Artificial Vocal Learning guided by Phoneme Recognition and Visual Information

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

[1]Technische Universität Dresden
[2]Affiliation not available

October 30, 2023

Abstract

This paper introduces a paradigm shift regarding vocal learning simulations, in which the communicative function of speech acquisition determines the learning process and intelligibility is considered the main measure of learning success. Thereby, a novel approach for artificial early vocal learning is presented that utilizes deep neural network-based phoneme recognition in order to calculate the speech acquisition objective function. This function guides a learning framework that involves the state-of-the-art articulatory speech synthesizer VocalTractLab as the motor-to-acoustic forward model. It is shown that in this way an extensive set of German phonemes consisting of most German consonants and all stressed vowels can be produced successfully. The synthetic phonemes were rated as highly intelligible by human listeners in a listening experiment. Furthermore, it is shown that visual speech information, such as lip and jaw movements can be extracted from video recordings and be incorporated into the learning framework as an additional loss component during the optimization process. It was observed that this visual loss did not increase the overall intelligibility of phonemes. Instead, the visual loss acted as a regularization mechanism that facilitated the finding of more biologically plausible solutions in the articulatory domain.

Artificial Vocal Learning guided by Phoneme Recognition and Visual Information

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*—This paper introduces a paradigm shift regarding vocal learning simulations, in which the communicative function of speech acquisition determines the learning process and intelligibility is considered the main measure of learning success. Thereby, a novel approach for artificial early vocal learning is presented that utilizes deep neural network-based phoneme recognition in order to calculate the speech acquisition objective function. This function guides a learning framework that involves the state-of-the-art articulatory speech synthesizer VocalTractLab as the motor-to-acoustic forward model. It is shown that in this way an extensive set of German phonemes consisting of most German consonants and all stressed vowels can be produced successfully. The synthetic phonemes were rated as highly intelligible by human listeners in a listening experiment. Furthermore, it is shown that visual speech information, such as lip and jaw movements can be extracted from video recordings and be incorporated into the learning framework as an additional loss component during the optimization process. It was observed that this visual loss did not increase the overall intelligibility of phonemes. Instead, the visual loss acted as a regularization mechanism that facilitated the finding of more biologically plausible solutions in the articulatory domain.**

***Index Terms*—Vocal learning simulation, articulatory speech synthesis, automatic phoneme recognition.**

I. Introduction

Articulatory synthesis is a promising candidate for future speech synthesis systems as this type of synthesis aims to mimic the speech generation process that happens within a human vocal tract during speech production. Thus, it has the potential to provide both natural sounding speech and, in contrast to current state-of-the-art neural synthesis systems, high flexibility and the ability to control every aspect of speech generation [1]. However, a major problem in using articulatory synthesis is its control, which is not known a priori, i.e. speech can only be generated with expert knowledge. Without such knowledge, the synthesizer can only be controlled randomly or according to certain patterns, whereas the acoustic consequences are observable. This is similar to the situation human vocal learners face when they start to explore their vocal tract. Consequently, computational simulations of vocal learning appear to be a promising tool in order to technically solve the control problem of articulatory synthesizers, as well as to answer questions in phonetics and child speech development.

A. Role of Visual Cues and Scientific Relevance

It is well known that congenitally blind children learn to speak without significant problems [2], while congenitally deaf children have difficulties learning to speak and require special training to obtain such ability [3]. This indicates that the main objective function that guides early vocal learning must be based on acoustic information rather than visual information. However, evidence was reported for sighted speakers to have a finer control over articulatory speech movements [4] and it was found that congenitally blind speakers show less lip rounding than speakers with normal vision [5]. These findings are in agreement with computational simulations that suggest that fine adjustments of the lip protrusion are necessary, e.g. to produce a clear vowel /u/ [6]. It is therefore reasonable to assume that computer simulations of the speech acquisition process can benefit from a multi-modal (audio-visual) observation space in terms of quality or efficiency. Nevertheless, it has not yet been demonstrated that natural, measured articulatory speech movements of the visible articulators can actually be incorporated into an appropriate simulation in order to learn an extensive set of phonemes.

With the present study the current state of research is extended by the following contributions: (i) A set of German vowels and syllables was generated via vocal learning simulation using the state-of-the-art [1] articulatory synthesizer VocalTractLab (VTL) [7]. Thereby, a novel method was used, which incorporates phoneme recognition as the objective function. (ii) Jaw and lip movement related information corresponding to vowels and syllables was extracted from audio-visual data and used in the vocal learning simulation to test the impact of visual information on the learning process. (iii) The resulting synthetic speech was evaluated both in terms of intelligibility as quantified by human listeners and in terms of the biological plausibility of the resulting articulatory states.

II. Methods

A. Artificial Vocal Learning

In the context of this study and in general, let an *artificial vocal learning scenario* be defined as follows: (i) Vocal learning is performed by an *agent*, which is an entity that has access to a *motor space* and an *observation space*, the latter of which encodes both the (acoustic) consequences of actions executed

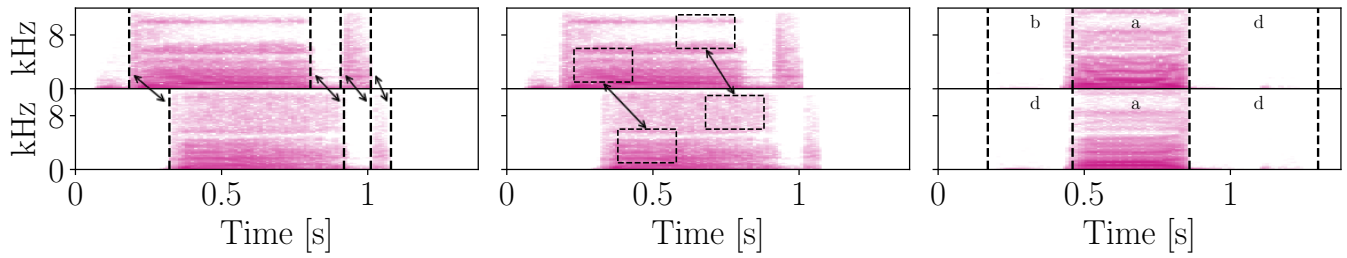

Fig. 1: Time (left plots) and frequency (middle plots) differences between two realizations of the word "bad" uttered by a female (top plots) and a male (bottom plots) speaker. The right plot shows realizations of "bad" (top) and "dad" (bottom) by a single speaker uttered with two distinct f_0 contours. If the former utterance was to be imitated and the latter an attempt of imitation, the major part of spectral difference would not come from the (incorrect) consonant but from the (actually correct) vowel part due to the strongly unequal distribution of the spectral information in the individual phoneme segments.

within the agent's motor space and other motor spaces, i.e. speech by external speakers. (ii) The vocal learning process is characterized by the agent trying to acquire motor space states that correspond to certain observation states. If such observation state originate from external speakers, one may speak of *imitative learning*. A successful imitation preserves the communicative intent of the utterance, but may change its acoustic realization. On the other hand, if the observation state of interest is a result of a motor command initiated by the agent itself, an {action, consequence} pair can be obtained. In that case, one may speak of the *acoustic-to-motor inverse*. Such an inverse provides the possibility for a true re-synthesis of the observation state. (iii) The learning process must be explorative, which means it is un- or semi-supervised, as it can only be guided by observables. The action states of action-consequence pairs produced by possible teachers, e.g. external speakers, however, are mostly hidden. Hence, vocal learning can only be guided by acoustic information, visual information and sensory feedback. Approaches like direct inversion based on action-consequence pairs crafted by experts, i.e. *copy synthesis* as done in [8]–[10] may be pragmatic and expedient but such methods can not be referred to as vocal learning. Note however that it would be legitimate in this frame to establish a direct inversion between actions and consequences if the respective pairs were previously determined by exploration, e.g. as proposed in [11], [12].

