Toward Expressive Timbre Modeling: Convergence of DDSP and Neural Voice Synthesis

Henrique Maia Lins Vaz[1,2] **, João Pedro Mendes de Oliveira**[1,2]

[1]Universidade Federal de Juiz de Fora (UFJF)

[2]Gambioluteria – da programação orientada à gambiarra ao entalhe da luteria pós-digital

henrique.maia.vaz@ufjf.br, joaopedro.oliveira@estudante.ufjf.br

Abstract. *This article investigates the convergence between Differentiable Digital Signal Processing (DDSP) and neural singing voice synthesis approaches, focusing on the modeling and control of vocal timbre. By integrating explainable acoustic structures with deep learning, DDSP-based models offer interpretable, efficient, and expressive synthesis, overcoming limitations of purely neural architectures such as WaveNet or Tacotron. Analyzing models like the Neural Parametric Singing Synthesizer (NPSS), the article proposes a conceptual and technical framework that highlights recent advances in the explicit separation of pitch, timbre, and prosody. The study discusses implementations using differentiable LPC and FIR filters, hybrid vocoders, harmonic + noise synthesis, and strategies for timbre transfer and singing voice conversion (SVC). It argues that hybridization between DSP and neural networks marks a new methodological frontier for vocal synthesis — one that is more transparent, controllable, and suitable for expressive, pedagogical, and interactive applications.*

1. Introduction

Audio synthesis has undergone significant transformations over the past century, ranging from the use of electromechanical instruments to techniques based on machine learning. More recently, Differentiable Digital Signal Processing (DDSP) has emerged as a novel methodology. It enables the backpropagation of loss function gradients through traditional digital signal processing (DSP) modules, such as oscillators, filters, and reverberators. These modules are implemented in a way that is compatible with automatic differentiation. This approach enables more interpretable and expressive synthesis compared to methods based purely on neural networks

The term DDSP refers to a hybrid approach that combines neural networks with signal processing principles, aiming to overcome the limitations of traditional models [1]. The central idea is to incorporate inductive biases grounded in acoustic knowledge — such as spectral modeling, additive synthesis, and noise filtering — directly into the neural architecture. Figure 1 provides an overview of the structure of a DDSP-based sound synthesis system. The diagram presents the complete processing flow, from input signals to parameter optimization.

This work explores how DDSP can enhance vocal timbre modeling by providing unprecedented dynamic control over acoustic parameters. This advantage is demonstrated through experiments with sung voice data, where

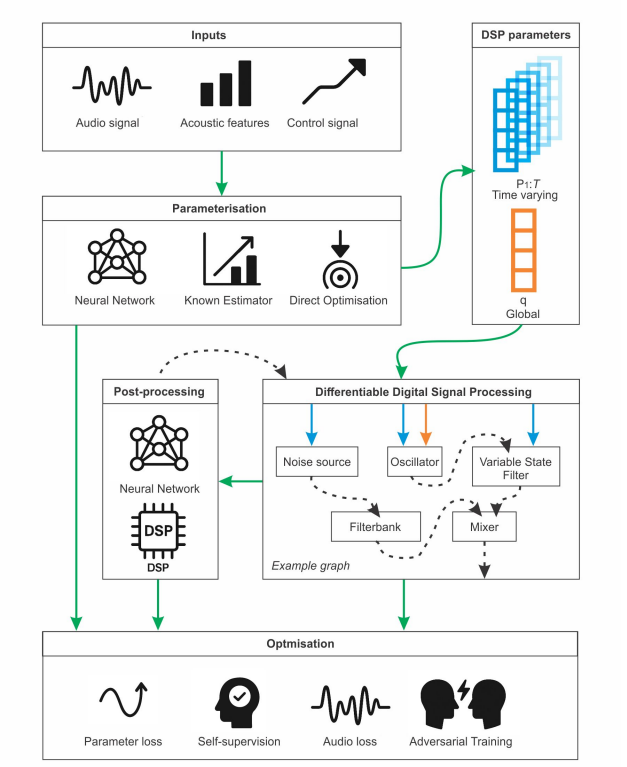

Figure 1: The architecture of a DDSP-based sound synthesis system (adapted from [2]).

limitations of traditional neural models — such as robustness and adaptability to limited data — are overcome. The objective is to establish a critical dialogue between existing approaches by proposing a framework that prioritizes not only sound quality but also the expressiveness and interpretability of the synthesis process.

In addition to reviewing state-of-the-art approaches, this article is positioned as a critical study that foregrounds timbre as a central organizing principle in differentiable digital signal processing (DDSP) for singing voice synthesis (SVS) and singing voice conversion (SVC). Alongside the review, we introduce original contributions: (i) a conceptual framework highlighting timbre, (ii) a comparative analysis bridging DDSP-based architectures and hybrid vocoders, and (iii) a final synthesis of practical applications in education, stylized synthesis, and interactive performance. These contributions aim to complement the reviewed literature with new perspectives for both research and artistic practice.

2. Background and prior approaches

Deep learning–based vocal synthesis has seen significant advancements through models such as WaveNet [3], Tacotron [4], and GANSynth [5]. Although these systems achieve high audio fidelity, they still present several limitations:

1. A requirement for large amounts of training data;
2. Lack of parametric control (e.g., the inability to independently adjust timbral attributes);
3. Inference instabilities (e.g., prosody breakdowns in Tacotron and the emergence of spectral artifacts);
4. Poor generalization to new domains with limited data.

