

DIFFERENTIABLE ARTICULATORY COPY-SYNTHESIS OF BIPHONIC SINGING

Mateo Cámara, María Pilar Daza-Llin, Fernando Marcos-Macías, and José Luis Blanco

Signal Processing Applications Group
Information Processing and Telecommunications Centre
Universidad Politécnica de Madrid, Spain
mateo.camara@upm.es

ABSTRACT

Sygyt is a Tuvan style of biphonic singing in which a low vocal drone is sustained while a high harmonic is selectively amplified in the 1–3 kHz region. Copy-synthesizing this effect remains challenging for articulatory models, since it requires fine control of narrowly focused resonances that standard low-dimensional tract parameterizations cannot easily reproduce. We address this problem with a sublingual second source, cubic B-spline tract parameterization, and spatially varying learnable damping, optimized end-to-end by gradient descent from audio. On 20 segments from two independent sygyt datasets (5 singers, 10 pitches), the proposed model reduces log-spectral distance by 30–38% relative to an articulatory baseline, with the largest gains concentrated in the overtone region. Cepstral-envelope analysis further shows more accurate recovery of the merged formant structure characteristic of sygyt production. The model also outperforms a DDSP harmonic-plus-noise baseline with direct per-harmonic spectral control, suggesting that explicit acoustic structure is a useful inductive bias for overtone-singing copy-synthesis.

1. INTRODUCTION

Biphonic singing, known as *khoomei* in Tuva, is a vocal technique that enables a single performer to produce two simultaneous percepts: a sustained fundamental drone and a selectable melodic overtone shaped through vocal tract control. Distinct styles include *sygyt* (high-pitched whistling overtones), *kargyraa* (low-register nonlinearities), and others. This paper focuses on *sygyt*, which is especially associated with narrow-band spectral focusing, often described as formant merging, in the 1–3 kHz region [1]. Fig. 1 illustrates this behavior, showing a prominent narrowband energy ridge superimposed on the fundamental drone.

Conventional synthesis approaches to overtone singing have relied on manually tuned formant synthesis [2], coupled-oscillator models [3, 4], and source-filter decomposition [5]. These methods yielded valuable insights but depend heavily on handcrafted control signals, expert analysis, or style-specific assumptions. Meanwhile, articulatory models based on Kelly–Lochbaum waveguides offer a physically meaningful representation of voice production, only recently have such models been optimized inversely for copy-synthesis from recorded speech and singing [6, 7]. A key question

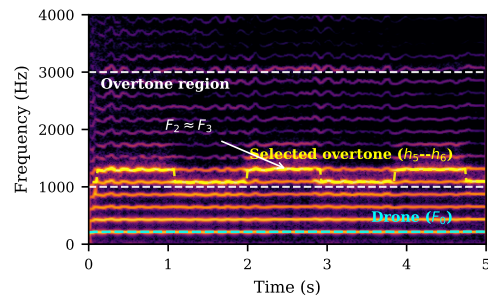


Figure 1: Spectrogram of a *sygyt* performance: h_5 – h_6 overtone selection. The singer sustains a drone at f_0 (cyan) while shaping the vocal tract to selectively amplify a single overtone (yellow). The 1–3 kHz overtone region is outlined in white.

that remains open is whether gradient-based optimization of a differentiable articulatory model can recover tract shapes that both reproduce recorded *sygyt* acoustics and align with the established formant-focusing mechanism.

We address this question using differential models to produce high-quality synthetic records to match the originals (Articulatory Copy Synthesis). We confront differential digital signal processing (DDSP) and an extended differentiable articulatory Kelly–Lochbaum model [8]. The proposed formulation introduces a sublingual secondary source, a cubic B-spline tract parameterization, spatially varying learnable damping, and an overtone-salience loss that emphasizes harmonic consistency in the target overtone region. The goal is to fit *sygyt* spectra and to determine a physically grounded model that matches acoustic features while maintaining interpretable parameters.

We evaluate the methods on 20 studio-recorded *sygyt* segments drawn from two independent datasets (5 singers, 10 pitches), comparing a Pink Trombone-style articulator-chain baseline [7], the proposed B-spline formulation, and a DDSP baseline. The proposed model improves substantially over the articulatory baseline and yields recovered tract configurations that are acoustically interpretable and qualitatively consistent with formant-focusing behavior in *sygyt*. Additional experiments on independent recordings suggest robustness across performers and pitches.

The main contributions of this work are as follows:

- We formulate *sygyt* copy-synthesis from recorded audio as a differentiable optimization problem.
- We introduce an extended Kelly–Lochbaum model that features a sublingual secondary source, smooth tract control, varying damping, and an overtone-aware objective.

- We demonstrate, on two independent datasets, that this formulation outperforms a standard articulatory baseline and matches the performance of a DDSP-based approach, with supporting subjective preference tests.
- We show that the optimized solutions reproduce the spectro-temporal signature of *sygyt*. Ablation results indicate that the sublingual source is the dominant contributor. Subjective tests and objective metrics support our results.

2. RELATED WORK

Magnetic resonance imaging (MRI) has played a central role in revealing the articulatory configurations underlying the highly selective resonances of overtone singing. Fig. 2 provides evidence showing synchronized midsagittal MRI data (left) alongside acoustic analysis (right) for a recorded segment of *sygyt* [1]. These measurements provide the link between vocal-tract shapes and the spectra for the sounds that characterize this singing style.

Earlier studies sought to formalize these relationships between articulation and acoustics. For instance, Laver [9] related vocal-tract configurations to long-term average spectra (LTAS) to mitigate noise effects and spurious variations. Subsequent work employed principal component analysis (PCA) to link spectral deviations with articulatory and resonant variations [10]. Such approaches are particularly revealing for overtone singing styles like *sygyt*, where the acoustic output depends on highly constrained resonant structures and fine posterior tongue adjustments. However, these studies remain primarily descriptive, lacking the controllable or differentiable synthesis frameworks required for direct inverse parameter estimation.

2.1. Acoustic Analysis of Biphonic Singing

Early acoustic studies sought to explain how overtone singing isolates and amplifies individual harmonics. Bloothoof *et al.* [11] interpreted the overtone as arising from closely spaced formants, while Klingholz [12] reported substantial formant bandwidth reduction, with resonances narrowing to approximately 20 Hz under pharyngeal constriction. Adachi and Yamada [13] further showed that *sygyt* production can be explained by the convergence of the second and third formants ($F_2 \approx F_3$), thereby creating a narrow spectral peak that selectively reinforces one harmonic component.

