ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3315-6701-9/26/$31.00 ©2026 IEEE | DOI: 10.1109/ICASSP55912.2026.11460533

LEARNING VOCAL-TRACT AREA AND RADIATION WITH A PHYSICS-INFORMED WEBSTER MODEL

*Minhui Lu**

Queen Mary University of London
Centre for Digital Music
minhui.lu@qmul.ac.uk

Joshua D. Reiss

Queen Mary University of London
Centre for Digital Music
joshua.reiss@qmul.ac.uk

ABSTRACT

We present a physics-informed voiced backend renderer for singing-voice synthesis. Given synthetic single-channel audio and a fundamental–frequency trajectory, we train a time-domain Webster model as a physics-informed neural network to estimate an interpretable vocal-tract area function and an open-end radiation coefficient. Training enforces partial differential equation and boundary consistency; a lightweight DDSP path is used only to stabilize learning, while inference is purely physics-based. On sustained vowels (/a/, /i/, /u/), parameters rendered by an independent finite-difference time-domain Webster solver reproduce spectral envelopes competitively with a compact DDSP baseline and remain stable under changes in discretization, moderate source variations, and about ten percent pitch shifts. The in-graph waveform remains breathier than the reference, motivating periodicity-aware objectives and explicit glottal priors in future work.

***Index Terms*—** Vocal-tract acoustics, PINNs, Webster equation, differentiable DSP, singing-voice synthesis

1. INTRODUCTION

Modern singing-voice synthesis (SVS) is often organised as a two-stage pipeline: a front end predicts control trajectories (e.g., f_0, phonetic content, loudness), and a back end renders audio. The back end is commonly a neural vocoder or an end-to-end generator [1–4]. While such renderers can be highly natural, their high-capacity parameters tend to entangle pitch, timbre, and articulation, limiting fine-grained control and diagnosis [5] and often requiring large corpora and retraining for new singers or styles.

A complementary route is to exploit classical vocal acoustics as an explicit control surface. For voiced sounds, a 1D time-domain Webster model with an open-end radiation boundary offers an interpretable *source–tract–radiation* decomposition, where tract geometry and boundary parameters directly shape formants and spectral tilt [6,7]. However, practical use hinges on calibrating the vocal-tract area function and boundary conditions from audio, which remains challenging and solver-dependent [8, 9].

These considerations motivate a middle ground: retain mechanistic structure while learning difficult parameters from audio. Two complementary paradigms support this: (i) differentiable rendering modules optimised with audio losses, as in Differentiable Digital Signal Processing (DDSP) [10, 11], and (ii) physics-informed neural networks (PINNs) that enforce governing equations and boundary conditions via residual penalties [12]. Related "physics-informed DDSP" hybrids have begun to embed physical operators or constraints within differentiable audio pipelines [13]; however, such systems typically retain fixed boundary/termination modelling and do not target solver-independent validation of recovered physical controls. For voice, physics-informed synthesis has been demonstrated in restricted steady regimes (e.g., one-period solutions with fixed radiation circuits) [14]. Yet, a practical time-domain vocal-tract backend still lacks (1) *joint* recovery of tract geometry with an explicit learnable radiation boundary, and (2) solver-independent validation that recovered parameters act as transferable physical controls rather than discretisation artefacts.

We train a time-domain Webster PINN to estimate a continuous tract area $A(x)$ and a Robin (open-end) radiation coefficient ζ from single-channel sustained vowels, given $f_0(t)$. Training combines Partial Differential Equations and Boundary Conditions (PDE/BC) residuals with audio/probe losses; a lightweight DDSP path is used only as a stabiliser and is removed at inference. To reduce inverse-crime risk, we evaluate by exporting $(\hat{A}, \hat{\zeta})$ to an *independent* explicit finite-difference time-domain (FDTD) discretisation of the Webster equation and computing objective envelope and periodicity metrics on the resulting post-rendered waveform [15].

This paper makes three contributions:

1. **Webster PINN with learnable radiation:** joint estimation of $A(x)$ and a Robin radiation coefficient ζ from single-channel sustained voiced audio in a controlled synthetic setting (given $f_0(t)$).
2. **Training-only differentiable supervision, physics-only inference:** audio/probe losses and an optional auxiliary DDSP stabiliser are used during optimisation, while inference remains purely physics-based.
3. **Out-of-graph evaluation:** solver-independent post-rendering with an independent FDTD–Webster implementation to test transfer under discretisation and source mismatches.

