EFFICIENT EXPLORATION OF ARTICULATORY DIMENSIONS

Paul Konstantin Krug[1], *Peter Birkholz*[1], *Branislav Gerazov*[2], *Daniel Rudolph van Niekerk*[3], *Anqi Xu*[3], *Yi Xu*[3]

[1]*Institute of Acoustics and Speech Communication, Technische Universität Dresden, Germany*
[2]*Faculty of Electrical Engineering and Information Technologies, Ss. Cyril and Methodius University in Skopje, Republic of North Macedonia*
[3]*Department of Speech, Hearing and Phonetic Sciences, University College London, United Kingdom*
paul_konstantin.krug@tu-dresden.de

Abstract: The key to a successful simulation of speech acquisition with a parametric articulatory synthesizer lies, inter alia, in the successful exploration of its articulatory dimensions. However, such an exploration (regardless of the respective algorithm) may be non-trivial due to the high dimensionality of a modeled vocal tract and the associated high probability of creating unnatural or humanly impossible vocal tract shapes. In this work, a method based on principal component analysis is used to reduce the scope of motor space of the articulatory synthesizer VOCALTRACTLAB. It is shown that such a technique can be used to increase the computational efficiency of vocal learning simulations and thus may help to establish better exploration-based acoustic-to-articulatory-inversion models.

1 Introduction

Simulations of early human vocal learning are of great importance because they may provide answers to questions both in the fields of phonetics and child development as well as in the area of motor learning [1]. Speech acquisition also represents a type of acoustic-to-articulatory inversion that has potentially great utility for both speech synthesis systems and speech recognition, e.g. in situations that only allow for a low amount of speech resources [2].
When it comes to simulating the imitative vocal learning process, articulatory synthesizers are particularly well suited because they allow full control over the articulatory dimensions, and thus the simulation of motor learning [1, 3]. In contrast to the direct learning of the motor trajectories as in [4], the early vocal learning scenario requires the objective function to be computable from observables. Since most articulators are unobservable, the objective function must be based on the consequences of the articulatory movements, i.e. acoustic information. Thus, a speech acquisition simulation can be seen as a reinforcement learning process, whereby an agent (the learner) tries to develop an action space based policy in order to produce an acoustic consequence similar to an observed reference. Such a simulation is technically challenging because acoustic similarity does not necessarily correspond to perceptual similarity due to the following reasons:

- Speech sounds to be imitated (references) and the produced imitations may be misaligned in terms of timing in a non-linear, time-dependent way.
- Acoustic templates and imitations may be misaligned in terms of frequencies in a non-linear way, which is due to the vocal tract differences among the speaker and the target-speaker. This is also known as the speaker-normalization problem.

	Description	Parameter	Minimum	Standard	Maximum	Unit
1	Hyoid position (horz.)	HX	0.0	1.0	1.0	cm
2	Hyoid position (vert.)	HY	-6.0	-4.75	-3.5	cm
3	Jaw position (horz.)	JX	-0.5	0.0	0.0	cm
4	Jaw angle	JA	-7.0	-2.0	0.0	deg.
5	Lip protrusion	LP	-1.0	-0.07	1.0	cm
6	Lip distance	LD	-2.0	0.95	4.0	cm
7	Velum shape	VS	0.0	0.0	1.0	
8	Velic opening	VO	-0.1	-0.1	1.0	cm^2
9	Tongue body (horz.)	TCX	-3.0	-0.4	4.0	cm
10	Tongue body (vert.)	TCY	-3.0	-1.46	1.0	cm
11	Tongue tip (horz.)	TTX	1.5	3.5	5.5	cm
12	Tongue tip (vert.)	TTY	-3.0	-1.0	2.5	cm
13	Tongue blade (horz.)	TBX	-3.0	2.0	4.0	cm
14	Tongue blade (vert.)	TBY	-3.0	0.5	5.0	cm
15	Tongue root (horz.)	TRX	-4.0	0.0	2.0	cm
16	Tongue root (vert.)	TRY	-6.0	0.0	0.0	cm
17	Tongue side elevation 1	TS1	0.0	0.0	1.0	cm
18	Tongue side elevation 2	TS2	0.0	0.0	1.0	cm
19	Tongue side elevation 3	TS3	-1.0	0.0	1.0	cm

Table 1 – The supra-glottal parameters of the articulatory synthesizer VTL.