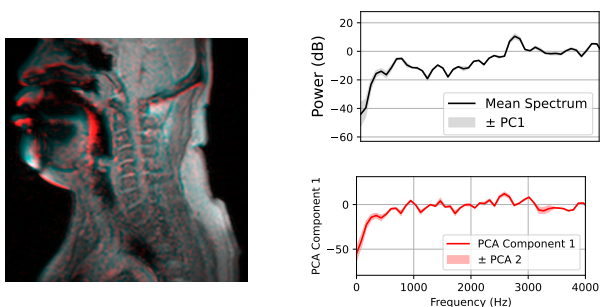


Figure 2: Visualizing articulatory and acoustic dynamics in *sygyt* [1]. **Left:** Average MRI frames (gray) overlaid with PC1 (red) to capture dominant temporal variations beyond the mean variance. **Right:** Spectrograms featuring LTAS (black) and acoustic PCA (red) displayed above for spectral comparison.

Subsequent work provided a more detailed articulatory interpretation of these acoustic observations. Kob [14] used impedance analysis to show that large overtone amplification can arise from the coincidence of resonances, while Bergevin *et al.* [1], combining dynamic MRI and studio-quality recordings, argued that *sygyt* can largely be explained by linear filtering, that is, by precise vocal-tract shaping rather than by an additional nonlinear sound source. Sundberg *et al.* [15] and Hefele *et al.* [16] further documented the relevant vocal-tract configurations and their acoustic consequences. These studies indicate that overtone singing, and *sygyt* in particular, relies on sharp resonant focusing and strong spectral sensitivity to small articulatory changes.

2.2. Synthesis of Overtone Singing

Several synthesis approaches have been proposed for overtone singing, although they differ substantially in their degree of physical grounding and parameter interpretability. Kob [2] achieved formant-based synthesis of *khoomi* by manually specifying formant frequencies and bandwidths, obtaining perceptually convincing results at the cost of expert parameter tuning. Tsai *et al.* [5] transformed conventional singing into overtone singing through source–filter decomposition and spectral-envelope modification, but without recovering an explicit vocal-tract configuration.

Articulatory synthesis offers a more interpretable alternative; control variables correspond directly to vocal production. Classical articulatory models [17, 18, 19, 20] simulate realistic vocal-tract acoustics and have been successfully used in speech production studies. Story’s TubeTalkerPkg [18], for instance, was employed by Bergevin *et al.* [1] to study formant focusing through forward synthesis from prescribed tube-area functions. Similarly, the Pink Trombone real-time synthesizer [21] implements an interactive Kelly–Lochbaum-style waveguide model. These systems are primarily designed for forward control and exploration rather than inverse estimation. In overtone singing, particularly in *sygyt*, small changes in tract geometry can produce changes in overtone placement and bandwidth [1, 12, 13], degrading our results.

2.3. Differentiable Sound Synthesis

Recent work in differentiable audio synthesis [22, 23] has shown that gradient-based optimization can be combined with structured synthesis models to perform copy-synthesis while preserving meaningful control parameters. Most of this literature has focused on neural synthesizers, differentiable signal processors, or hybrid physical–neural architectures. By contrast, articulatory waveguide models expose parameters with direct physical interpretation, such as cross-sectional areas, damping coefficients, and glottal source characteristics.

Applying differentiable optimization to overtone singing remains challenging because of the strongly underdetermined inverse mapping from sound to articulation. The target acoustics depend on narrow, high-Q resonances that are highly sensitive. Prior articulatory models have supported forward analysis and formant-targeted synthesis, but we found no evidence of gradient-based copy-synthesis of recorded overtone singing using a differentiable articulatory model with physically interpretable parameters. Likewise, we found no prior computational study directly testing whether gradient-based optimization can recover the formant-merging configurations reported in acoustic analyses of *sygyt* [13, 1]. This gap motivates the present work.

3. METHOD

This section describes the proposed *sygyt* copy-synthesis, including the DDSP baseline and the physically-informed differentiable waveguide architecture.

3.1. DDSP baseline

As a non-physical reference with direct spectral control, we include a DDSP-style synthesizer [22]. The model generates audio from multiple oscillators. Starting on the primary f_0 track, we sum 60 sinusoidal harmonics with learnable amplitudes, plus filtered noise from a 257-bin linear-phase FIR filter (Hann window, FFT size 512, 50% overlap). Each harmonic amplitude is independently optimized by gradient descent with respect to the target recording, thereby giving the model direct control over the energy at each frequency. Each frame has 318 free parameters (60 harmonic amplitudes + 1 overall amplitude + 257 noise-filter coefficients). The DDSP synthesizer operates in signal space rather than physical space. This baseline provides a complementary perspective to the psychologically-informed, which must discover spectral structure through tract geometry.

3.2. Differentiable source-filter model

Extended waveguide. Our synthesizer extends the differentiable model and efficient JAX implementation of VocalTrax [24]. We included three architectural additions: a nasal tract, a sublingual second glottal source, and learnable spatially varying damping. The vocal tract is discretized into three coupled waveguide sections for the **oral tract**—44 cylindrical segments from glottis to lips (following [24])—the **nasal tract**—28 segments from velum to nostrils—, and the **sublingual tube**—15 segments connecting the secondary source to the oral tract; segment counts follow the discretization in [24] and the simulation runs at 16 kHz.

Two **three-way junctions** couple the waveguide sections. At the sublingual junction (oral segment 9), three branches meet: the oral tract splits into a proximal section (segments 0–8) and a distal section (segments 10–43), connected to the sublingual tube. At the velum junction (segment 17), the oral and nasal tracts are coupled.

Propagation. Wave propagation follows the standard Kelly-Lochbaum scattering equations. At each segment boundary i , leftward (L) and rightward (R) traveling waves interact via reflection coefficient $r_i = (A_i - A_{i+1}) / (A_i + A_{i+1})$, where A_i is the cross-sectional area of segment i . In the standard Kelly-Lochbaum model, damping is applied uniformly across all segments. We extend this to a **learnable damping** $d_i \in [0.99, 0.9999]$ for the oral and sublingual cavity: $R_i \leftarrow R_i \cdot d_i$, $L_i \leftarrow L_i \cdot d_i$.