Table 1 positions this work within SVS renderers and related parameter-estimation lines, highlighting *learned radiation* and *solver-independent post-render evaluation*.

2. PHYSICS-INFORMED VOICED RENDERER

Figure 1 gives an overview of the proposed physics-informed voiced renderer. Given spatio-temporal coordinates (x, t) (with $x \in [0, L]$) and a fundamental-frequency trajectory $f_0(t)$—and, during training only, a reference waveform $y(t)$—DualNet predicts the acoustic velocity potential $\psi(x, t)$, a static vocal-tract area function $\hat{A}(x)$, and a learnable open-end radiation coefficient $\hat{\zeta}$. A differentiable Webster rendering path maps lip pressure to a waveform $\hat{y}(t)$, enabling audio/probe losses during training alongside PDE/BC residuals. At inference, the method is physics-only: it renders $\hat{y}(t)$ from

*This work was co-funded by the UKRI Centre for Doctoral Training in Artificial Intelligence and Music and Queen Mary University of London.

Table 1. Taxonomy of prior lines vs. this work (positioning, not a performance comparison).

Approach	Mechanism	Rad.	Temporal	Interp.	Render
Neural vocoders (GAN/Flow/Diffusion) [3, 16, 17]	none (implicit)	implicit	time-domain	low	in-graph
DDSP-based SVS (e.g., GOLF) [10, 11]	DSP prior	fixed	time-domain	medium	in-graph
Physics-informed DDSP hybrids (e.g., piano) [13]	DSP + constraints	fixed	time-domain	medium	in-graph
Inverse filtering / classical AAI [18–20]	analytic (filter)	fixed/implicit	steady-state	high (A)	inversion-only
Grad.-based analysis-by-synthesis [21]	numeric solver	fixed	steady-state	high (A)	in-loop (same solver)
PINN voice (1-period, fixed radiation) [14]	PINN (PDE/BC)	fixed	1-period	medium	in-graph
Current work	**PINN (Webster)**	**learned (Robin ζ)**	**time-domain**	**high** $(A(x), \zeta)$	**post-render (indep. impl.)**

Notes: "Mechanism" indicates how structure is imposed (implicit, DSP prior, analytic, numeric solver, or PINN with PDE/BC residuals). Interpretability is qualitative and refers to physically meaningful controls ($A(x)$, radiation, source/filters). "Render" denotes whether audio/spectra are produced in-graph, within an optimisation loop using the same solver, or via solver-independent post-rendering (ours). AAI: acoustic-to-articulatory inversion.

the predicted $(\psi, \hat{A}, \hat{\zeta})$ without any reference-based losses or auxiliary DDSP path. For solver-independent evaluation, $(\hat{A}, \hat{\zeta})$ are exported to an *independent* FDTD–Webster implementation for post-render metric computation (Sec. 3).

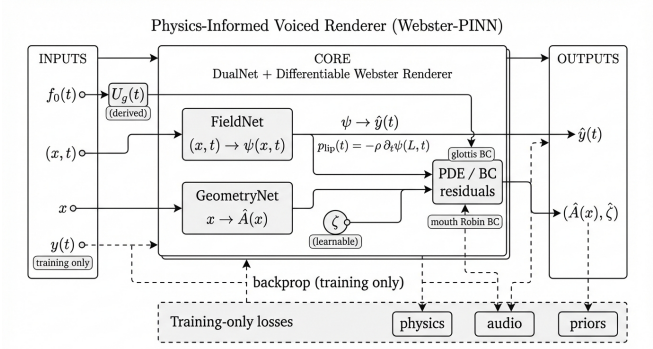

Fig. 1. Overview of the physics-informed voiced renderer. DualNet predicts $(\psi, \hat{A}, \hat{\zeta})$ and a differentiable Webster rendering path produces $\hat{y}(t)$ for reference-based losses during training (inference is physics-only). Solid arrows denote forward signal flow in the renderer; dashed arrows denote training-only loss/backprop connections (e.g., using $y(t)$), which are removed at inference. For solver-independent evaluation (not shown), $(\hat{A}, \hat{\zeta})$ are exported to an independent FDTD–Webster solver for post-render assessment.

