EURASIP Journal on Audio, Speech, and Music Processing

EMPIRICAL RESEARCH Open Access

Parameter optimisation for a physical model of the vocal system

Mateo Cámara[1,2*], José Luis Blanco[1,2] and Joshua D. Reiss[3]

Abstract

This study explores optimisation techniques for refining articulatory parameters in the Pink Trombone, a simplified physical speech synthesiser, to accurately emulate male and female vocal tract characteristics in non-speech sounds. We employ black-box and grey-box approaches, leveraging a genetic optimiser and Mel-spectrogram representations to infer articulatory configurations from human recordings via direct spectral comparison. Optimisation is performed over time windows to ensure temporal coherence, introducing modifications to SOTA objective metrics. We integrate grey-box strategies, incorporating pYIN for fundamental frequency estimation and a ResNet-based neural network as a neural codebook to enhance the optimisation process. Our findings confirm the synthesiser's ability to replicate human vocalisations, achieving superior performance over existing techniques in subjective evaluations. We refined the perceptual metric ViSQOL, providing a calibrated framework for future auditory assessments in physical speech synthesis. These contributions establish a methodology for articulatory parameter estimation, improving synthesis quality and expanding vocalisation modelling and analysis applications.

Keywords Analysis-by-synthesis, Acoustic-to-articulatory inversion, Articulatory copy synthesis, Pink Trombone, Procedural audio

1 Introduction

Articulatory synthesis simulates human vocal production by computationally modelling the physiological components of the vocal tract, such as the tongue, lips, and vocal folds, to produce human vocalisations [1–3]. Articulatory synthesis is one of the fundamental approaches to speech generation. It enhances realism and enables the modelling of non-linguistic human sounds—such as vowels, yawns, laughter, and growls—allowing their use as natural audio effects[1] produced by the human body. These sounds are critical for applications like expressive sound design, virtual avatars, and biomedical research, where precise control over vocal tract dynamics is essential [5, 6]. Unlike data-driven methods that learn statistical mappings from large datasets, articulatory synthesis provides explicit, interpretable control over articulatory parameters. It is uniquely suited for modelling the nuanced mechanics of non-verbal vocalisations [7].

[1] Artistically speaking, the British Broadcasting Corporation defines Audio Effects as any sound that is not speech or music [4]. Therefore, the human vocal tract can also create Audio Effects.

*Correspondence:
Mateo Cámara
mateo.camara@upm.es
[1] Grupo de Aplicaciones del Procesado de Señal, Universidad Politécnica de Madrid, Madrid, Spain
[2] Information Processing and Telecommunications Center, Universidad Politécnica de Madrid, Madrid, Spain
[3] Centre for Digital Music, Queen Mary University of London, London, UK

A central challenge in articulatory synthesis is obtaining accurate articulatory data. Direct measurement techniques like electromagnetic articulography (EMA) or magnetic resonance imaging (MRI) are invasive, costly, and typically restricted to speech-centric research in controlled laboratory settings [3, 8–11]. While most acoustic-to-articulatory inversion (AAI) systems focus on speech and rely on EMA data to train speaker-specific models, such approaches are ill-suited for synthesising non-linguistic sounds, which often involve extreme vocal tract configurations. Instead, model inversion techniques like analysis-by-synthesis (AbS) circumvent direct physiological measurements by estimating articulatory configurations directly from acoustic signals, effectively reversing the speech production process [12, 13].

In our prior work [14], we implemented an AbS framework using the Pink Trombone (PT) [15] synthesiser[2], a real-time physical model of the vocal tract. We selected PT over alternatives like VocalTractLab [16], Maeda [17], or Melmenstein [18] for three reasons: (1) its computational efficiency enables rapid exploration of extreme articulatory configurations (e.g. exaggerated yawns, tense growls), (2) its interactive graphical interface provides intuitive visual feedback for designing, refining non-verbal sounds, akin to tuning audio effects, visualising the inferred settings, and (3) its open-source implementation facilitates reproducibility. Our earlier study systematically evaluated an informed selection of *optimisation algorithms* (e.g. Genetic, Particle Swarm, Least Squares...), *acoustic features* (e.g. Mel, Multiscale Fourier Transforms, Cepstrum...), and *cost functions* to determine the most effective configuration for AAI. While the results confirmed the PT's potential for parameter estimation, synthesised non-speech sounds exhibited abrupt transitions and lacked perceptual realism compared to natural human vocalisations.

The present work builds upon these findings to refine AbS, focusing on synthesising non-linguistic human sounds. Figure 1 describes the process, from pre-processing human and synthesised audio through spectral analysis and parameter estimation (left) to optimisation (centre) and final synthesis (right). Building on the best-performing configurations identified in our previous work, we aim to enhance temporal coherence and perceptual realism, addressing key limitations observed:

- **Limited vocal diversity:** the default configuration of the PT is optimised for a male vocal tract, restricting its ability to model female or diverse vocal characteristics. To overcome this, we modified the PT model to dynamically adjust vocal tract length and formant scaling ratios based on anatomical studies, allowing for an accurate representation of both male and female voices.
- **Post-processing dependency:** prior approach required post-processing filtering to artificially smoothen discontinuities in synthesised outputs. We refined the optimisation windowing strategy and cost function to ensure smoother articulatory transitions, eliminating the need for external filtering.
- **Computational inefficiency and suboptimal convergence:** the previous optimisation process required extensive iterations, leading to slow performance and potentially suboptimal solutions. We addressed this by integrating a ResNet-based neural network to provide a starting point for the optimisation of the articulatory parameters, functioning as a *neural codebook*. Unlike traditional codebooks based on sampling, such as the approach in [19], which rely on a predefined discretised set of examples and require extensive searches, our method leverages neural inference. This warming strategy significantly accelerates convergence, reduces computational overhead, and enhances parameter optimisation efficiency.
- **Limitations in black-box and white-box approaches:** black-box optimisation lacked precision, while white-box methods were overly constrained by the PT model [14, 20], reducing generalisability. We implemented and tested a hybrid grey-box optimisation strategy that balances flexibility and accuracy.

The remainder of this paper is organised as follows. Section 2 outlines the literature review, focusing on AAI with AbS and Articulatory synthesisers. Section 3 details our methodology for optimising parameters in the Pink Trombone. Section 4 presents the experiments and datasets used. Section 5 describes the results from both objective and subjective evaluations. The paper concludes with a discussion and conclusions in Sects. 6 and 7, respectively.

2 Literature review

2.1 Acoustic to articulatory inversion with analysis by synthesis

AAI traditionally estimates vocal tract configurations from speech signals using paired articulatory–acoustic datasets [3, 8–11, 21]. However, AAI frameworks leveraging AbS circumvent the dependency on invasive or labour-intensive articulatory recordings (e.g. MRI, EMA) by iteratively refining synthetic articulatory parameters to match target acoustic signals. This approach integrates articulatory synthesisers as forward models, enabling physiologically plausible inversion while avoiding