This provides fine-grained Q-factor control: values closer to unity ($d \rightarrow 0.9999$) preserve wave energy and sustain sharper resonances (higher Q-factor), while lower values ($d \rightarrow 0.99$) increase energy dissipation and broaden formant bandwidth (lower Q-factor). The optimizer can selectively sharpen formants near the overtone frequency while attenuating competing resonances elsewhere. Physically, d_i models frequency-dependent losses from viscous boundary layers, heat conduction, and soft tissue absorption, which vary along the vocal tract due to differences in tissue composition, surface area, and constriction geometry [17].

The scattering at each three-way junction depends on the areas

of the three connecting segments, A_L, A_R, A_B :

$$\Sigma = A_L + A_R + A_B, \quad r_X = \frac{2A_X - \Sigma}{\Sigma}, \quad X \in \{L, R, B\} \quad (1)$$

Sources: tone and overtone. The glottal and sublingual sources provide the excitations to the waveguide. The latter focuses on the overtone frequency f_{ot} , injected into the vocal tract via the sublingual junction. This models the hypothesis that diphonic singing involves a secondary vibrating structure below the tongue [3].

Sources use a Liljencrants-Fant (LF) waveform with learnable parameters: **amplitude** $a \in [0.1, 1.0]$ —relative to primary source—, **tenseness** $\tau \in [0.1, 1.0]$ —controls spectral slope—, **open quotient offset** $\Delta OQ \in [-0.15, +0.15]$ —, **spectral tilt offset** $\Delta tilt \in [-2, +2]$.

The secondary f_{ot} is extracted from the target audio as described in Section 3.4 and held fixed during optimization. Crucially, the sublingual tube introduces a *second resonant cavity* parallel to the oral tract, whose interaction with the main tract at the three-way junction produces additional spectral features beyond what a single-tube model can achieve.

3.3. Tract Parameterizations

We compare two strategies for controlling the vocal tract. DOFs stand for Degrees of Freedom assigned in the models involving the parameters for the waveguide:

Articulator chain (~ 13 DOFs/frame). Following the classical configuration for the Kelly-Lochbaum [7], the tract shape is controlled through a list of physiologically-informed articulators: tongue position and diameter (2), throat constriction (1), lip rounding (1), velum opening (1), sublingual coupling (1), and global damping (1). This parameterization guarantees human-like shapes by construction but limits the optimizer’s ability to form the narrow resonances required for overtone focusing. The nasal waveguide is not parameterized but included in the model to incorporate nasalization effects and velum movement.

Cubic B-spline spatial basis (~ 70 DOFs/frame). Unconstrained per-segment optimization produces physically implausible discontinuities. Following an MRI study of vocal-tract geometry during overtone singing, we represent both the oral and sublingual diameter profiles, as well as their damping profiles, using separate sets of cubic B-spline control points placed at anatomically motivated landmarks (glottis, sublingual junction, velum, and lips). While the cubic basis increases the number of coefficients, it ensures C^2 continuity and remains physically plausible.

Let $\alpha_A \in \mathbb{R}^{K_A}$ and $\alpha_d \in \mathbb{R}^{K_d}$ be the diameter and damping coefficients respectively, where $K = 20$ (empirically selected attending to our analysis of the implementation in [24]) are the B-spline control points placed at uniform knot spacing. Segment-wise profiles are obtained by evaluating the B-spline bases at segment centers:

$$\mathbf{A} = B_A \alpha_A, \quad \mathbf{d} = \text{clip}(B_d \alpha_d, 0.99, 0.9999), \quad (2)$$

where B_A and B_d are the basis matrices (cubic B-spline bases of size $N \times K_A$ and $N \times K_d$ evaluated at the N segment centers). Area and damping are applied locally during synthesis but controlled by smooth B-spline parameterizations rather than independent per-segment variables.

Both parameterizations share identical primary and secondary source models (5 each), as well as the same loss functions and

optimizer settings. The only difference is how the vocal tract shape and damping are parameterized.

3.4. Computation

Tones extraction. Both the primary fundamental frequency f_0 and the overtone frequency f_{ot} are extracted from the target audio in a preprocessing step shared by all synthesis conditions. The primary f_0 is estimated using the YIN algorithm [25].

The overtone frequency is identified via spectral peak analysis on a high-resolution STFT ($n_{\text{fft}} = 4096$). For each voiced frame, harmonics H_2 – H_{10} of the primary f_0 are examined: the *enhancement* of harmonic H_k is defined as the difference between its measured energy and the energy expected under a natural spectral rolloff of 6 dB per octave. A harmonic is flagged as the active overtone when its enhancement exceeds 6 dB; if several harmonics exceed the threshold, the one with the largest enhancement is selected.

To ensure temporal coherence, two mechanisms prevent erratic harmonic switching: (1) hysteresis requiring >3 dB improvement before changing the active harmonic, and (2) Savitzky-Golay smoothing (window = 11, order = 2) of the harmonic-number track. Both f_0 and f_{ot} trajectories are held fixed during optimization — the synthesis methods differ only in how they use them.

Optimization. Copy-synthesis optimizes tract shape and source parameters to minimize the reconstruction loss:

$$\mathcal{L} = \mathcal{L}_{\text{STFT}} + \lambda_{\text{mel}} \mathcal{L}_{\text{mel}} + \lambda_{\text{harm}} \mathcal{L}_{\text{harm}} + \lambda_{\text{energy}} \mathcal{L}_{\text{energy}} + \lambda_{\text{ot}} \mathcal{L}_{\text{ot}} \quad (3)$$

In our experiments, we set all loss weights to unity ($\lambda_{\text{mel}} = \lambda_{\text{harm}} = \lambda_{\text{energy}} = \lambda_{\text{ot}} = 1.0$).

The multi-resolution STFT loss $\mathcal{L}_{\text{STFT}}$ uses FFT sizes {512, 1024, 2048} with both spectral convergence (Frobenius norm) and log-magnitude (L1) terms [22]. The mel loss $\mathcal{L}_{\text{mel}}$ operates on 80-band mel spectrograms, providing perceptually-weighted frequency resolution. The harmonic energy loss $\mathcal{L}_{\text{harm}}$ matches energy at integer multiples of f_0 to preserve the harmonic structure critical for overtone singing. The overtone salience loss $\mathcal{L}_{\text{ot}}$ penalizes deviation between the synthesized and target overtone:

$$S_{\text{ot}}(t) = 10 \log_{10} \frac{\sum_{f \in B_h} |X(f, t)|^2}{\sum_{f \in B_{\pm}} |X(f, t)|^2 + \epsilon} \quad (4)$$

where B_h is a frequency band of half-width 50 Hz centered on the target harmonic $h(t) \cdot f_0(t)$, $h(t) \in \{2, \dots, 10\}$ is the active overtone index, and $B_{\pm}$ covers the two adjacent harmonics at $(h \pm 1) f_0(t)$ with the same 50 Hz half-width; $\epsilon = 10^{-10}$ avoids division by zero. The overtone-salience loss is then $\mathcal{L}_{\text{ot}} = \|S_{\text{ot}}^{\text{target}} - S_{\text{ot}}^{\text{synth}}\|_2^2$.