DualNet has two heads: a compact SIREN field network maps (x, t) to the velocity potential $\psi(x, t)$ [22], and a geometry MLP maps x to a positive area function $\hat{A}(x)$ via a softplus output. A global scalar $\hat{\zeta}$ parameterises the mouth Robin boundary; spatial/temporal derivatives used in PDE/BC residuals are obtained via automatic differentiation.

2.1. Governing equations and boundary conditions

The model uses the velocity potential $\psi(x, t)$ and a static tract area $A(x)>0$ along $x \in [0, L]$ (with tract length L), predicted by DualNet as $\hat{A}(x)$. Pressure, particle velocity, and volume velocity are

$$\begin{aligned} p(x,t) &= -\rho\,\partial_t\psi(x,t), \\ v(x,t) &= \partial_x\psi(x,t), \\ U(x,t) &= A(x)\,\partial_x\psi(x,t), \end{aligned} \tag{1}$$

with air density ρ and sound speed c. The time-domain Webster equation [6] is

$$\frac{1}{c^2}\partial_{tt}\psi \;-\; \frac{1}{A(x)}\,\partial_x\big(A(x)\,\partial_x\psi\big) \;=\; 0. \tag{2}$$

At the mouth ($x{=}L$), a radiation (Robin) boundary is used:

$$\partial_x\psi(L,t) + \zeta\,\partial_t\psi(L,t) = 0, \tag{3}$$

consistent with low-frequency open-end radiation models for Webster tubes (low-ka) [7]. At the glottis ($x{=}0$), a volume-flow boundary is imposed:

$$U(0,t) = \alpha\,U_g(t), \tag{4}$$

where $U_g(t)$ is a periodic glottal flow derived from $f_0(t)$ when used; otherwise a weak envelope prior can be applied. Here α is a scalar amplitude calibration (implementation uses $\alpha = c \cdot u_{\text{scale}}$) to match the excitation scale used by the renderer. A deliberately minimal parameterisation—a static $A(x)$ and a single radiation scalar ζ—is adopted to preserve identifiability from single-channel steady voiced audio in the Webster regime [23].

2.2. Physics losses

Let $X_d = \{(x_i, t_i)\}_{i=1}^{N}$ denote interior collocation points. The physics objective includes PDE and boundary residuals and geometric regularisation (with $\mathbb{E}$ denoting empirical averages over the indicated sets):

$$\mathcal{L}_{\text{PDE}} = \mathbb{E}_{(x,t)\in X_d}\Big[\big(\tfrac{1}{c^2}\partial_{tt}\psi - \tfrac{1}{A}\partial_x(A\,\partial_x\psi)\big)^2\Big], \tag{5}$$

$$\mathcal{L}_{\text{BC,mouth}} = \mathbb{E}_t\big[(\partial_x\psi + \zeta\partial_t\psi)^2\big]_{x=L}, \tag{6}$$

$$\mathcal{L}_{A''} = \mathbb{E}_x\left[(\partial_{xx}A(x))^2\right], \qquad \mathcal{L}_{\text{geom}} = \mathbb{E}_x\left[\phi\big(A(x)\big)\right], \tag{7}$$

where ϕ softly penalises A outside $[A_{\min}, A_{\max}]$ and anchors $A(0) \approx A(L) \approx 1$. If U_g is used, an additional glottal boundary loss is applied:

$$\mathcal{L}_{\text{BC,glot}} = \mathbb{E}_t\Big[\big(\partial_x\psi(0,t) - \alpha\,U_g(t)/A(0)\big)^2\Big]. \tag{8}$$

For the grouped objective in Sec. 2.4, we define $\mathcal{L}_{\text{PDE/BC}} = \mathcal{L}_{\text{PDE}} + \mathcal{L}_{\text{BC,mouth}} + \mathcal{L}_{\text{BC,glot}}$ (with $\mathcal{L}_{\text{BC,glot}}$ omitted when U_g is not used), and include geometric regularisers in $\mathcal{L}_{\text{prior}}$.