These challenges motivate the pursuit of hybrid approaches. Time-domain models, such as WaveNet and SampleRNN [6], generate audio autoregressively, which limits computational efficiency due to sequential processing and reduces interpretability of intermediate parameters. Conversely, frequency-domain approaches, such as Tacotron and GANSynth, suffer from phase loss (resulting in phasiness artifacts), spectral leakage caused by limited filter bank resolution, and difficulty capturing microvariations — such as portamento or non-stationary oscillations — that are crucial for auditory perception.

In this context, DDSP proposes a hybrid strategy: neural networks predict parameters for synthesizers based on differentiable DSP modules — such as harmonic oscillators and finite impulse response (FIR) filters. This approach combines the flexibility of machine learning with the interpretability of acoustic models.For example, it enables the controllable adjustment of the fundamental frequency $f_0(n)$ for vibrato or the filtering of residual noise—capabilities not feasible in purely neural models. This method is based on the harmonic + noise model, where the output signal $x(n)$ is generated by a bank of oscillators, which can be represented as

$$x(n) = \sum_{k=1}^{K} A_k(n).sin(\phi_k(n)) \tag{1}$$

where $A_k(n)$ represents the amplitude and $\phi_k(n)$the instantaneous phase. The phase is derived from the frequency, as shown by the equation

$$\phi_k(n) = 2\pi \sum_{m=0}^{n} f_k(m) + \phi_{0,k} \tag{2}$$

where $\phi_{0,k}$ is the initial phase and $f_k(m)$ the instantaneous frequency accumulated over time, represented as $f_k(m) = k.f_0(m)$.

The frequency of each harmonic is determined by multiplying the fundamental frequency, $f_k(n) = k.f_0(n)$. The amplitude is given by

$$A_k(n) = A(n).c_k(n) \tag{3}$$

where $A(n)$ controls the global loudness and $c_k(n)$ defines the normalized spectral distribution (timbre envelope), such that $\sum_k c_k(n) = 1$. This structure allows for explicit separation and control of perceptual aspects: loudness, via $A(n)$, pitch, via $f_0(n)$, and spectral distribution or timbre, via $c_k(n)$.

DDSP can also apply temporal filters — such as linear time-varying (LTV) and FIR — via convolution in the frequency domain

$$Y_l = H_l.X_l \Rightarrow y_l = \text{IDFT}(Y_l) \tag{4}$$

where H_l is the frequency response learned by the network and applied to each frame X_l. This enables modules such as differentiable reverberation, subtractive synthesis, and acoustic environment transformation.

The key advantage of this approach is that, by incorporating interpretable modules directly within neural models, it becomes possible to apply perceptually relevant loss functions, such as the multi-scale spectral loss

$$L_i = \|S_i - \hat{S}_i\|_1 + \alpha\|logS_i - log\hat{S}_i\|_1 \tag{5}$$

where $\hat{S}_i$ is the spectrogram of the synthesized audio and α is a weighting factor (usually set to 1). The total loss is given by $L_{reconstr} = \sum_i L_i$.

The differentiable structure of DSP modules enables the use of composite loss functions that integrate multiple objectives — for example, spectral fidelity, accuracy in reconstructing acoustic parameters (pitch, loudness), and timbral regularization, as illustrated by the representation

$$L_{total} = \lambda_{audio}L_{audio} + \lambda_{params}L_{params} + ... \tag{6}$$

where each λ represents the weight assigned to its respective loss function.

While purely neural models suffer from phase loss and lack of parametric control, DDSP employs differentiable additive synthesis to preserve physically plausible harmonic relationships and enable fine real-time adjustments. For the focus of this article — the convergence between DDSP and neural voice synthesis with an emphasis on timbre — this structure provides a pathway that previous models were unable to achieve transparently.

3. Hybrid architectures and the NPSS as a bridge between vocoders and DDSP

The pathways of neural vocal synthesis have shown a clear trend toward the explicit separation of acoustic attributes — such as pitch, phonetic content, and timbre — in pursuit of greater control, expressiveness, and interpretability in generation systems. The Neural Parametric Singing Synthesizer (NPSS) [7] represents a significant milestone in this trajectory by structuring vocal synthesis within a modular architecture, anticipating concepts later consolidated in DDSP-based models. NPSS is a neural network-based vocal synthesis system that models timbre and musical expression from natural recordings by combining:

- an autoregressive model (based on a modified WaveNet) to generate acoustic features;

- a parametric vocoder (WORLD) for final waveform synthesis;
- the explicit separation of pitch ($F0$), timbre (harmonic/aperiodic), and phonetic timing.

The NPSS architecture is composed of three main components that work in an integrated manner. The "timbre model", based on a modified Multistream WaveNet architecture, generates three interdependent acoustic streams:

1. A harmonic envelope represented by 60 Mel-Frequency Spectral Coefficients (MFSC);
2. An aperiodic envelope with 4 coefficients;
3. A binary voiced/unvoiced (V/UV) decision.