1) Related Work: Numerous papers on the simulation of speech acquisition have been published in recent years [6], [12]–[21], see e.g. [22] for a more in-depth comparative review. Some of these studies deal explicitly with phonetic and child development issues, while the motor optimisation problem itself is of secondary importance. Other studies, however, focus on the level of motor learning and acoustic-to-articulatory inversion. This is usually done by *goal-directed babbling* (which means that the explorative process is not totally random but is driven towards a target state by some kind of loss or reward function) and involves the implicit or explicit creation of a mapping between motor space and observation space [18]. In the explicit case, neural networks are often trained for direct inversion from consequence to action, or a composite consisting of a trained inverse and a trained forward model is used [12], which is usually referred to as *distal learning* [11]. Both the defined motor and observation spaces may differ greatly among the mentioned studies. While the motor-to-consequence models are mostly articulatory synthesizer-based frameworks with varying degrees of realism, such as VOCAL LINEAR ARTICULATORY MODEL [23], DIRECTIONS INTO VELOCITIES OF ARTICULATORS [16] or VTL [7], the observation spaces are mostly based on spectral acoustic features such as formants, spectrograms, mel spectrograms [24], mel-frequency cepstral coefficients (MFCC) [25], or abstracted features obtained by embedding or dimension reduction of acoustic or spectral input [21]. The performance of vocal learning models is then usually evaluated by distance metrics defined in the observation spaces, such as formant differences or spectral distances. Sometimes it is also evaluated by the quality of the motor trajectories or distances in the motor space [26], although this is rather difficult because the true motor trajectories are usually unknown. In some cases, subjective auditory impressions are mentioned for evaluation purposes, but none of the listed works reported systematic listening tests with human listeners evaluating intelligibility.

With this in mind, in the context of the definition of artificial vocal learning given here, it can be said that the aforementioned works are incomplete in the broadest sense or ignore important conceptual prerequisites. I.e. it is often assumed that the goal of vocal learning is imitation through acoustic matching [12]. However, this may be a fundamental misconception, for the following technical and conceptual reasons: First of all, as shown in Figure 1 three main technical issues occur when trying to calculate differences between spectral features. (i) There may be non-linear time distortions among the goal speech and imitated speech. (ii) There will be intrinsic frequency mismatches between goal and imitated utterances due to differences in the vocal tract geometries of the target speaker and imitating speaker. This is often referred to as *speaker normalization problem*. (iii) There will be spectral weighting issues that occur from the widely differing amount of spectral information among different phonemes even if target and imitating utterances originate form the same speaker, e.g. see right plot in Figure 1. All these issues contribute to the fact that there is no correlation between the spectral difference and actual perceptual difference in the general case. While these problems may be dismissed as cosmetic, since they

could be circumvented through complicated engineering, such as time and frequency warping and spectral weighting etc., these problems are actually a symptom of a deeper conceptual problem: Acoustic matching is not the goal of vocal learning, but rather the goal of (true) *re-synthesis*. In order to achieve such a re-synthesis with an articulatory model, the underlying vocal tract geometry of the target speech material must be known prior to exploration, e.g. either derived from magnetic resonance imaging (MRI) scans, or somehow determined from acoustics. That a mapping from acoustic material to this geometry can be established is conceivable, but an open question for future research and it is clear that this procedure has little to do with human vocal learning, since humans develop their own vocal tract and do not copy those of others. Instead, it is reasonable to assume that the goal of vocal learning is to acquire motor states that fulfill a *communicative function*. This may be motivated in the context of evolution, as humans apparently developed the complicated process of speech in order to be understood by others. Successful communication requires this and successful communication is a basic prerequisite for human survival. Nevertheless, in terms of artificial vocal learning this means: (i) Imitation is not the goal but may be a path, i.e. the fundamental paradigm of vocal learning may not be described by "*How can I reproduce an utterance I just heard?*", but by "*How can I produce an utterance in a way that I am understood?*". (ii) The main measure of vocal learning success should therefore be intelligibility. In previous works, however, this measure usually plays no role. A notable exception in this regard is the work of Rasilo and Räsänen [19], where intelligibility is in fact included in the objective function. However, their work involves human subjects who guide the learning process as "caregivers" and thus is not fully automated, which is inconvenient.

2) Approach: This work presents a simple and elegant solution to these technical and conceptual problems. By using automatic speech recognition (ASR) with a recurrent deep neural network, acoustic time series inputs can be transformed into probability distributions representing the input utterances. The fundamental advantage over previous work is the implicit speaker normalization gained by training on multi-speaker data. At the same time, the computation of the loss function is considerably simplified by the fact that it is now only based on probability vectors which can be directly compared with each other. On the other hand, this means that a separately trained speech recognition model is used for the learning process.

ASR models may be used in two different ways within the vocal learning framework. (i) A model may be used to encode target and imitative utterances into respective probability distributions. Subsequently a distance between both vectors would be calculated and used a loss in the learning process. (ii) No target utterances are used explicitly and the model is used to encode the acoustic signals uttered by the agent only. Subsequently, the encodings are evaluated against the probability unit vectors, which represent the phonetic or word-level categories that the ASR model was trained to map to.

In this work only the second option was used, as this scenario ideally guarantees that the categorical communicative function of learned utterances is equal to the desired phonetic identities. In this study a single-phoneme recognition model (described in Section II-B) was used in order to guide the vocal learning process. Synthetic utterances were produced using VTL with its standard speaker model, accessed via the PYTHON front-end VTL-PYTHON[1]. VTL is an articulatory synthesizer that provides a one-dimensional aero-acoustic simulation of sound propagation within the human vocal tract, whereby the vocal tract shape is described through its tube cross-sectional area function. The simulation can be controlled via a parameterized three-dimensional vocal tract model that was derived from MRI data, as well as three types of glottis models: a geometric glottis, a triangular glottis and a two-mass model. Throughout this work, the geometric glottis model was used. While VTL provides high-level control such as phoneme-to-speech via articulatory presets representing the German phoneme inventory derived from MRI data, the synthesizer also allows direct control over the motor level, which is a prerequisite for simulating speech acquisition. In the configuration used here, VTL provides 19 supra-glottal parameters , see Table I. While the vocal tract dynamics in high-level control are always governed by the TARGET-APPROXIMATION-MODEL [27]–[29] (TAM), low-level control can in principle be executed arbitrarily, e.g. by Dynamic Movement Primitives [30], as done in [21]. However, throughout this work the TAM was used exclusively to drive the VTL synthesis on the motor level.

