Speech Communication 147 (2023) 51–62

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Simulating vocal learning of spoken language: Beyond imitation

Daniel R. van Niekerk [a],[*], Anqi Xu [a], Branislav Gerazov [b], Paul K. Krug [c], Peter Birkholz [c], Lorna Halliday [d], Santitham Prom-on [e], Yi Xu [a]

[a] Department of Speech, Hearing and Phonetic Sciences, University College London, United Kingdom
[b] Faculty of Electrical Engineering and Information Technologies, UKiM, Skopje, Republic of North Macedonia
[c] Institute of Acoustics and Speech Communication, Technische Universität Dresden, Germany
[d] Medical Research Council Cognition and Brain Sciences Unit, University of Cambridge, United Kingdom
[e] Department of Computer Engineering, King Mongkut's University of Technology Thonburi, Thailand

ARTICLE INFO

Keywords:
Speech production
Articulatory speech synthesis
Speech acquisition
Computational phonetics
Early vocal learning
Canonical babbling

ABSTRACT

Computational approaches have an important role to play in understanding the complex process of *speech acquisition*, in general, and have recently been popular in studies of *vocal learning* in particular. In this article we suggest that two significant problems associated with imitative vocal learning of spoken language, the *speaker normalisation* and *phonological correspondence* problems, can be addressed by linguistically grounded auditory perception. In particular, we show how the articulation of consonant–vowel syllables may be learnt from *auditory percepts* that can represent either individual utterances by speakers with different vocal tract characteristics or ideal phonetic realisations. The result is an optimisation-based implementation of *vocal exploration* – incorporating semantic, auditory, and articulatory signals – that can serve as a basis for simulating vocal learning beyond imitation.

1. Introduction

Investigating how children learn to speak is a promising path towards understanding the human speech system, language acquisition, and cognition. With increases in computational power and advances in the speech sciences, computational approaches have emerged as an important means of modelling this complex process (Dupoux, 2018). This is evident in recent work on vocal learning in particular. Simulations of early-stage babbling (Serkhane et al., 2007; Nam et al., 2013) have proposed mechanisms that may explain the typical distributions of early vocalisations in pre-linguistic learners (Davis and MacNeilage, 1995). Late-stage or *canonical babbling* is associated with the emergence of spoken language, including vowels and reduplicated consonant–vowel (CV) syllables (Oller and Eilers, 1988), and has been simulated as an *imitative* or *goal-directed* process (Bailly, 1997; Howard and Huckvale, 2005; Howard and Messum, 2007; Philippsen et al., 2014; Philippsen, 2021; Rasilo et al., 2013; Rasilo and Räsänen, 2017). Despite recent progress, auditory-acoustic imitation is confronted with two significant obstacles: the *speaker normalisation* (Howard and Huckvale, 2005; Rasilo and Räsänen, 2017) and *correspondence* (Philippsen, 2021; Messum and Howard, 2015) problems.

The normalisation problem refers to the difficulty of comparing phonologically equivalent utterances by different speakers. Differences in the vocal tract size and shape between infants and adults (especially of different sexes) result in significant acoustic differences, in particular the frequency-scaling of formants (Wakita, 1977; Zhan and Waibel, 1997). The question of correspondence concerns how utterances are associated with phonological units that are used to construct messages carrying linguistic meaning. This is related to the *symbol grounding* problem in the field of robotics which involves constructing mappings between invariant sensory inputs from the environment (Harnad, 1990).

If late-stage vocal exploration is goal-oriented and results in the articulation of phonological units, these problems may be addressed by relying on language oriented auditory-perceptual goals instead of auditory-acoustic imitation. This would be compatible with established facts about the developmental progression in infants; language oriented perception of vowels typically emerges by the age of 6 months and precedes the onset of canonical babbling (Kuhl, 2004). That is, some phonological aspects of auditory perception *start to develop before the production of simple utterances* that could be reformulated by a caregiver (Rasilo et al., 2013; Messum and Howard, 2015). A recent computational study has confirmed that distributional learning from auditory input alone may partially explain this process (Schatz et al., 2021). However, auditory signals may also be combined early on with inputs from the other senses towards semantic grounding (Frank et al., 2014) as proposed by Harnad (1990) (Harnad, 1990).

