Articulatory Synthesis of Speech and Diverse Vocal Sounds via Optimization

Luke Mo*, **Manuel Cherep***, **Nikhil Singh***, **Quinn Langford**, **Pattie Maes**
MIT
{lukemo,mcherep,nsingh1,langford,pattie}@mit.edu

vocaltrax.media.mit.edu

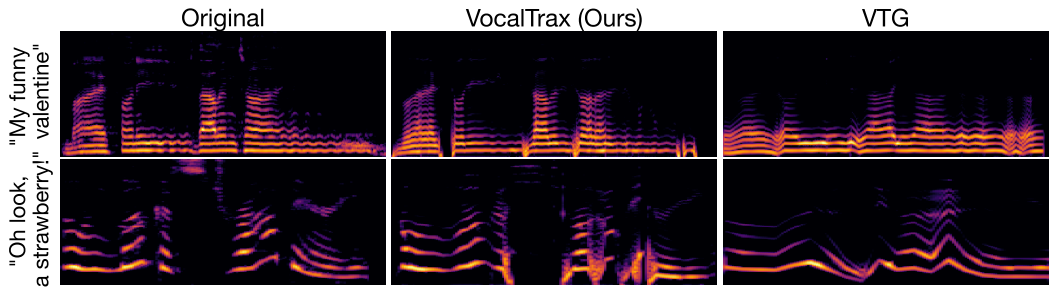

Figure 1: Spectrograms showing two target vocalizations with reconstruction via our approach (*VocalTrax*) and prior work [1]. (**Top**) a clip of Frank Sinatra singing *My Funny Valentine*. (**Bottom**) original speech audio from the popular "Oh Look, A Strawberry" meme.

Abstract

Articulatory synthesis seeks to replicate the human voice by modeling the physics of the vocal apparatus, offering interpretable and controllable speech production. However, such methods often require careful hand-tuning to invert acoustic signals to their articulatory parameters. We present *VocalTrax*, a method which performs this inversion automatically via optimizing an accelerated vocal tract model implementation. Experiments on diverse vocal datasets show significant improvements over existing methods in out-of-domain speech reconstruction, while also revealing persistent challenges in matching natural voice quality.

1 Introduction

The human voice presents a formidable challenge for computational modeling, with its complex physiology and acoustics. Articulatory speech synthesis [2], which aims to replicate this complexity by simulating the vocal tract's physical properties, has long been a prized goal of speech technology since it results in interpretable and controllable synthesis. While traditional approaches can be laboriously programmed to construct longer segments [3], they struggle to match the richness and variability of natural speech and even text-to-speech models [4]. As such, there is a need for acoustic-to-articulatory inversion methods that generalize to realistic speech, song, and other vocalizations.

This paper introduces *VocalTrax*, an optimization-based method for matching arbitrary vocal signals to articulatory parameters using the Pink Trombone[1] (PT) articulatory voice synthesizer. At the core of this is a fast, flexible implementation of PT which allows formulating the sound matching problem as an optimization task. Given an input voice clip, this approach iteratively optimizes the articulatory

*Equal contribution.
[1] https://dood.al/pinktrombone/

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

parameters of the synthesizer to match the acoustic qualities of a reference clip. This flexibility allows it to tackle a broad range of speech phenomena, from sustained vowels to non-linguistic vocalizations and even singing voices.

Overall, this work contributes:

1. An accelerated implementation of the Pink Trombone (PT) articulatory synthesizer in JAX, enabling efficient differentiation and optimization (planned for open-source release).
2. An approach to reconstructing arbitrary vocal signals, with pre-estimation of only the F_0s and otherwise end-to-end optimization, expanding the range of vocalizations that can be accurately synthesized.
3. Experiments on challenging out-of-domain (real voice) data showing that this approach is significantly more capable than existing gradient-based optimization approaches for vocal tract area function estimation, which are largely limited to synthesizing vowel sounds.

2 Related Work

We have been trying to implement machines and anatomical models to emulate human speech for centuries [5]. Computational methods for modeling the human vocal tract to synthesize speech are known as articulatory synthesis [2]. These methods consist of controlling aspects such as the tongue or lips to shape the vocal tract and generate speech by simulating the airflow. These simulations recover the parameters of the tract, which can be controlled and interpreted to assist with pronunciation [6], speech disorders [7], and speech recognition [8].

An important part of articulatory synthesis consists of modeling the human vocal tract [9–11]. One way of tackling this problem is gathering and training on the combination of speech and corresponding biometric data [12–15]. Another approach is what is known as analysis-by-synthesis, where parameters are iteratively refined to match the sound target [16, 17, 2]. This can be done using zero-order optimization techniques [18–25], but it usually requires more iterations to converge and it's harder to scale. Gradient methods are better at this—and previous research has leveraged neural networks to solve this task [26–33]—but require large training datasets.

