JOINT OPTIMIZATION OF ANATOMICAL AND GESTURAL PARAMETERS IN A PHYSICAL VOCAL TRACT MODEL

Christopher Liberatore, Ricardo Gutierrez-Osuna

Department of Computer Science and Engineering, Texas A&M University, College Station, TX

ABSTRACT

We describe a method for adapting a physical vocal tract model's anatomical and gestural parameters using acoustic information to match a target speaker. Physical vocal tract models are hard to adjust to match a speaker, as doing so requires information which is difficult to capture, such as X-Ray or MRI information. We propose an analysis-by-synthesis approach to adjust the parameters of the VocalTractLab (VTL) physical vocal tract model, optimizing on an acoustic distance objective function. We compare our method with one which does not adjust anatomy parameters, just gestural parameters, and find that the proposed method results in a net improvement. We also test our method's ability to recreate a synthetic speaker for which the ground truth parameters are known, and find that the method can reproduce the speaker if parameters pertaining to teeth and lips are fixed.

Index Terms—optimization, vocal tract model, gestures, speaker inversion

1 INTRODUCTION

Physical models of the vocal tract rely on detailed representations of the tract for synthesis of acoustics. They simulate the anatomy and articulatory gestures by modeling vocal tract features as two or three-dimensional meshes, computing the cross-sectional volume on the midsagittal plane, and approximating this volume as a series of connected tubes. The modeled volume then filters an excitation signal generated by a simulated glottis. One such vocal tract model is VocalTractLab (VTL) [1, 2], which presents the vocal tract as a set of seven distinct 3-dimensional meshes, a separate glottis model, and a model of vocal tract dynamics.

Tuning such vocal tract models to match the voice quality of a particular speaker requires manual adjustment of the parameters, which usually require access to specialized tools such as MRI or X-Ray imaging [3]. A method that could adapt a model's parameters to estimate a speaker's underlying anatomical and articulatory parameters without requiring such imaging would make physical models more accessible to multiple applications of speech synthesis, speech recognition, or speaker recognition.

Previously, Birkholz and Kröger [3] matched VTL anatomy and gestural parameters by manually-aligning the model parameters from X-Ray or MRI information. In later work [4], the authors used VTL to examine articulatory differences between a child speaker and adult speaker by adjusting the anatomical and gestural parameters. They parameterized the anatomy of VTL by scaling the anatomy meshes. The authors scaled the anatomy and gesture parameters of a model of a German male to a reference 11-year-old boy and found that the scaled model produced different formants compared to the expected scaled formants, suggesting that children's gestures are not simply scaled representations of the corresponding adult shapes. To recreate the process by which humans acquire speech production knowledge, Prom-on et al. [5] demonstrated a method to optimize VTL to learn articulatory configurations of a vocal tract from acoustic features. Using an analysis-by-synthesis approach, they minimized the sum-squared-error of the synthesized Mel Frequency Cepstral Coefficients (MFCC) against target MFCCs. The authors showed it was possible to begin from a random vocal tract configuration and iteratively optimize the model until an acceptable configuration was reached with similar acoustics. In a follow-up study [6], the authors showed that this method had cross-language potential, using a similar distal method to learn articulatory configurations of Thai vowels.

Relation to prior work: Our proposed method is similar to the optimization methods in these prior studies. However, it differs in that we seek to learn not only the articulatory configurations for particular speech sounds but also the overall anatomy of the speaker via analysis-by-synthesis. We find that including an anatomical optimization step prior to optimizing specific vowel shapes reduces the mean acoustic error when compared to optimizing vowel shapes alone. To test the method's ability to reproduce a speaker, we created a synthetic speaker based upon the default VTL model and optimized the anatomy parameters against acoustics from this model. We found that our method could find the synthetic model parameters best if teeth and lip parameters were not included in optimization.

The remainder of the paper is outlined as follows. First, we review VTL physical model and our analysis-by-synthesis optimization process. Following this, we discuss and review the results of our method on vowel samples from three speakers and from a synthetic speaker. Finally, we discuss future refinements of our proposed method.

2 METHODS

2.1 Vocal tract model

Our work is based on VocalTractLab (VTL), a physical vocal tract model developed by Peter Birkholz [1, 2]. VTL

represents the vocal tract using seven wireframe meshes, and controls the articulation of these surfaces with 23 tract parameters. VTL also includes a synthesizer that filters a source signal from an included glottis model through the vocal tract by extracting a tube model from a specific vocal tract configuration.