All conditions are optimized with Adam using a learning rate of $\eta = 5 \times 10^{-3}$, preceded by a 100-step warm-up from 5×10^{-4} , gradient clipping at a maximum norm of 1.0, and a 500 iterations budget. The warm-up stage is important for reducing early instability caused by strongly nonlinear waveguide dynamics. The full pipeline is differentiable via JAX, enabling gradient computation for the sample-by-sample waveguide simulation, scattering junctions, and LF source generation. All experiments are run on CPU (AMD Ryzen 7), requiring approximately 30 minutes for a 500-iteration optimization at 16 kHz. Table 1 summarizes the shared experimental protocol.

Table 1: Experimental protocol. Components above the mid-rule are shared across all conditions; components below differ.

Component	Waveguide	DDSP
Target input	Audio recording	
f_0 extraction	YIN + overtone det. (fixed)	
Optimizer	Adam ($\eta = 5 \times 10^{-3}$, warmup)	
Iterations	500	
$\mathcal{L}_{\text{STFT}} + \mathcal{L}_{\text{mel}} + \mathcal{L}_{\text{energy}}$	✓	
Synthesis engine	Kelly-Lochbaum waveguide	Harmonic + noise
Parameterization	Articulator / B-spline	60 harm. amps + 257-bin FIR
Pitch structure	Dual ($f_0 + f_{ot}$)	Single (f_0)
$\mathcal{L}_{\text{ot}}$ (overtone)	✓	—

4. EXPERIMENTAL RESULTS

4.1. Dataset and Setup

We evaluate copy-synthesis on 20 recordings of Tuvan *sygyt* (overtone singing) drawn from two independent studio-recorded datasets. The first comprises 10 recordings from the HFA Overtone Singing Preview dataset [26], spanning 10 pitches from F3 to Eb4 and covering varied harmonic patterns, including arpeggios, intervals, scales, and glissandi; each segment is 5.0 s long. The second comprises 10 excerpts from Bergevin et al. [1], featuring four Tuvan singers (T1–T4) recorded in a sound booth. Both datasets were originally sampled at 96 kHz and downsampled to 16 kHz for synthesis, with mean segment durations of 5.0 s (HFA) and 4.6 ± 1.7 s (Bergevin).

Copy-synthesis is particularly suitable when dealing with short recordings as it focuses on the specifics of individual frames. Each audio file is optimized independently from a single random initialization, without restarts, for up to 500 Adam iterations. For each recording, the reported metrics are computed at the iteration with the lowest training loss. All reported values are means $\pm$ standard deviations across recordings within each dataset.

We evaluate log-spectral distance (LSD), spectral Pearson correlation (SpCorr: correlation between flattened log-magnitude spectrograms of target and synthesized signals), PESQ (Perceptual Evaluation of Speech Quality; ITU-T P.862, MOS 1–5) [27], CD-PAM (Concatenated Deep Perceptual Audio Metric; learned distance metric, lower=better) [28], ViSQOL (Virtual Speech Quality Objective Listener; spectro-temporal similarity, MOS-like 1–5) [29], plus mean subjective ratings from MUSHRA tests [30] for overall quality (Q1) and harmonic similarity (Q2); as well as the energy ratio overtone salience, and harmonic prominence ratio (HPR, defined as the per-frame ratio of peak to mean magnitude inside the 1–3 kHz band, averaged over time) in the overtone region. The MUSHRA-inspired listening test was run with 23 listeners (3 expert, 2 trained, 18 novice) on 6 segments (3 per dataset) in a blind A/B/C rating with randomized order.

4.2. Pitch and Overtone Detection

The dual- f_0 extraction pipeline described in Section 3.4 successfully detects diphonic content in all 20 recordings. Across voiced frames, 99% are classified as diphonic (enhancement > 6 dB), with a mean enhancement of 37.6 dB. The detected overtone trajectories correspond to harmonic numbers ranging from H_2 to H_{10} , consistent with the expected *sygyt* overtone range [1]. These extracted primary- f_0 and overtone trajectories are held fixed across all synthesis conditions.

Table 2: Copy-synthesis results on 20 segments of Tuvan sygyt from two independent datasets. Objective metrics: mean $\pm$ std across N segments. Best in **bold**. LSD and SpCorr are signal-level metrics; PESQ (ITU-T P.862.2; wideband MOS-LQO), CDPAM (learned perceptual distance, lower=closer), and ViSQOL (audio-mode MOS-LQO, 1–5) are objective perceptual metrics. Q1 (overall quality) and Q2 (harmonic similarity) are subjective scores (mean $\pm$ 95% CI, 0–100 scale, 23 listeners $\times$ 3 segments per dataset). DOFs are total trainable parameters per audio segment (per-frame tract coefficients plus global source parameters). The per-frame counts (13 articulator, 70 B-spline) appear in Sec. 3.3.