2.3. Differentiable audio and probes

Lip pressure is computed as $p_{\text{lip}}(t) = -\rho\,\partial_t\psi(L,\, t+\tau)$ and rendered as $\hat{y}(t) = p_{\text{gain}}\, p_{\text{lip}}(t)$ with learnable gain p_{gain} and time shift τ. For windowed comparison with $y(t)$, the audio objective combines a multi-resolution STFT loss [24] and a log-mel envelope loss (RMS-normalised), optionally augmented with a weak full-utterance STFT and a small time-domain term. We also compute lightweight differentiable probes (formants $F_{1..3}$ and a harmonic spectral envelope H_{env}) and use them as low-weight auxiliary guidance/diagnostics during training. Quantitative evaluation relies on objective envelope/periodicity metrics and solver-independent post-rendering.

2.4. Auxiliary DDSP renderer (training only)

To stabilise mid-stage optimisation, the probe-derived H_{env} can be mapped to an auxiliary DDSP-style additive synthesiser [10] to produce a teacher waveform for envelope regularisation. This path is enabled only during training and removed at inference.

2.5. Overall objective

The total loss is written in grouped form as

$$\mathcal{L} = \lambda_{\text{phys}}\,\mathcal{L}_{\text{PDE/BC}} + \lambda_{\text{aud}}\,\mathcal{L}_{\text{audio}} + \lambda_{\text{probe}}\,\mathcal{L}_{\text{probe}} + \lambda_{\text{prior}}\,\mathcal{L}_{\text{prior}}. \quad (9)$$

Weight schedules and normalisation details are specified in the released code. During optimisation we use a simple staged weighting schedule (warm-up and ramps) to stabilise training.

3. TRAINING AND EVALUATION PROTOCOL

Data and models. One model is trained per sustained vowel (/a/, /i/, /u/) at 16 kHz. Reference waveforms $y(t)$ are synthesised by a standalone explicit FDTD discretisation of the Webster PDE in Eq. (2), driven via a Rosenberg glottal-flow boundary and terminated by a Robin lip-radiation boundary of the form in Eq. (3) (with $\zeta_{\text{ref}}{=}0.06$ in the reference solver) [9, 25]. Each utterance lasts $\approx$ 0.8 s with nominal pitch anchors {/a/: 200 Hz, /i/: 240 Hz, /u/: 180 Hz}.
Baseline. To contextualise spectral-envelope fit, a compact DDSP-only harmonic additive synthesiser driven by $f_0(t)$ and loudness (RMS) is used as a non-physics baseline.
Metrics. We report multi-resolution STFT error (mSTFT; lower is better), log-spectral distance (LSD, dB), formant MAE (Hz), and harmonic-to-noise ratio (HNR, dB). Signals are aligned by cross-correlation prior to windowed metrics.
Out-of-graph post-render evaluation (reduced inverse-crime risk). Training uses a differentiable in-graph Webster PINN (Sec. 2), whereas evaluation uses a *separate* explicit FDTD–Webster implementation. To test whether recovered parameters act as transferable physical controls rather than discretisation artefacts, we export $(\hat{A}(x), \hat{\zeta})$ and $f_0(t)$ to the independent solver and compute metrics on the post-rendered waveform. No gradients or discretisation operators are shared between the training graph and the post-render code, so post-render metrics reflect solver-transfer rather than in-graph fitting.
Robustness tests. We assess sensitivity by post-rendering under controlled mismatches: (i) discretisation (grid size / CFL); (ii) source and propagation factors (e.g., damping, Rosenberg shape, aspiration level); (iii) pitch shifts ($\pm 10\%$) applied to $f_0(t)$; and (iv) small perturbations of ζ. Code and audio examples are available at the project page.[1]

[1] https://minhuilu.github.io/webster-pinn-svs/

	PINN (post-render)		DDSP-only		PINN (in-graph)	
Vowel	mSTFT ↓	LSD ↓	mSTFT ↓	LSD ↓	mSTFT ↓	LSD ↓
/a/	1.292	6.704	2.749	15.881	6.046	24.711
/i/	3.295	15.634	2.097	13.219	6.363	27.437
/u/	1.846	9.186	2.988	15.452	6.413	27.382

Table 2. Envelope fit to the reference (lower is better). Columns: post-render (PINN→independent FDTD), DDSP-only, and in-graph PINN. One canonical reference per vowel.