The default VTL model is based on a German speaker [3], and includes many gestural targets (called "shapes" in VTL parlance). Of the 23 shape parameters in VTL, we were concerned with only of 18 of these parameters, shown in Table 1. The remaining 5 parameters, which we do not consider here, affect dynamics (i.e. motion) or the tongue root position, the latter of which VTL can infer from other tongue parameters. Following Birkholz and Kroger [4], we parameterize the anatomy with 13 scaling parameters describing the scaling of the mesh of a default anatomy model; see Table 1. The API for VTL allows two synthesis methods: one which outputs acoustics of an entire series of gestures, and one which allows for the user to specify the value of each articulator in 5-ms intervals. The latter method also returns the cross-sectional tube areas in addition to synthesized acoustics, and is the method we use here.

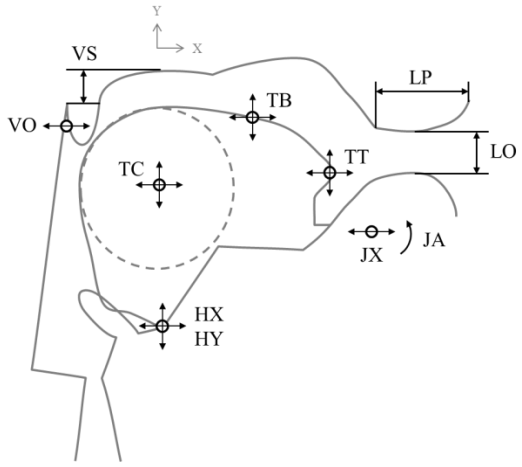
2.2 Analysis-by-synthesis

Our proposed analysis-by-synthesis approach consists of a two-tiered, iterative optimization method. The first tier treats the speaker as a collection of shapes associated with anatomy parameters, and optimizes the parameters associated with anatomy in this context. The second tier examines individual vowel shape configurations, optimizing them atomically. After shape optimization, we began another iteration of optimization, starting again with the anatomy. We approached the optimization in this two-tiered fashion as it would be difficult to attempt to optimize the anatomy and shape parameters simultaneously.

For both tiers of optimization, we used pattern search (PS), a form of direct search, to optimize the parameters for both anatomy and individual vowel shapes. PS is a form of direct search that does not require a local gradient to optimize objective function parameters. PS is also robust to noise that otherwise negatively affects methods requiring gradient computation [7]. Though PS does not require a local gradient, it behaves in a similar manner to gradient descent; an in-depth discussion is presented in [8].

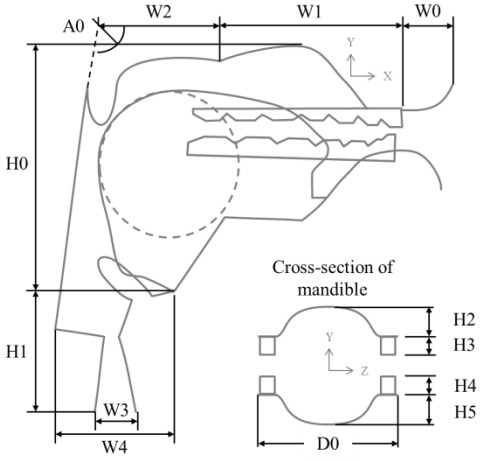
PS navigates the error surface by creating a mesh of nearby points by permeating the current solution on each vector in a set of basis vectors. Typically, this basis consists of the unit vectors of each dimension of the parameter space. PS evaluates each of these permeations by the objective function in what is called a "polling" step. PS chooses the point with the greatest reduction in the objective function as the center of the mesh for the next iteration. Following a poll, PS scales the mesh in response to

Figure 1: VocalTractLab shape parameters. All parameters are in centimeters except for *VS* and *VO*, which are unitless, bounded [0,1], and *JA*, which is in degrees. (From [3])



Param	Description	Param	Description
HX, HY	Hyoid	*TTX, TTY*	Tongue tip
JX	Jaw position	*TBX, TBY*	Tongue body
JA	Jaw angle	*TCX, TCY*	Tongue center
LP	Lip protrusion	*TS1-4*	Tongue side height
LD	Lip distance	*VS*	Velum shape
		VO	Velic opening

Figure 2: parameterization of VTL anatomy mesh. This parameterization represents various scaling dimensions of the default vocal tract model. All units of these parameters are in centimeters, except for *A0*, which is in degrees. (From [4])