Instead, differentiable vocal tract models can be used to get the best of both worlds: gradient optimization and no requirement for training data. In prior work [1], a differentiable mapping between control parameters and the PT synthesizer is optimized by gradient descent for sound matching vowel sounds. Our approach follows this concept, extending its capabilities beyond vowel sounds generated by PT; synthesizing realistic speech, song, and other vocalizations.

3 Methods

The Pink Trombone (PT) is a widely used articulatory speech synthesizer, composed of time-invariant models of the glottal flow derivative (GFD) and vocal tract V. The source GFD is filtered through V, synthesizing the output. To perform end-to-end sound matching, we split our audio input into frames, estimate the fundamental frequency F_0 for each frame, and optimize the vocal tract and GFD parameters for all frames simultaneously. We use a simple objective function involving computing the L_2 distance between the log-mel spectrograms of the target and synthesized audio.

3.1 Glottal Flow Derivative

Pink Trombone uses a simplified Liljencrants-Fant (LF) model of the GFD waveform [34]. The LF model is composed of two parameters, the fundamental frequency F_0 and tenseness T, representing the degree of vocal effort. White noise proportional to $1 - \sqrt{T}$ is added to the GFD waveform. We estimate the fundamental frequency (F_0) of each frame using CREPE [35]. The tenseness T for each frame is optimized alongside the vocal tract parameters.

3.2 Vocal Tract

The GFD waveform is filtered through the vocal tract, allowing for the articulation of consonant and vowel sounds. PT uses the Kelly-Lochbaum [36] piecewise cylindrical vocal tract model,

composed of a sequence of 44 segments of increasing distance from the glottis with cross-sectional areas $A_1, A_2, \ldots, A_{44}$. At each segment junction, the forward and reversed waves are reflected and propagated as described by the scattering coefficients:

$$k_i = \frac{A_i - A_{i-1}}{A_i + A_{i-1}} \; \forall_i \in \{2, \ldots, 44\} \tag{1}$$

To aim for physiologically plausible vocal tracts, we used a simplified physical vocal tract model to determine the diameters $d_1, d_2, \ldots, d_{44}$ shared across all frames. At each frame, two types of transformations are applied: the tongue and two constrictions [1]. The tongue, defined by two parameters, tongue diameter (t_d) and tongue position (t_p), modifies the base diameter into a sinusoidal shape, mimicking the behavior of the human tongue. One lip and one tract constriction, defined by parameters c_l and c_t scale the base diameters of the subset of diameters furthest from and closest to the glottis, respectively, by a factor of $1 - c_l$ and $1 - c_t$. To simplify the gradient-based optimization approach, we keep the constriction indices set at 12 and 39.

3.3 Optimization

We use a common mel spectrogram representation of the audio signals, and define our objective $\mathcal{L}$ as the L_2 distance between the target (T) and synthesized (S) audio:

$$\mathcal{L}(\text{T}, \text{S}) = \| \log(|\text{MELSPEC(T)}|) - \log(|\text{MELSPEC(S)}|) \|_2 \tag{2}$$

We minimize $\mathcal{L}$ over our parameter space using the AdamW [37] optimizer (with $\gamma = 0.01$), and use a box projection to keep the parameters $\in [0, 1]$. We use a normalized parameter space, back-transformed to each parameter's respective range as needed as has been done in other synthesis packages [38]. We initialize the diameters using the canonical values [1], and other parameters to 0.5 (middle) except T (tenseness coefficients) to 1, to minimize unnecessary noise at the beginning of the optimization. Unlike prior work [1], we do not use inverse filtering to recover any coefficients, and instead perform end-to-end optimization of the full apparatus (except for pre-estimated F_0s).

4 Results

We evaluated *VocalTrax* against Vocal-Tract-Grad [1] and ground truth using automated metrics and human evaluations on multiple datasets. Table 1 shows results of automated evaluations on three datasets: TIMIT [39] (subset), AudioMNIST [40] (subset), and VIVAE [41]. We used match error rate (MER) for TIMIT[2] and accuracy otherwise. Given the distribution shift between target audio and even relatively high quality reconstructions, we complement the automated evaluation with a human evaluation. Table 2 shows human accuracy responses. Importantly, all our evaluations are on *out-of-domain* data (i.e. data not synthesized with vocal tract models which can be perfectly reconstructed given the same tract model, but rather recordings of real speech).