The generation process follows an autoregressive structure modeled by the equation

$$p(x_t|x_{<t}, c) = \Pi_{i=1}^{N} p(x_{t,i}|x_{<t}, c) \tag{7}$$

where x_t represents the feature vector at time t, and c includes control inputs such as pitch and phoneme information. The output uses a Constrained Gaussian Mixture (CGM) model defined as

$$p(x) = \sum_{k=0}^{3} \omega_k N(x; \mu_k, \sigma_k^2) \tag{8}$$

with the components μ_k, σ_k, and ω_k derived from free parameters

$$\sigma_k = \omega e^{(/\alpha/\gamma_s - 1)k} \tag{9}$$

$$\mu_k = \xi + \sum_{i=0}^{k-1} \sigma_k \gamma_u a \tag{10}$$

$$\omega_k = \frac{\alpha^{2k}\beta^k \gamma_\omega}{\sum_{i=0}^{K-1} \alpha^{2i}\beta^i \gamma_\omega} \tag{11}$$

where ξ is the location, ω the scale, α the skewness, ands β the shape parameter. The constants $\gamma_s = 1.1$, $\gamma_u = 1.6$, $\gamma_\omega = 1/1.75$ are empirically tuned.

The interdependence among the streams is carefully designed: the aperiodic stream depends on the harmonic one, while the V/UV decision considers both. The "pitch model", also based on WaveNet, employs data augmentation through pitch transposition

$$\hat{f}_0 = f_0 . 2^{p_{pshift}/12} \tag{12}$$

$$pshift \sim U(pshift_{min}, pshift_{max}) \tag{13}$$

and includes a sophisticated post-processing step that corrects deviations using a weighted average considering note position (via a Tukey window) and the derivative of F_0.

The "phonetic timing model" operates in two stages: a feedforward network predicts note onset deviations, while a convolutional network estimates phoneme durations. A final heuristic adjustment ensures precise alignment with the note durations in the musical score. This tripartite architecture allows NPSS to achieve an optimal balance between flexibility, naturalness, and computational efficiency in vocal synthesis.

The NPSS system employs the WORLD vocoder (D4C version) for both analysis and synthesis, operating at a sampling rate of 32 kHz and a hop time of 5 ms. During the feature extraction phase, the vocal signal is decomposed into its essential components. In the synthesis stage, the system reconstructs the voice signal from the generated parameters, according to the equation

$$x(n) = Vocoder(F_0[n], h[n], a[n], v[n]) \tag{14}$$

where $F_0[n]$ represents the pitch contour, $h[n]$ the harmonic envelope, $a[n]$ the aperiodic envelope, and $v[n]$ the binary V/UV decision.

NPSS has established itself as a fundamental precursor to the DDSP paradigm, introducing key concepts that would later be refined. Its main contribution was the explicit separation of pitch and timbre — while DDSP implements this through additive synthesis, NPSS had already developed a timbre control system using MFSC coefficients conditioned on f_0. Both systems share the crucial feature of acoustic interpretability, relying on vocoder parameters as their foundation: NPSS uses the WORLD vocoder, whereas DDSP opts for FIR or LPC filters. A significant architectural difference lies in the generation mechanism: NPSS employs an autoregressive approach based on WaveNet, which, although powerful, faces limitations in terms of parallelization. DDSP, in turn, evolved toward feedforward models (RNN or Transformer), achieving greater computational efficiency without sacrificing synthesis quality. This evolution demonstrates how NPSS paved the way for more advanced systems, while still remaining relevant in the development of interpretable neural vocal synthesis.

4. Interpretable Modeling of Vocal Timbre with DDSP: Methods and Applications

In monophonic voice synthesis, DDSP enables the generation of expressive sounds from a compact acoustic representation. In the case of the human voice, the model is based on a decomposition of the signal into harmonic and noise components, allowing for precise and differentiable control of acoustic parameters over time. The explicit manipulation of acoustic parameters through differentiable additive synthesis is described by the Harmonic + Noise model

$$x(n) = HC_{(Harmonic\ Component)} + NC_{(Noise\ Component)} \tag{15}$$

$$HC = \sum_{k=1}^{N} A_k(n) sin\left(2\pi k \sum_{m=0}^{n} f_0(m) + \phi_k(n)\right) \tag{16}$$

$$NC = \sum_{l=1}^{L} h_l(n) * r(n) \tag{17}$$

where $A_k(n)$ is the amplitude of the k-th harmonic at time n (controlling timbre), $f_0(m)$ is the time-varying fundamental frequency, $\phi_k(n)$ is the phase of each harmonic, $h_l(n)$ are convolutional filters that shape the noise spectrum, and $r(n)$ is white (or colored) noise used as input to the convolution $h_l(n) * r(n)$.