A complete overview of the vocal learning framework used in this study is given in Figure 2. On the left side, the optimization procedure itself is visualized, which is performed using the Whale Optimization Algorithm [31] (WOA), see Section II-D. At each time step of the optimization, this algorithm receives a single value as input (loss) and outputs a parameter vector containing the respective state of the articulatory variables to be optimized. Subsequently, supra-glottal states are then tested with regards to an externally set constriction constraint. This means specifically, the minimum of the tube area function $\mathrm{T_{min}}$ corresponding to the respective articulatory state is calculated. Supra-glottal states are referred to as *open*, if $\mathrm{T_{min}} \geq 0.3\,\mathrm{cm}^2$, *tight*, if $0.3 > \mathrm{T_{min}} > 0.001\,\mathrm{cm}^2$ or *closed*, if $\mathrm{T_{min}} \leq 0.001\,\mathrm{cm}^2$. Successful learning of clear vowels requires, for example, open vowel tract states, while learning fricatives requires tight states, since the fricative noise sources in the VTL simulation are only activated beyond a certain level of narrowness. Plosives require a closure within the vocal tract, i.e. a closed state. If the calculated constriction does not match the constriction required by the phonetic category being learned, a large loss value of 100 (arbitrary) is directly returned to the optimization algorithm, bypassing the residual chain of processes. This constraint is justified by the computationally expensive synthesis of a state. As a consequence, the computational efficiency of the simulation is increased. Nevertheless, if a state does fulfill the constriction constraint, a *motor score* is calculated. This is a set of parameter curves describing the temporal deformation of the 3D vocal tract model and the dynamics of the geometric glottis within VTL. Starting from the motor score, VTL can calculate the time evolution of the one-dimensional tube cross-sectional area function and finally a synthetic speech waveform. Then mel-spectrograms are calculated from the audio signal, which

[1]https://github.com/paul-krug/VocalTractLab-Python

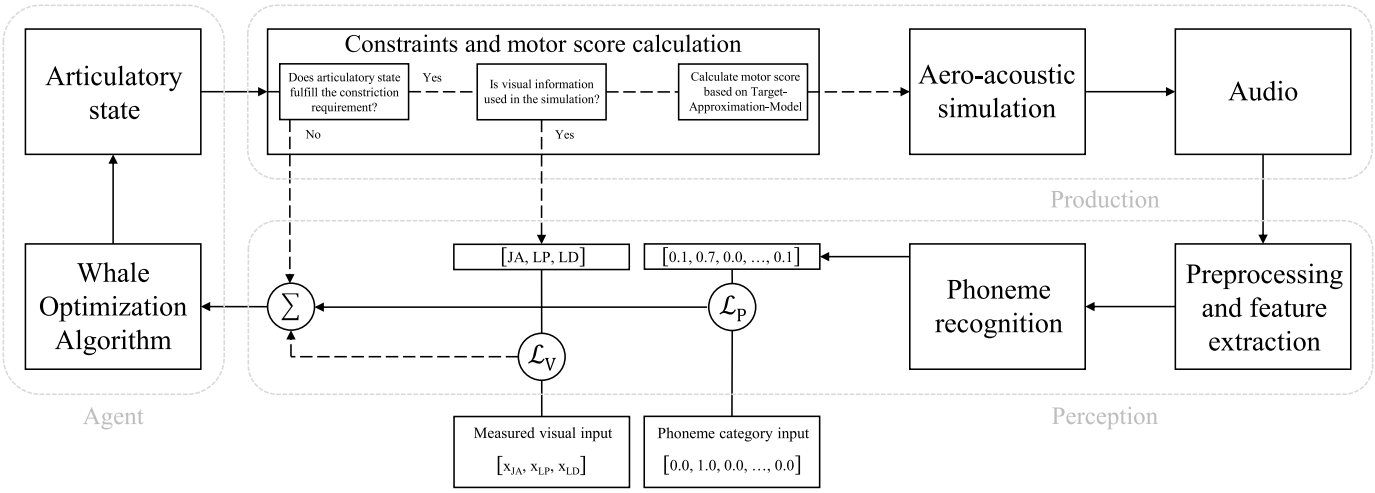

Fig. 2: Schematic block diagram of the implemented framework for artificial vocal learning guided by phoneme recognition.

serve as input features for the phoneme recognition model. This model in turn outputs a probability distribution that describes the predicted phonetic identity of the input utterance. The similarity of the probability vector to a corresponding previously externally determined unit vector (intent of the agent) is then calculated via categorical cross-entropy for each phoneme j to be included in the loss function during the optimization process:

$$\mathcal{L}_{\mathrm{P}} = \sum_{j} \left(-\sum_{i=1}^{n_C} y_{ij} \cdot \log(\hat{y}_{ij}) \right), \tag{1}$$

which whereby the number of phonetic categories is $n_C = 37$, and y_{ij} and $\hat{y}_{ij}$ denote the $i-$th component of the phonetic identity vector corresponding to a phoneme j and the related phoneme recognition model output vector, respectively. The phoneme loss $\mathcal{L}_{\mathrm{P}}$ is then passed to the optimization algorithm which closes the process loop. In the case where visual information is used, however, a second loss value is calculated from the three visually accessible VTL parameters jaw angle, lip protrusion and lip distance. For this purpose, the corresponding values determined by the optimization algorithm are extracted and compared with measured values obtained from video recordings of speech movements (see Section II-E for details) via the mean-square error (MSE):

$$\mathcal{L}_{\mathrm{V}} = \frac{1}{3} \sum_{i \in V} (x_i - \hat{x}_i)^2, \tag{2}$$

where $V = \{\mathrm{JA}, \mathrm{LP}, \mathrm{LD}\}$ describes the set of visually accessible VTL parameters and x_i and $\hat{x}_i$ denote the measured values and the values proposed by the optimizing agent, respectively. Finally a total loss is obtained from the sum:

$$\mathcal{L} = \mathcal{L}_{\mathrm{P}} + \mathcal{L}_{\mathrm{V}}. \tag{3}$$

The described framework was implemented in the PYTHON programming language and published open source[2].

[2]https://github.com/paul-krug/artificial-vocal-learning

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

TABLE I: Supra-glottal VTL control parameters.

B. Phoneme Recognition

A deep recurrent neural network with the architecture introduced in [32] was used as the phoneme recognition model. This model consists of five consecutive bi-directional gate recurrent unit layers (Bi-GRU) with 256 neurons each (with *tanh* activation functions), followed by a dense layer of 37 neurons with *softmax*, see [32]. Each dimension of the 37-dimensional output corresponds to a single phoneme category. In this way, the model acts as an encoder that maps the input time series directly to a phoneme probability distribution. The model was trained on single phoneme samples with a preceding and succeeding temporal context of $\tau_{\mathrm{C}} = 32\,\mathrm{ms}$ extracted from the combined German KIEL and BITS-US corpora, as described in [32]. Categorical crossentropy was used as the loss function during the training process. Logarithmized mel-scaled spectrograms with 80 frequency bands were used as input features. For their calculation the underlying audio samples were resampled to 16 kHz, and for the subsequent short time Fourier transformation window length of 256 samples

(16 ms) and a hop length of 40 samples (2.5 ms) were used.