- Measuring the distance between acoustic representations (such as spectrograms) of a reference and imitations requires careful normalization and a weighting of specific parts such as consonants and vowels.

These issues make it difficult, if not impossible, to directly compare template and imitated utterances, e.g. by calculating the spectral distance between both. This even applies in the case of copy synthesis, where the same vocal tract is used for references and imitations. While these problems may be solved with the help of neural networks, which can translate the time series input into percept-vectors, another issue persists:

- Any search or optimization of speech parameters suffers from the curse of high dimensionality, as articulatory synthesizers typically have numerous degrees of freedom.

This issue may be addressed by constraining articulatory parameters to specific values or ranges that are motivated by phonetic knowledge [2]. However, such constraints may not generalize among different articulatory synthesizers and, even more importantly, introduce explicit expert knowledge into the simulation that a real learner may not have. In this work, principal components (PCs) are used to span a subspace of natural vocal tract shapes. Several ways are shown in which this space can be used to circumvent the problem of unnatural vocal tract configurations. While this study is specific to the state-of-the-art [5] articulatory synthesizer VOCALTRACTLAB (VTL) [6] version dev-2.4, the proposed techniques are general approaches that can potentially be extended to other parametric articulatory synthesizers as well.

2 Methods

2.1 VocalTractLab

VTL is an articulatory speech synthesizer that provides a realistic human vocal tract model, based on magnetic resonance imaging (MRI) data. The synthetic speech is generated via a

one-dimensional aero-acoustic simulation of the vocal tract dynamics. Three different types of glottis models are available. While the VTL allows for model-based high level controls, such as phoneme-to-speech [7] via so called *gestural scores*, a direct low-level operation of the individual articulatory parameters, which is the prerequisite for the simulation of speech acquisition, is also possible. The current VTL vocal tract model is controlled by 19 supra-glottal parameters (see Table 1) and either 6 or 11 glottal parameters, depending on which glottis model is selected. In this work the geometric glottis model with 11 parameters is used, which is the VTL default. However, glottal parameters are not the subject of this study.

2.2 Principal Component Space

VTL provides 69 predefined vocal tract shapes that represent most of the German phonemes. These shapes were derived from MRI data and thus, are biologically plausible. Principal component analysis (PCA) models can be trained on a specific number of the 19 dimensions of the predefined shapes. For general purposes it is reasonable to exclude the parameters VO, TRX, TRY for the following reasons: The velum opening parameter VO is orthogonal to all other dimensions as it can be independently controlled in order to introduce nasality to a certain speech sound. As long as no nasal sounds should be created it can (and should) be set to its default value of -0.1 (velum closed). The tongue root parameters TRX and TRY can be set to any arbitrary value because the VTL allows for an automatic calculation of the parameters. Only if the vocal tract model geometry itself was modified, e.g. by scaling the model, a manual adjustment of these parameters would be needed. Further, the parameter TS3, which controls the side elevation of the tongue tip may be excluded because it has no significant impact on the plausibility of a given shape. However, in case that this parameter is excluded, it must be fixed or resampled after PC decoding. This may be acceptable if the goal of the PC transformation is a general sampling over certain parameter ranges. However, if the goal is to integrate the PCA model in an active optimization strategy, including TS3 may be more convenient.
A trained PCA model can be used in several ways: (i) Tract states $\mathbf{v}_i$ can be sampled within the full vocal tract (VT) parameter space, encoded into the PC space and then decoded again (VT$\rightleftarrows$PC). (ii) PC states $\mathbf{p}_j^*$ can be sampled within the PC space directly and then decoded into the VT space (PC$\rightarrow$VT). While the implementation of the first method is straight forward, one needs to define the boundaries of the sampling range in case of the second method since the PC space is unbounded. Useful boundaries may be obtained in the following way: A large set of vocal tract states is sampled uniformly from the respective parameter ranges. The set gets transformed into a PC space. The individual PC values will be Gaussian distributed in every dimension. Boundaries for uniform sampling methods may then be defined as $[\boldsymbol{\mu}_{\mathrm{PC}} \pm \mathrm{n}_\sigma \cdot \boldsymbol{\sigma}_{\mathrm{PC}}]$, see Figure 1 (left plot). The parameter n_σ has a strong influence on resulting vocal tract state distributions sampled from the PC space (Figure 1, middle plot) and can be tuned as desired (Figure 1, right plot).