4.1 Automated Evaluations

	TIMIT [39] (MER ↓)	AudioMNIST [40] (Acc ↑)	VIVAE [41] (Acc ↑)
Ground Truth	6.4	73.5	29.2
VocalTrax (Ours)	82.9	20.0	18.4
VTG [1] @ 1024	99.5	9.7	17.7
VTG [1] @ 2048	99.4	12.2	17.3
VTG [1] @ 4096	99.6	10.7	15.6

Table 1: Results from automated evaluations. TIMIT uses the match error rate.

In our automated evaluations, we use three datasets to capture different capabilities. AudioMNIST [40] focuses on spoken numbers, which are simple and brief excerpts but do contain semantic information. Given the scale of this dataset, we use a stratified random sample of 600 test-set

[2]We use MER because word error rate is sensitive to insertions, and thus brief uninformative responses like "thanks for watching" (a common Whisper hallucination given incoherent inputs) result in inflated performance.

examples to evaluate across the different methods. By contrast, VIVAE [41] focuses on paralinguistic vocalizations, which do not contain any words but convey affective information. For both datasets, we evaluate using the ARCH [42] benchmark protocol, modified to train on *real* data, and test on resynthesized (or ground truth) data. This ensures a realistic evaluation, wherein models are not trained specifically on the resynthesized speech and can adapt their representations accordingly. Both datasets are for multi-class classification (10 for AudioMNIST and 6 for VIVAE respectively). For AudioMNIST, given the scale (30,000 clips), we do a train-test (instead of cross-validated) evaluation. Finally, we evaluate on a more challenging task: longer-range, higher-vocabulary speech synthesis. We sub-sample 100 clips from TIMIT [39], which contain multi-word phrases or sentences, and aim to resynthesize these fully. We use 2000 optimization iterations for TIMIT to account for its complexity, vs. 1000 for others.

Results are shown in Table 1 for ground truth test set data, our method, and Vocal-Tract-Grad [1]. For the latter, we evaluate it at multiple matched hop and frame lengths: 1024 (ours), 2048, and 4096 (their original). Since Vocal-Tract-Grad focuses on vowel synthesis, AudioMNIST and especially TIMIT are likely to be quite challenging for it. Overall, we observe that our method is able to deliver improved reconstructions, judged by their classification and transcription performance, over these baselines. However, for TIMIT and AudioMNIST, our results remain distant from the ground truth results due to the significant distribution shift in addition to reconstruction artifacts present.

4.2 Human Evaluations

	AudioMNIST [40]		VIVAE [41]	
	Acc ↑	Conf ↑	Acc ↑	Conf ↑
Ground Truth	100.0 (0.0)	4.9 (0.0)	47.8 (3.7)	3.7 (0.1)
VocalTrax (Ours)	48.7 (2.9)	2.9 (0.1)	23.9 (3.2)	2.5 (0.1)
VTG [1] @ 1024	11.0 (1.8)	1.5 (0.1)	14.4 (2.6)	1.7 (0.1)

Table 2: Results from human evaluations (N=10 participants, each rating 30 AudioMNIST [40] and 18 VIVAE [41] samples per source). We show both response accuracy and confidence, each with standard errors (in parenthesis), computed directly from the sample.

To complement automated evaluations, we ran a listening study (results are shown in Table 2). We used subsets of AudioMNIST and VIVAE in this study, focusing on (1) how accurately listeners could identify the category the reconstruction (or original example) belongs to, and (2) how confident listeners were about their choices. We recruited 10 participants via Prolific, and estimated that the study took about 20 minutes to complete. The study was determined by our IRB to be exempt. Participants listened and responded to 90 total AudioMNIST clips (stratified random sample of 3 clips per digit category, and the same 30 for each of ground truth, ours, and Vocal-Tract-Grad [1]) and 54 total VIVAE clips (similarly, 3 per affect category, and the same 18 across the 3 sources).

We modeled accuracy using a mixed-effects logistic regression for each dataset, with random intercepts for digit (AudioMNIST) or category (VIVAE) and for participants. Then, we conducted pairwise post-hoc contrasts. For AudioMNIST, participants were significantly more accurate identifying *VocalTrax*-synthesized digits compared to Vocal-Tract-Grad (odds ratio = 10.7, p<.0001). This was also true for VIVAE, though with a more modest difference (odds ratio = 1.96, p=.019). These p-values were adjusted using the Benjamini-Hochberg correction for pairwise tests. For both datasets, participants were also more confident in classifying our reconstructions. Participants were less accurate and confident with our reconstructions compared to the ground truth clips, suggesting significant opportunities to further improve reconstructions of challenging, out-of-domain samples.

