doi: https://doi.org/10.1016/0167-6393(82)90017-6. URL https://www.sciencedirect.com/science/article/pii/0167639382900176.

Matthias Mauch and Simon Dixon. Pyin: A fundamental frequency estimator using probabilistic threshold distributions. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 659–663, 2014. URL https://api.semanticscholar.org/CorpusID:3287290.

Hong Nguyen, Sean Foley, Kevin Huang, Xuan Shi, Tiantian Feng, and Shrikanth S. Narayanan. Speech2rtmri: Speech-guided diffusion model for real-time mri video of the vocal tract during speech. *ArXiv*, abs/2409.15525, 2024. URL https://api.semanticscholar.org/CorpusID:272832441.

Sankaran Panchapagesan and Abeer Alwan. A study of acoustic-to-articulatory inversion of speech by analysis-by-synthesis using chain matrices and the maeda articulatory model. *The Journal of the Acoustical Society of America*, 129 4:2144–62, 2011. URL https://api.semanticscholar.org/CorpusID:18781420.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, 2022. URL https://api.semanticscholar.org/CorpusID:252923993.

Man Sondhi and J. Schroeter. A hybrid time-frequency domain articulatory speech synthesizer. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 35(7):955–967, 1987. doi: 10.1109/TASSP.1987.1165240.

Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *NeurIPS*, 2020.

Yiheng Xie, Towaki Takikawa, Shunsuke Saito, Or Litany, Shiqin Yan, Numair Khan, Federico Tombari, James Tompkin, Vincent Sitzmann, and Srinath Sridhar. Neural fields in visual computing and beyond. *Computer Graphics Forum*, 41, 2021. URL https://api.semanticscholar.org/CorpusID:244478496.

Anqi Xu, Daniel R. van Niekerk, Branislav Gerazov, Paul Konstantin Krug, Peter Birkholz, Santitham Prom-on, Lorna F. Halliday, and Yi Xu. Artificial vocal learning guided by speech recognition: What it may tell us about how children learn to speak. *Journal of Phonetics*, 105:101338, 2024. ISSN 0095-4470. doi: https://doi.org/10.1016/j.wocn.2024.101338. URL https://www.sciencedirect.com/science/article/pii/S0095447024000445.