JOINT ESTIMATION OF SOURCE AND FILTER PARAMETERS FOR SPEAKER ADAPTATION IN ARTICULATORY SPEECH SYNTHESIS

Tianyi Zhang, Peter Birkholz

Institute of Acoustics and Speech Communication, Technical University Dresden, Dresden, Germany
tianyi.zhang2@tu-dresden.de, peter.birkholz@tu-dresden.de

Abstract: Magnetic Resonance Imaging (MRI) data of the vocal tract were used to reproduce the anatomy and articulation of two new speakers for the articulatory speech synthesizer VocalTractLab. Due to the limited MRI resolution and the artificially sustained phonemes during the scans, the vocal tract shapes reconstructed from the MRI data usually did not fully correspond to the natural articulatory configurations. Therefore, we introduced a strategy for the joint optimization of the vocal tract model and vocal fold model parameters to match the synthesized vowels with their naturally produced counterparts. We used genetic algorithms (GA) and particle swarm optimization (PSO) to minimize the root mean squared error (RMSE) between the mel-frequency-scaled true spectral envelopes of the natural and synthetic vowel sounds. PSO achieved the best results for 65% of the vowels, reducing the RMSE from 15 dB to 3 dB, while the best results for the remaining vowels were obtained by the GA. Furthermore, informal listening tests confirmed a substantial improvement in the quality of the synthetic vowels.

1 Introduction

In recent years, neural network-based speech synthesis methods [1, 2, 3] have achieved remarkably high naturalness of synthetic speech. These models can even capture the acoustic characteristics of real speakers and reproduce, to a large extent, their prosody, timbre, and voice quality. Despite their remarkable capabilities, neural-based synthesis models exhibit limited interpretability and controllability in terms of speech production.

In contrast to neural speech synthesis methods, articulatory speech synthesis [4, 5, 6] simulates speech production using an articulatory model of the vocal system [7, 8, 9, 10], an aeroacoustic model [11, 12], and a control model [13, 14, 15]. This has a range of applications like testing phonetic hypotheses, speech synthesis with precisely controllable parameters, and the visual representation of articulatory processes. However, a main problem with articulatory synthesizers is to adjust the model parameters to produce phonemes that sound natural or even like those of a specific real speaker [16]. This study shows how the parameters of the vocal tract and the vocal fold models can be jointly optimized to closely reproduce the vowels of real speakers by utilizing VocalTractLab (VTL) [6], which features not only a comprehensive articulatory speech synthesis model but also an accessible API, thereby providing an advantageous platform for experimentation.

Previous work on vowel parameter adjustment in articulatory speech synthesis has followed several distinct methodological paradigms. Early studies relied on trial-and-error procedures or on MRI- or X-ray-based articulatory adjustments with subsequent manual refinement [17]. Later approaches introduced automatic refinement strategies, in which vocal tract

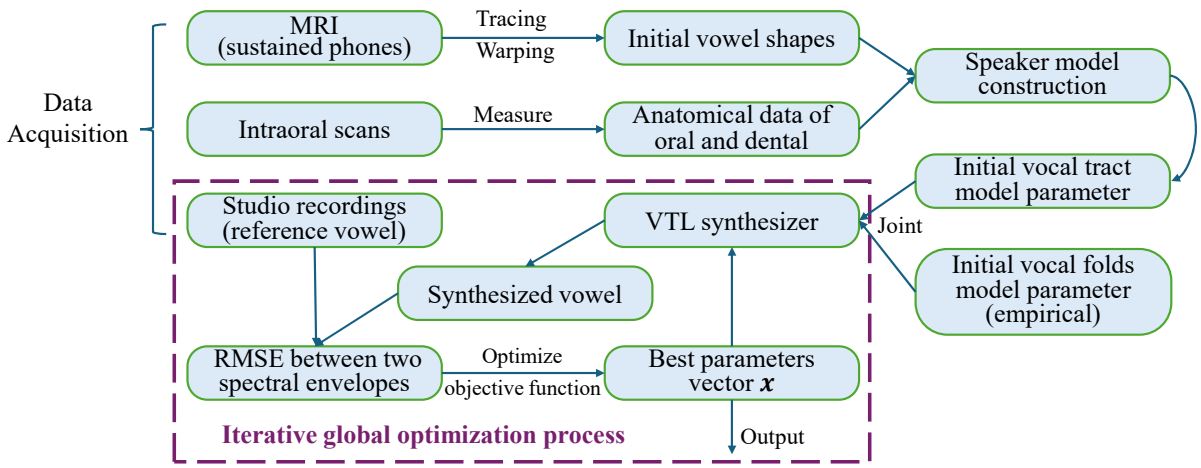

Figure 1 – Overview of the proposed workflow for speaker-specific vocal tract model construction and optimization of control parameters.

parameters derived from MRI, X-ray, or EMA data were optimized with respect to acoustic criteria, typically focusing on the reproduction of formant frequencies [13, 18, 19, 20]. In parallel, articulatory–acoustic inversion methods were proposed to estimate vocal tract configurations directly from acoustic observations, often within an analysis-by-synthesis framework [21, 22, 23].