C. Vocal Learning Simulations

Within the scope of this study, sets of motor states corresponding to the German tense vowels /a, e, i, o, u, E, 2, y/[3] and German consonants /p, t, k, b, d, g, f, v, s, z, S, j, C, x, R, m, n, l/ in the context of /a/ were acquired via vocal learning simulations using the previously presented framework. Although the system is in principle capable of learning all 37 phoneme categories the recognizer was trained on, only tense vowels and the listed consonants were considered in order to simplify the subsequent listening experiment design. Naïve listeners are usually not familiar with categories such as lax vowels or consonants such as /Z, N/ and the question of visual information can be addressed without this aspect. The phoneme /h/ was excluded, since it would not involve the optimization of supra-glottal parameters. During the vowel learning, single static articulatory parameter vectors were optimized including the supra-glottal parameters within the limits as defined in Table I, excluding VO, TRX and TRY. VO was set to $-0.1\,\mathrm{cm}^2$, as an optimization of this parameter is only needed if nasality is desired. TRX and TRY can be set to arbitrary values, as VTL allows for an automatic calculation of these values if the standard speaker file is used [26]. For the glottal parameters, the modal voice quality settings of the geometric glottis in VTL-Python were used. The VTL motor score then consisted of a single articulatory TAM target vector and the target duration was set to 200 ms which is long enough to produce a meaningful utterance and at the same time short enough to ensure computational efficiency of the simulation. The calculation of the VTL motor score is more complicated in the case of consonants, because they have to be embedded in a syllable. This means each parameter dimension features two consecutive articulatory targets, one for the consonant and one for the vowel. Following the idea of Krug et al. [26], consonant related states were acquired individually, but in acoustic accompaniment with a following vowel, which means only consonant related parameters were optimized, while both the acoustic realizations of the consonant and vowel contribute to the total phoneme loss. Compared to the joint optimization of a consonant with a vowel, e.g. as done in [25], this process has the advantage of higher computational efficiency due to the much smaller scope of the motor space by the reduced number of required target parameters. The 16 supra-glottal parameters (as described earlier) were then optimized in case of all consonant learning simulations. The parameter VO was included in the optimization for the nasals /m, n/. For the voiced consonants a single modal target described the glottal dynamics, except for /R/, which required aspiration from the glottis in order to sound plausible. Therefore, the glottis was slightly opened by changing the lower and upper rest displacement of the vocal folds (XB and XT, respectively) from their modal setting to 0.05 cm. In that case, glottal dynamics would be described by two consecutive glottal targets similar to the supra-glottal domain, whereby the glottal target onset times were synchronous with the supra-glottal target onset times. In the case of the voiceless consonants, however, the supra-glottal and glottal onset times must be set asynchronously, otherwise glottis-induced artifacts may occur in the acoustics due to implausible voice-onset times. The onset times of the glottal vowel targets were set to -30 ms, $+50$ ms and $+60$ ms relative to the onset time of the supra-glottal vowel target for /f, s, S, C, x/, /p, t/ and /k/, respectively. The supra-glottal target durations were 50 ms and 150 ms in case of the consonant and vowel targets, respectively. For the voiceless plosives the duration of the respective vowel target was set to 225 ms, which was needed due to the larger voice-onset time. For the voiceless consonants, the glottal parameters XB and XT were set to 0.1 cm, the chink area (CA) was set to $0.1\,\mathrm{cm}^2$ and the relative amplitude (RA) was set to 0. Finally, the acoustic window, which corresponds to the input of the phoneme recognition system needed to be defined. This is not trivial, because even though the articulatory target boundaries are known, the acoustic phoneme boundaries are not. With the used TAM time constant of 12 ms the acoustic signal is following the articulatory target onset with a delay of approximately $\tau_\mathrm{D} = 50$ ms. The acoustic window for a specific phoneme was then reasonably estimated from the respective target boundaries plus $\tau_\mathrm{D} \pm \tau_\mathrm{C}$.

[3]For phonetic symbols, X-SAMPA notation is used throughout this work.

D. Optimization Method

The vocal learning process described in this study is understood as an optimization problem, where the goal is to achieve an optimal state in a high-dimensional motor space corresponding to a minimization of an observable objective function. Such a high-dimensional search problem may be solved by gradient-free, metaheuristic optimization algorithms of which many have been published in recent years. In order to find an algorithm well suited for the vocal learning process, a number of candidate algorithms were tested in advance. The PYTHON library PYMETAHEURISTIC was used for this purpose. The algorithms defined in [31], [33]–[36] were used to find optimal articulatory states corresponding to the vowels /a, e, i, o, u, E, 2, y/ using the vocal learning framework previously described, without the visual loss component. The algorithms' hyperparameters were not specifically tuned. The optimizations were performed 20 times for each vowel. Each run was stopped after 100 synthesis steps within the optimization. Both the phoneme loss as well as the computation time were monitored. The results for the algorithms are shown in Figure 3. It can be seen that the runs calculated via WOA gave the overall smallest loss values, as well as the lowest computation time. Hence, it was selected as the optimization algorithm in the following experiments. First, however, the WOA hyperparameters *hunting party* and *spiral parameter* were optimized in a grid search over the values [100, 300] in steps of 100 and [0.0, 1.0] in steps of 0.1, resulting in optimal values of 200 and 0.5, respectively.

E. Visual Data Acquisition

The calculation of the visual loss introduced in Equation 2 requires the input of phoneme category-related visual target parameters. Such parameters were derived from video recordings of a speaker uttering respective speech sounds.

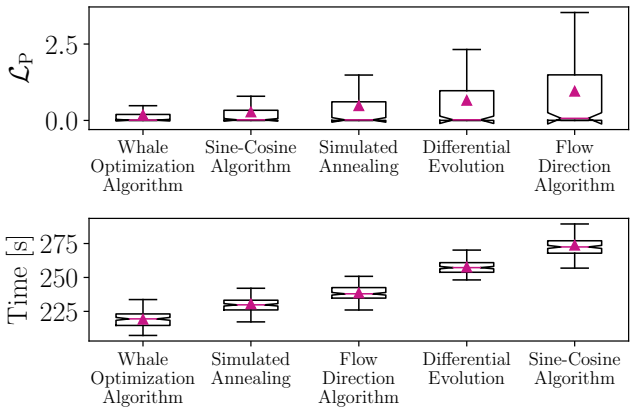

Fig. 3: Results for different metaheuristic algorithms.

To accomplish this, an audio visual data set was recorded containing vowels and syllables spoken by a 26 year old native German male speaker. Since the visual measurements had to be normalized to the dimensions of VTL for the calculation of $\mathcal{L}_V$, it was sufficient to record a single speaker. Multiple speakers would have been normalized to the same ranges and sufficient variability among the parameter distributions was already generated for a single speaker, due to the intra-speaker variance during the phoneme and viseme production. The previously mentioned vowels and consonants in the context of vowel /a/ were recorded individually 10 times each. Additionally the facial extreme positions, e.g. jaw and lips fully closed/open, as well as lips fully spread/rounded were recorded multiple times. The subject was required to stand still and in a fixed position in order to avoid movements of the recorded face in three-dimensional space, such as rotations of the head, which would complicate the subsequent calculation of distances based on the video material. The video data was recorded with a resolution of 1080x1920 pixels at a frame rate of 120 frames per second on an Apple Inc. iPhone 11, audio was thereby recorded at a sample rate of 48 kHz. The separate audio stream was used exclusively for the manual segmentation of the speech material. Based on the segmentation the relevant video frames were extracted. Subsequently, 68 facial landmarks (following the Multi-PIE [37] or IBUG [38] standard) were extracted from each frame using a convolutional pose machine [39] in the exact same way as described in [40]. The model was trained using supervision-by-registration [40] because this technique allows temporally coherent trajectories to be determined across consecutive video frames. Thereby, the intrinsic stability of the landmark predictions is enhanced by using optical flow as a loss function in addition to the landmark detection loss [40]. The training material consisted of the 300-W landmark data set [38], [41] for the landmark detection loss and the recorded video material for the optical flow-based loss.