5 Conclusion

VocalTrax demonstrates how end-to-end optimization can improve articulatory speech reconstruction of acoustic signals. Our JAX implementation of Pink Trombone and reconstruction approach can rapidly reconstruct a variety of vocal signals, which we hope will open up possibilities in speech analysis, therapy, and voice conversion. However, the quality gap between such synthetic and natural speech persists. Future work should focus on refining vocal tract models, incorporating perceptual factors, and expanding to more complex vocal phenomena.

Acknowledgements

The project that gave rise to these results received the support of a fellowship from "la Caixa" Foundation (ID 100010434). The fellowship code is LCF/BQ/EU23/12010079.

References

[1] David Südholt, Mateo Cámara, Zhiyuan Xu, and Joshua D Reiss. Vocal tract area estimation by gradient descent. *arXiv preprint arXiv:2307.04702*, 2023.

[2] Korin Richmond. Estimating articulatory parameters from the acoustic speech signal. *Annexe Thesis Digitisation Project 2017 Block 11*, 2002.

[3] Andrey Anikin. Soundgen: An open-source tool for synthesizing nonverbal vocalizations. *Behavior research methods*, 51:778–792, 2019.

[4] Paul Konstantin Krug, Simon Stone, and Peter Birkholz. Intelligibility and naturalness of articulatory synthesis with vocaltractlab compared to established speech synthesis technologies. *Proc. SSW*, 11:102–107, 2021.

[5] Brad H Story. History of speech synthesis. In *The Routledge Handbook of Phonetics*, pages 9–33. Routledge, 2019.

[6] Li Liu, Gang Feng, and Denis Beautemps. Inner lips feature extraction based on clnf with hybrid dynamic template for cued speech. *EURASIP Journal on Image and Video Processing*, 2017:1–15, 2017.

[7] Junhong Zhao, Hua Yuan, Wai-Kim Leung, Helen Meng, Jia Liu, and Shanhong Xia. Audiovisual synthesis of exaggerated speech for corrective feedback in computer-assisted pronunciation training. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8218–8222. IEEE, 2013.

[8] Li Liu, Gang Feng, Denis Beautemps, and Xiao-Ping Zhang. Re-synchronization using the hand preceding model for multi-modal fusion in automatic continuous cued speech recognition. *IEEE Transactions on Multimedia*, 23:292–305, 2020.

[9] Gunnar Fant. The lf-model revisited. transformations and frequency domain analysis. *Speech Trans. Lab. Q. Rep., Royal Inst. of Tech. Stockholm*, 2(3):40, 1995.

[10] Khalil Iskarous, Louis Goldstein, Douglas H Whalen, Mark Tiede, and Philip Rubin. Casy: The haskins configurable articulatory synthesizer. In *International Congress of Phonetic Sciences, Barcelona, Spain*, pages 185–188, 2003.

[11] Peter Birkholz. Modeling consonant-vowel coarticulation for articulatory speech synthesis. *PloS one*, 8(4):e60603, 2013.

[12] Jianwu Dang and Kiyoshi Honda. Estimation of vocal tract shapes from speech sounds with a physiological articulatory model. *Journal of Phonetics*, 30(3):511–532, 2002.

[13] Peng Liu, Quanjie Yu, Zhiyong Wu, Shiyin Kang, Helen Meng, and Lianhong Cai. A deep recurrent approach for acoustic-to-articulatory inversion. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4450–4454. IEEE, 2015.

[14] Cheol Jun Cho, Peter Wu, Abdelrahman Mohamed, and Gopala K Anumanchipalli. Evidence of vocal tract articulation in self-supervised learning of speech. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[15] Peter Wu, Shinji Watanabe, Louis Goldstein, Alan W Black, and Gopala K Anumanchipalli. Deep speech synthesis from articulatory representations. *arXiv preprint arXiv:2209.06337*, 2022.

[16] Bishnu S Atal, Jih Jie Chang, Max V Mathews, and John W Tukey. Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique. *The Journal of the Acoustical Society of America*, 63(5):1535–1555, 1978.

[17] Victor N Sorokin, Alexander S Leonov, and Alexander V Trushkin. Estimation of stability and accuracy of inverse problem solution for the vocal tract. *Speech Communication*, 30(1):55–74, 2000.

[18] Janne Riionheimo and Vesa Välimäki. Parameter estimation of a plucked string synthesis model using a genetic algorithm with perceptual fitness calculation. *EURASIP Journal on Advances in Signal Processing*, 2003:1–15, 2003.