Despite substantial differences in methodology, these approaches share a common characteristic: vocal tract parameters are treated as the primary optimization targets, while vocal fold parameters are either fixed, determined by rule-based procedures, or adjusted independently from the vocal tract. As a consequence, the acoustic optimization is largely restricted to formant frequencies or related low-dimensional spectral features.

In the present study, we propose the *joint* estimation of vocal tract and vocal fold model parameters, aiming to reproduce not only the formant frequencies of natural vowels, but also their overall spectral characteristics in order to better capture voice quality features. To achieve this, we propose a method that involves the calculation of the mel-scaled true spectral envelope [24] from both original and synthesized vowel audio signals. The deviation between a natural and synthetic vowel is represented by the root mean squared error (RMSE) between these envelopes. We apply global optimization algorithms [25, 26] to find parameter combinations minimizing this error.

2 Method

2.1 Overview of the proposed workflow

Figure 1 illustrates the proposed workflow for speaker-specific vocal tract model construction and joint optimization of vocal tract and vocal fold control parameters. The process begins with the acquisition of multimodal anatomical and acoustic data, including MRI scans, intraoral measurements, and studio recordings of sustained vowels. Based on these data, an initial speaker model is constructed by reproducing the vocal tract anatomy and initializing vowel-specific articulatory configurations. Speech synthesis is then carried out using the VocalTractLab synthesizer, and the synthesized vowels are evaluated by computing the RMSE between their mel-scaled spectral envelopes and those of the corresponding natural vowels. An iterative global optimization process is subsequently applied to jointly refine the vocal tract and vocal fold parameters, resulting in optimized parameter vectors for each vowel. The subsequent sections provide detailed descriptions of the individual components of this workflow.

2.2 Overview of VocalTractLab

As illustrated in the workflow in Figure 1, speech synthesis is performed using the VocalTract-Lab (VTL) system, which consists of four main components: the vocal tract model [13], the vocal folds model [16], the aerodynamic-acoustic simulation model [6], and the control model [27, 28].

The vocal tract is modeled as a 3D geometry controlled by 19 articulatory parameters, while the vocal folds model employs 10 parameters. For the aerodynamic-acoustic simulation, the vocal system is represented in terms of short cylindrical tube segments, each defined by its length, cross-sectional area, and axial position. Acoustics and aerodynamics are simulated via a transmission-line model with lumped elements, synthesizing speech by propagating pressure waves through this dynamically adapting structure.

The control model maps phoneme sequences and pitch targets to a gestural score, where each phoneme is associated with temporally ordered articulatory gestures (e.g., vocal fold vibration, lip closure). These gestures generate parameter time functions for the vocal tract and folds, including German-specific articulatory configurations.

2.3 Subjects and data acquisition

Previously, there was only a single speaker model of a male German speaker in VTL. In this study, we recruited two additional native German speakers (47-year-old male, 185 cm; and 41-year-old female, 174 cm) to provide data for new speaker model prototypes. Volumetric Magnetic Resonance Imaging was used to capture the vocal tract of the subjects while they produced a set of sustained speech sounds with the same MRI protocol as in [29]. The set of speech sounds consisted of 16 German vowels, including both tense and lax vowels, namely /a, e, i, o, u, ɛː, ø, y, ɛ, ɪ, ɔ, ʊ, œ, ʏ, ə, ɐ/. In addition, the consonants /p, t, k, f, s, ʃ, ɕ, x, l, m, n, ŋ/ were included and produced in the context of each of the corner vowels /a, i, u/. The MRI-scanning duration was 14 s per speech sound, during which the participant had to sustain the articulation. For all phonemes produced in the MRI scanner, clean audio recordings were made in a separate recording session in a sound-proofed audio studio with a studio microphone (Microtech M930) and an audio interface (TASCAM UH-7000). The recordings were made as mono-channel waveforms with Audacity (`www.audacityteam.org`) at a sampling rate of 48 kHz. The subjects were sitting on a chair and sustained each phoneme with flat intonation for about 7 s. Additionally, we used an intraoral scanner to acquire detailed 3D shapes of the lower and upper jaws.