This paper targets physics-informed parameter recovery and solver-transfer validation rather than full-system SVS. Accordingly, a lightweight DDSP-only baseline is used to contextualise envelope fit, while broader benchmarking against complete SVS pipelines and black-box vocoders (e.g., WORLD [26], NSF [27], WaveNet/MelGAN/HiFi-GAN [1–3]) is deferred to future work under a unified evaluation harness.

4. RESULTS

Our evaluation follows the workflow in Fig. 1 and the protocol in Sec. 3: we test whether the recovered controls $(\hat{A}, \hat{\zeta})$ (i) transfer to an out-of-graph forward implementation (post-render validation), (ii) expose systematic failure modes of the in-graph differentiable renderer (the "periodicity gap"), and (iii) behave as stable yet potentially non-identifiable parameters under steady voiced supervision. We do not benchmark full SVS pipelines; instead we evaluate parameter transfer and envelope/periodicity fidelity under controlled mismatches.

4.1. Post-render validation: recovered controls transfer beyond the training graph

Exporting $(\hat{A}, \hat{\zeta})$ to an independent FDTD–Webster solver preserves the target spectral envelope, indicating that the recovered parameters are not tied to the particular training graph representation. Table 2 compares post-rendered audio from the out-of-graph solver, a DDSP-only baseline, and the in-graph PINN renderer (all envelope metrics are computed *against the same reference*). On /a/ and /u/, post-render reduces LSD by roughly 6–9 dB relative to DDSP-only, while also substantially improving over in-graph rendering. For /i/, post-render trails DDSP-only but still improves markedly over the in-graph output, suggesting that parameter transfer holds even when the vowel is challenging for envelope fitting.

4.2. The periodicity gap: in-graph rendering is systematically more aperiodic

Despite successful envelope transfer under post-render evaluation, voiced periodicity reveals a systematic discrepancy between in-graph and out-of-graph rendering. Table 3 reports median frame-wise HNR: post-render nearly matches the reference periodicity on /a/ and /u/ and remains within $\sim$1.4 dB on /i/, indicating quasi-periodic voicing when driven by the reference f_0. DDSP-only is also close to the reference but slightly lower on /i/. In contrast, the in-graph PINN output exhibits substantially reduced HNR across vowels (about 2–4 dB), consistent with perceptual breathiness; we refer to this as the *periodicity gap*. This points to a structural under-constraint: PDE/BC residuals plus short-time *envelope*-focused losses can admit solutions that match smooth spectral

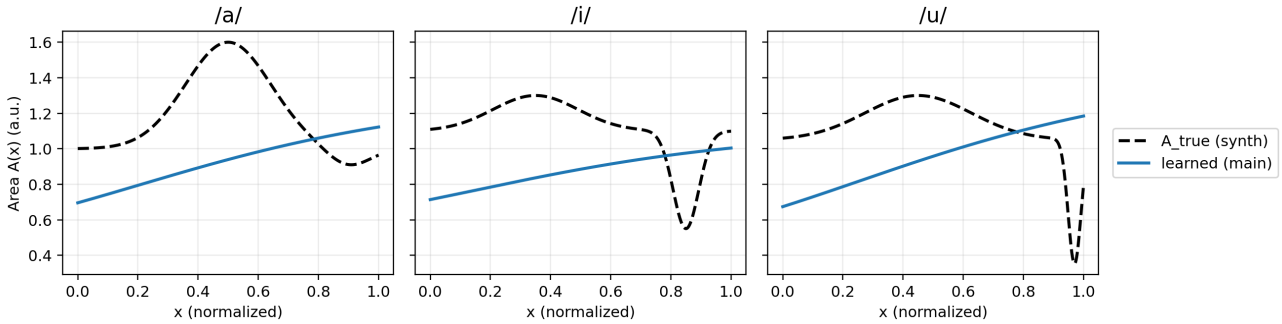

Fig. 2. Recovered area functions $\hat{A}(x)$ (normalised units). Here x increases from the glottis (0) to the lips (1). The solutions capture broad vowel-dependent trends (e.g., anterior constriction for /i/ and a narrower mouth end for /u/), while fine-scale details remain ambiguous under single-channel steady voiced supervision.

Vowel	Ref.	PINN (post-render)	DDSP-only	PINN (in-graph)
/a/	8.439	8.449	8.434	2.827
/i/	9.225	7.806	6.833	4.243
/u/	7.901	7.803	7.664	2.284

Table 3. Periodicity (HNR, dB; median). Columns (left→right): Reference, PINN post-render, DDSP-only, PINN in-graph.