	DOFs	N	LSD (dB) $\downarrow$	SpCorr $\uparrow$	PESQ $\uparrow$	CDPAM $\downarrow$	ViSQOL $\uparrow$	Q1 $\uparrow$	Q2 $\uparrow$
<i>HFA [26] — 1 singer, 10 pitches (F3–Eb4)</i>									
Artic. chain	19	10	13.84 $\pm$ 0.54	0.71 $\pm$ 0.04	1.17 $\pm$ 0.15	3.06 $\pm$ 3.34	2.81 $\pm$ 0.53	13.6 $\pm$ 3.7	25.3 $\pm$ 4.5
DDSP	$\sim$ 100k	10	10.99 $\pm$ 0.54	0.82 $\pm$ 0.01	1.20 $\pm$ 0.13	3.56 $\pm$ 3.71	3.81 $\pm$ 0.21	42.4 $\pm$ 5.0	50.3 $\pm$ 4.9
B-spline	86	10	9.64 $\pm$ 0.29	0.86 $\pm$ 0.02	1.37 $\pm$ 0.34	2.36 $\pm$ 3.07	3.60 $\pm$ 0.22	44.8 $\pm$ 5.5	52.9 $\pm$ 5.8
<i>Bergevin [1] — 4 singers (T1–T4)</i>									
Artic. chain	19	10	14.53 $\pm$ 0.93	0.66 $\pm$ 0.05	1.10 $\pm$ 0.05	0.73 $\pm$ 0.41	2.88 $\pm$ 0.21	11.6 $\pm$ 3.6	15.7 $\pm$ 3.9
DDSP	$\sim$ 100k	10	10.71 $\pm$ 0.72	0.83 $\pm$ 0.03	1.26 $\pm$ 0.13	0.77 $\pm$ 0.55	3.69 $\pm$ 0.33	35.4 $\pm$ 5.5	45.1 $\pm$ 5.4
B-spline	86	10	9.04 $\pm$ 0.46	0.88 $\pm$ 0.01	1.58 $\pm$ 0.35	0.60 $\pm$ 0.34	3.85 $\pm$ 0.23	38.5 $\pm$ 5.3	46.3 $\pm$ 5.6

4.3. Overall Quality of Copies

Table 2 summarizes copy-synthesis performance across all 20 recordings for three methods: articulator chain, DDSP, and B-spline waveguide. All metrics favor the B-spline, outperforming the articulator chain and, to a lesser extent, the baseline DDSP.

LSD shows B-spline reduces error by 30% on HFA recordings and 38% on Bergevin recordings, relative to articulator chain. B-spline also achieves lower LSD than DDSP on both datasets, with the lowest values on all 20 individual segments —see Fig. 3 confronting LSD measures for B-spline with the articulator chain and DDSP methods on the HFA (blue) and Bergevin (red) datasets. The results yield large effect sizes (standardized differences in group means, with Cohen’s $d = 2.1$ for HFA and $d = 1.9$ for Bergevin) with strong practical significance.

Objective perceptual metrics corroborate these findings on PESQ and CDPAM. On ViSQOL B-spline and DDSP largely outperform the articulator chain. Subjectively, B-spline scored highest for overall quality and overtone similarity across datasets, followed by DDSP and the articulator chain. These subjective results corroborate the objective perceptual metrics, confirming that the spectral advantages of the B-spline waveguide are perceptible to human listeners.

Friedman tests confirm significant method effects ($p < 10^{-4}$). B-spline and DDSP significantly outperform articulator chain (Wilcoxon $p < 0.001$, Cliff’s $\delta > 0.7$). B-spline and DDSP do not differ significantly under MUSHRA ($p = 0.19$); B-spline does, however, achieve consistently lower objective error (LSD, cepstral peak) than both baselines.

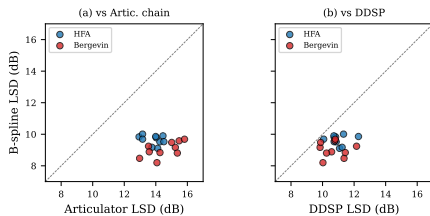


Figure 3: Per-segment LSD comparison. Each point corresponds to one of the 20 recordings. Points below the diagonal indicate B-spline superiority. All 20 segments favor B-spline.

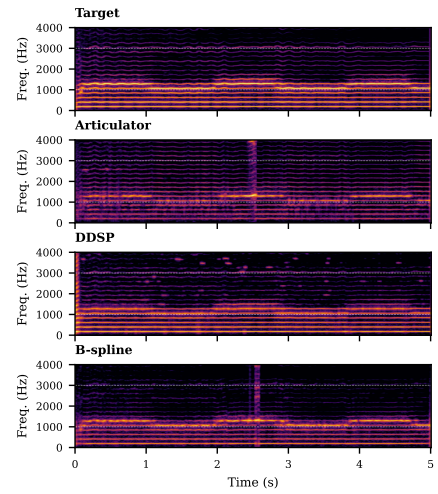


Figure 4: Spectrogram comparison (0–4 kHz) of the same recording ($h_5 \leftrightarrow h_6$). Rows: target, articulator chain, DDSP, and B-spline. Dashed lines indicate the 1–3 kHz overtone region. The short broadband transient at the start of the DDSP row is an initialization artifact of the harmonic-plus-noise oscillator bank and is excluded from all reported metrics.

4.4. Acoustic Analysis

Overtone analysis. Table 3 reports per-segment absolute errors to the target in the 1–3 kHz overtone band. B-spline yields the lowest error in eR , S_{ot} and HPR across both datasets, with overall errors of $|\Delta eR| = 0.12 \pm 0.07$, $|\Delta S_{ot}| = 0.96 \pm 0.70$ dB and $|\Delta HPR| = 5.5 \pm 2.0$, clearly below articulator (0.36, 2.18 dB, 10.8) and DDSP (0.35, 1.70 dB, 12.4). DDSP edges B-spline in $|\Delta \text{SpCorr}_{OT}|$ (0.12 vs. 0.15 overall), consistent with its direct per-harmonic amplitude control; Fig. 4 visualises the corresponding $h_5 \leftrightarrow h_6$ overtone ridge. DDSP thus matches the local spectrum better, but B-spline preserves the global prominence and spectral balance characteristic of sygyt.

Formant analysis. Cepstral envelope analysis shows that B-spline more accurately reproduces the target formant-merging pattern, a hallmark of sygyt production in which F_2 and F_3 converge to am-

Table 3: Overtone-region errors $|\Delta|$ to the per-segment target (mean $\pm$ 1 std across segments, lower is better). $|\Delta\text{SpCorr}_{\text{OT}}| = 1 - \text{SpCorr}_{\text{OT}}$; remaining columns are $|x_{\text{synth}} - x_{\text{target}}|$ for the Bergevin energy ratio eR (1–2/0–8 kHz), overtone salience S_{ot} (dB) and HPR. Best in **bold**.