Param	Description	Param	Description
H1	Larynx length	*W0*	Lip width
H2	Palate height	*W1*	Hard palate length
H3	Upper molars	*W2*	Soft palate length
H4	Lower molars	*W3*	Vocal fold length
H5	Mandible height	*W4*	Larynx width
D0	Palate depth	*A0*	Oral-Pharyngeal angle

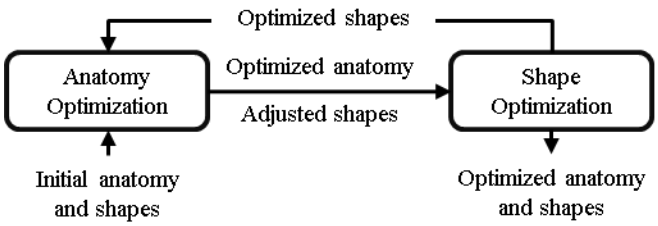


Figure 3: block diagram of optimization method. For Anatomy-Shape optimization, we exited the optimization loop after 2 iterations.

observations in response to the error surface. In the event of a successful poll, the mesh size is expanded by a constant factor (bounded by a maximum size parameter) to speed up convergence. However, in the case of an unsuccessful poll, the mesh is contracted. PS converges once the mesh size contracts to below a specified size.

In our case, the objective function is the acoustic difference between a target vowel and the synthesized vowel, which we compute as the sum-squared error between the mean of the synthesized MFCCs and the mean of the target vowel's MFCCs. Specifically, we synthesized 150 ms of the vowel acoustics, ignoring the first 25 ms due to a documented transient in the signal, and use the remaining 125 ms. We computed $MFCC_{1\text{-}24}$ using RASTAMAT [9] over 35 ms windows, sliding at 10 ms intervals, and drop the first coefficient as we wish to ignore the synthesis energy. We use no liftering in MFCC computation.

To prevent the optimization algorithm from choosing a state that would generate turbulent airflow, we added a penalty factor in our acoustic error measure if the minimum tube area is less than 0.25 cm^2. Our error function for a given vowel can be expressed in the following form:

$$E_{vowel} = \begin{cases} \rho + \Sigma(\hat{c} - c)^2 & \text{if} \min(\boldsymbol{t}) < 0.25 \\ \Sigma(\hat{c} - c)^2 & \text{otherwise} \end{cases} \quad \textbf{(1)}$$

where n is the nth vowel, ρ is the penalty factor, $\hat{\boldsymbol{c}}$ are the mean MFCCs of the synthesized acoustics, $\boldsymbol{c}$ are the mean MFCCs of the target acoustics, and $\boldsymbol{t}$ is a vector containing the cross-sectional areas of the tube model in cm^2 units. We found it most effective if ρ is set to a value large enough that it dominated the error of nearby articulatory configurations, creating a constraint on the search space. VTL exhibited variability in synthesis, resulting in noise in the objective function, with a standard deviation of approximately $\sigma = 0.25$.

Anatomy		Shape	
Initial size	5%	*Initial size*	0.1σ
Scale factor	2	*Scale factor*	2.1544
Max size	5%	*Max size*	1σ
Converge	1.25%	*Converge*	0.01σ

Table 1: pattern search parameters. Anatomy optimization has step sizes which start at 5% of the total range of each parameter, and converge when the size is below 1.25%. We chose the scaling factor of the Shape optimization to provide 7 steps, between 0.01σ and 1 σ, where σ is the standard deviation of the associated search vector, as computed from the vowels included in the VTL speaker model.

2.2.1 Anatomy Optimization

In the first optimization step, we optimize the anatomy parameters on the sum of the errors (as computed from eq. (1)) across all the vowels we seek to optimize. The anatomy parameters define the scaling of various meshes of the vocal tract model, but they also affect the upper and lower bounds for each articulator in the vocal tract—as anatomy parameters change, so do the bounds of articulators associated with those parameters. Because each anatomy parameter has different bounds with varied parameter ranges, we adjust each step size to be a proportion of the range of that parameter. The parameters for pattern search are shown in Table 1.

To capture this fact in our optimization process, whenever anatomy parameters are updated we also scaled the vowel shapes linearly with the changes in these bounds:

$$\hat{s}_i = (l_i - s_i)\left(\frac{\hat{u}_i - \hat{l}_i}{u_i - l_i}\right) + \hat{l}_i \quad \textbf{(2)}$$

where $\boldsymbol{i}$ is the $\boldsymbol{i^{th}}$ shape parameter, $\boldsymbol{l_i}$and $\boldsymbol{u_i}$ are the old lower and upper bounds, and $\hat{\boldsymbol{l}}_{\boldsymbol{i}}$and $\hat{\boldsymbol{u}}_{\boldsymbol{i}}$ are the new lower and upper bounds.