The raw detected landmarks were processed as follows. The landmarks have a standard numbering, e.g. see [38], hence, individual landmarks are identified by the numbers 1 to 68. Four observable pixel-coordinate-based distances Ω_i ($i \in \{\mathrm{JVD}, \mathrm{LHD}, \mathrm{LVD}, \mathrm{IOD}\}$, see Figure 4) were calculated: a

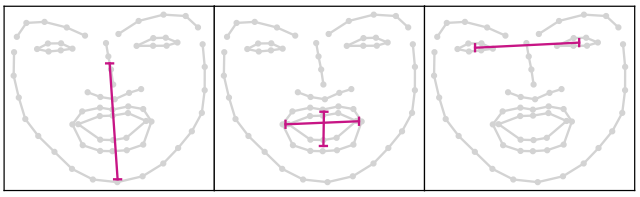

Fig. 4: Exemplary plot of observable distances on obtained landmarks. Right: JVD. Middle: LHD and LVD. Left: IOD.

horizontal lip distance (LHD) from the left corner of the mouth (center of landmarks 49 and 61) to the right corner (center of landmarks 55 and 65), a vertical lip distance (LVD) from the upper lip (center of landmarks 51, 52, 53, 62, 63, 64) to the lower lip (center of landmarks 57, 58, 59, 66, 67, 68), a vertical jaw distance (JVD) from the chin (center of landmarks 8, 9, 10) to the nose (center of landmarks 28, 29, 30, 31) and an inter-ocular distance (IOD) from the left eye (center of landmarks 37, 38, 39, 40, 41, 42) to the right eye (center of landmarks 43, 44, 45, 46, 47, 48). Thereby, distances were calculated between the centers of coordinate ensembles in order to increase robustness against detection noise. Normalized distances $\hat{\Omega}_j = \Omega_j \cdot \Omega_{\mathrm{IOD}}^{-1}$, $j \in \{\mathrm{JVD}, \mathrm{LHD}, \mathrm{LVD}\}$ were calculated. The division by IOD was done in order to account for small drifts of the subject along the camera axis, which may change the overall size of the recorded face. The VTL parameters $x \in \{\mathrm{JA}, \mathrm{LP}, \mathrm{LD}\}$ are then calculated from $j \in \{\mathrm{JVD}, \mathrm{LHD}, \mathrm{LVD}\}$, respectively, via a linear min-max-scaling, which is appropriate in case of LD, and a valid approximation in case of JA and LP:

$$x(t) = m \cdot \frac{|\hat{\Omega}_j(t) - \hat{\Omega}_j^{\min}|}{|\hat{\Omega}_j^{\max} - \hat{\Omega}_j^{\min}|} + b, \tag{4}$$

whereby $\hat{\Omega}_j^{\min}$ and $\hat{\Omega}_j^{\max}$ denote the normalized minimum and maximum distances measured from the facial extreme positions. The slope m and offset b are determined by:

$$m = \delta_x \cdot (|x_{\max} - x_{\min}| + \alpha_x), b = \begin{cases} x_{\min} - m, & \text{if } \delta_x = -1 \\ x_{\max} - m, & \text{if } \delta_x = 1. \end{cases} \tag{5}$$

Thereby, $x_{\min}$ and $x_{\max}$ denote the minimal and maximum VTL values of the respective dimension x. The factor δ_x was introduced to preserve the correct sign in the specific dimensions, e.g. in case of the observables JVD and LHD large measured values (which mean open jaw or spread lips, respectively) correspond to negative JA and LP values, respectively. Hence, $\delta_{\mathrm{JA,LP}} = -1$ and $\delta_{\mathrm{LD}} = 1$. The constant α_x allows for an additional dimension specific rescaling, which was used for the consonants only. Thereby, α_{LD} was set to 0.05 to ensure that the lip distance is negative for the labial closures /p, b, m/. Further, α_{JA} was set to 5.0 to ensure that the JA values are close to zero in case of /s, z/. In all other cases $\alpha_x = 0$. Figure 5 shows the distributions of measured visual parameters for the different phonemes as a boxplot. The median values of respective distributions were used as visual target parameters during optimizations with the proposed vocal learning framework.

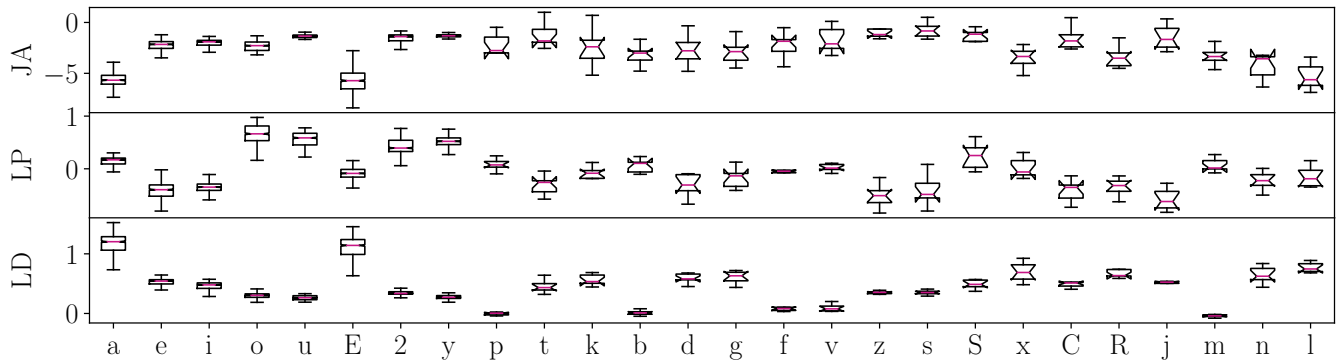

Fig. 5: Measured distributions for the visually accessible VTL parameters JA, LP, LD.

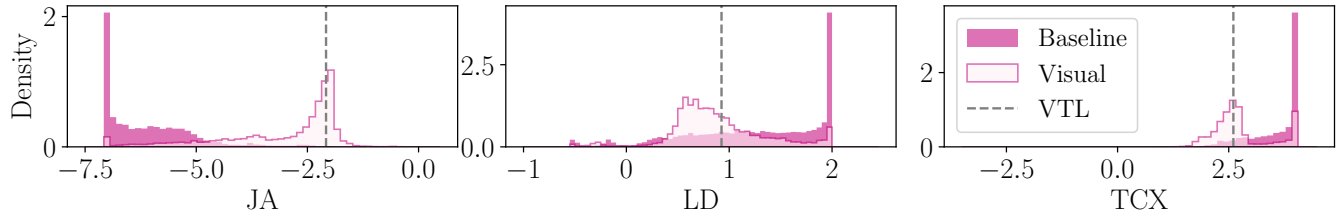

Fig. 6: Articulatory distributions of states corresponding to the vowel category /i/.

F. Experiments

First, vowel and consonant learning simulations were carried out as described in Section II-C. The simulations were repeated 100 times for each of the 26 phonemes, as the exploration based optimization process is non-deterministic. Hence different outcomes were obtained for each run, which provided an adequate statistical basis for the subsequent evaluation. Each run was stopped after a total amount of 1000 steps that actually involved synthesis. During optimization each state that whose acoustic outcome was identified as the desired phoneme category was saved for subsequent articulatory analyses. For the purpose of consonant learning, the state with the lowest phoneme loss corresponding to the vowel category /a/ was selected from the set of solutions obtained from the vowel learning experiment and used as the fixed vowel state during the syllable production. Both experiments, vowel and consonant learning, were then repeated with the additional $\mathcal{L}_{\mathrm{V}}$ loss component to test the impact of visual information on the learning success. All parameter settings were identical to the experiments without visual information.