Dataset	Method	$ \Delta\text{SpCorr}_{\text{OT}} \downarrow$	$ \Delta eR \downarrow$	$ \Delta S_{\text{ot}} \text{ (dB)}\downarrow$	$ \Delta\text{HPR} \downarrow$
HFA	Artic. chain	0.21 $\pm$ 0.05	0.36 $\pm$ 0.12	1.83 $\pm$ 1.07	14.16 $\pm$ 7.77
	DDSP	0.10 $\pm$ 0.02	0.46 $\pm$ 0.11	1.68 $\pm$ 0.84	15.69 $\pm$ 3.39
	B-spline	0.12 $\pm$ 0.03	0.12 $\pm$ 0.05	0.88 $\pm$ 0.66	5.83 $\pm$ 2.31
Berg.	Artic. chain	0.48 $\pm$ 0.21	0.36 $\pm$ 0.16	2.53 $\pm$ 1.28	7.52 $\pm$ 3.98
	DDSP	0.14 $\pm$ 0.08	0.23 $\pm$ 0.20	1.73 $\pm$ 1.07	9.16 $\pm$ 4.62
	B-spline	0.18 $\pm$ 0.12	0.11 $\pm$ 0.08	1.04 $\pm$ 0.73	5.13 $\pm$ 1.47
All	Artic. chain	0.34 $\pm$ 0.21	0.36 $\pm$ 0.14	2.18 $\pm$ 1.23	10.84 $\pm$ 7.01
	DDSP	0.12 $\pm$ 0.06	0.35 $\pm$ 0.20	1.70 $\pm$ 0.97	12.43 $\pm$ 5.20
	B-spline	0.15 $\pm$ 0.09	0.12 $\pm$ 0.07	0.96 $\pm$ 0.70	5.48 $\pm$ 1.97

ply a single overtone [1] (example illustrated in Fig. 5). Its C^2 -continuous tract profiles place the merged peak closer to the target than the articulator chain, whose coarse tongue-position control limits its ability to form the narrow constrictions required for precise formant placement. This formant-merging behavior is not explicitly imposed by the loss function. Rather, it emerges from optimization and is consistent with the formant-focusing account described by Bergevin *et al.* [1] (see Section 2.1). Quantitatively, the cepstral peak location of the B-spline envelope tracks the target peak with the lowest error of the three methods (see below), which we read as evidence of more accurate formant merging.

Across all 20 recordings, B-spline achieves a mean formant-peak error of 28 Hz (measured as the absolute deviation between the cepstral-envelope peak locations of target and synthesis in the 1–3 kHz band) relative to the target cepstral peak location in the 1–3 kHz overtone region, compared with 222 Hz for the articulator chain and 120 Hz for DDSP. Its formant peak prominence (12.0 dB) also closely matches that of the target recordings (12.4 dB), suggesting that the optimizer captures the location of the merged peak and part of its spectral prominence.

4.5. Learning Process for Tract Parameterizations

We examine learned parameters from the physically-grounded baselines, B-spline and articulator chain, excluding DDSP due to its signal-domain harmonic parameterization, which does not yield

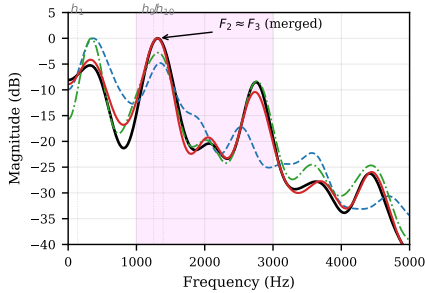


Figure 5: Cepstral spectral envelopes ($N_{\text{cep}} = 30$) for the Bergevin T1 segment at $t = 1.9$ s (midpoint). Curves: target (black, solid), articulator chain (blue, dashed), DDSP (green, dash-dot), B-spline (red, solid).

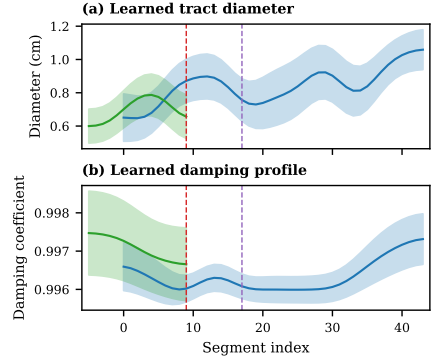


Figure 6: Learned B-spline profiles for a representative segment (A3 interval): oral and sublingual diameters (top) and dampings (bottom) along the tract. Solid lines: oral (blue), sublingual (green); dashed verticals: sublingual junction (red), velum (purple); shaded bands: ± 1 SD across segments.

tract configurations amenable to acoustic interpretation. The articulator chain exposes a single global damping coefficient rather than a per-segment damping profile, so equivalent damping panels cannot be drawn for it; we focus on the B-spline profiles below.

Learned Tract Shape and Damping. The B-spline parameterization yields tract profiles for diameter (top) and damping (bottom), such as the one in Fig. 6. These foster a physically meaningful acoustic interpretation of the oral (blue) and sublingual (green) waveguides. Junctions are marked with dashed lines. The oral tract exhibits pronounced constriction in segment 9 (sublingual junction), consistent with the position of the posterior tongue observed in MRI studies of sygyt production [13]. This configuration emerges end-to-end from gradient descent without explicit anatomical constraints. This pattern holds in 16 of 20 optimized segments (80%), covering the overtones h_5 to h_{10} .

The damping profile ($d \in [0.99, 0.9999]$, bottom part of Fig. 6) exhibits complementary spatial structure. Lower values ($d \rightarrow 0.99$) near the sublingual junction increase energy dissipation, effectively isolating the proximal tract from the anterior oral cavity. Higher values ($d \rightarrow 0.9999$) within the anterior resonance cavity sustain narrowband, high- Q resonances (i.e., high quality factor $Q = f_c / \Delta f$, for resonance centre frequency f_c (not the voice f_0) and bandwidth Δf) characteristic of the 1–3 kHz overtone region. The sublingual branch maintains uniformly low damping, consistent with reduced acoustic coupling during overtone production.

Together, these profiles suggest a coherent mechanism for sygyt spectral focusing: tract constriction aligns relevant formants, junction damping limits energy backscatter, and the anterior cavity supports overtone resonance—all without introducing physiology-specific loss terms into the training losses.

Convergence. Despite substantially more learnable parameters (70 vs. 13 DOFs/frame), B-spline optimization converges at rates comparable to the articulator chain baseline (Fig. 7). Through the initial 300 iterations, B-spline maintains lower LSD and higher spectral correlation throughout the trajectory. These dynamics indicate that the flexible tract representation facilitates discovery of acoustically interpretable configurations within equivalent computational budgets. Optimization is offline (~ 30 min per 5 s segment on CPU, RTF $\approx 360\times$); real-time use is out of scope.