[*] Corresponding author.
E-mail address: d.vniekerk@ucl.ac.uk (D.R. van Niekerk).

https://doi.org/10.1016/j.specom.2023.01.003
Received 9 June 2022; Received in revised form 8 January 2023; Accepted 20 January 2023
Available online 23 January 2023

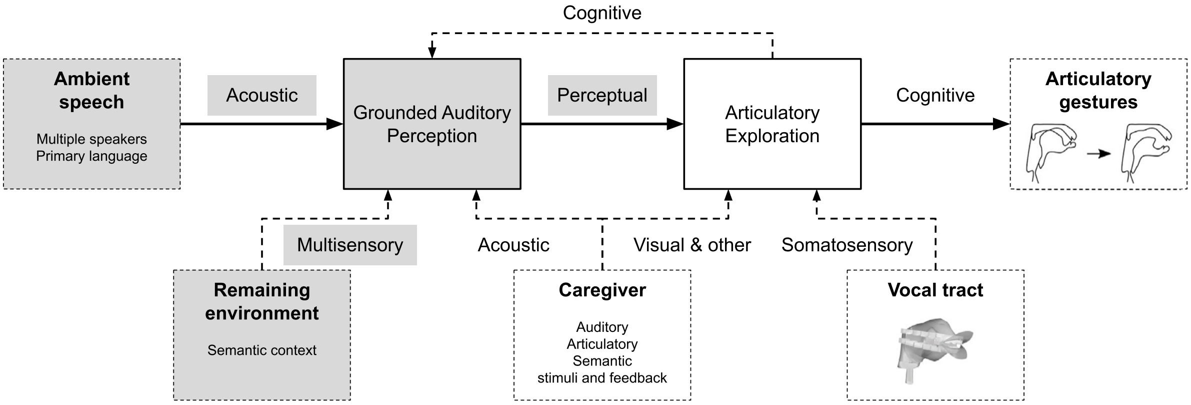


Fig. 1. Sources of information (dashed-line blocks) and the nature of sensory signals (between blocks) that are applicable to articulatory exploration. The shaded areas illustrate how grounded auditory perception can integrate information that has not been included in previous vocal learning simulations.

Fig. 1 presents a simplified view of the sources of information and the nature of sensory signals that are applicable to articulatory exploration. The shaded areas show how grounded auditory perception can naturally facilitate the inclusion of information sources needed to address the speaker normalisation and correspondence problems. That is, by combining the semantic context and ambient speech data from multiple speakers, the perceptual mapping can perform implicit speaker normalisation and produce a representation with phonological correspondence. By contrast, the unshaded areas represent aspects that have been discussed in previous work based on acoustic imitation (Bailly, 1997; Howard and Huckvale, 2005; Howard and Messum, 2007; Philippsen et al., 2014; Philippsen, 2021; Rasilo et al., 2013; Rasilo and Räsänen, 2017). This includes the possible role of caregiver interaction to provide semantic and articulatory information (Rasilo et al., 2013; Messum and Howard, 2015; Murakami et al., 2015) and the importance of somatosensory feedback (Tourville and Guenther, 2011). While computational works exist that construct a phonological perceptual space as a basis for vocal learning (Philippsen, 2021; Kröger et al., 2009, 2014; Barnaud et al., 2019), these simulations differ materially from the proposal in Fig. 1 in that they construct the perceptual space either through self-organisation of utterances generated by the learner or stimuli from a single external speaker who serves as teacher or "master agent". This does not consider the effect that (1) multiple speakers in ambient speech stimuli and (2) perceptual development before the onset of vocal exploration may have on learning.

In this work, we simulate *articulatory exploration*, or babbling, driven by auditory-perceptual goals which include the sources of information highlighted in Fig. 1. We do not model the ontogenesis of the perceptual mapping but construct a functional equivalent based on an existing speech corpus to investigate the use of perceptual representations, *auditory percepts*, as basis for goal-directed vocal exploration. Specifically, we seek to answer the following questions:

1. Can auditory percepts be used to reproduce individual utterances by different speakers regardless of their vocal tract characteristics?
2. If the encoding of the relevant phonological units are known, can auditory percepts be used to find the articulation of ideal phonetic realisations? That is, to produce utterances that represent phonologically appropriate generalisations over the speech inputs they are derived from.