In practice, the manipulation of spectral coefficients $c_k(t)$ (which act as filters over the harmonics), normalized by $\sum_k c_k(n) = 1$, enables the shifting of formant peaks (F_1, F_2) to higher or lower regions of the spectrum. This allows, for example, the conversion of an adult voice into a child-like voice by simulating a shorter vocal tract through adaptive filters applied to the harmonic component. Expressive variations such as vibrato and belting can be modeled directly. Vibrato, for instance, is represented as a periodic modulation of the fundamental frequency

$$\Delta f_0(t) = A sin(2\pi f_{vib} t) \tag{18}$$

where A defines the depth and f_{vib} the rate of the vibrato. Belting, on the other hand, can be simulated by reinforcing the upper harmonics in the series, increasing the coefficients $A_k(n)$ for $K \geq 5$, resulting in a more intense and "brighter" sound. The main advantage of DDSP lies in its interpretability: parameters such as f_0 and $c_k(n)$ can be adjusted in real time —something that is not possible with models like WaveNet or Tacotron, which operate in unstructured latent spaces.

By combining oscillators, harmonics, noise generators, and convolution, DDSP proves to be a viable architecture for reconstructing monophonic vocalizations. The separation of f_0, loudness, and spectral envelope enables models to learn independent and controllable representations — an essential feature for expressive voice manipulation.

4.1. Vocal Timbre Transfer: Differentiable Source-Filter Models and Hybrid DDSP Vocoders

Vocal timbre transfer aims to transform the timbral identity of a source voice while preserving properties such as textual content, melody, and dynamics. Deep learning-based models have increasingly employed differentiable source-filter architectures, in which the excitation source (harmonic voice and glottal noise) is separated from a filter representing the vocal tract.

In the implementation using linear predictive coding (LPC), the coefficients are estimated by a neural network

$$H(z) = \frac{1}{1 - \sum_{i=1}^{p} a_i z^{-i}} \tag{19}$$

where a_i are stable LPC coefficients (ensured by reflection conditions $|k_i| < 1$). Alternatively, using FIR, a filter bank can be implemented with learned frequency responses $H_l(e^{j\omega})$, such that

$$H_l(e^{j\omega}) = MLP(z) \tag{20}$$

where z is a latent vector.

An Long Short-Term Memory (LSTM) encoder, being a type of recurrent neural network (RNN) specialized in processing temporal sequences, is often used to map input spectrograms to acoustic parameters relevant for synthesis, such as normalized harmonic amplitudes ($A_k(n) = A(n).c_k(n)$, $\sum_k c_k(n) = 1$) and FIR filter coefficients applied to the noise component. In this context, vocal timbre transfer is achieved by replacing the filter $H(z)$ — which represents the vocal tract of the source voice — with a filter trained on the spectral characteristics of the target voice. This approach offers advantages such as precise spectral control, with explicit adjustment of formants, and computational efficiency due to its frequency-domain execution (e.g., via FFT), enabling real-time synthesis.

Approaches inspired by classical acoustic models enable explicit and controlled manipulation of vocal timbre through hybrid vocoders such as DDSP-VAE (variational autoencoder) or Neural Source-Filter [8], which integrate DSP modules with supervised learning to synthesize voices with high spectral fidelity and interpretable control. This allows for timbre transformation between speakers, vocal identity changes, or expressive stylizations (such as adding "brightness" or "opacity" to the voice), achieving a high degree of perceptual naturalness.

In order to enhance expressivity and robustness while retaining timbral control, recent approaches have sought to combine the acoustic interpretability of DDSP with the generative capacity of adversarial architectures. One prominent example is NANSY++ [9], a hybrid neural vocoder that integrates DDSP-inspired source-filter decomposition with adversarial training and vector-quantized latent embeddings.

It separates pitch, content, and timbre into distinct, disentangled representations, allowing for high-fidelity voice conversion and timbre transfer. The harmonic + noise source signal is decoded via differentiable modules, while the target timbre is captured using a quantized embedding space that can be swapped or interpolated. This allows not only realistic synthesis but also expressive, interpretable control over vocal identity — aligning closely with the goals of hybrid DDSP models.

The NANSY++ architecture operates in three main stages. The first is in timbre embedding extraction, where a convolutional neural network (CNN) processes reference spectrograms to generate a latent vector $t \in \mathbb{R}^d$. The second is in conditioned DDSP synthesis, where the synthesizer receives t together with acoustic parameters (f_0, A) to produce an initial signal

$$x(n) = DDSP(f_0, A, t) + GAN_{residual}(n) \tag{21}$$

where the Generative Adversarial Network (GAN) generates residuals to capture micro-variations not modeled by the DDSP. The third is in optimization, where a composite loss function is minimized

$$L_{\text{total}} = \underbrace{\lambda_1 \|S - \hat{S}\|}_{\text{Spectral loss}} + \underbrace{\lambda_2 E[\log D(\hat{x})]}_{\text{Adversarial loss}} + \underbrace{\lambda_3 \|t_{\text{src}} - t_{\text{tgt}}\|_2}_{\text{Timbre regularization}} \tag{22}$$

with S being the magnitude of the target spectrogram and D a discriminator based on HiFi-GAN [10]. For zero-shot voice conversion tasks, NANSY++ replaces t_{src} with t_{tgt} (extracted from 1 second of target audio), achieving a Mean Opinion Score (MOS) of 4.2 compared to 3.5 for AutoVC [11].

The implementation of hybrid DDSP vocoders

presents two main challenges:

1. Infinite impulse response (IIR) filter instability: unregularized LPC coefficients may result in poles outside the unit circle, compromising filter stability. This issue is addressed in [12] through the normalization of reflection coefficients using the transformation $k_i = tanh(k_i)$;
2. Joint training complexity: the combined training of DDSP and GAN modules requires careful tuning of the composite loss function weights λ_i to maintain training stability and synthesis quality.