2.3 Uniform Vocal Tract Exploration

A set of 10^5 random *open* vocal tract states was drawn from a uniform distribution within the limits of the allowed range (see Table 1) of the respective articulatory parameters. Thereby, *open* means that the minimum cross-sectional area (see left plot in Figure 2) of the corresponding tube is $\mathrm{T}_{\mathrm{Min}} \geq 0.3\,\mathrm{cm}^2$. Such states produce vowels when excited with a modal voice. The states were subsequently encoded and decoded using a PCA model, which was trained on 15 of the total 19 dimensions of the predefined shapes, excluding the parameters VO, TRX, TRY and TS3 for the reasons explained earlier. The number of PC dimensions was chosen to be 7, which explain approximately 95 % of the observed variance in the training data.

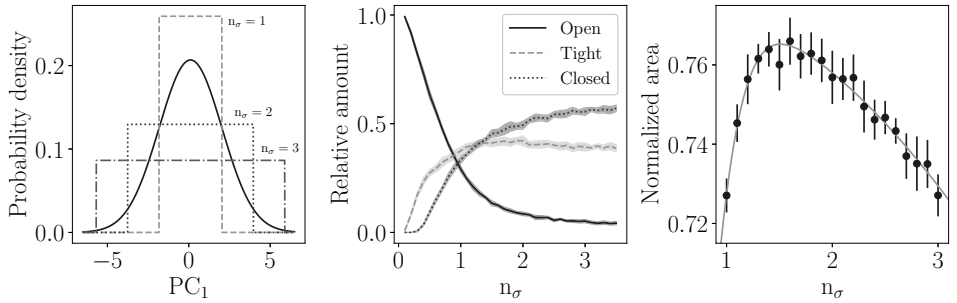

Figure 1 – Left: States sampled uniformly from the VT space, will be normal distributed in the PC space (solid line). When sampling from the PC space, boundaries of the uniform distributions may be set to $[\boldsymbol{\mu}_{\mathrm{PC}} \pm \mathrm{n}_\sigma \cdot \boldsymbol{\sigma}_{\mathrm{PC}}]$. Here, the distributions for $\mathrm{n}_\sigma = 1, \ldots, 3$ are shown for a single dimension PC_1 as a dashed, dotted and dash-dotted line, respectively. **Middle:** Amount of open ($\mathrm{T}_{\mathrm{Min}} \geq 0.3\,\mathrm{cm}^2$), tight ($0.3\,\mathrm{cm}^2 > \mathrm{T}_{\mathrm{Min}} > 0.001\,\mathrm{cm}^2$) and closed states ($\mathrm{T}_{\mathrm{Min}} \leq 0.001\,\mathrm{cm}^2$) as a function of the sampling hyper parameter n_σ. **Right:** Normalized area of histogram entries that lie within the convex hull formed by predefined VTL vowels in f_{R1}-f_{R2} space, visualized as a function of n_σ. A scaled skew normal fit is shown as a solid line.