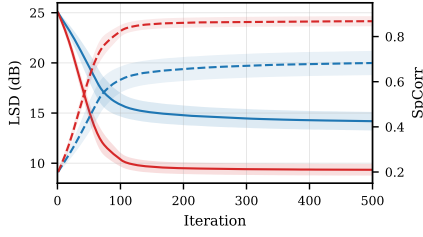


Figure 7: Convergence analysis on both physical-model synthesis, articulator chain (blue) and B-spline (red), averaged across all segments (shading: $\pm 1 \sigma$). Solid lines show LSD (left axis, dB $\downarrow$); dashed lines show SpCorr (right axis, $\uparrow$).

5. ABLATION STUDY

To disentangle the contributions of the sublingual second source and spatially variable damping, we perform a 2×2 factorial ablation of the B-spline condition across all 20 recordings (Table 4). This ablation targets architectural components; a separate ablation of the loss terms $\mathcal{L}_{\text{mel}}$, $\mathcal{L}_{\text{harm}}$, $\mathcal{L}_{\text{energy}}$, $\mathcal{L}_{\text{ot}}$ is left to future work. Removing the sublingual source increases LSD more strongly (+1.0 dB) than removing spatially variable damping (+0.1 dB), indicating that the secondary excitation mechanism is the stronger contributor under this experimental setup. Spatially variable damping still provides a consistent but smaller improvement. The minimal condition, with neither sublingual source nor spatially variable damping, performs worst overall. The combined pattern also suggests only a limited interaction between both components, as each provides a largely separable benefit.

6. DISCUSSION

Our model includes a sublingual tube that, while not anatomically human, provides a controllable physical system for studying the acoustics of diphonic singing. We treat the sublingual branch as an *acoustic abstraction*, not as a literal anatomical claim: it is one way to introduce the additional spectral degree of freedom that Bergevin et al. [1] attribute to precise linear filtering of a single source. The 2×2 factorial ablation (Section 5) indicates that the sublingual source is the stronger contributor under the present experimental setup, with a substantially larger effect on LSD than spatially variable damping. This suggests that the secondary excitation mechanism has a larger impact on the reported reconstruction metrics than fine-grained Q-factor control alone. Spatially variable damping provides a slight secondary improvement, suggesting that tract shaping and formant focusing remain relevant

Table 4: 2×2 factorial ablation (all 20 segments, 500 iterations). Sub. = sublingual second source; Damp. = per-segment damping. Values are mean $\pm$ std across recordings.

Condition	Sub.	Damp.	LSD (dB) $\downarrow$	SpCorr $\uparrow$	SpCorr $\text{OT}\uparrow$
Full	✓	✓	9.34$\pm$0.49	0.86$\pm$0.02	0.85$\pm$0.09
No sublingual	–	✓	10.32 $\pm$ 0.56	0.82 $\pm$ 0.03	0.81 $\pm$ 0.13
No damping	✓	–	9.48 $\pm$ 0.48	0.86 $\pm$ 0.02	0.85 $\pm$ 0.09
Minimal	–	–	10.45 $\pm$ 0.61	0.82 $\pm$ 0.03	0.81 $\pm$ 0.14

even when the secondary source dominates the metric gains. This is a relevant result that we wanted to track for future work.

The DDSP comparison tests whether physical modeling remains useful relative to an unconstrained signal-space baseline. DDSP offers direct per-harmonic amplitude control without any tract-based inductive bias, yet the waveguide achieves lower reconstruction error overall. This suggests that explicit acoustic structure is beneficial in this copy-synthesis setting. More specifically, the waveguide constrains the solution to spectra that are realizable through tube resonances and source–filter interaction, while still allowing the optimizer to discover tract configurations that are interpretable in acoustic terms. We therefore view the advantage of the physical model not as proof of a specific physiological mechanism, but as evidence that a structured waveguide representation is a useful bias for overtone-singing reconstruction.

6.1. Limitations

Six key limitations for *sygyt* copy-synthesis warrant discussion: (1) moderate absolute LSD (9–15 dB) reflects the intrinsic difficulty of matching overtone-singing spectra with a simplified physical model, but relative 30–38% gains over baseline show efficacy; (2) sublingual source is only an approximation to complex aeroacoustic interactions in diphonic singing (+1.0 dB ablation gain), where some secondary pitches may arise without distinct vibrating structures; (3) while $N = 20$ curated segments from two independent datasets (5 singers, 10 pitches) effectively demonstrate proof-of-concept across varied performers and pitches—with low per-recording standard deviation ($\sigma_{\text{LSD}} < 0.93$ dB)—and prioritize pattern diversity (arpeggios, glissandi, steady tones) over volume, future work should extend our analysis; the present scale prioritizes diversity over volume, aligning with the niche nature of high-quality *sygyt* data; (4) *sygyt*-specific scope excludes other *khoomei* styles like *kargyraa* or uncured field recordings for broader robustness; (5) fixed pitch extraction limits glissandi but achieves 86% correlation, fixable via joint optimization; plus uncharacterized real-time potential post-optimization; (6) DDSP uses a neural encoder vs. our per-frame parameters (both run to convergence with the same loss), and porting our sublingual mechanism to DDSP requires non-trivial architectural changes – encoder-matched DDSP and an SVS extension with unfrozen f_0 /overtone are left to future work.

7. CONCLUSION

We have presented copy-synthesis methods for *sygyt*-style singing, based on differentiable DSP and a differentiable Kelly–Lochbaum waveguide synthesizer extended for diphonic singing. The model combines a sublingual second glottal source with a cubic B-spline spatial parameterization. The B-spline basis uses a compact set of control points for oral-tract diameter and damping, yielding smooth tract profiles and enabling spatially structured learnable damping for fine-grained control of resonant behavior.

Copy-synthesis experiments from two independent datasets (5 singers, 10 pitches) show that the B-spline parameterization reduces LSD by 30–38% relative to the articulator-chain baseline, while also outperforming a DDSP baseline in overall LSD and global spectral correlation. The gain is concentrated in the 1–3 kHz overtone region, where the model achieves an average improvement of 4.7 ± 1.8 dB over the articulator chain and reproduces the target merged-formant peak with a mean error of 28 Hz. The

learned tract profiles also admit a plausible acoustic interpretation, with constrictions and localized damping structure consistent with overtone-focused spectral shaping. A 2×2 factorial ablation further shows that the sublingual second source is the stronger contributor under the present setup, while spatially structured damping provides a smaller complementary benefit.