Mismatch	Median ΔLSD (dB) $\downarrow$	Median ΔHNR (dB) $\downarrow$
Discretization (grid/CFL)	0.287	0.013
Source (β, Oq/Cq)	0.554	0.025
Pitch $\pm 10\%$	1.541	0.481

Table 4. Robustness of learned controls. Medians across vowels and non-baseline settings; absolute deltas.

envelopes while failing to pin down pitch-synchronous excitation and harmonic sharpness. Addressing this gap likely requires adding pitch-synchronous/harmonic-structured objectives and/or an explicit glottal-source prior, rather than further tuning of envelope losses alone.

4.3. Learned $A(x)$ and ζ: transferable controls but not uniquely identifiable under steady vowels

Figure 2 shows that $\hat{A}(x)$ remains smooth, positive, and properly anchored; however, the recovered shapes tend to simplify local constrictions (most notably for /i/), reflecting the known non-uniqueness of acoustic-to-articulatory inversion from single-channel steady voiced audio [23]. We therefore interpret $\hat{A}(x)$ as a *spectrally-equivalent control parameterisation* under the assumed Webster+Robin model rather than a uniquely identifiable anatomical reconstruction.

The learned Robin coefficient $\hat{\zeta}$ is tightly clustered across vowels (last-epoch mean 0.127 ± 0.001).[2] The cross-solver post-render results indicate that $\hat{\zeta}$ functions as a transferable boundary control, while the steady-vowel setting also allows $\hat{\zeta}$ to absorb residual modelling mismatch (e.g., source assumptions and high-frequency decay). Since training also includes a weak regulariser on ζ, $\hat{\zeta}$ should be interpreted as an *effective* boundary parameter and is not expected to match ζ_{ref}.

4.4. Robustness to controlled mismatches

To test whether $(\hat{A}, \hat{\zeta})$ behave as reusable physical controls beyond a single numerical setting, we fix them and post-render under controlled deviations (Table 4). Discretisation changes (grid/CFL) induce only small metric drifts, suggesting practical invariance to solver resolution within stable regimes [9]. Moderate source/propagation variations (damping β and Rosenberg (O_q, C_q)) also cause limited drift, consistent with a source–filter view where tract and radiation primarily shape the envelope. Pitch perturbations of $\pm 10\%$ yield larger envelope drift as harmonics shift relative to the envelope, while periodicity changes remain moderate. Overall, these tests support that $(\hat{A}, \hat{\zeta})$ act as stable controls under reasonable forward-model variations.

4.5. Audio examples and qualitative observations

Audio examples for reference, post-render, and DDSP-only outputs are provided on the project page (Sec. 3). In-graph audio is omitted from the demo because it does not provide additional perceptual insight beyond the objective evidence in Table 3. Qualitatively, post-render is closest to the reference in both envelope and periodicity, whereas DDSP-only remains competitive but characteristically brighter in this low-data regime.

5. CONCLUSION

We presented a time-domain Webster PINN with learnable open-end radiation that estimates a vocal-tract area function and a Robin radiation coefficient from single-channel sustained vowels. Exporting the recovered controls to an independent FDTD–Webster solver enables solver-transfer evaluation: post-rendered audio matches target spectral envelopes competitively and is stable under discretisation and moderate source/pitch mismatches. The in-graph differentiable rendering path remains more aperiodic, motivating periodicity-aware objectives and explicit glottal-source priors. Broader benchmarking against complete SVS systems and additional physics-informed baselines is deferred to future work.

[2]Under the Robin boundary $\partial_x \psi + \zeta\, \partial_t \psi = 0$, with $p = -\rho\, \partial_t \psi$ and $U = A\, \partial_x \psi$, an effective low-frequency radiation impedance can be approximated as $Z_{\mathrm{rad}} \approx \rho/(A\zeta)$. In our runs $\hat{\zeta}$ converges to a narrow band around 0.127, consistent with weak radiation (high reflection magnitude) at low ka [6, 7, 9].

6. REFERENCES

[1] Aaron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, Koray Kavukcuoglu, et al., "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, vol. 12, pp. 1, 2016.