Future research directions include:

1. Replacement of GANs with diffusion models: diffusion-based approaches, such as DiffSVC [13], offer enhanced robustness in modeling residual components;
2. Neural filter synthesis: instead of predefined DSP-based filters, neural networks can directly learn a nonlinear mapping for $H(z)$ [14].

This methodological evolution retains the mathematical interpretability of DDSP while integrating recent advances in generative modeling, establishing hybrid vocoder architectures as a promising paradigm for next-generation voice synthesis systems.

4.2. Singing Voice Synthesis (SVS) and Singing Voice Conversion (SVC)

SVS and SVC are frontier tasks in the field of neural audio synthesis. SVS aims to generate realistic vocal performances from symbolic scores (e.g., MIDI) and lyrics. DDSP-based models enable this synthesis by mapping musical control parameters — $f_0(n)$ and loudness $A(n)$ — to acoustic parameters used by a harmonic + noise synthesizer.

The modular architecture further allows for explicit timbral control through differentiable LPC or FIR filters. The LPC filter assumes that a signal (such as the human voice) can be approximated by an autoregressive (AR) model, in which each sample is a linear combination of previous samples plus a residual noise term (the excitation signal), expressed as

$$x(n) = \sum_{i=1}^{p} a_i.x(n-i) + e(n) \tag{23}$$

where $x(n)$ is the current sample, a_i are the LPC coefficients that define the filter, p is the filter order (i.e., the number of past samples used), and $e(n)$ is the residual error (excitation).

In the context of SVS and SVC, LPC filters are typically employed to model the vocal tract resonances, providing an efficient parametric representation of the spectral envelope. This same principle underlies their integration into hybrid DDSP-based vocoders, where LPC is reformulated in a differentiable manner and combined with neural modules. In this setting, the LPC filters no longer serve only as classical analysis tools but also act as trainable components within an end-to-end pipeline, enabling explicit timbral control while benefiting from gradient-based optimization. This continuity highlights a conceptual bridge: both SVS/SVC systems and DDSP hybrids rely on the same source–filter decomposition, yet the latter extend it by embedding the filter estimation into a differentiable, learnable architecture. We propose here a conceptual link between LPC filters used in SVS/SVC tasks and those implemented in hybrid DDSP vocoders, which to our knowledge has not been explicitly discussed in the literature.

Regarding the FIR filter, its output depends solely on a linear combination of past inputs (with no feedback), and its impulse response has finite duration. Its transfer function is given by

$$y[n] = \sum_{k=0}^{N} b_k.x[n-k] \tag{24}$$

where b_k are the filter coefficients (which define the frequency response), and N is the filter order (higher values allow for more precise spectral control).

However, in neural synthesis tasks (such as SVS/SVC), all operations must be differentiable to enable training via backpropagation. A traditional FIR filter is not differentiable with respect to its coefficients, but neural network–compatible versions address this issue. One common approach is to implement the FIR as a linear operation, expressing the convolution as a matrix multiplication in which the coefficients b_k are learnable and adjusted through gradient-based optimization. Recent studies [15] have proposed differentiable oscillators with phase continuity to avoid auditory artifacts (e.g., glitches or pitch jitter), along with parametric filters for real-time timbral control.

The goal of SVC is to transform the vocal performance of a source singer — preserving melody and prosody — so that it sounds as if it were sung by a target singer. The main challenge lies in preserving expressive features such as vibrato, articulation, and dynamics, while simultaneously adapting the timbre. DDSP-based models have explored autoencoder architectures combined with differentiable vocoders, such as a differentiable implementation of the WORLD vocoder [16]. The loss function typically employed in this context combines terms for spectral reconstruction, pitch coherence, and phase smoothness

$$L_{total} = \lambda_1 \|S - \hat{S}\|_1 + \lambda_2 \|f_0 - \hat{f}_0\|_2 + \lambda_3 \|\phi - \hat{\phi}\|_3 \tag{25}$$

where S is the target spectrogram, f_0 is the fundamental frequency (pitch), and ϕ is the phase spectrum. Timbre transfer is performed explicitly by replacing the timbre vector $t_{src} \rightarrow t_{tgt}$, enabling the transformation of vocal identity. This approach is exemplified by the NANSY++ model, which applies latent factor disentanglement —separating pitch, linguistic content, and timbre — through specialized neural networks.

Differentiable FIR and LPC filters exhibit distinct characteristics that make them suitable for different applications. FIR filters stand out for their simplicity, as they operate without feedback and are effective in modeling arbitrary frequency responses, making them particularly useful

in equalization tasks. In contrast, LPC filters, due to their feedback nature (as IIR filters), are more efficient in modeling the spectral structure of the voice, especially in the accurate representation of formants, which makes them particularly valuable for vocal synthesis. In many DDSP-based models, these two approaches are combined in a hybrid fashion, leveraging the strengths of each: FIR filters are used to shape harmonic content, while LPC filters control formants, resulting in a more flexible and high-quality synthesis.