[2] https://dood.al/pinktrombone/

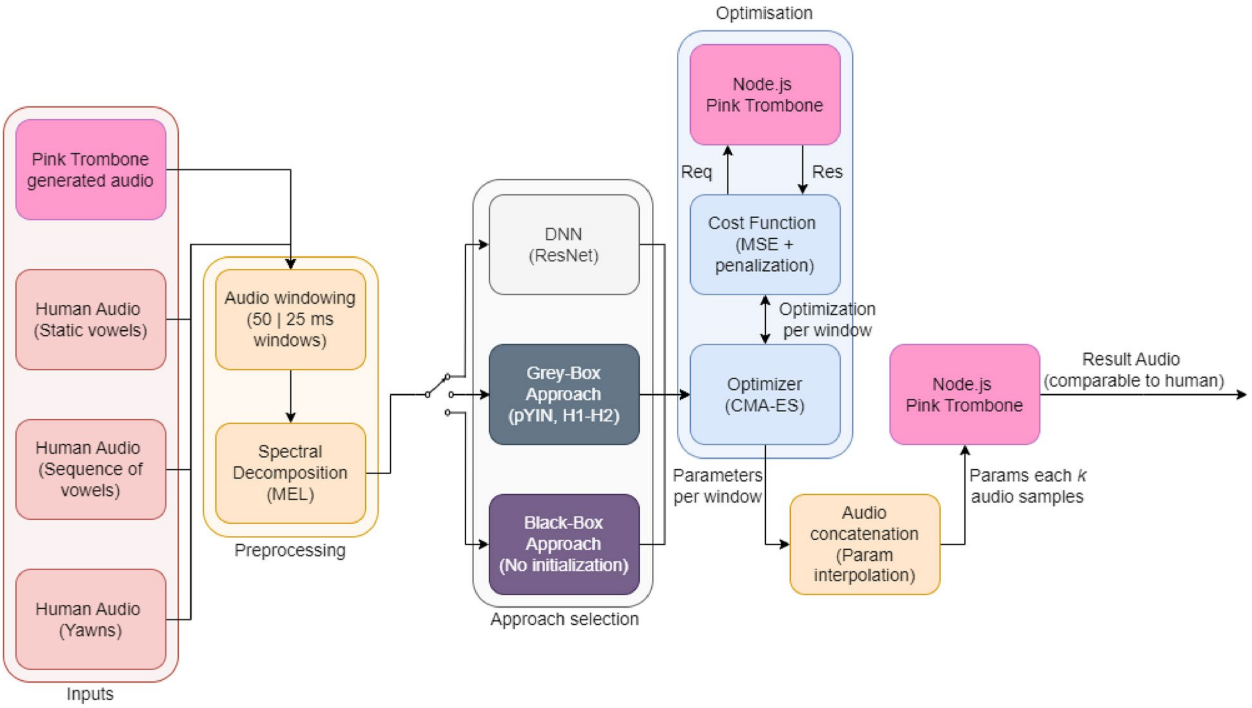
Optimisation
Node.js Pink Trombone
Req
Res
Cost Function (MSE + penalization)
Optimization per window
Optimizer (CMA-ES)
Pink Trombone generated audio
Human Audio (Static vowels)
Human Audio (Sequence of vowels)
Human Audio (Yawns)
Inputs
Audio windowing (50 | 25 ms windows)
Spectral Decomposition (MEL)
Preprocessing
DNN (ResNet)
Grey-Box Approach (pYIN, H1-H2)
Black-Box Approach (No initialization)
Approach selection
Parameters per window
Audio concatenation (Param interpolation)
Node.js Pink Trombone
Params each *k* audio samples
Result Audio (comparable to human)

Fig. 1 Flowchart of the conducted tests. On the left, the inputs consist of audio records. Records undergo a fixed preprocessing, including spectral decomposition and temporal segmentation, before being fed into the parameter optimisation models. The optimisation process is conducted per window, minimising the defined cost function. Ultimately, the estimated parameters need to be interpolated to synthesise the final audio output

the speaker-dependent limitations of recorded data [22, 23]. For instance, Prom-on et al. [23, 24] optimised vowel configurations in VocalTractLab [16] using stochastic gradient descent, starting from neutral articulatory positions to ensure biomechanical realism. Their work demonstrated that synthesisers could replicate reference formants and EMA trajectories with high precision, as validated by acoustic metrics (e.g. spectral correlation) and perceptual tests.

A critical advancement in AAI-AbS lies in the adoption of metaheuristic optimisation algorithms. Fairee et al. [25] replaced gradient descent with Particle Swarm Optimisation (PSO) to accelerate parameter search for Thai vowels, while Gao et al. [26] employed genetic algorithms to estimate gestural scores for German words, optimising both timing and stiffness parameters to model coarticulation. Neural networks have further enhanced generalisation capabilities: Gao et al. [27] trained Long Short-Term Memory (LSTM) models on VocalTractLab-generated data, augmented with vocal tract length and pitch variations, achieving robust cross-linguistic performance. Similarly, Sun and Wu [28] combined convolutional bidirectional LSTMs with the Tube Resonance Model (TRM) [29], iteratively refining mel-spectrograms through a self-supervised learning framework.

The inherent non-uniqueness of the inversion problem—where multiple articulatory configurations can produce similar acoustics—remains a central challenge. Panchapagesan and Alwan [19] addressed this issue by embedding regularisation terms into loss functions to penalise deviations from neutral vocal tract positions. Likewise, Dang and Honda [22] incorporated physiological constraints, such as the relationship between formant frequencies and tongue positioning, to guide the inversion process. Other work by Aryal and Gutierrez-Osuna [30] further demonstrated that statistical synthesisers could resolve ambiguities through probabilistic mappings of Mel-Frequency Cepstral Coefficients (MFCCs) to articulatory parameters. Despite these advances, computational complexity remains a challenge—especially for a sentence-level inversion—needing initialisation strategies such as rule-based gestural scores [31] or adaptive regularisation [26].

A recent innovation in this framework is the application of gradient descent optimisation on the Pink Trombone synthesiser—a real-time, browser-based tool offering intuitive control over vocal tract parameters. A previous contribution [14] already demonstrated how to train deep learning structures and run optimisation strategies for PT to match human speech sounds. On the same testbed, Südholt et al. [20] demonstrated that gradient descent can effectively refine articulatory configurations on Pink Trombone, while Mo et al. [32] similarly exploited this approach with improved objective results using a JAX version of the PT. Both studies observed that even subtle modifications in tract configurations produce significant acoustic variations, emphasising the tool's capacity for rapid hypothesis testing and iterative refinement.

2.2 Articulatory synthesis models

Modern articulatory synthesisers vary in complexity, anatomical accuracy, and usability. The Maeda synthesiser [17], parameterised by seven vocal tract variables (e.g. lip aperture, tongue body position), has been widely used for vowel and diphthong synthesis [19]. Its low-dimensional parametrisation facilitates efficient control and inversion but limits the phonetic detail that can be reproduced. In contrast, VocalTractLab [16] offers high anatomical fidelity by modelling up to 18 articulatory parameters (e.g. jaw position, velum opening) to simulate dynamic gestures and coarticulation [23, 27]. Although its rigorous biomechanical framework supports precise replication of EMA trajectories, this increased detail comes at the cost of greater computational intensity.

While less anatomically detailed than its counterparts, the PT synthesiser [15] excels in accessibility and real-time interaction. It is based on a Graphical User Interface (GUI), which controls the tract shape and glottal source characteristics. The graphical interface allows users to adjust articulators while receiving instantaneous acoustic feedback. A snapshot of the GUI is shown in Fig. 2. This simplicity makes Pink Trombone ideal for rapid prototyping and educational applications, even though it sacrifices some physiological granularity compared to VocalTractLab or Maeda. Several recent works have leveraged the agility of PT to optimise vocal tract configurations, demonstrating its utility for exploratory AAI-AbS studies [14, 32–34]. Collectively, these models underscore a trade-off between computational complexity and practical usability; with the Pink Trombone democratising articulatory synthesis, enabling broader experimentation without requiring extensive expertise.

3 Methodology

3.1 Vocal tract adaptation for increased diversity

Starting from the original implementation of the PT, we introduced several changes and optimisations for improved flexibility and accelerated processing.

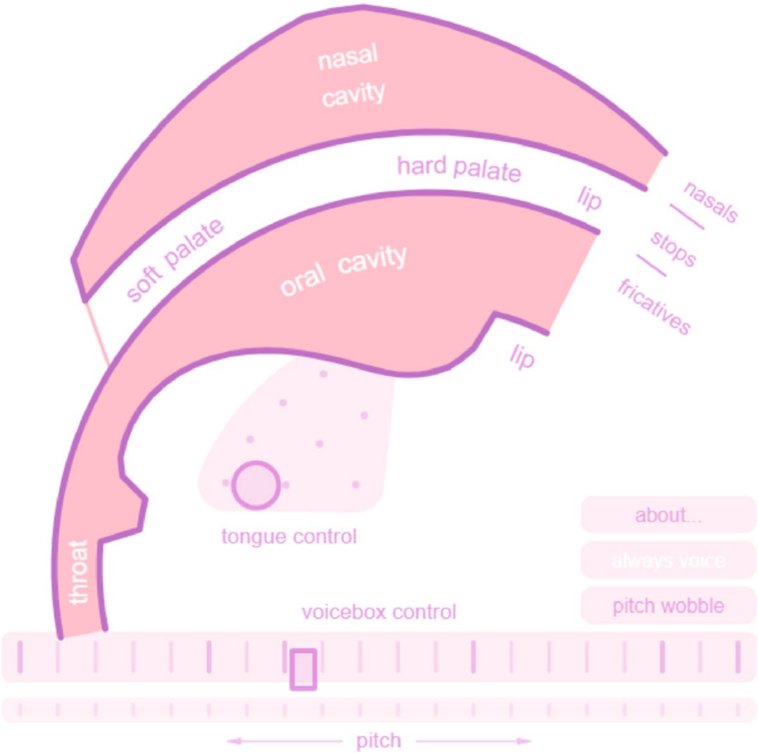

Fig. 2 Interactive user interface of the PT synthesiser displaying the clickable vocal tract, allowing for real-time structure manipulation. Below the tract, a control box is available to adjust pitch and voiceness. The screenshot was taken from Pink Trombone, under MIT license [15]

3.1.1 The Pink Trombone

The PT is a real-time articulatory speech synthesiser based on the Kelly-Lochbaum (KL) model [35], which simulates one-dimensional wave propagation through the vocal tract. Sound originates at the glottis and propagates through cylindrical segments of varying cross-sectional areas, where reflections and turbulence shape the output. Excitation is implemented via Rosenberg's model, simulating airflow modulation by the vocal folds [36]. A rigorous mathematical description of the PT can be found in [20].