2.4 Creation of speaker models

2.4.1 Reproduction of vocal tract anatomy

Consistent with the approach in [13], the 3D geometry of rigid vocal tract structures (teeth, palate, jaw) and reference contours for movable parts (velum, larynx) were defined using anatomical parameters. The hard palate and teeth were reproduced from the intraoral 3D scans of the human counterparts using predefined landmarks and dimensions such as height and width of the individual teeth. Furthermore, three velum configurations (see contours in the top-left of Fig. 2a) were represented by point sets based on MRI images: the highest position with a fully closed velo-pharyngeal port (black contour), based on /s, ʃ/; a lower intermediate position, also with a closed port (dashed contour), based on /a/; and the lowest position with a fully open port (gray contour), derived from /m/. Similarly, the contour of the epilarynx was defined for a narrow state (as in /a/) and for a wide state (as in /i/) from the respective MRI images, between

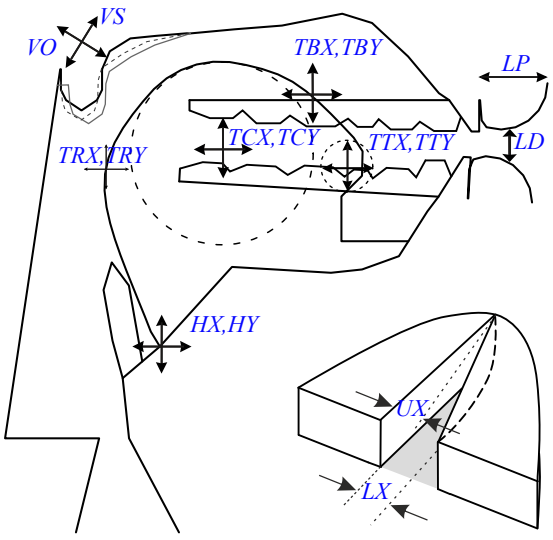

Figure 2 – (a) 2D sagittal view of the vocal tract with the parameters shown in blue; (b) geometric vocal fold model.

Table 1 – Overview of the vocal tract model parameters.

Name	**Abbr.**	**Unit**	$[x^{\min}, x^{\max}]$	**Optimized?**
Horiz hyoid pos.	HX	-	[0, 1]	✓
Vertical hyoid pos.	HY	cm	[-5.9, -4.8]	✓
Horiz jaw pos.	JX	cm	[-0.5, 0]	
Jaw angle	JA	deg	[-7, 0]	✓
Lip protrusion	LP	-	[-1, 1]	✓
Lip distance	LD	cm	[-2, 4]	✓
Velum shape	VS	cm	[0, 1]	✓
Velic opening	VO	cm^2	[-0.1, 1.5]	
Tongue body X	TCX	cm	[-3, 4]	✓
Tongue body Y	TCY	cm	[-3, 1]	✓
Tongue tip X	TTX	cm	[-1.5, 5.5]	✓
Tongue tip Y	TTY	cm	[-3, 2.5]	✓
Tongue blade X	TBX	cm	[-3, 4]	✓
Tongue blade Y	TBY	cm	[-3, 5]	✓
Tongue root X	TRX	cm	[-4, 2]	✓
Tongue root Y	TRY	cm	[-6, 0]	✓
Tongue side 1	TS1	-	[-1, 1]	✓
Tongue side 2	TS2	-	[-1, 1]	✓
Tongue side 3	TS3	-	[-1, 1]	✓

which the actual epilarynx shape is interpolated with the HX control parameter.

In addition, the tongue shape was parameterized by tip radius, body radius, and root position, while the uvula was modeled as a cone-cylinder combination with specified length, width, and height that were also estimated from the MRI data. All components were integrated into a unified vocal tract model, where fixed structures provide stability and movable parts enable the dynamic shaping required for speech.

2.4.2 Reproduction of vocal tract shapes for vowels

To reproduce the vocal tract shapes of vowels, the MRI vocal tract contours were traced in the midsagittal plane using spline functions in Inkscape (`www.inkscape.org`). Additionally, the MRI-slice located five layers (9 mm) lateral to the midsagittal plane was used to trace the contour of the tongue side. Since the MRI scans were conducted in two separate sessions and the speakers involuntarily adjusted their head posture during the process, we applied a warping transformation—analogous to that described in [13]—to all traced vocal tract contours to ensure a consistent angle between the hard palate and the rear pharyngeal wall.