2.2.2 Shape optimization

Following anatomy optimization, we optimize each vowel shape using pattern search against the objective function in eq. (1). To capture correlations in the positioning of vowel shapes, we perform Principal Components Analysis (PCA) on the 23 vowel shapes included in VTL. PCA extracts an orthogonal basis that represents the dimensions of maximum variance by computing the eigenvectors of the covariance matrix of a dataset. As we have more vowels than dimensions, the covariance matrix has full rank and we extract 18 components. We scale these components by their observed standard deviations (σ) and use them as the basis for the PS mesh; the unit for each vector then becomes one standard deviation along that component. The parameters were chosen so as to allow for features as small as 1 mm to be captured by the PS mesh. The remaining PS parameters are summarized in Table 1.

3 EXPERIMENTS AND RESULTS

We extracted vowel samples from three male speakers (JW11, JW12, and JW15) in the University of Wisconsin XRay Microbeam (XRMB) Database [10]. These speakers were selected as they were all male speakers, as we wanted to minimize any cross-gender effects, as VTL's default model is based on a male German speaker. We used 5 vowels from task TP14 and manually selected segments which had the most constant formant trajectories, extracting MFCCs from these segments. We also extracted the average pitch over that same segment using STRAIGHT analysis

		JW11		JW12		JW15	
		ASO	**SO**	**ASO**	**SO**	**ASO**	**SO**
ae	[æ]	**19.7**	20.6	32.9	**32.4**	**26.8**	38.9
ah	[ɑ]	**39.7**	48.9	**14.7**	27.9	**27.9**	72.6
ee	[i]	**29.8**	30.7	**34.2**	51.5	**57.4**	66.2
eh	[ɛ]	**18.9**	33.3	**21.3**	29.1	**18.3**	24.4
oo	[ʊ]	**31.4**	32.7	**33.3**	54.9	38.3	**34.7**
Total		139.5	166.2	136.4	195.8	168.7	236.8

Table 3: detailed results of the first experiment, comparing AS optimization (ASO) with Shape optimization (SO). While the sum of the errors was reduced in all cases using the AS method over the SO baseline, the reduction was not uniform across all vowels. The first column is the XRMB prompt in task TP14, the second being the corresponding IPA vowel label.

[11]. The acoustic targets for each vowel were encapsulated as the mean MFCCs and mean pitch value of the segment. During optimization, synthesized vowels used this mean pitch value.

To provide a baseline for comparing the performance of our proposed algorithm, we performed an experiment where we compared our proposed method with a shape-only optimization method. We used the default anatomy parameters for the shape-only method, and in both instances, the initial configurations for all vowels were the same.

3.1 Results

The proposed algorithm reduced the sum of errors over the baseline, shape-only optimization; however, not every vowel was improved—see Table 3. Some shapes, such as /ɑ/ in JW15 and /i/ in JW12, were significantly improved by including an anatomy optimization step, but /æ/ in JW12 and /ʊ/ in JW15 were not.

To verify the ability of pattern search to reach known anatomy configurations, we created a synthetic speaker with longer hard and soft palate parameters, and ran anatomy optimization on the synthetic speaker. We used four vowels representing the extrema of the vowel space (/i/, /a/, /ɑ/, /u/), as well as *schwa*, in the optimization. We found that when we included all anatomy parameters in the optimization process, before reaching the known global minimum, PS would adjust lip width (W0) or tooth height parameters (H2, H3), before continuing to adjust the hard and soft palate lengths. This would result in the search algorithm converging at a local minimum to different parameters than the known ground-truth speaker. When we tested the optimization method fixing the dental, lip, and jaw parameters (W0, D0, H2-H5) only allowing optimization on the palate, larynx, and pharynx parameters, we found that the search algorithm was able to find a similar configuration as the synthetic speaker—see Table 2. This suggests that the method is capable of navigating the error surface and finding known ground-truth values, but that it is sensitive to the parameters of the model being optimized, or the mesh size for each parameter. Further refinements to the anatomy optimization procedure, either by choosing more precise mesh sizes or optimizing different anatomy parameters separately may improve results.

	All	**Limited**	**Target**
W1	5.15 cm	5.13 cm	5.1 cm
W2	2.69 cm	3.09 cm	3.1 cm
Error	13.86	0.74	

Table 2: effect of dimensions on anatomy optimization using a synthetic speaker. A synthetic speaker was made based upon the default VTL model, but with hard and soft palate (W1 and W2) lengthened—see target parameters in the last column. "All" is the result of anatomy optimization on all dimensions of the anatomy parameterization, while "Limited" fixed the values of the lip, jaw, and dental parameters. The error metric is derived in the same manner as equation (1).