We answer these questions by subjecting the outputs of the simulation to recognition experiments (see Section 4) to obtain a quantitative measure of success and show that both of these outcomes are possible. The empirical results confirm that goal-directed vocal exploration can succeed on the basis of reproducing low-dimensional auditory percepts. The perceptual mapping used in this work may also be a natural way of aggregating auditory experience to support autonomous exploration and vocal learning from memory.

2. Approach

We view articulatory exploration in abstract terms such that it is also directly applicable in the field of articulatory speech synthesis (Van Niekerk et al., 2022). The process is formulated as an optimisation task using the *VocalTractLab* (VTL) articulatory synthesiser (Birkholz, 2005, 2013) to produce candidate utterances. The objective function combines articulatory specifications and an auditory-perceptual mapping to evaluate articulatory gestures. An outline of the approach is illustrated in Fig. 2 and motivated in the following subsections.

2.1. Articulatory exploration

In our simulation, goal-directed exploration is the process of minimising auditory and articulatory losses to discover a phonologically appropriate utterance. The central block in Fig. 2 is the global optimisation task

$$\boldsymbol{u}^* = \underset{\boldsymbol{u}\in \boldsymbol{U}_\theta}{\arg\min}\, L(\boldsymbol{u}, Q, \theta) \tag{1}$$

of finding an articulatory gesture $\boldsymbol{u}^*$ that minimises the loss function L which is dependent on the speaker vocal tract θ and the auditory-perceptual mapping Q described in Section 2.2. This is done by the optimisation algorithm sampling each gesture $\boldsymbol{u}$ from the articulatory space $\boldsymbol{U}_\theta$ which is also determined by the speaker model.

2.2. Auditory perceptual objectives

We construct an auditory-perceptual mapping that is functionally equivalent to the proposal in Fig. 1, by means of two practical restrictions. Firstly, in the absence of raw stimuli from ecological environments as proposed in Dupoux (2018), we make use of a transcribed speech recognition corpus. This data source satisfies both of the conditions to construct a grounded mapping: it contains the linguistic (semantic) context associated with the speech signal and utterances from multiple speakers. Secondly, the mapping is derived before the onset of the simulation and not updated continuously throughout the process. These restrictions should not affect the validity of the conclusions regarding the questions raised in Section 1 and can, in principle, be relaxed to expand the scope of future work.

Furthermore, we are not concerned with modelling the emergence of a phonological representation, but focus on the task of vocal exploration. With this aim, we adopt the syllable as unit of perception for the following reasons: (1) It provides the necessary context to account for signal variation known from acoustic–phonetic descriptions of coarticulation (Adriaans, 2018; Liu et al., 2022). (2) This would correspond directly to articulatory representations that we adopt in this work on articulatory-phonetic grounds. That is, a single percept would