Then, the constructed speaker model was loaded into VTL, and the vocal tract model shape for each phoneme was initialized by visually aligning it with the MRI-based reference contour, yielding the *initial vowel shape* for subsequent optimization. Hence, each initial vowel shape is represented by one vector of vocal tract parameter values (see Table 1).

2.4.3 Vocal fold model

The vocal fold model is represented by a parameterized glottal geometry (see Figure 2b). It is controlled by ten parameters, including the fundamental frequency, upper and lower vocal fold rest displacement, and chink area. Based on existing studies on vocal fold parameter dimensions [16, 30, 31, 32], all vocal fold parameters involved in the optimization were empirically restricted to certain value ranges, while the remaining ones were fixed. See Tab. 2 for configuration details.

In addition, static parameters such as vocal fold rest thickness, and rest length were adjusted based on gender-specific anatomical characteristics. For the female speaker model, values were

Table 2 – Overview of the vocal fold model parameters.

Name	Abbr.	Unit	$[x^{\min}, x^{\max}]$	$x^{(0)}$	$[l_b, u_b]$
Fundamental frequency	f_0	Hz	[40, 600]	f_0^*	-
Subglottal pressure	P_{sub}	dPa	[0, 20000]	8000	-
Lower displacement	LX	mm	[-0.5, 3]	0.2	[0, 1]
Upper displacement	UX	mm	[-0.5, 3]	0.2	[0, 1]
Chink area	CA	cm^2	[-0.25, 0.25]	0	-
Phase lag	PL	deg	[0, 180]	70	[60, 80]
Relative amplitude	RA	-	[-1, 1]	1	-
Double pulsing	DP	-	[0, 1.0]	0	-
Pulse skewness	PS	-	[-0.5, 0.5]	0	[-0.2, 0.2]
Flutter	FL	%	[0, 100]	25	-

f_0^* takes the measured value from the recording of the participant.
The symbol "-" in the $[l_b, u_b]$ column indicates the parameter is fixed as a constant.

set to 0.4 cm and 1.2 cm, respectively; for the male speaker model to 0.45 cm and 1.6 cm.

3 Optimization

3.1 Objective function

Although participants tried to produce stable vowel articulations during the MRI scanning, the intense acoustic noise generated by the scanner severely disrupted auditory feedback. This led to inaccurate vocalizations. Further errors might be introduced during contour tracing, warping transformations, and the anatomical measurements of the upper and lower jaws. Any of these errors may lead to changes in the derived vocal tract area functions and consequently cause deviations in the formants compared to the reference vowel.

To mitigate such effects, we applied optimization algorithms to minimize the spectral deviation between the synthesized and reference vowels. The success of this process relies on a well-designed objective function. We used the root-mean-squared error (RMSE) between mel-scaled log spectral envelopes, which capture not only local spectral features (e.g., formant frequencies and bandwidths) but also global spectral tilt. This representation helps reduce sensitivity to small f_0 differences, which may distort results if raw magnitude spectra are used. Moreover, the use of log envelopes and a mel-frequency scale up to 4000 Hz mimics human auditory perception.

More specifically, the spectral envelope was computed based on cepstral smoothing according to the method in [24] from a 2048-sample frame of the audio signal (48 kHz sampling rate), using a Hamming window. The lifter function $w(\tau)$ in the quefrency domain was defined as:

$$w(\tau) = \begin{cases} 1, & \tau \leq \tau_1 \\ \frac{1}{2} + \frac{1}{2}\cos\left(\pi \frac{\tau - \tau_1}{\tau_2 - \tau_1}\right), & \tau_1 < \tau \leq \tau_2 \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where τ denotes the quefrency, and $\tau_1 = 0.5 \cdot T_0$, $\tau_2 = 0.9 \cdot T_0$. Here, the fundamental period T_0 is estimated from the input frame using an autocorrelation-based method. This lifter produced a spectral envelope that closely follows the harmonic peaks in the magnitude spectrum of input segments, as illustrated in Figure 4, for both male and female voices.