Additionally, 10^5 random open tract states were drawn from a uniform distribution within in the PC space using the same PCA model. The limits of the uniform distributions were determined by generating two dimensional f_{R1}-f_{R2} distributions. Thereby, the first and second tube resonances f_{R1} and f_{R2} were derived from the respective volume-velocity transfer functions (see right plot in Figure 2). These resonances are closely related to the first two formants of the vowels that would be produced if the corresponding vocal tract states were excited. The f_{R1}-f_{R2} distributions were generated ten times as histograms (100 bins in each dimension), for each $\mathrm{n}_\sigma \in \{1, 1.1, 1.2, \ldots, 3\}$. The area covered by bins with more than one entry that lie within the convex hull formed by the f_{R1}-f_{R2} data of the predefined VTL vowel shapes (see Figure 4) was calculated and divided by the total area of the convex hull. The resulting normalized area is visualized in Figure 1 (right plot) as a function of n_σ. The parameter n_σ was chosen so that it maximizes the normalized area. The optimal value was found to be $\mathrm{n}_\sigma = 1.5$ in this case.[1]
In both cases of the PCA method application, the excluded parameters VO, TRX, TRY were set to their default values and TS3 is set to a random value sampled from the uniform distribution in the TS3 parameter range.
The three resulting vocal tract state distributions were compared in terms of their f_{R1}-f_{R2} tube resonance distributions.

2.4 Simulation of Goal-directed Babbling

In order to test the general effectiveness of the PCA approach, a simple simulation of goal-directed babbling was performed. Thereby, an acoustic reference matrix, which was an $(80 \times \mathrm{n}_F)$ dimensional log-mel scaled spectrogram (whereby n_F denotes the number of spectrogram time frames) of a reference utterance was to be approximated by a vocal learning agent following a specific policy. By using the VTL speaker as the reference voice, both the timing of the speech signals and the (potential) vocal tract length differences are under control. The spectral weighting issue is solved by only optimizing a single articulatory target at once. Hence, the high number of dimensions remains the only important variable that influences the success of

[1] It is important to note that the optimal value for n_σ needs to be re-calculated for each new configuration of a PCA model, and each sampling method, e.g. Gaussian distributed sampling needs a different value than uniformly distributed sampling.

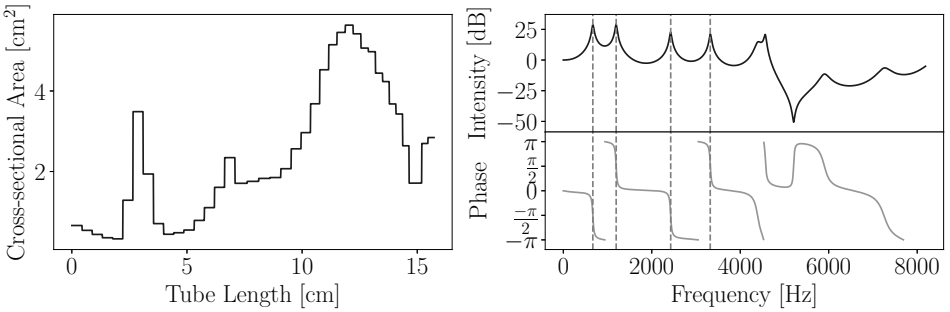

Figure 2 – Vocal tract state related functions calculated for the predefined VTL shape /a/. **Left:** Cross-sectional area of the tube as a function of the tube length. **Right:** The magnitude and phase spectra of the volume-velocity transfer function are visualized in the top and bottom plots, respectively.

the acoustic-to-articulatory-inversion.

2.4.1 Vowel Learning

In case of vowel imitations the goal state was formed by a single articulatory vector. The vowels /a, e, i, o, u, ø, y, ɛ, ɪ, ɔ, ʊ, œ, ʏ, ɐ/ were generated with VTL and used as references. The acoustic references, as well as the imitations were synthesized using a modal voice glottis state. Synthetic audio samples had a duration of 400 ms.

A constrained greedy random search (CGRS) strategy was used as the vocal learning policy. The algorithm is visualized in Figure 3. Starting from a search space position $\mathbf{v}_i$ (whereby the initial position $\mathbf{v}_0$ is the predefined /ə/ state in this case), a number of $n_D = 100$ potential states is sampled around $\mathbf{v}_i$ using normal distributed noise with dimension-specific standard deviations. The noise vector is denoted as δ, whereby each noise dimension corresponds to 10 % of the respective vocal tract parameter range. The n_D roll outs are constrained to be open vocal tract states. Subsequently, the spectral acoustic loss ($\mathcal{L}_S$) values between a reference and the n_D roll outs is calculated using the mean squared Euclidean distance between the reference and imitation log-mel spectra. If the respective minimal loss is smaller than the loss corresponding to the current state $\mathbf{v}_i$, the state with minimal loss becomes the next search space state $\mathbf{v}_{i+1}$. Else, if no new minimum is found, the search space radius is iteratively enlarged and n_D new states are sampled and evaluated. These procedures are repeated until a maximum number of iterations (in this case 15) is reached or an early stopping criterion (if the search radius is enlarged five times in a row) is reached.