Neural source-filter architectures have become one of the most widely adopted approaches in SVS and SVC tasks, where differentiable FIR and LPC filters are commonly used as components of the filter block. In this model, an excitation signal $x[n]$ — typically harmonic or noise — is processed by a nonlinear filter f_θ, parameterized by a neural network, which introduces additional frequency components due to its nonlinearity. The output $y[n]$ is given by

$$y[n] = f_\theta(x[n], z[n]) \tag{26}$$

where $z[n]$ is a conditioning signal that modulates the filter response. This formulation is analogous to waveshaping synthesis [17], where the nonlinearity of f_θ generates new harmonic components from a purely harmonic input signal.

The neural filter block often employs an architecture inspired by non-autoregressive WaveNet, using dilated convolutions, gated activations, and residual connections, such that

$$f_\theta(x[n]) = \textit{ResBlock}(\textit{DilatedConv}(x[n])) + x[n] \tag{27}$$

given a harmonic excitation

$$x[n] = \sum_{k=1}^{K} \alpha_k \textit{sin}(2\pi k f_0 + n) \tag{28}$$

the nonlinear filter f_θ introduces harmonic distortion, yielding

$$y[n] = f_\theta \left(\sum_{k=1}^{K} \alpha_k \textit{sin}(2\pi k f_0 + n) \right) \tag{29}$$

where f_θ is implemented as a stack of residual layers with nonlinear activations.

Neural Source-Filter Models face significant challenges, particularly regarding optimization and generalization. One of the main obstacles lies in the optimization of frequency-related parameters — as in the case of frequency modulation (FM) synthesis — where the loss surface is non-convex, exhibiting multiple local minima that hinder convergence via gradient descent [18]. This property makes it difficult to fine-tune oscillator frequencies, limiting the effectiveness of gradient-based training. Moreover, the strong inductive bias toward harmonic signals inherent to these models restricts their applicability to more complex sounds, such as non-harmonic or polyphonic audio, in which the spectral structure does not follow simple harmonic patterns.

A persistent challenge in SVS and SVC lies in the non-convexity of optimization when estimating or generating frequency-related parameters such as f_0 trajectories or formant positions. Recent works have proposed strategies to mitigate this issue, including smooth parameterization of pitch contours, regularization through differentiable sinusoidal models, and multi-scale loss functions that stabilize training across harmonic and sub-harmonic ranges. From our perspective, integrating such strategies into hybrid DDSP vocoders could provide a principled way to balance flexibility with stability, allowing frequency parameters to be optimized within a tractable and musically meaningful space.

5. The Voice to Come

The convergence of DDSP and Neural Vocal Synthesis represents a turning point in the field of audio synthesis, enabling more precise and interpretable control of vocal timbre — an essential yet historically elusive dimension of sonic experience. As discussed, DDSP models allow the integration of the formal rigor of physical-parametric models with the adaptability of deep learning methods, offering an inductive bias that supports not only expressiveness but also computational efficiency and training stability.

The DDSP paradigm offers not only more explainable architectures but also new ways of sonic control and manipulation, as seen in SVS/SVC applications. Recent advances — such as vocoders based on differentiable IIR filters, fully trainable LPC models, and wavetables parameterized by timbral embeddings — illustrate a clear movement toward hybrid models that combine DSP robustness with neural flexibility. This transition becomes even more promising in the face of challenges such as generalization across speakers or singers, expressive real-time synthesis, and timbral personalization in interactive and participatory contexts.

When considering current limitations — such as the difficulty in estimating highly inharmonic filters, the instability of recursive filters, and the opacity of latent models — the field now turns to the development of alternatives to automatic differentiation, more efficient and participatory architectures, and the incorporation of interdisciplinary knowledge into synthesis models. The emergence of diffusion-based models (e.g., DiffSVC, DiffSinger) reinforces this trend, suggesting a new generation of synthesizers capable of preserving phase coherence, semantic modularity, and fine-grained timbral control.

Thus, the current landscape points to a new kind of synthesis — not only sonic, but epistemological: one in which scientific methods and artistic practices converge to recreate, manipulate, and investigate the voice as a physical, cultural, and computational phenomenon. In this hybrid territory, mathematical rigor meets expressive uncertainty. The challenge ahead is more than computational — it is political, cultural, and sensitive. Because to synthesize is also to imagine.

The practical and artistic potential of DDSP and neural synthesis extends far beyond technical benchmarks, as demonstrated in the author's electroacoustic works. In Modeling Study: Naqshe Khani [19], Mimosa Pudica [20],

De Silenti Natura [21] [22], and 8 FRAMES [23], these architectures were employed not merely for replication but for the stylization and hybridization of vocal and instrumental identities. By combining neural voice synthesis, physical modeling, and articulatory synthesis within audiovisual contexts, these works serve as proof-of-concept for the expressive and cross-modal capabilities of DDSP-based approaches. Figure 2 presents some notations for the Modeling Study: Naqshe Khani.

Finally, Table 1 presents practical applications of the convergence between DDSP and neural vocal synthesis with an emphasis on timbre, reinforcing the relevance of this topic in both artistic and academic contexts, and highlighting its technological potential.