The synthesiser employs a two-layer control system. The primary layer governs tongue position and diameter, while the secondary layer introduces tract constrictions, allowing for localised narrowing. Figure 3 shows a picture of the vocal tract and a schematic of its functional parts, from the generation of the excitation (bottom) up to the final sound produced (right), traversing the vocal tract.

3.1.2 Modifications of the PT

Despite its flexibility, PT's default configuration is tailored for a male vocal tract, limiting its applicability for diverse voice synthesis. The standard model assumes a vocal tract length of approximately 16 cm, divided into 44 segments. To enable more inclusive synthesis, we introduced a female-specific adaptation by shortening the tract length to 14 cm and adjusting segment distributions accordingly [37]. Table 1 details the modified parameters, which ensure a physiologically plausible vocal tract representation.

The proposed modifications broaden PT's applicability beyond its original male-oriented design, enabling more representative modelling of vocal diversity while maintaining computational efficiency. Other explorations following this approach may be covered in future contributions.

3.1.3 Node.js version of the PT

To facilitate structured parameter optimisation and large-scale computational modelling, we transitioned PT from its GUI to a Node.js-based environment. This adaptation allows direct programmatic control, batch processing, and seamless integration with machine learning techniques, avoiding the impracticalities of manual GUI-based adjustments. The GUI remains valuable for qualitative validation, demonstration, and educational purposes.

3.2 Eliminating post-processing dependency

Expanding from the insights originated from [14], we aim to minimise the post-processing of the optimisation outputs. The goal is to serve the optimised values directly to

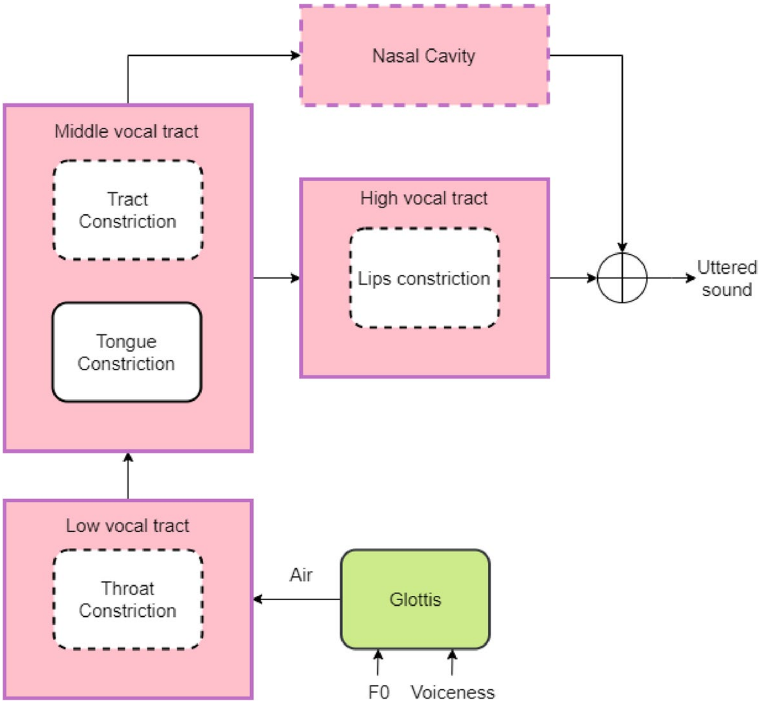

Fig. 3 Flowchart of the PT vocal synthesis structure, indicating the sequential flow from the glottis to the mouth; including nasal cavity. Dotted boxes represent optional features within the synthesiser

Table 1 Pink Trombone parameters and their bounds depending on the gender. The lip index is located in $M=44$ for a male and $M=38$ for a female. The Throat index is located in $M=12$ for both

Parameter	Male bounds	Female bounds
Pitch (Hz)	70–180	140–220
Voiceness	0.5–1	0.5–1
Tongue Index	12–29	8–25
Tongue Diam. (cm)	2.0–3.5	1.5–3.0
Lips Diam. (cm)	0.5–1.7	0.5-1.3
Constr. Index	20–40	14–34
Constr. Diam. (cm)	0.5–2.0	0.5–1.5
Throat Diam. (cm)	0.5–2.0	0.5–1.5

the synthesiser, using suitable analysis windows, focusing on the relevant spectral information, and improving the cost function for the optimisations.

3.2.1 *Temporal windowing for parameter evolution*

The optimisation process follows the pipeline illustrated in Fig. 1. Input waveforms undergo spectral decomposition and segmentation into short time windows, where suitable synthesiser parameters are inferred. These parameters are iteratively optimised, with PT generating new audio outputs that are compared against the original input to minimise error.

A critical aspect of this process is the selection of an appropriate window size. While longer windows (e.g. 100 ms) provide more stable estimates, they introduce artefacts by blending rapid articulatory changes, leading to spurious vocalic transitions. Conversely, excessively short windows may degrade spectral resolution and increase computational complexity. To balance these factors, we experimented with window sizes of 25 ms and 50 ms, optimising the trade-off between temporal resolution and spectral accuracy. We assume quasi-stationary behaviour for the PT parameters within each segment, simplifying the optimisation as a time-fixed process.

3.2.2 *Low-pass spectral representation*

To refine parameter estimation, we employed mel-spectrogram representations, which provided superior performance over short-time Fourier transform (STFT) and multiresolution spectrograms in early tests. Additionally, we introduced a frequency cap at 8000 Hz for error computation rather than using the full 24,000 Hz bandwidth dictated by PT's default sampling rate of 48 kHz. This decision is motivated by the fact that frequencies beyond 8000 Hz contribute minimally to the perceptual quality of speech and are not well-controlled by the PT synthesis model [33, 38]. This spectral constraint improves the robustness of parameter estimation, particularly in real-world conditions where input recordings may not be perfectly captured.

3.2.3 *Cost function for smooth parameter transitions*

A key limitation of previous methods was the reliance on post-processing to artificially smooth parameter variations. To overcome this, we introduced a penalisation term in the cost function that discourages abrupt changes between consecutive analysis windows, similar to [19]. This penalisation ensures articulatory parameters evolve naturally, maintaining speech continuity without requiring external smoothing techniques.

The new cost function is defined as

$$\mathrm{E}_1 = \frac{1}{N}\sum_{i=0}^{N}(\mathrm{Mel}(S_\mathrm{o}[i]) - \mathrm{Mel}(S_\mathrm{s}[i]))^2, \tag{1}$$

$$\mathrm{E}_2 = \frac{1}{M}\sum_{j=0}^{M}\max{}^2\left\{|P_\mathrm{p}[j] - P_{\mathrm{c}+1}[j]| - \alpha, 0\right\}, \tag{2}$$

$$\mathrm{E} = \mathrm{E}_1 + \beta \cdot \mathrm{E}_2, \tag{3}$$

where N is the length of the Mel filterbank, S_o is the target signal, S_s is the synthesised signal, M is the total number of PT parameters, P_p and P_{p+1} are the articulatory parameters at consecutive windows, β is a weighting hyperparameter, and α defines the threshold for penalisation. We fixed the value of α at 10% and optimised β to achieve the best balance between spectral accuracy and smooth articulatory transitions.

To initialise the optimisation windows, we employ a neural network that provides an initial estimation of the parameters, effectively acting as a codebook to guide the optimisation process. Additionally, a consensus with the previous window is used to ensure temporal coherence and smooth parameter transitions.

This formulation ensures that minor parameter variations are tolerated, allowing organic speech modulation, while large deviations are penalised to maintain consistency. This advancement eliminates the need for post-processing filtering.