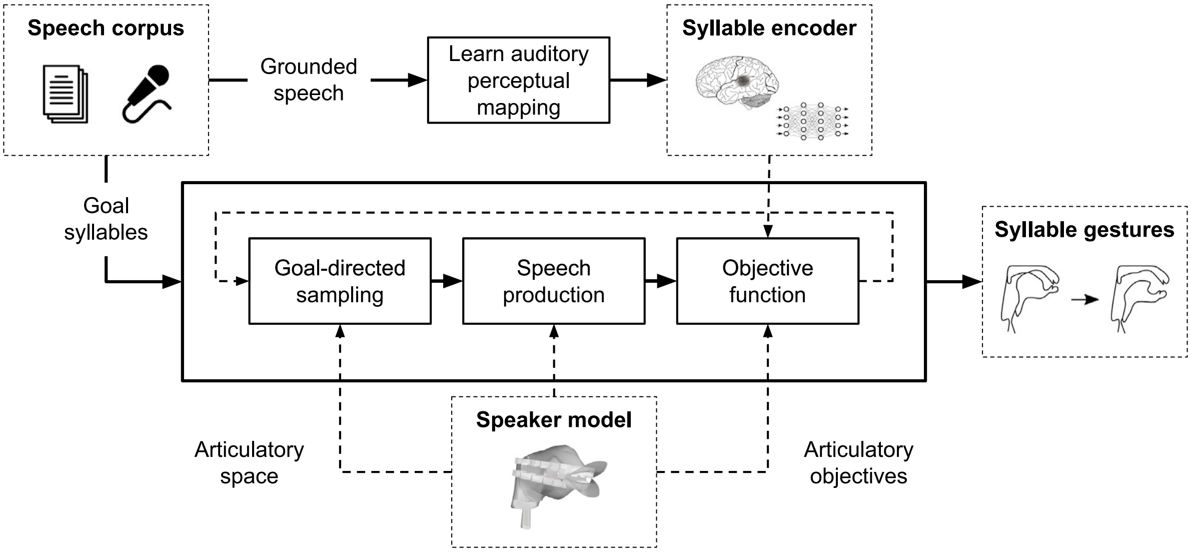


Fig. 2. A process for discovery of phonological articulatory gestures. The central exploration task is goal-directed and relies on articulatory sampling, speech production, and auditory and articulatory objectives (Section 2). The speech production component is implemented using VTL and is responsible for generating two different types of output given the speaker model and sampled articulatory targets: (1) synthesised audio representing a candidate utterance which is processed by the syllable encoder (Section 3.1) to evaluate the auditory perceptual objective, and (2) the vocal tract tube areas and transfer function which is used to implement articulatory objectives (Section 3.3).

correspond to a fixed set of *articulatory targets* per syllable (refer to Section 2.4). (3) Syllables are plausible early perceptual units that may be involved in overcoming the *segmentation problem* (Jusczyk, 1997; Räsänen et al., 2018) and require further attention in computational studies of language acquisition (Schatz et al., 2021; Räsänen, 2012).

The result is a *syllable encoder* that consumes a pre-segmented acoustic signal and produces an *auditory percept* vector or embedding that is used as objective during optimisation.

2.3. Articulatory objectives

Articulatory information that forms part of the optimisation goal in our simulation can be described as either *somatosensory* or *regularisation* objectives.

Somatosensory objectives represent specifications, obtainable from non-auditory signals, that the learner can consciously monitor through somatosensory feedback (Nasir and Ostry, 2006) once a correspondence between external stimuli and imitating actions is established (Brass and Heyes, 2005). For example, it is known that sighted vocal learners may benefit from visual examples of articulation (Mills, 1988) from which they may learn to expect the somatosensory feedback associated with lip closure. We do not model the process of establishing *articulatory correspondence* between the different senses and somatosensory targets but simply substantiate their inclusion.

Regularisation objectives may originate in physiologically motivated processes (Serkhane et al., 2007; Nam et al., 2013) or constraints (Oohashi et al., 2013) that are not under conscious control, but nevertheless determine the typical articulatory solution space. Regularisation objectives are explicit in our work since we rely on a general optimisation algorithm with uniform priors and VTL only provides basic physical restrictions. That is, the synthesiser does not implicitly enforce articulatory coordination or implement higher-level organisation such as the direct control of constriction that may be relevant in natural speech production (Saltzman and Munhall, 1989).

2.4. Speech production

To produce articulatory trajectories, we use the target-approximation model (TAM) (Xu and Wang, 2001) implemented using a 5th-order critically damped linear system (Birkholz et al., 2018). This model is implemented in VTL, allowing the complete specification of trajectories – and synthesis of utterances – from a set of *articulatory targets*, durations for each segment, and time constants which correspond to "articulatory effort" (Birkholz, 2007).

This compact parameterisation of articulatory dynamics combined with assumptions of synchronisation (Birkholz et al., 2011) has enabled the discovery of simple CV syllables using derivative free optimisation by significantly reducing the associated degrees of freedom (Xu et al., 2019; Van Niekerk et al., 2020). Furthermore, articulatory targets may be more relevant phonologically than complete trajectories (Turk and Shattuck-Hufnagel, 2020) and provide a means of decoupling temporal parameters that vary over different prosodic forms and speech rates (Birkholz et al., 2011).