[2] Kundan Kumar, Rithesh Kumar, Thibault De Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre De Brebisson, Yoshua Bengio, and Aaron C Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis," *Advances in neural information processing systems*, vol. 32, 2019.

[3] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," vol. 33, pp. 17022–17033, 2020.

[4] Peng Bai, Meizhen Zheng, and Xiaodong Shi, "A survey of singing voice synthesis," pp. 19–30, 2023.

[5] Ben Hayes, Jordie Shier, György Fazekas, Andrew McPherson, and Charalampos Saitis, "A review of differentiable digital signal processing for music and speech synthesis," *Frontiers in Signal Processing*, vol. 3, pp. 1284100, 2024.

[6] David T. Blackstock, *Fundamentals of physical acoustics*, John Wiley & Sons, 2000.

[7] Oriol Guasch, Marc Arnela, and Arnau Pont, "Resonance tuning in vocal tract acoustics from modal perturbation analysis instead of nonlinear radiation pressure," *Journal of Sound and Vibration*, vol. 493, pp. 115826, 2021.

[8] Malte Kob, *Physical modeling of the singing voice*, Ph.D. thesis, Aachen, Techn. Hochsch., Diss., 2002.

[9] Stefan Bilbao and Alberto Torin, "Numerical modeling and sound synthesis for articulated string/fretboard interactions," *Journal of the Audio Engineering Society*, 2015.

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

[11] Chin-Yun Yu and György Fazekas, "Golf: A singing voice synthesiser with glottal flow wavetables and lpc filters," *Transactions of the International Society for Music Information Retrieval*, vol. 7, no. 1, 2024.

[12] Maziar Raissi, Paris Perdikaris, and George E Karniadakis, "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations," *Journal of Computational physics*, vol. 378, pp. 686–707, 2019.

[13] Riccardo Simionato, Stefano Fasciani, and Sverre Holm, "Physics-informed differentiable method for piano modeling," *Frontiers in Signal Processing*, vol. 3, 2024.

[14] Kazuya Yokota, Takahiko Kurahashi, and Masajiro Abe, "Physics-informed neural network for acoustic resonance analysis in a one-dimensional acoustic tube," *The Journal of the Acoustical Society of America*, vol. 156, no. 1, pp. 30–43, 2024.

[15] Jari Kaipio and Erkki Somersalo, "Statistical inverse problems: Discretization, model reduction and inverse crimes," *Journal of Computational and Applied Mathematics*, vol. 198, no. 2, pp. 493–504, 2007.

[16] Ryan Prenger, Rafael Valle, and Bryan Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," pp. 3617–3621, 2019.

[17] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro, "DiffWave: A versatile diffusion model for audio synthesis," 2021.

[18] H. Wakita, "Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms," *IEEE Transactions on Audio and Electroacoustics*, vol. 21, no. 5, pp. 417–427, 1973.

[19] Juergen Schroeter and Man Mohan Sondhi, "Techniques for estimating vocal-tract shapes from the speech signal," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 1, pp. 133–150, 1994.

[20] Louis-Jean Boë, Pascal Perrier, and Gérard Bailly, "The geometric vocal tract variables controlled for vowel production: Proposals for constraining acoustic-to-articulatory inversion," *Journal of Phonetics*, vol. 20, no. 1, pp. 27–38, 1992.

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

[22] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein, "Implicit neural representations with periodic activation functions," vol. 33, pp. 7462–7473, 2020.

[23] Brad H. Story, "Technique for "tuning" vocal tract area functions based on acoustic sensitivity functions," *The Journal of the Acoustical Society of America*, vol. 119, no. 2, pp. 715–718, 2006.

[24] Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim, "Parallel Wavegan: A Fast Waveform Generation Model Based on Generative Adversarial Networks with Multi-Resolution Spectrogram," pp. 6199–6203, 2020.

[25] A. E. Rosenberg, "Effect of Glottal Pulse Shape on the Quality of Natural Vowels," *The Journal of the Acoustical Society of America*, vol. 49, no. 2B, pp. 583–590, 1971.

[26] Masanori Morise, Fumiya Yokomori, and Kenji Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions on Information*, vol. E99-D, no. 7, pp. 1877–1884, 2016.

[27] Xin Wang, Shinji Takaki, and Junichi Yamagishi, "Neural source-filter waveform models for statistical parametric speech synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 402–415, 2020.