As a result of these optimisations, the optimised PT parameters shall provide better results and evolve consistently over time. In Fig. 4, we illustrate how the different improvements affect the computed parameters (only tongue diameter illustrated), comparing the true articulator trajectory (black) with estimations from different methods. The Grey Box Optimisation (blue) follows the ground truth closely with minimal deviations, while the Old Method without filtering (green) exhibits higher variability and transient inconsistencies. The NN Prediction

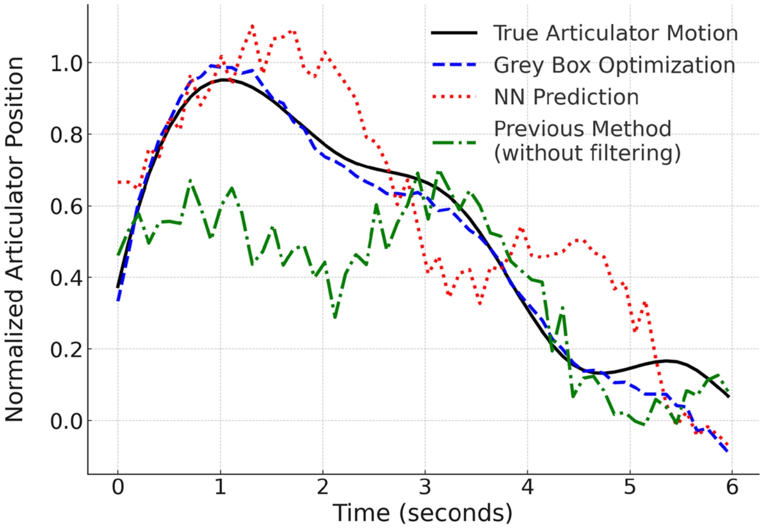

Fig. 4 Articulator trajectory over time, comparing the ground truth (black) with estimations from Grey Box Optimisation (blue), the NN Prediction (red), and the result with the previous method from [14] (green)

(red) shows erratic behaviour with significant deviations from the expected trajectory, yet its fast inference makes it a helpful initialisation strategy compared to an uninformed starting point.

3.3 Computational inefficiency and suboptimal convergence

The optimisation process for the PT required many iterations to converge, leading to increased computational cost and, in some cases, suboptimal solutions due to uninformed initialisation. While effective, traditional black-box optimisation methods operate without prior knowledge of the search space, making convergence unpredictable and susceptible to local minima, as reported in literature [19, 39]. To address this, we introduce a deep learning-based initialisation strategy, conceptualised as a *neural codebook*. This approach leverages a trained model to provide an informed starting point, accelerating convergence while improving optimisation reliability.

We implemented a deep neural network (DNN) inspired by the ResNet architecture [40], incorporating 14 convolutional layers with residual connections every two layers. ResNet architectures have demonstrated strong performance in speech synthesis tasks [41], making them a suitable choice for parameter estimation in this context. The network operates on a single mel-frequency spectrum instance, computed on the low-passed input (restricted to 8 kHz and represented as a 128-dimensional vector) and map it to the eight articulatory parameters of PT (see Table 1). It employs one-dimensional convolutions with a filter size of 16 and an initial filter count of 64, which doubles from the third residual block onward.

To train the model, we used a dataset of one million synthetic sounds using the PT, sampled at 48 kHz. Two separate datasets were generated: one containing sounds synthesised using parameter ranges corresponding to male vocal tract characteristics and another for female configurations (see Table 1). This separation allows the model to specialise in each vocal configuration, improving parameter estimation accuracy.

To enhance training stability and prevent overfitting, batch normalisation is applied at each layer, and a dropout rate of 20% is incorporated. The training dataset is split into 80% for training, 10% for testing, and 10% for validation, with a batch size of 32. The Adam optimiser [42] is employed with a learning rate of 0.0001 to ensure efficient gradient-based learning. Training continues until the validation loss stabilises, ensuring optimal generalisation. The implementation was developed in PyTorch 2.0.1.

3.4 Balancing exploration and control: a grey-box evolutionary approach

Following our choice for the CMA-ES, we propose a grey-box approach towards estimating and initialising the optimisation process.

3.4.1 CMA-ES for robust articulatory parameter optimisation

The Covariance Matrix Adaptation Evolution Strategy (CMA-ES) [43] is employed to optimise PT parameters by iteratively refining a multivariate normal distribution over candidate solutions. Unlike traditional black-box techniques, CMA-ES adapts dynamically through:

- A *maximum-likelihood update* of the distribution mean, favouring previously successful parameter configurations.
- An *adaptive covariance matrix*, which improves search efficiency by prioritising promising directions while controlling step sizes to avoid instability.

These mechanisms enable CMA-ES to effectively handle PT optimisation's non-linear and non-convex nature, outperforming alternative heuristic and gradient-based methods [44, 45]. We used a population size of 10 individuals for the CMA-ES optimiser.

While CMA-ES ensures robust search capabilities, relying solely on evolutionary optimisation can still result in slow convergence or local minima trapping, particularly without prior knowledge about the articulatory space. To mitigate it, we introduce a grey-box initialisation scheme.

3.4.2 Grey-box approach: direct estimation and dynamic initialisation

The grey-box approach precomputes certain parameters using heuristic methods, ensuring direct correspondence with PT's internal controls and reducing the dimensionality of the optimisation process. Instead of treating all parameters as free variables, we explicitly calculate those that can be determined algorithmically:

- **Fundamental frequency via pYIN:** the pYIN algorithm [46] estimates f_o from the waveform, segmenting it temporally to assign PT's glottal frequency parameter.
- **Harmonic difference for voiceness estimation:** the spectral amplitude difference between the first and second harmonic components of the glottal source, denoted as $H_1 - H_2$, serves as an indicator of vocal fold tension. This relationship, empirically established in [47], follows a linear model:

$$H_1 - H_2 = -7.6 + 11.1 \cdot R_d \tag{4}$$

where R_d is a glottal shape parameter. Since PT does not expose R_d as a control parameter, but rather a "tenseness" parameter T, the mapping is adjusted using:

$$T = 1 - \frac{R_d}{3} \tag{5}$$

This transformation allows us to estimate voiceness in PT based on spectral characteristics directly related to the phonation process.

Beyond direct parameter estimation, we refine the optimisation process by integrating a dynamic initialisation scheme within CMA-ES, balancing prior knowledge with adaptive search strategies:

- For each time window (except the first), the optimiser is initialised based on a consensus between the best-known parameters from the previous window and the predictions from the neural network. This ensures temporal consistency while leveraging the network's ability to infer plausible articulatory configurations.
- For the first window, where no prior optimisation results exist, parameters are initialised exclusively using the neural network's predictions, providing a structured starting point.

By combining direct parameter estimation with informed initialisation, this approach harmonises the exploratory capabilities of CMA-ES with prior articulatory constraints, overcoming the limitations of purely black-box or white-box methodologies. The integration of prior information acknowledges the natural smoothness of vocal tract transitions across consecutive windows, reducing abrupt parameter shifts and improving synthesis coherence.

4 Experiments

We conducted a series of experiments that analysed objective and subjective quality metrics to determine whether the new methodology improves the method described in [14].

4.1 Datasets

We utilised a dataset comprising human-recorded sounds, which included:

- Records of the five vowels of Spanish sustained by male and female speakers.
- Twenty vowel sequences, equally divided between male and female voices. Including
- Ten yawns of varying duration.
- Two datasets, each containing one million samples generated randomly using the PT, one for male and another for female voices, exclusively for training purposes of the DNN.

All recordings were sampled at 48 kHz, normalised, and stored in mono format with 16-bit depth.

4.2 Optimisation procedures

Experiments compared optimisation techniques, initialisation strategies, and the impact of hyperparameters. Key analyses included:

- *Comparison of optimisation methods:* evaluating our proposed approach against the prior method [14].
- *Initialisation strategies:* comparison of black-box optimisation (without prior knowledge) against grey-box methods, which incorporate an F_0 estimator, a voiceness estimator, and DNN-based initialisation. Additionally, evaluation of the DNN to assess its effectiveness relative to the optimisation-based approaches.
- *Penalisation factor (β):* determining the optimal value to balance smoothness and accuracy in parameter estimation.
- *Sound type:* analysing performance across static vowels, vowel sequences, and yawns.

4.3 Objective evaluation

We employed Virtual Speech Quality Objective Listener (ViSQOL) [48] as a perceptual-approximating metric to quantify synthesised sound quality relative to human recordings. This metric is used only for evaluation, as the optimisation is computed only by direct spectral difference. ViSQOL is a full-reference objective metric designed to model human auditory perception by using a modified similarity measure to compare spectro-temporal patches between reference and test signals. It was originally developed to evaluate telecommunication speech quality and has been adapted for general speech and audio quality assessment.

To enhance its applicability to PT synthesis, we fine-tuned ViSQOL by retraining its support vector regressor (SVR) using perceptual test data collected from our experiments. This refinement aligns ViSQOL's output with subjective human judgments, improving its relevance for assessing synthesised speech. The fine-tuning process was systematically documented and follows the recipe the original ViSQOL developers proposed. This ensures that the metric accurately captures the perceptual characteristics relevant to our synthesis framework.

Additionally, we evaluated the synthesised outputs using another perceptual metric: Perceptual Evaluation of Speech Quality (PESQ) [49]. However, its results did not exhibit sufficient variability to be considered a reliable measure for this type of evaluation.