Table 1: Practical Applications of the Convergence between DDSP and Neural Vocal Synthesis with an Emphasis on Timbre

Application Area	Description	Technological Potential
Music Education	Simulation of various vocal timbres for pedagogical purposes (e.g., child, operatic, guttural).	Real-time timbre control via sliders or MIDI-controllable parameters.
Stylized Synthesis	Creation of voices with invented or hybrid vocal identities (e.g., "digital baroque voice").	Parametric modeling of formants, vibrato, and spectral envelope.
Interactive Performance	Expressive vocal control through gestures, sensors, or graphical interfaces during live shows.	Mapping physical input to DDSP parameters (f_0, loudness, timbre).
Generative Composition	Use of neural networks to generate singing in various styles and languages.	Integration with symbolic notation systems and f_0/text encoders.
Voice Restoration or Stylization	Restoration of degraded singing recordings or stylistic transformation (e.g., historical voice).	Timbre transfer between recordings and reference models using SVC.
Vocal Accessibility	Creation of personalized synthetic voices for individuals with speech impairments.	Use of vocal identity embeddings combined with text/melody input.

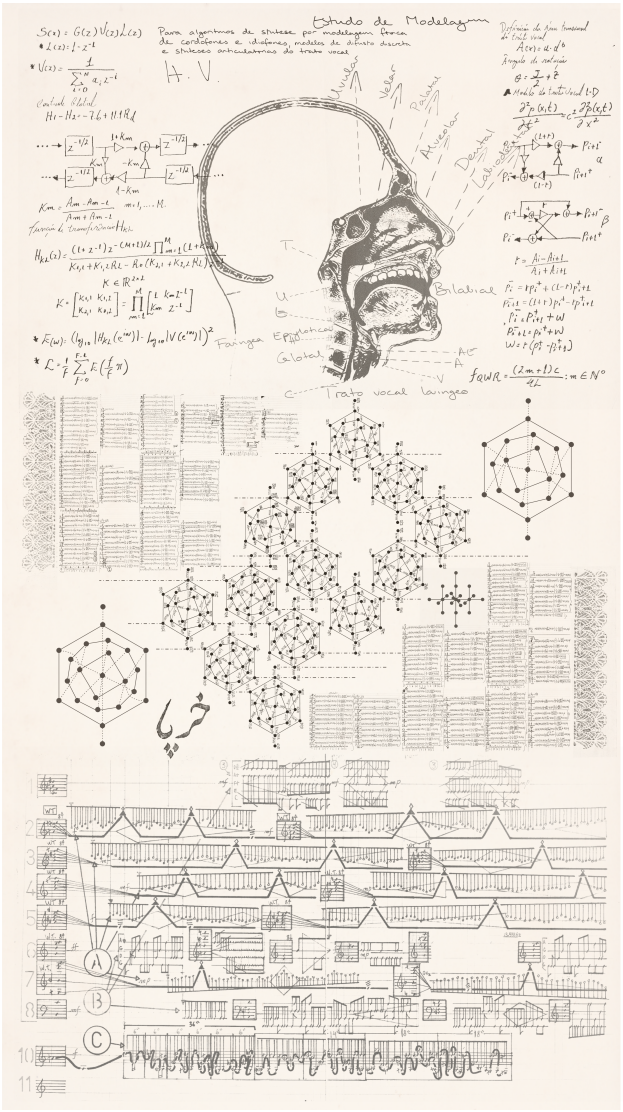

Figure 2: Notations for the Modeling Study: Naqshe Khani.

References

[1] Jesse Engel, Lamtharn (Hanoi) Hantrakul, Chenjie Gu, and Adam Roberts. DDSP: Differentiable digital signal processing. In *International Conference on Learning Representations*, 2020. Available at: https://openreview.net/forum?id=B1x1ma4tDr. Accessed on: 27 May 2025.

[2] Ben Hayes, Jordie Shier, György Fazekas, Andrew McPherson, and Charalampos Saitis. A review of differentiable digital signal processing for music speech synthesis, 2023. Available at: https://arxiv.org/abs/2308.15422. Accessed on: 27 May 2025.

[3] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. WaveNet: A generative model for raw audio. In *9th ISCA Workshop on Speech Synthesis Workshop (SSW 9)*, page 125, 2016. Available at: https://www.isca-archive.org/ssw_2016/vandenoord16_ssw.html#. Accessed on: 27 May 2025.

[4] Yuxuan Wang, R.J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying

Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. Tacotron: Towards end-to-end speech synthesis. In *Interspeech 2017*, pages 4006–4010, 2017. Available at: https://www.isca-archive.org/interspeech_2017/wang17n_interspeech.html#. Accessed on: 27 May 2025.

[5] Jesse Engel, Kumar Krishna Agrawal, Shuo Chen, Ishaan Gulrajani, Chris Donahue, and Adam Roberts. GANSynth: Adversarial neural audio synthesis. In *International Conference on Learning Representations*, 2019. Available at: https://openreview.net/forum?id=H1xQVn09FX. Accessed on: 27 May 2025.

[6] Soroush Mehri, Kundan Kumar, Ishaan Gulrajani, Rithesh Kumar, Shubham Jain, Jose Sotelo, Aaron Courville, and Yoshua Bengio. SampleRNN: An unconditional end-to-end neural audio generation model. In *International Conference on Learning Representations*, 2017. Available at: https://openreview.net/forum?id=SkxKPDv5xl. Accessed on: 27 May 2025.