[19] Crispin Cooper, Damian Murphy, David Howard, and Alexander Tyrrell. Singing synthesis with an evolved physical model. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4):1454–1461, 2006.

[20] Olaf Schleusing, Tomi Kinnunen, Brad Story, and Jean-Marc Vesin. Joint source-filter optimization for accurate vocal tract estimation using differential evolution. *IEEE transactions on audio, speech, and language processing*, 21(8):1560–1572, 2013.

[21] Yingming Gao, Simon Stone, and Peter Birkholz. Articulatory copy synthesis based on a genetic algorithm. In *INTERSPEECH*, pages 3770–3774, 2019.

[22] Naotake Masuda and Daisuke Saito. Quality-diversity for synthesizer sound matching. *Journal of Information Processing*, 31:220–228, 2023.

[23] Mahmoud A Ismail. Vocal tract area function estimation using particle swarm. *J. Comput.*, 3(6):32–38, 2008.

[24] Manuel Cherep, Nikhil Singh, and Jessica Shand. Creative text-to-audio generation via synthesizer programming. *arXiv preprint arXiv:2406.00294*, 2024.

[25] Mateo Cámara, Zhiyuan Xu, Yisu Zong, José Luis Blanco, and Joshua D Reiss. Optimization techniques for a physical model of human vocalisation. *arXiv preprint arXiv:2309.14761*, 2023.

[26] Leonardo Gabrielli, Stefano Tomassetti, Stefano Squartini, Carlo Zinato, et al. Introducing deep machine learning for parameter estimation in physical modelling. In *Proceedings of the 20th international conference on digital audio effects*, 2017.

[27] Matthew John Yee-King, Leon Fedden, and Mark d'Inverno. Automatic programming of vst sound synthesizers using deep networks and other techniques. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2(2):150–159, 2018.

[28] Pramit Saha and Sidney Fels. Learning joint articulatory-acoustic representations with normalizing flows. *arXiv preprint arXiv:2005.09463*, 2020.

[29] Hayato Shibata, Mingxin Zhang, and Takahiro Shinozaki. Unsupervised acoustic-to-articulatory inversion neural network learning based on deterministic policy gradient. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 530–537. IEEE, 2021.

[30] Marco A Martínez Ramírez, Oliver Wang, Paris Smaragdis, and Nicholas J Bryan. Differentiable signal processing with black-box audio effects. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 66–70. IEEE, 2021.

[31] Peter Wu, Li-Wei Chen, Cheol Jun Cho, Shinji Watanabe, Louis Goldstein, Alan W Black, and Gopala K Anumanchipalli. Speaker-independent acoustic-to-articulatory speech inversion. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[32] Jianrong Wang, Jinyu Liu, Longxuan Zhao, Shanyu Wang, Ruiguo Yu, and Li Liu. Acoustic-to-articulatory inversion based on speech decomposition and auxiliary feature. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4808–4812. IEEE, 2022.

[33] Mateo Cámara, Fernando Marcos, and José Luis Blanco. Decoding vocal articulations from acoustic latent representations. *arXiv preprint arXiv:2406.14379*, 2024.

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

[35] Jong Wook Kim, Justin Salamon, Peter Li, and Juan Pablo Bello. Crepe: A convolutional representation for pitch estimation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 161–165. IEEE, 2018.

[36] John L. Kelly and Carol C. Lochbaum. Speech synthesis. *Proceedings of the Fourth International Congress on Acoustics*, 1962.

[37] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[38] Manuel Cherep and Nikhil Singh. Synthax: A fast modular synthesizer in jax. In *Audio Engineering Society Convention 155*. Audio Engineering Society, 2023.

[39] John S Garofolo. Timit acoustic phonetic continuous speech corpus. *Linguistic Data Consortium, 1993*, 1993.

[40] Sören Becker, Johanna Vielhaben, Marcel Ackermann, Klaus-Robert Müller, Sebastian Lapuschkin, and Wojciech Samek. Audiomnist: Exploring explainable artificial intelligence for audio analysis on a simple benchmark. *Journal of the Franklin Institute*, 2023.

[41] Natalie Holz, Pauline Larrouy-Maestri, and David Poeppel. The variably intense vocalizations of affect and emotion (vivae) corpus prompts new perspective on nonspeech perception. *Emotion*, 22(1):213, 2022.

[42] Moreno La Quatra, Alkis Koudounas, Lorenzo Vaiani, Elena Baralis, Paolo Garza, Luca Cagliero, and Sabato Marco Siniscalchi. Benchmarking representations for speech, music, and acoustic events. In *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, 2024.