4.4 Subjective evaluation

A perceptual study was conducted with 33 participants (no reported hearing impairments) using the Go Listen platform [50]. The survey included:

- *Vowel intelligibility:* participants identified synthesised vowels without reference.
- *Vowel sequences:* evaluated intelligibility across different β values.
- *Similarity assessment:* compared synthesised sounds to human recordings, assessing articulatory fidelity.

Each participant evaluated 12 of 24 stimuli to mitigate auditory fatigue, averaging 7.5 min per test. Results validated the optimisation framework and informed adjustments to perceptual metrics for articulatory synthesis.

5 Results

This section presents the results obtained from objective and subjective evaluations of the synthesised speech. The subjective tests were conducted using the configurations that achieved the highest performance in objective assessments.

5.1 SVR-adapted ViSQOL analysis

Figure 5 displays the distribution of ViSQOL MOS scores (1 to 5 values) across all experiments. These scores follow

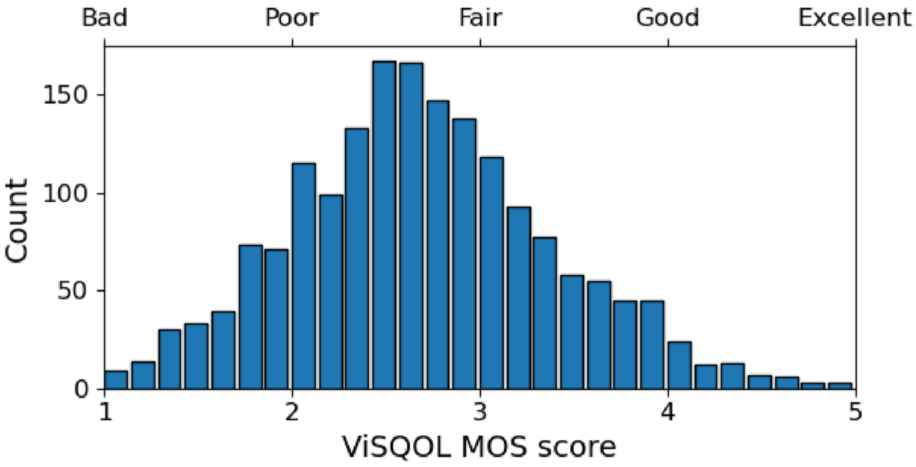

Fig. 5 Distribution of objective quality metrics across all experiments. ViSQOL MOS scores approximate a normal distribution

a normal-like distribution, reflecting a balanced spread of synthesis quality. This behaviour stems from the fine-tuning of the ViSQOL mapper, which aligns objective ViSQOL scores with subjective MOS ratings. The adjustment process consists of mapping perceptual test results to ViSQOL scores, ensuring consistency between the two metrics. Given that our subjective evaluations were designed to yield a normal distribution of MOS values, the corresponding ViSQOL scores now follow a similar pattern. This indicates that the mapper effectively captures perceptual differences, reinforcing ViSQOL's reliability as an objective assessment tool for synthesised speech quality.

The evaluation of ViSQOL scores under different experimental settings provides insights into system performance. Figure 6 categorises results by experiment type, distinguishing between black-box and grey-box approaches, as well as the previous baseline [14]. An ANOVA test revealed significant differences ($p < 0.001$), indicating a clear distinction between methods. While statistical significance confirms detectability, the actual improvement is observed in the ViSQOL scores themselves, which consistently favour the proposed approach over the baseline.

Figure 7 presents results grouped by initialisation strategy. While both the black-box method and grey-box reached similar performance levels across all sound types, the primary distinction lies in convergence time. The grey-box approach, leveraging the neural network as a neural codebook, reduced optimisation time by 37% on average. Although the black-box method eventually reaches the same solution, the neural network initialisation mitigates the risk of local minima, facilitating a more efficient search. However, the network alone does not reach the optimal solution, particularly for vowels and their sequences, likely due to its training being limited to Pink Trombone-generated data, which may not fully generalise to human recordings.

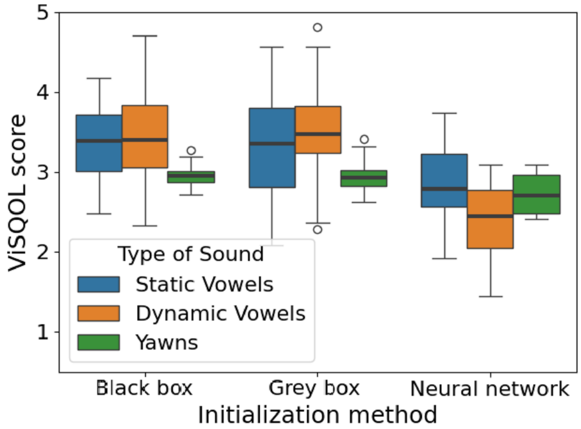

Fig. 7 ViSQOL score by *method and DNN*

The impact of the penalisation factor is illustrated in Fig. 8. While an optimal β value is expected to lie between 1 and 2, the results do not show a statistically significant distinction within this range. The observed performance remains comparable across these values, making it challenging to identify a precise sweet spot. Moderate constraints on temporal variations may enhance sound quality, but the optimal penalisation level cannot be conclusively determined.

Finally, Fig. 9 categorises ViSQOL scores by sound class. Results indicate that vocal sounds are generally better reconstructed than yawns, with vowel sequences achieving higher scores than static vowels. This finding suggests that ViSQOL computations may favor the dynamics of vowel sequences over isolated vowels, which is consistent with [51] due to the richness of the time-frequency patterns of the former over the latter.

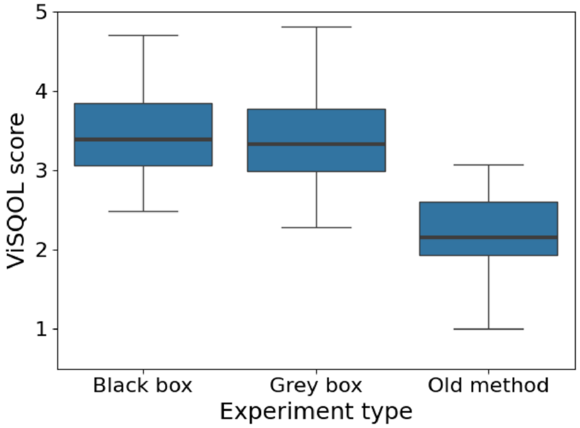

Fig. 6 ViSQOL score per *approach*, box type

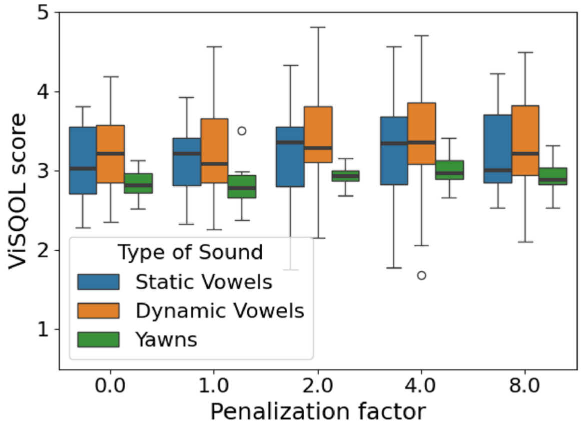

Fig. 8 ViSQOL score by *penalisation factor*

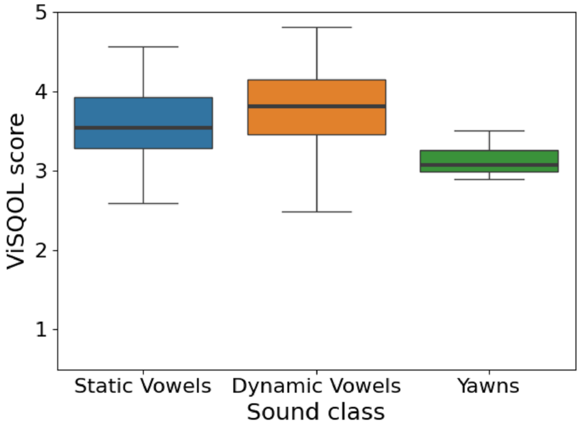

Fig. 9 ViSQOL score by *sound class*

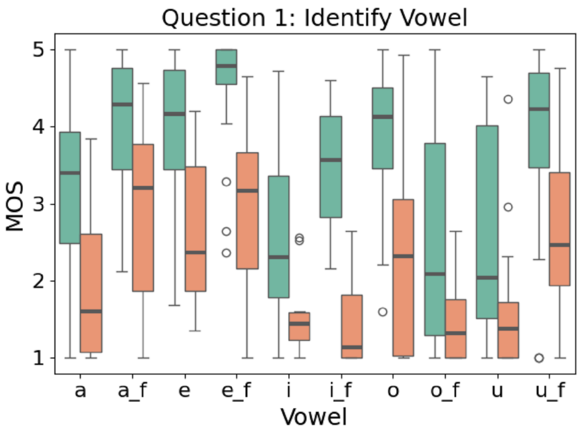

Fig. 10 MOS for vowel identification across different methods. Each colour refers to an experiment variation: green to the New Method, red to the Old Method. The suffix "_f" indicates the use of the female model, while its absence corresponds to the male model

5.2 Subjective evaluation: perceptual tests

The perceptual evaluation assessed the technical accuracy and perceptual validity of the synthesised sounds.