In the following, the spectral envelopes, converted to a mel-frequency scale, are denoted as h_{ref} and $h_{\text{syn}}(x)$ for the reference vowel and the synthetic vowel. The vector x contains both the vocal tract and vocal fold model parameters. The optimization consisted of finding the parameter x that minimized the RMSE between h_{ref} and $h_{\text{syn}}(x)$. To avoid generating physiologically implausible vocal tract shapes, the initial parameter values were taken from the initial (MRI-

based) vowel shapes. Since too small cross-sectional areas in the vocal tract may cause critical constrictions and result in frication noise, a large penalty was applied whenever the minimum cross-sectional area $A_{\min}(x)$ fell below A_{thresh} (30 mm^2). Additionally, we introduced a regularization term that computes the squared Euclidean distance between the optimized vocal tract parameters $x_{1:19}$ and the initial shape parameters $x^{(0)}_{1:19}$, weighted by a penalty coefficient λ. This term helps to limit excessive deviations from the original anatomical configuration. The objective function is therefore defined as follows:

$$x = \arg\min_{x \in B} \left(L_{\text{h}}(x) + L_{\text{a}}(x) + L_{\text{r}}(x) \right) \tag{2}$$

$$L_{\text{h}}(x) = \sqrt{\frac{1}{N} \left\| h_{\text{ref}} - h_{\text{syn}}(x) \right\|_2^2} \tag{3a}$$

$$L_{\text{a}}(x) = \mu \cdot \mathbb{I}\left(A_{\min}(x) < A_{\text{thresh}} \right) \tag{3b}$$

$$L_{\text{r}}(x) = \lambda \cdot \left\| x_{1:19} - x^{(0)}_{1:19} \right\|_2^2 \tag{3c}$$

$$B = \{ x \in \mathbb{R}^n \mid l_b \le x \le u_b \}, \quad \begin{array}{l} l_b = \max\left(x^{(0)} - \delta,\, x^{\min}\right), \\ u_b = \min\left(x^{(0)} + \delta,\, x^{\max}\right) \end{array} \tag{4}$$

where δ denotes the maximum allowed search radius around the initial parameter vector $x^{(0)}$, N denotes the number of frequency bins of the spectral envelope, and n denotes the 29 dimensions of the x.

3.2 Optimization strategy

The optimization process consisted of two steps. In the first step, we performed a global unconstrained optimization across all vowels to determine the search ranges for the parameters in the 2nd optimization step. Due to the regularization constraining deviations from the initial shape, the optimized parameters remained near their initial values, and optimization led to clustering around promising regions. For each vocal tract parameter, we computed the maximal deviation from its initial value over all vowel shapes and defined it as the local search radius $\delta_i = \max |x_i - x_i^{(0)}|$. The new search range was then set to $[x_i^{(0)} - \delta_i, x_i^{(0)} + \delta_i]$, reducing the search space while preserving the likelihood of capturing optima. The second optimization step within these refined bounds yielded better objective values and improved perceptual voice quality.

3.3 Optimization algorithms

Due to the discrete and non-smooth nature of the objective function, conventional gradient-based methods are not usable for our problem. Therefore, we employed gradient-free heuristic optimization algorithms, namely genetic algorithms (GA) and particle swarm optimization (PSO), implemented using MATLAB R2024b's Optimization Toolbox.

For the GA, in each generation 80% of new individuals were generated by crossover using parents selected through the *Stochastic Uniform Selection* function, while the remaining 20% were produced by the *Gaussian mutation*, which was configured with a scale of 0.2 and a shrink value of 0.8 to ensure a fine-grained search capability without compromising global exploration. For the PSO, all default functions and parameter settings were used according to the recommendations of the toolbox, except that the dynamic range of the inertia weight was set to [0.4, 1.2] to prevent premature convergence to local optima. For both algorithms, 50% of the initial population was seeded with initial vocal tract and vocal fold parameters, while the remaining 50% was randomly initialized. The population size and maximum number

Table 3 – Average RMSE in dB between synthesized and reference vowels after each optimization step.

Speaker	Method	Step 1	Step 2	Initial
W01	GA	4.30	3.75	14.53
	PSO	3.51	**3.44**	
M01	GA	6.02	4.94	10.48
	PSO	4.70	**4.66**	