8. ACKNOWLEDGMENTS AND SUPP. MATERIALS

The authors would like to thank Christopher Bergevin for their valuable contribution to this work. Audio examples and supplementary materials are available on the companion website.¹ This work was supported by the Ministry of Economy and Competitiveness of Spain under grant PID2021-128469OB-I00, and «Ayuda Económica Para Personal Investigador Postdoctoral 2025» of «Fundación Santander».

9. REFERENCES

- [1] C. Bergevin, C. Narayan, J. Williams, N. Bhatt, A. Bhatt, and N. P. Todd, “Overtone focusing in biphonic Tuvan throat singing,” *eLife*, vol. 9, p. e50476, 2020.
- [2] M. Kob, “Analysis and modelling of overtone singing in the vocal tract,” in *Proc. Forum Acusticum*, 2002.
- [3] K. Sakakibara, T. Konishi, H. Imagawa, K. Kondo, E. Z. Murano, M. Kumada, and S. Niimi, “Observation of laryngeal movements for period-doubling voices using high-speed digital imaging,” in *Proc. Eurospeech*, 2001.
- [4] K. Sakakibara, T. Konishi, K. Kondo, E. Z. Murano, M. Kumada, H. Imagawa, and S. Niimi, “Growl voice in ethnic and pop styles,” in *Proc. ICMC*, 2002.
- [5] C.-G. Tsai, L. C. Wang, S.-F. Wang, Y.-W. Shau, T.-Y. Hsiao, and W. Auhagen, “Aggressiveness of the growl-like timbre: acoustic characteristics, musical implications, and biomechanical mechanisms,” *Music Perception*, vol. 31, no. 3, pp. 209–221, 2014.
- [6] D. Südholt, M. Cámara, Z. Xu, and J. D. Reiss, “Vocal tract area estimation by gradient descent,” in *Proc. 26th DAFx*, 2017, pp. 126 – 133.
- [7] M. Cámara, Z. Xu, Y. Zong, J. L. Blanco, and J. D. Reiss, “Optimization techniques for a physical model of human vocalisation,” in *Proc. 26th DAFx*, 2023, pp. 29–36.
- [8] J. L. Kelly and C. C. Lochbaum, “Speech synthesis,” in *Proc. 4th Int. Congress Acoust.*, 1962.
- [9] J. Laver, *The Gift of Speech: Papers in the Analysis of Speech and Voice*. Edinburgh University Press, 1991.
- [10] J. L. Blanco and J. Schoentgen, “Vocal tract settings in speakers with obstructive sleep apnea syndrome,” in *Proc. MAVEBA*, 2013.
- [11] G. Bloothoof, E. Bringmann, M. van Cappellen, J. B. van Luipen, and K. P. Thomassen, “Acoustics and perception of overtone singing,” *J. Acoust. Soc. Am.*, vol. 92, no. 4, pp. 1827–1836, 1992.
- [12] F. Klingholz, “Overtone singing: Productive mechanisms and acoustic data,” *J. Voice*, vol. 7, no. 2, pp. 118–122, 1993.
- [13] S. Adachi and M. Yamada, “An acoustical study of sound production in biphonic singing, xöömij,” *J. Acoust. Soc. Am.*, vol. 105, no. 5, pp. 2920–2932, 1999.
- [14] M. Kob, “Analysis and modelling of overtone singing in the sygyt style,” *Applied Acoustics*, vol. 65, no. 12, pp. 1249–1259, 2004.
- [15] J. Sundberg, B. Lindblom, and J. Hefele, “Voice source, formant frequencies and vocal tract shape in overtone singing,” *Logopedics Phoniatrics Vocology*, 2023.
- [16] J. Hefele, R. Eklund, and A. McAllister, “Polyphonic overtone singing: MRI analysis,” in *Proc. Fonetik*, 2019.
- [17] M. M. Sondhi and J. Schroeter, “A hybrid time-frequency domain articulatory speech synthesizer,” *IEEE Trans. ASSP*, vol. 33, no. 4, pp. 955–967, 1985.
- [18] B. H. Story, “Phrase-level speech simulation with an airway modulation model of speech production,” *Computer Speech & Language*, vol. 27, no. 4, pp. 989–1010, 2013.
- [19] J. Mullen, D. Howard, and D. Murphy, “Acoustical simulations of the human vocal tract using the 1d and 2d digital waveguide software model,” in *Proc. 7th DAFx*, 2004, pp. 311–314.
- [20] D. T. Murphy, M. Jani, and S. Ternström, “Articulatory vocal tract synthesis in supercollider,” in *18th DAFx*, 2015, pp. 307–313.
- [21] N. Thayer, “Pink Trombone,” <https://dood.al/pinktrombone/>, 2017.
- [22] J. Engel, L. H. Hantrakul, C. Gu, and A. Roberts, “DDSP: Differentiable digital signal processing,” in *Proc. ICLR*, 2020.
- [23] B. Hayes, J. Shier, G. Fazekas, A. McPherson, and C. Saitis, “Sinusoidal frequency estimation by gradient descent,” in *Proc. ICASSP*, 2023.
- [24] L. Mo, M. Cherep, N. Singh, Q. Langford, and P. Maes, “Articulatory synthesis of speech and diverse vocal sounds via optimization,” in *Audio Imagination. NeurIPS Workshop*, 2024.
- [25] A. de Cheveigné and H. Kawahara, “YIN, a fundamental frequency estimator for speech and music,” *J. Acoust. Soc. Am.*, vol. 111, no. 4, pp. 1917–1930, 2002.
- [26] Harmonic Frontier Audio, “Overtone singing preview dataset,” 2024, cC BY-NC 4.0.
- [27] ITU-T, “P.862 – perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs,” 2001.
- [28] P. Manocha, Z. Jin, R. Zhang, and A. Finkelstein, “CDPAM: Contrastive learning of disentangled perceptual audio metrics,” in *Proc. ICASSP*, 2021, pp. 286–290.
- [29] M. Chinen, F. S. C. Lim, J. Skoglund, N. Gureev, F. O’Gorman, and A. Hines, “ViSQOL v3: An open source production ready objective speech and audio quality metric,” in *Proc. QoMEX*, 2020, pp. 1–6.
- [30] ITU-R, “Recommendation BS.1534-3: Method for the subjective assessment of intermediate quality level of audio systems (MUSHRA),” 2014.

¹<https://mateocamara.com/khoomei-supp-materials>