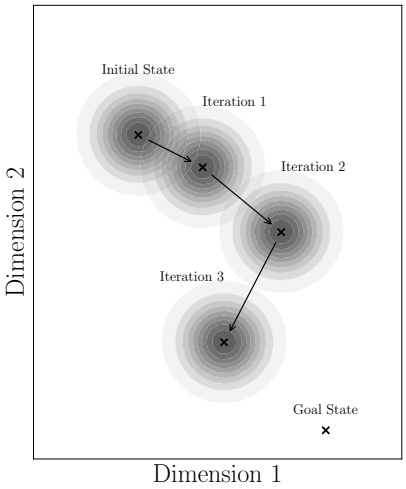

Figure 3 – Sketch of the CGRS algorithm in a simplified two dimensional search space. Exploration noise is indicated by blue density distributions.

Two vowel imitation experiments were performed. For the first one, ten CGRS runs with different random seeds, were conducted as described above for each reference vowel. The second experiment was the same, except that the states were transformed into the PC space and retransformed into the VT space during the constrain step. For this purpose, another PCA model

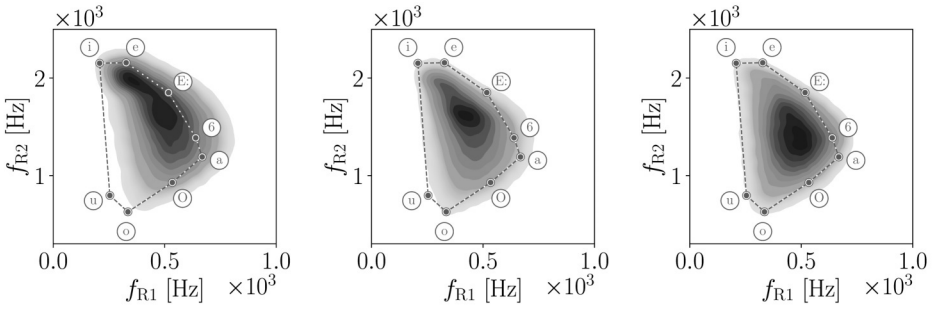

Figure 4 – Tube resonance distributions, visualized as Gaussian kernel density estimations based on histograms with 10^5 entries each. High density: blue, low density: yellow. Predefined vowel positions are indicated by the corresponding SAMPA symbols. **Left:** Shapes sampled in the VT parameter space. **Middle:** Shapes sampled in the VT space, encoded in the PC space and subsequently decoded. **Right:** Shapes sampled in PC space and decoded into the VT parameter space.

was trained on 16 of the total 19 dimensions of the predefined shapes, excluding the parameters VO, TRX, TRY. This was done to keep the parameter TS3 well defined during the decoding step which is important for the optimization process. The number of principal components was chosen to be 8, which explain approximately 95 % of the observed variance in this case.
In both experiments the articulatory loss $\mathcal{L}_\mathrm{A}$, defined as the mean squared Euclidean distance between the articulatory goal state and the current search space state, was monitored. Thereby, the dimensions VO, TRX and TRY were excluded from the calculation, since these dimensions were not part of the optimization process.