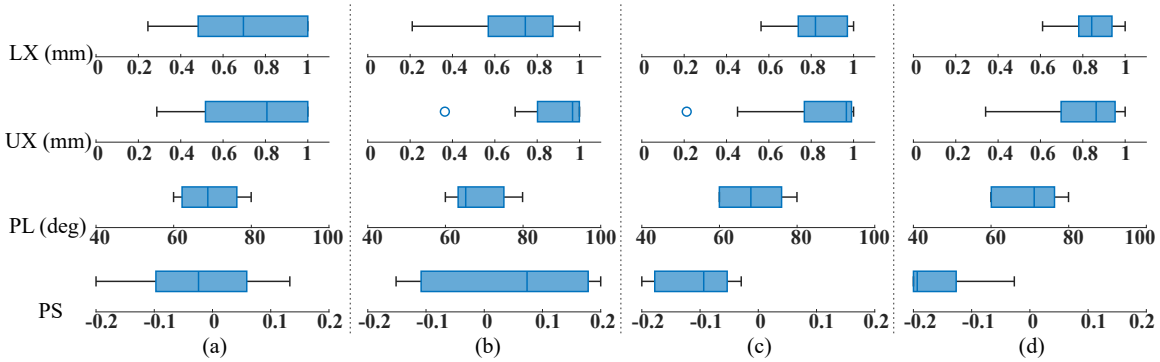

Figure 3 – Boxplots of the optimized vocal fold parameters; (a) M01 with GA; (b) M01 with PSO; (c) W01 with GA; (d) W01 with PSO.

of iterations were both set to 100. Additionally, an early stopping criterion was introduced, terminating the optimization if the loss variation remained below 0.1 dB for 15 consecutive iterations.

4 Results and discussion

The experimental results are summarized in Figures 3–5, and Table 3. Optimization of control parameters using heuristic algorithms substantially reduced the mean spectral envelope differences (in dB) between synthesized and reference vowels. As illustrated in Figure 4, the optimized vowel /a/ exhibits harmonic components closely matching those of the reference. Subjective listening tests further confirmed a clear improvement in vowel quality, often approaching that of the natural recordings. Audio samples of all reference, initial, and optimized vowels are available at `www.vocaltractlab.de/supplement-01`. Interestingly, some vowels with slightly higher loss values were perceived as more natural than those with lower loss, though both outperformed the unoptimized versions. A second optimization step further stabilized the parameter configurations and resulted in a further reduction of the average RMSE across vowels.

As shown in Figure 5, regularization effectively constrained vocal tract parameter shifts, and the vocal fold parameters (Figure 3) showed stable, concentrated distributions. Notably, speaker W01 exhibited larger rest displacements (LX, UX) of the vocal folds than M01, consistent with a breathier voice quality.

In summary, these results demonstrate that employing heuristic optimization algorithms to jointly search for vocal tract and vocal fold parameters is both feasible and effective for improving the perceptual quality of articulatory speech synthesis.

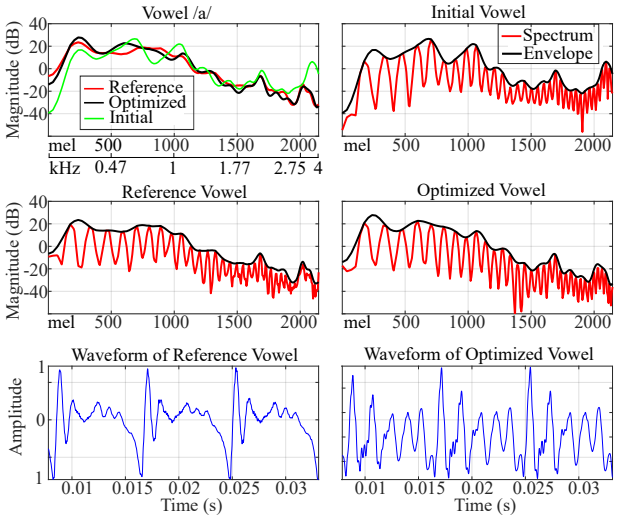

Figure 4 – Mel-frequency scaled spectral envelopes and waveform plots before and after optimization, in comparison to the reference vowel.

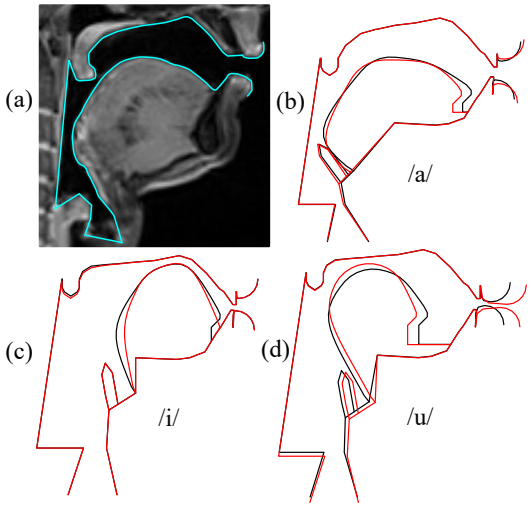

Figure 5 – (a) Traced contour (blue) of vowel /a/ based on MRI data; (b)–(d) M01's vocal tract shapes of /a/, /i/, /u/ before (black) and after optimization (red).