3. Implementation

3.1. Auditory perceptual mapping

To construct the syllable encoder we used supervised learning to find a mapping between speech signals of up to 1 second in duration to fixed-size vectors that encode each syllable type uniquely. The clean subset of the *Librispeech* speech recognition corpus (Panayotov et al., 2015) was annotated using the CMU pronunciation dictionary (Carnegie Mellon University, 2000), with a phoneset appropriate for American English, to extract CV syllables. Combinations of 3 consonants and 15 vowels (including diphthongs) were included in the selection to represent a complete set of vowels and minimal set of consonants for the experimental conditions described in Section 4 (Table 2). Using this dataset, containing 487 male and 453 female speakers, a recurrent neural network was trained to map a sequence of mel-frequency cepstral coefficient vectors (MFCCs) to a single output vector — defined as the concatenation of the one-hot encoded phonetic labels in the training data.

While the training data labels are categorical, representing ideal points in the output space, the model was trained using the mean squared error (MSE) loss to construct a regression model. Therefore it is interpreted as a mapping between acoustic realisations of CV syllables and points in an 18-dimensional continuous perceptual space (3 consonant and 15 vowel types). A point in this space is a *language oriented percept* $\boldsymbol{p}$ with components p_i where $i \in \{0, \ldots, N-1\}, N = 18$ and $p_i \in [0.0, 1.0]$. Furthermore, since the model was trained on multiple speakers from both sexes, it is expected to perform implicit