2.4.2 Consonant Learning

Consonant imitation experiments were performed using consonant-vowel syllables (all permutations between the consonants /b, d, g/ and the vowels /a, e, i, o, u/) as references. The initial consonant states were set to /ə/ and the vowel states were set to the respective goal vowel, so that only the consonant states needed to be optimized.
The optimization strategy was the same as for the vowel experiments, but the consonant states were constrained to be closed, which means that the minimum cross-sectional area of the corresponding tube is $\leq 0.001\,\mathrm{cm}^2$. As with the vowels, two experiments were conducted, first in the vocal tract space only and second with transformation into the PC space and the respective reverse transformation during the constrain stage. The same PCA model as for the vowels was used. Ten runs with different random seeds were performed for each syllable configuration.
Similar to the vowel experiments, the articulatory loss values between the consonant goal state and the search space states were monitored.

3 Results

Uniform Exploration

Figure 4 shows the distributions of the first and second tube resonance of each sampled open vocal tract state. From the left plot, it is visible that the f_R1-f_R2 distribution of shapes sampled uniformly from the VTL space are shifted far outside the convex hull formed by the natural positions of VTL vowel shapes derived from MRI data. Further, regions around the vowels /o, u/ are barely covered by the samples. When the samples get encoded and decoded using the

	CGRS Iterations		$\hat{\mathcal{L}}_\mathrm{S}$		$\hat{\mathcal{L}}_\mathrm{A}$	
	VT	VT $\rightleftarrows$ PC	VT	VT $\rightleftarrows$ PC	VT	VT $\rightleftarrows$ PC
Vowels	5	5	0.105	**0.089***	2.186	**1.294***
Consonants						
/b/	3	3	**0.008**	0.013	**1.3**	1.62
/d/	3	3	0.012	**0.009***	1.385	**0.595***
/g/	4	**3.5**	0.015	**0.01***	1.181	**0.739***

Table 2 – Results of the goal-babbling experiments. The table shows the median values obtained for CGRS iterations, the normalized acoustic loss $\hat{\mathcal{L}}_\mathrm{S}$ and the normalized articulatory loss $\hat{\mathcal{L}}_\mathrm{A}$. Better values are indicated by bold numbers. A star indicates a significant difference between the values obtained with the VT and VT$\rightleftarrows$PC methods. Thereby *significant* means $p < 0.05$ based on Mood's median test.

PCA model (middle plot), the resulting distribution has a higher similarity to the area formed by the predefined VTL vowels. Maximum similarity can be achieved by sampling in the PC space directly (right plot) using an optimized value for n_σ (here $\mathrm{n}_\sigma = 1.5$).

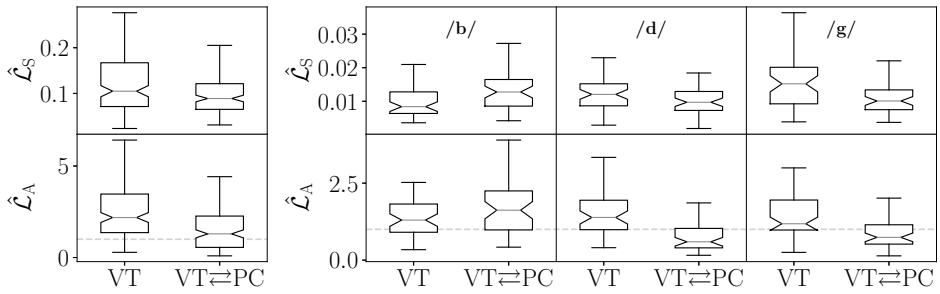

Figure 5 – Results of the goal-babbling experiments. The normalized acoustic and articulatory loss distributions are visualized in the top and bottom plots, respectively. A dashed line indicates the $\hat{\mathcal{L}}_\mathrm{A} = 1.0$ level in the bottom plots. **Left:** Summarized results for the vowels /a, e, i, o, u, ø, y, ɛ, ɪ, ɔ, ʊ, œ, ʏ, ɐ/ . **Right:** Results for the consonants, separated into /b, d, g/.