5 Limitations and future work

The spectral envelope calculated for a section of a vowel sound does not discriminate between the voiced (harmonic) and voiceless (anharmonic) components of the signal. Future work could include separating these components of vowel signals for individual optimization, extending the set of speaker models, and conducting formal perceptual evaluations of the synthetic voices.

6 Acknowledgements

This study was funded by the "Zentrales Innovationsprogramm Mittelstand (ZIM)" by the German Federal Ministry for Economic Affairs and Energy (BMWK), grant no. KK5049503FG3.

References

[1] NGUYEN, N.-S., T. V. T. TRAN, H.-N. HUYNH-NGUYEN, T.-S. HY, and V. NGUYEN: *Diflow-tts: Compact and low-latency zero-shot text-to-speech with factorized discrete flow matching*. 2026. URL `https://arxiv.org/abs/2509.09631`. `2509.09631`.

[2] JU, Z., Y. WANG, K. SHEN, X. TAN, D. XIN, D. YANG, Y. LIU, Y. LENG, K. SONG, S. TANG, Z. WU, T. QIN, X.-Y. LI, W. YE, S. ZHANG, J. BIAN, L. HE, J. LI, and S. ZHAO: *Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models*. 2024. URL `https://arxiv.org/abs/2403.03100`. `2403.03100`.

[3] VAN DEN OORD, A., S. DIELEMAN, H. ZEN, K. SIMONYAN, O. VINYALS, A. GRAVES, N. KALCHBRENNER, A. W. SENIOR, and K. KAVUKCUOGLU: *Wavenet: A generative model for raw audio*. *CoRR*, abs/1609.03499, 2016. URL `http://arxiv.org/abs/1609.03499`. `1609.03499`.

[4] ISKAROUS, K., L. GOLDSTEIN, D. H. WHALEN, M. TIEDE, and P. RUBIN: *CASY: The Haskins configurable articulatory synthesizer*. In *International Congress of Phonetic Sciences*, pp. 185–188. 2003.

[5] FELS, S., J. E. LLOYD, K. VAN DEN DOEL, F. VOGT, I. STAVNESS, and E. VATIKIOTIS-BATESON: *Developing physically-based, dynamic vocal tract models using ArtiSynth*. In *Proc. of the 7th International Seminar on Speech Production (ISSP 2006)*, pp. 419–426. Ubatuba, Brazil, 2006.

[6] BIRKHOLZ, P.: *3D-Artikulatorische Sprachsynthese*. Ph.D. thesis, University of Rostock, 2005.

[7] BADIN, P., G. BAILLY, L. REVÉRET, M. BACIU, C. SEGEBARTH, and C. SAVARIAUX: *Three-dimensional linear articulatory modeling of tongue, lips and face, based on MRI and video images*. *Journal of Phonetics*, 30, pp. 533–553, 2002.

[8] ELIE, B. and Y. LAPRIE: *Extension of the single-matrix formulation of the vocal tract: Consideration of bilateral channels and connection of self-oscillating models of the vocal folds with a glottal chink*. *Speech Communication*, 82, pp. 85–96, 2016.

[9] ARNELA, M., S. DABBAGHCHIAN, O. GUASCH, and O. ENGWALL: *MRI-based vocal tract representations for the three-dimensional finite element synthesis of diphthongs*. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(12), pp. 2173–2182, 2019.

[10] CRANEN, B. and J. SCHROETER: *Physiologically motivated modelling of the voice source in articulatory analysis/synthesis*. *Speech Communication*, 19, pp. 1–19, 1996.

[11] MAEDA, S.: *A digital simulation method of the vocal-tract system*. *Speech Communication*, 1, pp. 199–229, 1982.

[12] VAN DEN DOEL, K. and U. M. ASCHER: *Real-time numerical solution of Webster's equation on a nonuniform grid*. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(6), pp. 1163–1172, 2008.

[13] BIRKHOLZ, P.: *Modeling consonant-vowel coarticulation for articulatory speech synthesis*. *PLOS ONE*, 8(4), p. e60603, 2013.