Figure 10 shows the Mean Opinion Scores (MOS) for vowel identification across different synthesis methods collected in our perceptual tests. The new method consistently outperformed the previous approach. Notably, masculine [i] and [u] vowels were rated lower, likely due to pronunciation biases affecting Spanish-speaking participants. These nuances highlight the role of phonetic perception and linguistic background in evaluation.

Figure 11 examines the effect of the penalisation factor on vowel sequence perception. However, the observed variations fall within the margin of error, making it difficult to determine any clear trend or optimal β value. The differences between conditions are small, and no conclusive advantage can be attributed to any specific penalisation level.

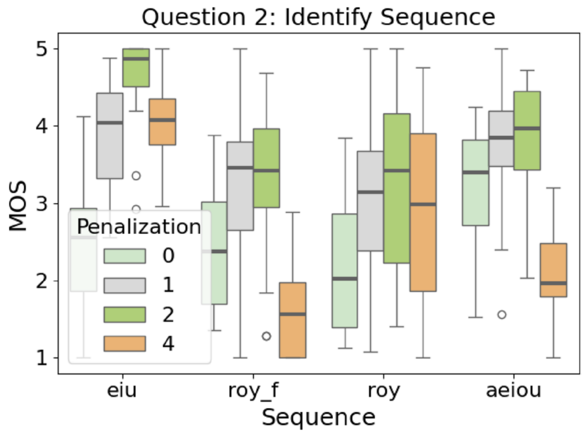

Fig. 11 MOS values obtained for the different vowel sequences recorded. The results are organised by penalisation factor, β. The suffix "_f" indicates the use of the female model, while its absence corresponds to the male model

The comparison of synthesised versus human sound samples is presented in Fig. 12. Grey-box initialisation produced results perceptually similar to human references, whereas neural network initialisation underperformed in most cases. Notably, while yawns were rated lower overall, they maintained a distinct and identifiable character, with informal feedback suggesting they elicited contagious yawning responses.

Figure 13 examines the correlation between ViSQOL and subjective MOS scores. A Pearson correlation of 69% confirms that the trained ViSQOL model closely aligns with perceptual evaluations, while the default ViSQOL and PESQ correlations remained significantly lower (30% and 20%, respectively). A Mann-Whitney test delivered $p = 0.7$ supporting the equivalence of both distributions, validating the trained ViSQOL SVR model.

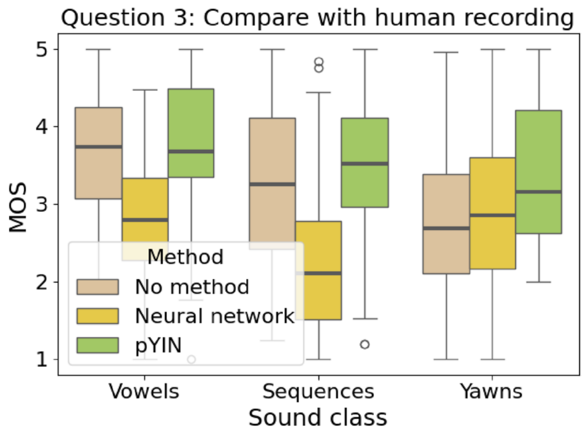

Fig. 12 Comparison of synthesised and human sounds from records on sustained vowels, vowel sequences, and yawns

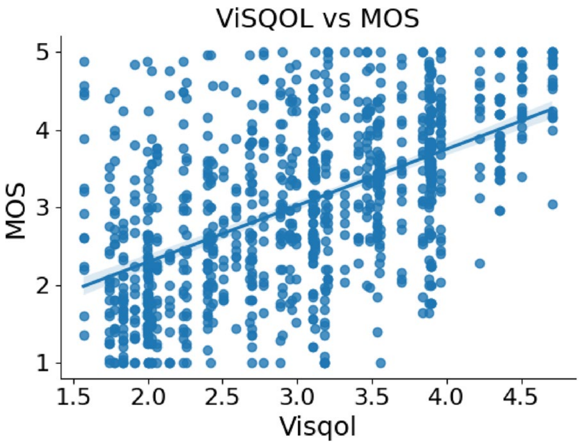

Fig. 13 Correlation between ViSQOL and MOS scores, including the correspondences between the two (dots) and a tendency line (line)

5.3 Ablation studies

To further validate our findings, ablation studies assessed the accuracy of fundamental frequency (f_o) and voiceness reconstruction compared to state-of-the-art (SOTA) methods (pYIN and $H1 - H2$).

Figure 14 (left) shows that f_o reconstruction was not significantly affected by the choice of initialisation method. However, the grey-box method resulted in near-optimal recovery, particularly for vowel sequences. Despite minor discrepancies for static vowels, over 50% of cases remained within human perceptibility thresholds.

Figure 15 demonstrates that penalisation factors between 2 and 4 yield optimal temporal coherence. Our results support this statement with f_o errors clustered around zero, confirming accurate reconstruction. In contrast, voiceness errors exhibited a more uniform distribution, likely influenced by background noise compensation and the non-linearity of the voiceness parameter.

5.4 Comparison with state-of-the-art

To benchmark our approach against existing SOTA methods, we compared its performance with the Speech Articulatory Coding (SAC) system presented in [52], which represents one of the most recent advances in the field of AAI. The SAC system employs a Transformer architecture to estimate vocal tract articulator positions from the speech waveform, subsequently enabling speech resynthesis using a vocoder.

For this comparison, we generated a dataset comprising 6000 unique English words synthesised using the PT, leveraging available phonetic transcriptions and a defined phoneme-to-articulator mapping. These synthesised waveforms were then processed by the SAC system to obtain its predicted articulator positions. To establish a common ground for comparison, we trained a Random Forest Regressor to map the SAC-derived articulator positions to the corresponding PT parameter space used by our method. This mapping allows us to evaluate both systems based on their ability to reconstruct target sounds represented within the PT framework, using the same audio files employed in our subjective evaluations. We focused the comparison on the final synthesised output quality rather than directly comparing articulator parameters, acknowledging that different articulatory configurations can produce perceptually equivalent

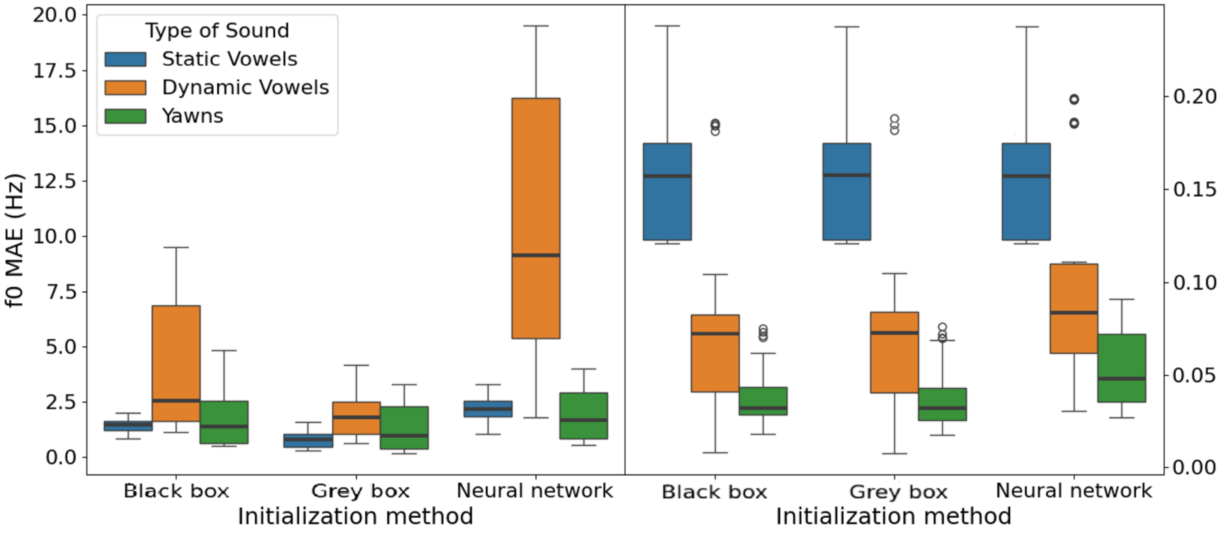

Fig. 14 Ablation study of f_o (left) and voiceness (right) grouped by *initialisation method* and focusing on the Mean Average Error (MAE)