The intelligibility of the generated samples was then assessed in a perceptual experiment. The selection of audio stimuli for such an experiment is non-trivial for the following reasons: (i) Since the size of the listening experiment should be kept small in order to allow the subjects to concentrate as much as possible during the entire participation, a representative assessment of all the states recorded during the optimizations is not possible due to the large articulatory and acoustic scope of these states. (ii) During optimization, a large number of individual (articulatory) solutions are found by the optimizing agent. Some of these solutions may preserve the intended phonetic category while others do not, e.g. due to recognizer misidentification. The individual solutions may also differ strongly in their biological plausibility. While it was found that a separation of these individual solutions in the high-dimensional articulatory space is in principle possible by dimensionality reduction and clustering, this technically challenging approach was left open for future work.

With this in mind, the solutions presented in the perceptual test were selected as follows: First, the $Q_{1.0}$, $Q_{0.75}$, $Q_{0.5}$, $Q_{0.25}$, $Q_{0.0}$ quantiles of the total loss distribution were calculated for each phoneme within both the data from optimizations without and with visual information. Then, for each phoneme, the five articulatory states whose corresponding total loss values were closest to the respective quantile were selected. That means e.g. the samples belonging to $Q_{1.0}$, $Q_{0.5}$ and $Q_{0.0}$ are the ones with the highest, the median and the lowest total loss. This procedure is motivated by the fact that the loss scale itself can be tested this way, i.e. whether, or, to what extent lower loss values are actually related to higher intelligibility or whether there exists a kind of overfit at very low loss values. In addition, a complete set of optimized phonemes was selected from the non-visual and visual data, based on subjective auditory impressions of the 100 samples with the lowest total loss from each run of the respective phonemes. This manual selection $\mathcal{M}$ was included to estimate the maximum achievable quality of the simulation. In addition, the MRI-based VTL preset states were tested as a baseline. The preset states of the phonemes were synthesized together with the same glottal states that were also used during the optimizations. Thus, the obtained motor scores for the VTL presets had exactly the same lengths and time constants, as well as glottal offsets in the case of unvoiced consonants, which is useful for comparison. All stimuli were newly syn-

thesized with vowel target durations of 300 ms and 250 ms for single vowels and syllables, respectively. Additionally 50 ms of silence were added to the beginning of each sample. Both modifications were done to allow for a more pleasant listening experience during the experiment. The perception experiment was carried out as an online multiple choice listening test. Thereby, participants heard one of the vowels or syllables at a time and had to choose which one they heard from the set of 8 vowels or the set of 16 syllables, respectively. Participants could also chose the category "other" in case they did not understand the given utterance. In total 20 subjects (13 male, 7 female) aged between 18 and 49 years (median: 28.5, mean: 29.6 ± 7.8) participated in the test. Participants were required to be German native speakers.

Beside the perceptual test, the optimization results were also analyzed within the domain of articulatory distributions. Figure 6 shows the distributions of articulatory parameters JA, LD, TCX obtained from states that were identified by the phoneme recognizer as the category /i/ during vowel learning. It can be seen that the distributions from optimizations with visual information are significantly closer to the biologically plausible values obtained from the respective VTL preset. This is expected to a certain degree for the visual parameters but it is an interesting observation in case of other parameters. To quantitatively test the degree of biological plausibility of the baseline and visual distributions, the mean absolute errors (MAE) between the distributions and the VTL preset values were calculated in each dimension. Then for a certain group G of articulatory parameters and group P of phonemes, coefficient can be calculated via:

$$C_G^P = \frac{1}{n_G}\frac{1}{n_P}\sum_{i\in G}\sum_{j\in P}\frac{\mathrm{MAE}_{ij}^{\mathrm{Visual}}}{\mathrm{MAE}_{ij}^{\mathrm{Baseline}}}. \tag{6}$$

For this purpose, phonemes were grouped into vowels, voiced and voiceless plosives, voiced fricatives plus lateral, voiceless fricatives, nasals, and a group containing all phonemes. VTL parameter groups were the visual parameters, all parameters, TCX only, as well as groups of *important dimensions*. The latter are of interest, since not all dimensions have an equal impact on the obtained phonetic categories, e.g. changes in an important dimension such as TCX may turn an /a/ into something else, while changes in a rather unimportant parameter such as VS may not. Parameter importances were determined by training a simple feed-forward neural network (three layers with 32 neurons each and *relu* activation function followed by a layer with 37 neurons and softmax activation function) to map between the 19 dimensional supra-glottal articulatory states and the 37 dimensional unit vectors representing the intended phonetic category that the respective articulatory state was optimized for. Performance was measured via F_1 score during 10-fold cross-validation. In each split, the permutation feature importance [42] was calculated in terms of F_1 score decrease on randomly shuffling the input matrix 10 times. Subsequently, the features were decreasingly ordered after their importance and the knee-point (point of maximum curvature) of the importance curve was determined using the Kneedle algorithm [43]. Features below the knee point were regarded as important dimensions.

Complete visualizations of all articulatory distributions, as well as all audio samples used in the listening experiment can be found in the supplementary materials[4].

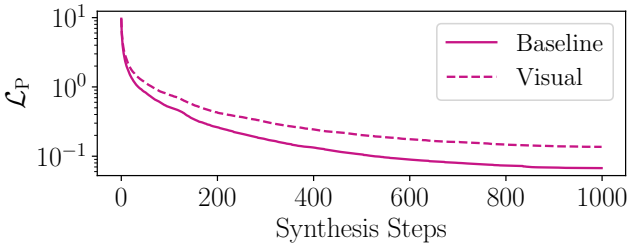

Fig. 7: Phoneme loss averaged across all categories and runs shown for the baseline and visual optimizations.

III. Results

Figure 7 shows the phoneme loss component of the total loss averaged across all optimization runs as a function of synthesis steps performed during optimization. The loss curves are shown for the baseline and optimization with additional visual information. It can be seen, that optimizations with visual information had systematically higher loss values than the baseline, which underlines the regularizing effect of the visual constraints. Even though Figure 7 shows an average, this pattern was observed consistently across phoneme categories. Figure 8 shows the recognition rates $\mathcal{R}$ calculated from the answers given by the participants of the perception experiment. Thereby, answers were separated into several stimuli groups that were tested during the experiment. It can be seen that the recognition rates of stimuli corresponding to Q_x are monotonically increasing when x is decreasing which validates the used phoneme loss. This effect is more prominent in the visual data, which shows a significant difference of recognition rates ($p < 0.05$ based on two-sided t-tests) between $Q_{0.75}$ and $Q_{0.0}$, whereas the baseline does not. Further, one can see that the average recognition rates for the manual selection of stimuli $\mathcal{M}$ are $(96.9 \pm 5.0)\,\%$[5] and $(94.2 \pm 7.6)\,\%$ for the baseline and visual stimuli, respectively. Hence, these are significantly higher than recognition rates corresponding to $Q_{0.0}$, which are $(75.0 \pm 6.8)\,\%$ and $(83.9 \pm 8.7)\,\%$, respectively. They also outperform the VTL baseline which was, on average, recognized correctly $(87.7 \pm 7.6)\,\%$ of the time. Overall, no significant difference was observed among the recognition rates for stimuli generated with and without visual information. From the articulatory analysis, TCX could be identified to be the most relevant VTL parameter for the model to discriminate between phoneme categories based on the articulatory state input vectors, see Table II. This result seems reasonable, given the strong impact of TCX on the tube area function. Furthermore, the visual parameters, especially lip distance, are often present among the important dimensions. Table II also shows the VTL preset distance coefficients calculated for different category groups and VTL parameter groups. For the groups of visual dimensions and all dimensions, all obtained coefficients are below 1.0, indicating that the visual distributions are closer to the VTL preset shapes than the

[4]https://github.com/paul-krug/visual-vocal-learning

[5]Given uncertainties describe the $1\,\sigma$ interval throughout this work.