[7] Merlijn Blaauw and Jordi Bonada. A neural parametric singing synthesizer modeling timbre and expression from natural songs. *Applied Sciences*, 7(12):1313, 2017. Available at: https://www.mdpi.com/2076-3417/7/12/1313. Accessed on: 27 May 2025.

[8] Ye-Xin Lu, Yang Ai, and Zhen-Hua Ling. Source-filter-based generative adversarial neural vocoder for high fidelity speech synthesis. In Ling Zhenhua, Gao Jianqing, Yu Kai, and Jia Jia, editors, *Man-Machine Speech Communication*, pages 68–80, Singapore, 2023. Springer Nature Singapore. Available at: https://link.springer.com/chapter/10.1007/978-981-99-2401-1_6. Accessed on: 27 May 2025.

[9] Hyeong-Seok Choi, Jinhyeok Yang, Juheon Lee, and Hyeongju Kim. NANSY++: Unified voice synthesis with neural analysis and synthesis. In *International Conference on Learning Representations*, 2023. Available at: https://openreview.net/forum?id=elDEe8LYW7-. Accessed on: 27 May 2025.

[10] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. HiFi-GAN: generative adversarial networks for efficient and high fidelity speech synthesis. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, number 1428 in NIPS'20, pages 17022 – 17033, Red Hook, NY, USA, 2020. Curran Associates Inc. Available at: https://dl.acm.org/doi/abs/10.5555/3495724.3497152. Accessed on: 27 May 2025.

[11] Eliya Nachmani and Lior Wolf. Unsupervised singing voice conversion. In *Interspeech 2019*, pages 2583–2587, 2019. Available at: https://www.isca-archive.org/interspeech_2019/nachmani19_interspeech.html#. Accessed on: 27 May 2025.

[12] Shahan Nercessian. End-to-end zero-shot voice conversion using a DDSP vocoder. In *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 1–5, 2021. Available at: https://ieeexplore.ieee.org/document/9632754. Accessed on: 27 May 2025.

[13] Jinglin Liu, Chengxi Li, Yi Ren, Feiyang Chen, and Zhou Zhao. DiffSinger: Singing voice synthesis via shallow diffusion mechanism. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11020–11028, 2022. Available at: https://cdn.aaai.org/ojs/21350/21350-13-25363-1-2-20220628.pdf. Accessed on: 27 May 2025.

[14] 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*, 28:402–415, 2020. Available at: https://ieeexplore.ieee.org/document/8915761. Accessed on: 27 May 2025.

[15] Da-Yi Wu, Wen-Yi Hsiao, Fu-Rong Yang, Oscar D Friedman, Warren Jackson, Scott Bruzenak, Yi-Wen Liu, and Yi-Hsuan Yang. DDSP-based singing vocoders: A new subtractive-based synthesizer and a comprehensive evaluation. In *Proceedings of the 23rd International Society for Music Information Retrieval Conference*, pages 76–83. ISMIR, 2022. Available at: https://archives.ismir.net/ismir2022/paper/000008.pdf. Accessed on: 27 May 2025.

[16] Nercessian Shahan. Differentiable world synthesizer-based neural vocoder with application to end-to-end audio style transfer. *Journal of the Audio Engineering Society*, (10661), May 2023. Available at: https://aes2.org/publications/elibrary-page/?id=22073. Accessed on: 27 May 2025.

[17] Ben Hayes, Charalampos Saitis, and Gyorgy Fazekas. Neural waveshaping synthesis. In *Proceedings of the 22nd International Society for Music Information Retrieval Conference*, pages 254–261. ISMIR, 2021. Available at: https://archives.ismir.net/ismir2021/paper/000031.pdf. Accessed on: 27 May 2025.

[18] Joseph Turian and Max Henry. I'm sorry for your loss: Spectrally-based audio distances are bad at pitch. In *NeurIPS 2020 Workshop ICBINB*, 2020. Available at: https://openreview.net/forum?id=Z4UwGkTRTes. Accessed on: 27 May 2025.

[19] Henrique Vaz. Estudo de modelagem: Naqshe khani. IV Encontro Internacional de Música Eletroacústica da SBME–UNESPAR, 2024. Available at: https://www.sbme.com.br/ola-mundo/. Accessed on: 31 Aug 2025.

[20] Henrique Vaz. Mimosa pudica. Festival Escuta Aqui!, SBME, 2024. Available at: https://www.sbme.com.br/escuta-aqui-2024/. Accessed on: 31 Aug 2025.

[21] Henrique Vaz. De silenti natura. Seleção de Músicas Experimentais, Académie Charles Cros, 2025. Available at: http://www.charlescros.org/Selection-Musiques-experimentales-2024-1077. Accessed on: 31 Aug 2025.

[22] Henrique Vaz. De silenti natura. Editado por Mappa Editions, 2024. Available at: https://mappa.bandcamp.com/album/de-silenti-natura. Accessed on: 31 Aug 2025.

[23] Henrique Vaz. 8 frames. Festival ARTELLIGENT, Caixa Cultural Salvador, 2024. Available at: https://www.instagram.com/p/DN6blggkVqI/. Accessed on: 31 Aug 2025.