Goal-directed Babbling

Figure 5 shows the distributions of the normalized acoustic and articulatory loss values in case of the vowels (left plot) and the consonants (right plot), which are separated into the different consonant classes. Normalization means in this case that the final values of the finished optimization are divided by the respective initial values. For the vowels, both the acoustic and articulatory losses are concentrated around lower values in the VT$\rightleftarrows$PC method compared to VT-only method. The distributions are significantly different based on two-sided Kolmogorov–Smirnov tests ($p = 0.01$ and $p = 2 \cdot 10^{-6}$ in case of $\hat{\mathcal{L}}_\mathrm{S}$ and $\hat{\mathcal{L}}_\mathrm{A}$, respectively).
No significant differences were observed between distributions of the consonant /b/. For the consonants /d, g/, the VT and VT$\rightleftarrows$PC distributions are significantly different ($\hat{\mathcal{L}}_\mathrm{S}$: $p = 0.02$, $\hat{\mathcal{L}}_\mathrm{A}$: $p = 5 \cdot 10^{-6}$ and $\hat{\mathcal{L}}_\mathrm{S}$: $p = 2 \cdot 10^{-4}$, $\hat{\mathcal{L}}_\mathrm{A}$: $p = 1 \cdot 10^{-3}$ in case of /d, g/, respectively).
Table 2 shows the median values for $\hat{\mathcal{L}}_\mathrm{S}$ and $\hat{\mathcal{L}}_\mathrm{A}$, as well as the median number of CGRS iterations that were performed until the early stopping criterion was reached. No significant differences were found among the CGRS iteration distributions.

4 Discussion

This work shows that the use of PCA in vocal learning simulations can make it easier to obtain high quality vowels and consonants compared to the case without. The study also shows that acoustic optimization does not imply an optimization of the articulatory states (at least not across all dimensions) in case of the VT method, since the normalized articulatory loss at the end of optimization is mostly above 1.0 (i.e., greater than at the beginning) for both vowels and consonants. It is likely that acoustically important dimensions [8] have been optimized, but unimportant dimensions may vary greatly between the articulatory goal states and imitations. This implies that a complete acoustic-to-articulatory inversion based on an acoustic loss only is difficult and ambiguous even in this simplest case of copy synthesis. However, it could be shown that in many of the acoustic optimizations with the PCA method, the articulatory states were also optimized. For vowels, about half of the $\hat{\mathcal{L}}_{\mathrm{A}}$ distribution is below 1.0, for the consonants /d/ and /g/ it is about three quarters. Only for the consonant /b/ there seems to be no advantage over the VT-only method. This may be due to the fact that this consonant is acoustically and articulatory more ambiguous, since it can be achieved by simply closing and opening the mouth, while the other two consonants require precise points of articulation.
In future studies, the performance gain through PCA should be validated in a more elaborated vocal learning simulation using natural speech utterances as references.

Acknowledgements

This work has been funded by the Leverhulme Trust Research Project Grant RPG-2019-241: “High quality simulation of early vocal learning".

References

[1] PHILIPPSEN, A.: *Goal-directed exploration for learning vowels and syllables: A computational model of speech acquisition. KI-Künstliche Intelligenz*, 35(1), pp. 53–70, 2021.

[2] VAN NIEKERK, D. R. ET AL.: *Finding intelligible consonant-vowel sounds using high-quality articulatory synthesis*. In *Proc. INTERSPEECH*, pp. 4457–4461. ISCA, 2020.

[3] PAGLIARINI, S. ET AL.: *Vocal imitation in sensorimotor learning models: A comparative review. IEEE Trans. Cogn. Dev. Syst.*, 13(2), pp. 326–342, 2021.

[4] GAO, Y. ET AL.: *Articulatory copy synthesis using long-short term memory networks*. In *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2020*, pp. 52–59. TUDpress, Dresden, 2020.

[5] KRUG, P. K. ET AL.: *Intelligibility and naturalness of articulatory synthesis with VocalTractLab compared to established speech synthesis technologies*. In *Proc. SSW 11*, pp. 102–107. 2021.

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

[7] KRUG, P. K. ET AL.: *Modelling microprosodic effects can lead to an audible improvement in articulatory synthesis. J. Acoust. Soc. Am.*, 150(2), pp. 1209–1217, 2021.

[8] XU, A. ET AL.: *Model-based exploration of linking between vowel articulatory space and acoustic space*. In *Proc. INTERSPEECH*, pp. 3191–3195. ISCA, 2021.