Group	F_1 $[10^{-2}]$	Important Dimensions	C_{TCX}	C_{Imp}	C_{Vis}	C_{All}
Vowels	96.0 ± 0.2	TCX, LP, LD, HY, JA, TCY	0.74 ± 0.28	0.85 ± 0.15	0.77 ± 0.20	0.94 ± 0.09
Plosives	97.4 ± 0.4	TCX, TTY	1.09 ± 0.55	1.07 ± 0.38	0.71 ± 0.27	0.94 ± 0.14
Plosives†	89.5 ± 0.4	TCX, TCY, TS2, LD, TTY, TTX, LP	1.15 ± 0.72	0.94 ± 0.21	0.67 ± 0.22	0.97 ± 0.15
Fricatives	97.5 ± 0.2	TCX	1.13 ± 0.61	1.13 ± 0.61	0.70 ± 0.18	0.99 ± 0.13
Fricatives†	95.4 ± 0.2	TCX, TTX, TTY, LD, JA, TS3, LP, TCY	1.01 ± 0.42	0.89 ± 0.14	**0.68 ± 0.16**	0.96 ± 0.11
Nasals	98.9 ± 0.1	TCX, VO, LD	1.04 ± 0.57	0.90 ± 0.29	0.56 ± 0.26	0.93 ± 0.17
All	80.6 ± 0.6	TCX, LD, TTY, JA, LP, TCY, TS3, VO, TTX, HY, TS2	0.98 ± 0.20	0.92 ± 0.07	**0.70 ± 0.09**	0.96 ± 0.05

TABLE II: Results form the articulatory analysis. Groups of voiceless phonemes are indicated by †. Listed F_1 scores refer to the accuracy of the described forward model measured via 10-fold cross-validation (without feature permutation).

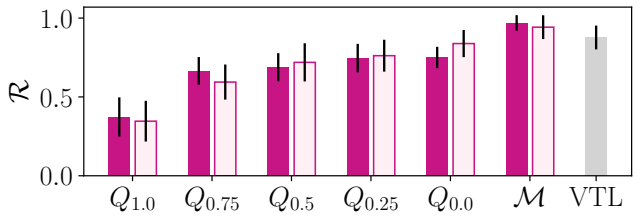

Fig. 8: Recognition rate results from the perception experiment with human listeners. Solid: baseline, outlined: visual stimuli.

baseline distributions, on average. For the group of important dimensions, the coefficients are largely below 1.0. For TCX they are below 1.0 only for the vowels. The coefficients for the visual dimensions are significant in case of the group of voiceless fricatives as well as the group of all phonemes in the sense that 1.0 lies outside $1.96\,\sigma$ interval around the respective measured values.

IV. Discussion

A. Conclusion

In this work a novel framework for vocal learning simulations was presented. The following key results were obtained from the experiments carried out:

- Single phoneme recognition constitutes a sufficient mechanism to formulate a loss function that correlates with the intelligibility quantified by human listeners.
- By using the said loss as an objective during artificial vocal learning, highly intelligible vowels and consonants embedded in corresponding CVs with the vowel /a/ could be generated.
- The main influence of visual information on the optimization process can be understood as regularization – leading to a higher degree of biological plausibility among the optimized states.

The last point is particularly interesting as it reveals the actual articulatory impact of the regularizing effect caused by the visual information. This effect is reasonable, due to the compensation possibilities through the different vocal tract dimensions. I.e. implausible configurations of certain dimensions can be compensated by implausible configurations of other dimensions in such a way that still results in highly intelligible speech. E.g., for the vowels in the baseline, a fully open jaw angle is preferred, which is implausible for vowels such as /i, e/. Since the visual information forces certain configurations, this limits the possibilities for compensation in other dimensions, so that more plausible configurations were found on average, as demonstrated by the obtained VTL distance coefficients.

B. Limitations and Future Work

This work has following limitations. First, consonants were only produced in context of vowel /a/. However, to generate continuous speech, the coarticulation model of the VTL uses consonants in the three different contexts /a, i, u/ [7]. It can be assumed that consonants in a single context are not sufficient to generalize to continuous speech. The generation of these further contexts is open for future work. With the vocal learning model presented here, the generation of consonants in any context is possible, but /a/ is the vowel that is easiest to generate. Consequently, consonants in the context of other vowels may require more simulation effort.
Another limiting factor is the quality of phoneme recognition, as worse models cause stronger confusion between individual phoneme categories. As a result, solutions are allowed during the simulation which do not correspond to the communicative intent. This problem will occur especially in the case of low-resource languages, where there is not much training material for phoneme recognition models. Whether this problem can be avoided by analyzing the articulatory distributions with the help of appropriate constraints would be conceivable. For example, consistency and minimal effort criteria could be used to select plausible solutions from the ensemble of correct and incorrect solutions a posteriori. Finding suitable criteria remains an open topic for future work.
Another problem often encountered in the simulation of syllables were articulatory artifacts or discontinuities that form between consonants and vowels, see Figure 9. As a consequence, the resulting syllables often sounded like clusters, e.g. /fa/ sounding like /fRa/ or /Sa/ sounding like /Sga/. The reason for the occurrence of these artifacts may be the single phoneme recognition. Due to the context independence, the following phoneme is rather unimportant and therefore the presence of an artifact is not evaluated negatively. On the articulatory level, however, the artifacts arise from the fact that the consonant states found predominantly match the vowel but are not completely appropriate. Whether the use of phoneme recognition systems trained on a larger acoustic intervals such as syllable-based or continuous phoneme recognition can cause a stronger rejection of auch states has to be tested in future work.

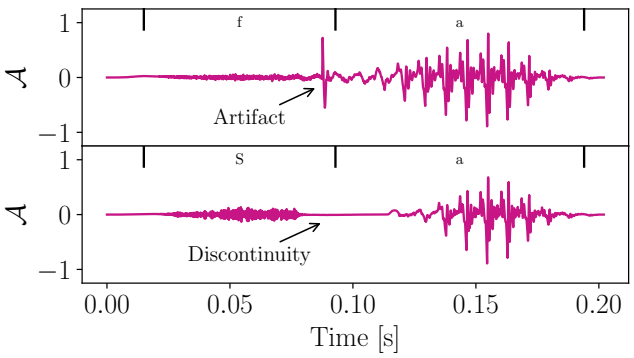

Fig. 9: Two exemplary audio amplitudes $\mathcal{A}$, representing syllables /fa/ (top) and /Sa/ (bottom). Different articulatory artifacts are visible.

Acknowledgments

This work has been funded by the Leverhulme Trust Research Project Grant No. RPG-2019-241: "High quality simulation of early vocal learning."