4 DISCUSSION AND FUTURE WORK

We found that our proposed optimization method, which adjusted both anatomy and gestural parameters, improved the observed total acoustic errors over just adjusting gestural parameters on a default anatomy model. However, even though the sum of errors was improved, not all vowels showed improvement. Additionally, we found that pattern search was able to find the known correct parameters for a synthetic speaker, provided we restricted the search space.

Problems with not all vowels showing an improvement over the shape-only optimization method may be explained by the anatomy parameterization: they simply scaled the default mesh and not finer details such as anatomy curvature which may be specific to a given speaker. Moreover, when comparing the optimization results to a synthetic speaker for which we knew the ground-truth parameters, including teeth and lip parameters in the optimization process hampered the search algorithm, forcing it into a local minimum.

4.1 Future Work

Our results suggest some additional ways to improve the performance of our proposed method. First, splitting up the anatomy optimization to focus on vocal tract parameters (ignoring lip and teeth parameters) may improve the optimization results. Evaluating the AS method on VCV and CVC sequences could further improve results, as consonant constrictions are more sensitive to the dimensions of the vocal tract. Further dimensionality reduction by optimizing on dimensions of known correlation in the vocal tract (e.g. principal components of the vocal tract area function [12]) would also serve to decrease the total optimization time.

Learning vocal tract parameters in this distal manner would be useful in synthesis and conversion applications. Accent conversion [13, 14], requires the synthesis of acoustics with a target speaker's voice quality, but using acoustics which that speaker has not uttered. Using this method to modify a physical vocal tract model to match a target speaker, one could simulate native phonemes with the target speaker's voice quality, potentially overcoming a current issue in accent conversion methods [14].

5 REFERENCES

[1] P. Birkholz, et al., "Construction and control of a three-dimensional vocal tract model," in IEEE International Conference Acoustics, Speech and Signal Processing (ICASSP), Toulouse, France, 2006, pp. 873-876.

[2] P. Birkholz. (2010). Vocal Tract Lab. Available: www.vocaltractlab.de

[3] P. Birkholz and B. J. Kröger, "Vocal tract model adaptation using magnetic resonance imaging," in 7th International Seminar on Speech Production, 2006, pp. 493-500.

[4] P. Birkholz and B. J. Kröger, "Simulation of vocal tract growth for articulatory speech synthesis," in International Congress Phonetic Sciences (ICPhS), Saarbrücken, Germany, 2007, pp. 377-380.

[5] S. Prom-on, et al., "Training an articulatory synthesizer with continuous acoustic data," in INTERSPEECH, 2013, pp. 349-353.

[6] S. Prom-on, et al., "Identifying underlying articulatory targets of Thai vowels from acoustic data based on an analysis-by-synthesis approach," EURASIP Journal on Audio, Speech, and Music Processing, vol. 2014, p. 23, 2014.

[7] A. J. Booker, et al., "A rigorous framework for optimization of expensive functions by surrogates," Structural optimization, vol. 17, pp. 1-13, 1999.

[8] V. Torczon, "On the convergence of pattern search algorithms," SIAM Journal on optimization, vol. 7, pp. 1-25, 1997.

[9] D. P. W. Ellis. (2005). PLP and RASTA (and MFCC, and inversion) in Matlab. Available: http://ee.columbia.edu/~dpwe/resources/matlab/rastamat/

[10] J. Westbury, et al., "X-ray microbeam speech production database," The Journal of the Acoustical Society of America, vol. 88, pp. S56-S56, 1990.

[11] H. Kawahara, et al., "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," Speech communication, vol. 27, pp. 187-207, 1999.

[12] P. Mokhtari, et al., "Principal components of vocal-tract area functions and inversion of vowels by linear regression of cepstrum coefficients," Journal of Phonetics, vol. 35, pp. 20-39, 2007.

[13] S. Aryal and R. Gutierrez-Osuna, "Reduction of non-native accents through statistical parametric articulatory synthesis," The Journal of the Acoustical Society of America, vol. 137, pp. 433-446, 2015.

[14] D. Felps, et al., "Foreign accent conversion through concatenative synthesis in the articulatory domain," Audio, Speech, and Language Processing, IEEE Transactions on, vol. 20, pp. 2301-2312, 2012.