[14] TOUTIOS, A. and S. NARAYANAN: *Simulating anticipatory coarticulation in vcv utterances with a gestural articulatory synthesizer*. In: International Seminar on Speech Production (ISSP) 2020, Providence, RI, 2020.

[15] ALEXANDER, R., T. SORENSEN, A. TOUTIOS, and S. NARAYANAN: *A modular architecture for articulatory synthesis from gestural specification*. *The Journal of the Acoustical Society of America*, 146(6), pp. 4458–4471, 2019.

[16] BIRKHOLZ, P., S. DRECHSEL, and S. STONE: *Perceptual optimization of an enhanced geometric vocal fold model for articulatory speech synthesis*. In *Proc. of the Interspeech*, pp. 3765–3769. 2019.

[17] MERMELSTEIN, P.: *Articulatory model for the study of speech production*. *The Journal of the Acoustical Society of America*, 53(4), pp. 1070–1082, 1973.

[18] GOODYEAR, C. C. and D. WEI: *Articulatory copy synthesis using a nine-parameter vocal tract model*. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-96)*, vol. 1, pp. 385–388. Atlanta, GA, USA, 1996.

[19] TOUTIOS, A., S. OUNI, and Y. LAPRIE: *Estimating the control parameters of an articulatory model from electromagnetic articulograph data*. *The Journal of the Acoustical Society of America*, 129(5), pp. 3245–3257, 2011.

[20] TOUTIOS, A., T. SORENSEN, K. SOMANDEPALLI, R. ALEXANDER, and S. S. NARAYANAN: *Articulatory synthesis based on real-time magnetic resonance imaging data.* In *Interspeech*, pp. 1492–1496. 2016.

[21] DANG, J. and K. HONDA: *Estimation of vocal tract shapes from speech sounds with a physiological articulatory model*. *Journal of Phonetics*, 30(3), pp. 511–532, 2002.

[22] OUNI, S. and Y. LAPRIE: *Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion*. *The Journal of the Acoustical Society of America*, 118(1), pp. 444–460, 2005.

[23] GAO, Y., P. BIRKHOLZ, and Y. LI: *Articulatory copy synthesis based on the speech synthesizer vocaltractlab and convolutional recurrent neural networks*. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32, pp. 1845–1858, 2024. doi:10.1109/TASLP.2024.3372874.

[24] RÖBEL, A. and X. RODET: *Efficient spectral envelope estimation and its application to pitch shifting and envelope preservation*. In *International Conference on Digital Audio Effects*, pp. 30–35. 2005.

[25] JONES, D. R., M. SCHONLAU, and W. J. WELCH: *Efficient global optimization of expensive black-box functions*. *Journal of Global Optimization*, 13, 455-492, 1998. URL `https://doi.org/10.1023/A:1008306431147. 1609.04747`.

[26] KENNEDY, J. and R. EBERHART: *Particle swarm optimization*. In *Proceedings of ICNN'95-International Conference on Neural Networks*, vol. 4, pp. 1942–1948. 1995.

[27] BIRKHOLZ, P.: *Control of an articulatory speech synthesizer based on dynamic approximation of spatial articulatory targets*. In *Interspeech 2007 - Eurospeech*, pp. 2865–2868. Antwerp, Belgium, 2007.

[28] BIRKHOLZ, P., B. J. KRÖGER, and C. NEUSCHAEFER-RUBE: *Model-based reproduction of articulatory trajectories for consonant-vowel sequences*. *IEEE Transactions on Audio, Speech and Language Processing*, 19(5), pp. 1422–1433, 2011.

[29] BIRKHOLZ, P., S. KÜRBIS, S. STONE, P. HÄSNER, R. BLANDIN, and M. FLEISCHER: *Printable 3d vocal tract shapes from MRI data and their acoustic and aerodynamic properties*. *Scientific Data*, 7(255), pp. 2052–4463, 2020. doi:10.1038/s41597-020-00597-w.

[30] ALKU, P. and E. VILKMAN: *A comparison of glottal voice source quantification parameters in breathy, normal and pressed phonation of female and male speakers*. *Folia Phoniatrica et Logopaedica*, 48, pp. 240–254, 1996.

[31] BARNEY, A., A. DE STEFANO, and N. HENRICH: *The effect of glottal opening on the acoustic response of the vocal tract*. *Acta Acustica united with Acustica*, 93(6), pp. 1046–1056, 2007.

[32] TITZE, I. R.: *A four-parameter model of the glottis and vocal fold contact area*. *Speech Communication*, 8, pp. 191–201, 1989.