References

[1] P. K. Krug *et al.*, "Intelligibility and naturalness of articulatory synthesis with VocalTractLab compared to established speech synthesis technologies," in *Proc. SSW 11*, 2021, pp. 102–107.
[2] M. Pereira and G. Conti-Ramsden, *Language Development and Social Interaction in Blind Children*, ser. Essays in developmental psychology. Psychology Press, 1999.
[3] D. K. Oller and P. F. MacNeilage, "Development of speech production: Perspectives from natural and perturbed speech," in *The production of speech*. Springer, 1983, pp. 91–108.
[4] L. Ménard *et al.*, "Production and perception of French vowels by congenitally blind adults and sighted adults," *J. Acoust. Soc. Am.*, vol. 126, no. 3, pp. 1406–1414, 2009.
[5] ——, "Acoustic and articulatory analysis of French vowels produced by congenitally blind adults and sighted adults," *J. Acoust. Soc. Am.*, vol. 134, no. 4, pp. 2975–2987, 2013.
[6] M. Murakami *et al.*, "Seeing [u] aids vocal learning: Babbling and imitation of vowels using a 3D vocal tract model, reinforcement learning, and reservoir computing," in *Proc. ICDL-EpiRob*, 2015, pp. 208–213.
[7] P. Birkholz, "Modeling consonant-vowel coarticulation for articulatory speech synthesis," *PloS ONE*, vol. 8, no. 4, p. e60603, 2013.
[8] A. K. Philippsen *et al.*, "Learning how to speak: Imitation-based refinement of syllable production in an articulatory-acoustic model," in *Proc. ICDL-EpiRob*, 2014, pp. 195–200.
[9] Y. Gao *et al.*, "Articulatory copy synthesis based on a genetic algorithm." in *Proc. Interspeech*, 2019, pp. 3770–3774.
[10] ——, "Articulatory copy synthesis using long-short term memory networks," *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2020*, pp. 52–59, 2020.
[11] M. I. Jordan and D. E. Rumelhart, "Forward models: Supervised learning with a distal teacher," *Cogn. Sci.*, vol. 16, no. 3, pp. 307–354, 1992.
[12] I. Howard and M. Huckvale, "Training a vocal tract synthesiser to imitate speech using distal supervised learning," in *Proc. SpeCom*, vol. 2, 2005, pp. 159–162.
[13] G. Bailly, "Learning to speak. Sensori-motor control of speech movements," *Speech Commun.*, vol. 22, no. 2-3, pp. 251–267, 1997.
[14] J. Serkhane *et al.*, "Infants' vocalizations analyzed with an articulatory model: A preliminary report," *J. Phon.*, vol. 35, no. 3, pp. 321–340, 2007.
[15] B. J. Kröger *et al.*, "Towards a neurocomputational model of speech production and perception," *Speech Commun.*, vol. 51, no. 9, pp. 793–809, 2009.
[16] J. A. Tourville and F. H. Guenther, "The DIVA model: A neural theory of speech acquisition and production," *Lang. Cognit. Process.*, vol. 26, no. 7, pp. 952–981, 2011.
[17] H. Nam *et al.*, "Computational simulation of CV combination preferences in babbling," *J. Phon.*, vol. 41, no. 2, pp. 63–77, 2013.
[18] A. K. Philippsen *et al.*, "Goal babbling of acoustic-articulatory models with adaptive exploration noise," in *Proc. ICDL-EpiRob*, 2016, pp. 72–78.
[19] H. Rasilo and O. Räsänen, "An online model for vowel imitation learning," *Speech Commun.*, vol. 86, pp. 1–23, 2017.
[20] I. S. Howard and P. Birkholz, "Modelling vowel acquisition using the Birkholz synthesizer," *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2019*, pp. 304–311, 2019.
[21] A. Philippsen, "Goal-directed exploration for learning vowels and syllables: A computational model of speech acquisition," *KI-Künstliche Intelligenz*, vol. 35, no. 1, pp. 53–70, 2021.
[22] S. Pagliarini *et al.*, "Vocal imitation in sensorimotor learning models: A comparative review," *IEEE Trans. Cogn. Develop. Syst.*, vol. 13, no. 2, pp. 326–342, 2020.
[23] S. Maeda, "Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model," in *Speech production and speech modelling*. Springer, 1990, pp. 131–149.
[24] P. K. Krug *et al.*, "Modelling microprosodic effects can lead to an audible improvement in articulatory synthesis," *J. Acoust. Soc. Am.*, vol. 150, no. 2, pp. 1209–1217, 2021.
[25] D. R. van Niekerk *et al.*, "Finding intelligible consonant-vowel sounds using high-quality articulatory synthesis," in *Proc. Interspeech*, 2020, pp. 4457–4461.
[26] P. K. Krug *et al.*, "Efficient exploration of articulatory dimensions," *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2022*, pp. 51–58, 2022.
[27] Y. Xu and Q. E. Wang, "Pitch targets and their realization: Evidence from Mandarin Chinese," *Speech Commun.*, vol. 33, no. 4, pp. 319–337, 2001.
[28] S. Prom-On *et al.*, "Modeling tone and intonation in Mandarin and English as a process of target approximation," *J. Acoust. Soc. Am.*, vol. 125, no. 1, pp. 405–424, 2009.
[29] P. Birkholz *et al.*, "Model-based reproduction of articulatory trajectories for consonant–vowel sequences," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 5, pp. 1422–1433, 2010.
[30] S. Schaal, "Dynamic movement primitives-a framework for motor control in humans and humanoid robotics," in *Adaptive motion of animals and machines*. Springer, 2006, pp. 261–280.
[31] S. Mirjalili and A. Lewis, "The whale optimization algorithm," *Adv. Eng. Softw.*, vol. 95, pp. 51–67, 2016.
[32] P. K. Krug *et al.*, "Articulatory synthesis for data augmentation in phoneme recognition," in *Proc. Interspeech (Accepted)*, 2022.
[33] S. Kirkpatrick *et al.*, "Optimization by simulated annealing," *Science*, vol. 220, no. 4598, pp. 671–680, 1983.
[34] S. Mirjalili, "SCA: A sine cosine algorithm for solving optimization problems," *Knowl.-Based Syst.*, vol. 96, pp. 120–133, 2016.
[35] R. Storn and K. Price, "Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces," *J. Glob. Optim.*, vol. 11, no. 4, pp. 341–359, 1997.
[36] H. Karami *et al.*, "Flow direction algorithm (FDA): A novel optimization approach for solving optimization problems," *Comput. Ind. Eng.*, vol. 156, p. 107224, 2021.
[37] R. Gross *et al.*, "Multi-PIE," *Image Vis. Comput.*, vol. 28, no. 5, pp. 807–813, 2010.
[38] C. Sagonas *et al.*, "300 faces in-the-wild challenge: The first facial landmark localization challenge," in *Proc. ICCVW*, 2013, pp. 397–403.
[39] S.-E. Wei *et al.*, "Convolutional pose machines," in *Proc. CVPR*, 2016, pp. 4724–4732.
[40] X. Dong *et al.*, "Supervision-by-registration: An unsupervised approach to improve the precision of facial landmark detectors," in *Proc. CVPR*, 2018, pp. 360–368.
[41] C. Sagonas *et al.*, "300 faces in-the-wild challenge: Database and results," *Image Vis. Comput.*, vol. 47, pp. 3–18, 2016.
[42] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
[43] V. Satopää *et al.*, "Finding a "kneedle" in a haystack: Detecting knee points in system behavior," in *Proc. DCSW*, 2011, pp. 166–171.