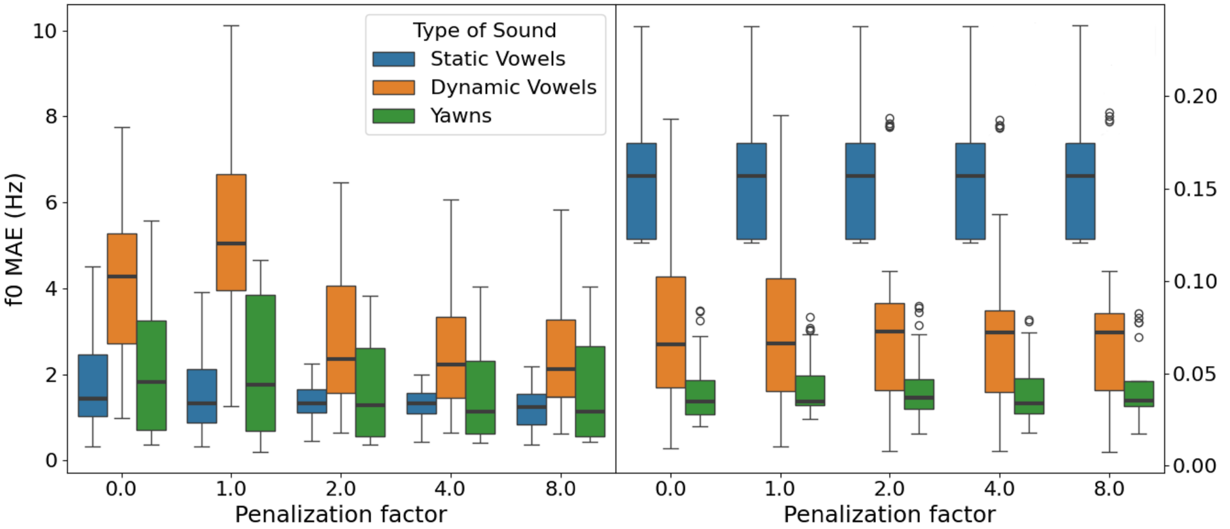

Fig. 15 Ablation study of f_0 (left) and voiceness (right) grouped by *penalisation factor* and focusing on the Mean Average Error (MAE)

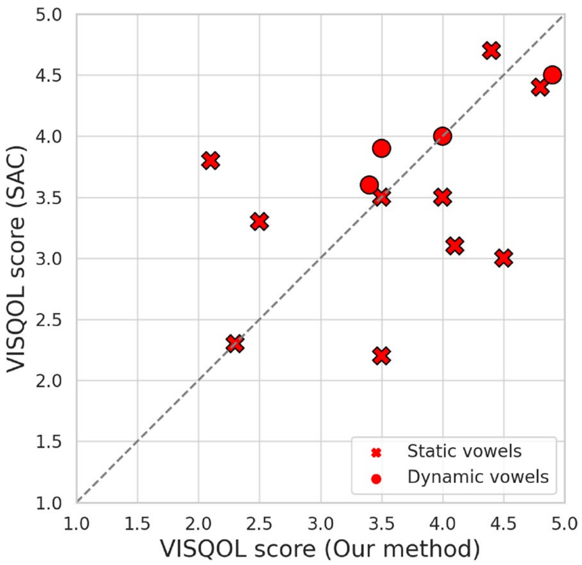

Fig. 16 Comparison of ViSQOL scores between our proposed method and the SAC SOTA system for static (red crosses) and dynamic (red circles) vowels. The dashed line represents identity $y = x$

utterances. Furthermore, yawn sounds were excluded from this specific comparison, as the SAC system was primarily trained on speech data, making a direct comparison potentially unfair.

The results of this comparison are presented in Fig. 16, which plots the ViSQOL scores obtained by our method against those achieved by the SAC system for static and dynamic vowels. Visually, the data points cluster around the identity line $y = x$, indicating comparable performance between the two methods across various sounds.

Quantitatively, our method achieved a mean ViSQOL score of 3.7, slightly higher than SAC's mean score of 3.6. Analyzing individual data points, our method yielded better or equal performance in approximately 80% of cases (60% better, 20% equal), while SAC performed better in the remaining 20% of cases. Despite these minor differences, the overall performance distributions suggest that the two methods are statistically equivalent in terms of output quality for the evaluated speech sounds. This equivalence validates our PT-based optimisation approach relative to a contemporary data-driven SOTA method. A key advantage of our methodology, however, lies in its inherent flexibility; being based on a physical model (PT), it is not restricted to specific training vocabularies and can potentially be applied to diverse sound types, languages, or speaker characteristics without retraining large neural models.

6 Discussion

This study refines optimisation techniques for articulatory synthesis using the PT, establishing a methodological framework for future research. Unlike purely data-driven AAI approaches, our method operates within a constrained physical model, requiring tailored optimisation strategies. Furthermore, our approach demonstrated statistically equivalent performance to the SAC SOTA system for vowel synthesis, validating its effectiveness while offering greater flexibility beyond word-based synthesis.

Key contributions include a gender-adaptive configuration that enhances inclusivity in articulatory synthesis, a fine-tuned ViSQOL metric for evaluating PT-generated speech based on perceptual data, and a grey-box optimisation strategy that accelerates convergence by 37%. While this approach significantly improves dynamic speech sequences, neural network initialisation remains constrained by the domain gap between PT-generated and real human speech. The successful comparison against SAC underscores the viability of optimising physical models like PT to achieve competitive results, particularly highlighting our method's advantage in not being inherently limited by the scope of training data, thus potentially extending to broader acoustic domains.

This work and prior studies establish a standardised foundation for articulatory model optimisation. The combination of perceptual validation, improved optimisation, and tailored evaluation tools, and demonstrated comparability with data-driven SOTA methods provides a robust basis for advancing physical-model-based speech synthesis.

7 Conclusion

This study validates CMA-ES as an effective optimisation method for articulatory synthesis using PT. Our solution successfully refined the parameters while maintaining smooth transitions without post-filtering. Subjective evaluations confirm that the optimised method produces perceptually preferred results over previous approaches. Objective comparisons further demonstrated that our method achieves performance statistically equivalent to the SAC SOTA system for vowel synthesis, lending strong support to its validity.

Introducing a grey-box optimisation strategy accelerates convergence and improves parameter stability, particularly for dynamic speech sequences. Additionally, adapting PT for gender-aware synthesis enhances its applicability, addressing a key gap in articulatory modelling.

A neural codebook was introduced to structure the optimisation process, improving parameter estimation efficiency. However, neural network initialisation remains limited by the domain mismatch between PT-generated and human speech. Fine-tuning ViSQOL as a PT-specific evaluation metric bridges the gap between subjective and objective assessments, ensuring perceptually relevant evaluations.

These refinements establish a solid methodological framework for future research in physical-model-based articulatory synthesis. The validation against a contemporary SOTA system, coupled with the inherent flexibility of the PT model, suggests promising avenues for future work. Future work should explore extending these techniques to more complex speech synthesis models and non-verbal sound production, ensuring continued advancements grounded in perceptual validation and leveraging the adaptability of our physics-based approach.

Acknowledgements
The authors would like to thank Zhiyuan Xu who wrote a first version of the Pink Trombone Node.js software.

Authors' contributions
MC made substantial contributions to the design of the work; acquisition, analysis, and interpretation of data; the creation of new software used in the work; drafted the work. JB made substantial contributions to the conception of the work; interpretation of data; drafted the work and substantively revised it. JR made substantial contributions to the conception of the work; interpretation of data, and substantively revised the drafted work.

Funding
Activities described in this contribution were partially funded by the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 101003750; and by the Ministry of Economy and Competitiveness of Spain under grant PID2021-128469OB-I00.

Data availability
The datasets generated and/or analysed during the current study are available in the GitHub repository, https://mateocamara.github.io/pink-trombone/. Due to their larger size, the PT datasets and DNN models are available upon request.

Declarations

Competing interests
The authors declare that they have no competing interests.

Received: 18 January 2024 Accepted: 16 June 2025
Published online: 14 July 2025

References

1. B.J. Kröger, *in Konferenz Elektronische Sprachsignalverarbeitung, Articulatory speech synthesis in the context of speech research and speech technology: Review and prospect* (TUDpress, Dresden, 2023), pp.173–180
2. P. Birkholz, S. Ossmann, R. Blandin, A. Wilbrandt, P.K. Krug, M. Fleischer, Modeling speech sound radiation with different degrees of realism for articulatory synthesis. IEEE Access **10**, 95008–95019 (2022)
3. B.J. Kröger, Computer-implemented articulatory models for speech production: A review. Front. Robot. AI **9**, 796739 (2022)
4. BBC, The BBC Year Book 1931. Chapter "The Use of Sound Effects" 194–197, Editorial: British Broadcasting Corporation, Savoy Hill, London. (1931)
5. M.M. Afsar, et al., Generating diverse realistic laughter for interactive art (2021), arXiv preprint arXiv:2111.03146
6. A. Anikin, Soundgen: An open-source tool for synthesizing nonverbal vocalizations. Behav. Res. Methods **51**, 778–792 (2019)
7. K. Richmond, Estimating articulatory parameters from the acoustic speech signal (Ph.D. thesis, University of Edinburgh, 2002)
8. Q. Fang, in *Proceedings of the 2024 International Conference on Generative Artificial Intelligence and Information Security*, On the performance of ema-synchronized speech and stand-alone speech in speech recognition and acoustic-to-articulatory inversion, Association for Computing Machinery, New York. pp. 162–166 (2024)
9. S. Azzouz, P.A. Vuissoz, Y. Laprie, Complete reconstruction of the tongue contour through acoustic to articulatory inversion using real-time mri data (2024). arXiv preprint arXiv:2411.02037
10. B.H. Story, I.R. Titze, E.A. Hoffman, Vocal tract area functions from magnetic resonance imaging. J. Acoust. Soc. Am. **100**(1), 537–554 (1996)
11. A. Toutios, S.S. Narayanan, in *INTERSPEECH*, Articulatory synthesis of french connected speech from ema data, Lyon. pp. 2738–2742 (2013)
12. B.S. Atal et al., Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique. J. Acoust. Soc. Am. **63**(5), 1535–1555 (1978)
13. K.N. Stevens, Remarks on analysis by synthesis and distinctive features. Models for the perception of speech and visual form (1967)

14. M. Cámara, et al., in Proceedings of the 26th international conference on digital audio effects, Optimization techniques for a physical model of human vocalisation, Denmark. pp 29-36 (2023)
15. N. Thapen. Pink trombone (2017)
16. P. Birkholz. Modeling consonant-vowel coarticulation for articulatory speech synthesis. PloS one **8**(4), e60603 (2013)
17. S. Maeda, Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. Speech Prod. Speech Model. **55**:131–149 (1990)
18. P. Mermelstein, Articulatory model for the study of speech production. J. Acoust. Soc. Am. **53**(4), 1070–1082 (1973)
19. S. Panchapagesan, A. Alwan, A study of acoustic-to-articulatory inversion of speech by analysis-by-synthesis using chain matrices and the maeda articulatory model. J. Acoust. Soc. Am. **129**(4), 2144–2162 (2011)
20. D. Südholt, et al., in *Proceedings of the 20th international conference on digital audio effects*, Vocal tract area estimation by gradient descent, Denmark. (2017)
21. Y. Laprie, M. Loosvelt, S. Maeda, R. Sock, F. Hirsch, in *InterSpeech-14th Annual Conference of the International Speech Communication Association-2013*, Articulatory copy synthesis from cine X-ray films, Association for Computing Machinery, New York. (2013)
22. J. Dang, K. Honda, Estimation of vocal tract shapes from speech sounds with a physiological articulatory model. J. Phon. **30**(3), 511–532 (2002)
23. S. Prom-on, P. Birkholz, Y. Xu, in *INTERSPEECH*, Training an articulatory synthesizer with continuous acoustic data, Association for Computing Machinery, New York. pp. 349–353 (2013)
24. S. Prom-on, P. Birkholz, Y. Xu, Identifying underlying articulatory targets of thai vowels from acoustic data based on an analysis-by-synthesis approach. EURASIP J. Audio Speech Music Process. **2014**, 1–11 (2014)
25. S. Fairee, B. Sirinaovakul, S. Prom-on, in *2015 12th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, Acoustic-to-articulatory inversion using particle swarm optimization (IEEE, 2015), pp. 1–6
26. Y. Gao, S. Stone, P. Birkholz, in *INTERSPEECH*, Articulatory copy synthesis based on a genetic algorithm, Graz. pp. 3770–3774 (2019)
27. Y. Gao, P. Birkholz, Y. Li, Articulatory copy synthesis based on the speech synthesizer vocaltractlab and convolutional recurrent neural networks. IEEE/ACM Trans. Audio Speech Lang. Process. **32** pp. 1845–1858 (2024)
28. Y. Sun, X. Wu, Embodied self-supervised learning by coordinated sampling and training (2020). arXiv preprint arXiv:2006.13350
29. L. Manzara, The tube resonance model speech synthesizer. Master's thesis, University of Calgary (2009)
30. S. Aryal, R. Gutierrez-Osuna, in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, Articulatory inversion and synthesis: towards articulatory-based modification of speech (IEEE, 2013), pp. 7952–7956
31. H. Nam, V. Mitra, M. Tiede, M. Hasegawa-Johnson, C. Espy-Wilson, E. Saltzman, L. Goldstein, A procedure for estimating gestural scores from speech acoustics. J. Acoust. Soc. Am. **132**(6), 3980–3989 (2012)
32. L. Mo, M. Cherep, N. Singh, Q. Langford, P. Maes, in *Audio Imagination: NeurIPS 2024 Workshop AI-Driven Speech, Music, and Sound Generation*, Articulatory synthesis of speech and diverse vocal sounds via optimization, Vancouver. (2024)
33. P. Saha, et al., Sound stream: Towards vocal sound synthesis via dual-handed simultaneous control of articulatory parameters. J. Acoust. Soc. Am. **144**(3_Supplement), 1907 (2018)
34. P. Saha, S. Fels, Learning Joint Articulatory-Acoustic Representations with Normalizing Flows (2020). arXiv e-prints arXiv:2005.09463
35. J.L.J. Kelly, C. Lochbaum, Speech synthesis. Proc. Fourth Int. Congr. Acoust. 1–4 (1962)
36. B.H. Story, A parametric model of the vocal tract area function for vowel and consonant simulation. J. Acoust. Soc. Am. **117**(5), 3231–3254 (2005)
37. U.G. Goldstein, An articulatory model for the vocal tracts of growing children (Ph.D. thesis, Massachusetts Institute of Technology, 1980)
38. ISO 8253-1:2010, Acoustics—Audiometric test methods. Part 1: Pure-tone air and bone conduction audiometry (2010)
39. M.R. Schroeder, Determination of the geometry of the human vocal tract by acoustic measurements. J. Acoust. Soc. Am. **41**(4B), 1002–1010 (1967)
40. K. He, X. Zhang, S. Ren, J. Sun, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Deep residual learning for image recognition, Las Vegas. pp. 770–778 (2016)
41. Y. Ma, Z. Ren, S. Xu, RW-Resnet: A novel speech anti-spoofing model using raw waveform (2021). arXiv preprint arXiv:2108.05684
42. D.P. Kingma, J. Ba, Adam: A method for stochastic optimization (2014). arXiv preprint arXiv:1412.6980
43. A. Auger, N. Hansen, in *2005 IEEE congress on evolutionary computation*, vol. 2, A restart CMA evolution strategy with increasing population size (IEEE, 2005), pp. 1769–1776
44. M.J. Yee-King et al., Automatic programming of vst sound synthesizers using deep networks and other techniques. IEEE Trans. Emerg. Top. Comput. Intel. **2**(2), 150–159 (2018)
45. S. Ruder, An overview of gradient descent optimization algorithms (2016). arXiv preprint arXiv:1609.04747
46. M. Mauch, S. Dixon, in *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pYIN: A fundamental frequency estimator using probabilistic threshold distributions, Florence. pp. 659–663 (2014)
47. G. Fant, The LF-model revisited. Transformations and frequency domain analysis. STL-QPSR **2**(3), p. 40 (1995)
48. M. Chinen, et al., in *twelfth international conference on quality of multimedia experience (QoMEX)*, ViSQOL v3: An open source production ready objective speech and audio metric (IEEE, 2020), pp. 1–6
49. 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)
50. D. Barry, et al., Go listen: an end-to-end online listening test platform. J Open Res. Softw. **9**(1), p. 20 (2021)
51. A. Hines, J. Skoglund, A.C. Kokaram, N. Harte, ViSQOL: an objective speech quality model. EURASIP J. Audio Speech Music Process. **2015**(1), 13 (2015). https://doi.org/10.1186/s13636-015-0054-9
52. C.J. Cho, P. Wu, T.S. Prabhune, D. Agarwal, G.K. Anumanchipalli, Coding speech through vocal tract kinematics. IEEE J. Sel. Top. Signal Process. **18**(8), pp. 1427–1440 (2024)

Publisher's Note