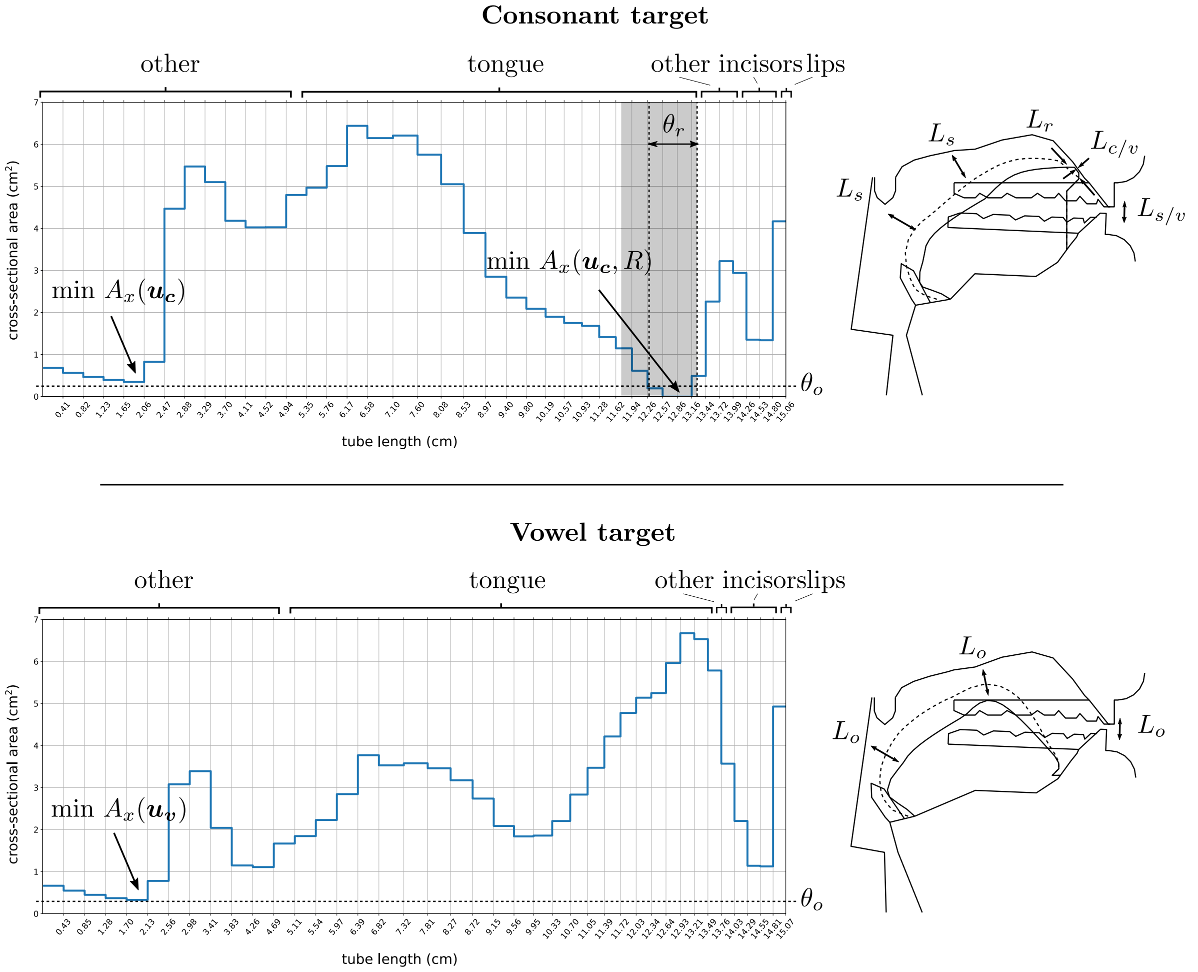


Fig. 3. An example of articulatory targets, top to bottom, producing the CV utterance [dae]. On the right are plots of the articulatory parameters with associated tube area function on the left. The influence of individual loss terms are illustrated on the articulatory plots. The tube area plots illustrate some of the thresholds and values used in the loss terms, as detailed in Appendix A.2. VTL provides the identity of the articulator associated with different sections of the tube area function (labelled along the top) (Birkholz, 2014). This allows applying functions to specific articulators. For example, in the top plot, $\min A_x(\boldsymbol{u}_c, R)$ detects a closure in the shaded section where $R = \{tongue_{tip}\}$.

speaker normalisation, that is, the output space should be speaker-invariant. See Appendix A.1 for a more complete description of the dataset, model architecture and signal processing involved.

This forms an auditory goal-space used to specify a goal percept $\boldsymbol{q}$ which can be compared to each trial $\boldsymbol{p}$, obtained during exploration, using a metric such as the Euclidean distance:

$$L_p(\boldsymbol{p}, \boldsymbol{q}) = \|\boldsymbol{p} - \boldsymbol{q}\|_2. \tag{2}$$

In our experiments (Section 4), the values of $\boldsymbol{q}$ are obtained either by mapping a specific acoustic realisation to the perceptual space (Experiment 1) or by specifying ideal syllables in terms of their one-hot encoding (Experiment 2). For example, an ideal target $\boldsymbol{q}$ for the syllable /bæ/, as in "bad", would correspond to a vector $[1, 0, 0, 1, 0, \dots, 0]$ given that /b/ is the first component for the 3 consonants and /æ/ the first component for the 15 vowels.

3.2. Speech production

For speech synthesis, VTL[1] was used to realise articulatory targets with the "JD2" male speaker and geometric glottis model (Birkholz et al., 2019). While an experimental scaled down "child" vocal tract model does exist, it is not guaranteed to be as realistic as JD2 which is based on magnetic resonance imaging (MRI) data of a real speaker and the adult male speaker model suffices to answer the questions posited in Section 1.

Since we focused on investigating the upper vocal tract parameters and the production of voiced speech, the glottal parameters were kept constant at the preset values for "modal voice" with the exception of the *chink area* and *relative amplitude* which were optimised to allow control of the voice onset time (VOT) and voice quality of the consonant (Abramson, 1977). All of the upper vocal tract parameters were optimised, except the *velum opening* (VO) which was kept closed – since we did not account for utterances with nasality (Table 2) – and the *tongue root* (TRX, TRY) parameters which were derived from the tongue body values (Krug et al., 2022). The temporal aspects of an utterance were determined by a preset fixed duration for each of the two segments and the target-approximation trajectories for the transition between consonant and vowel targets controlled by two free parameters — one time constant each for the glottal and upper vocal tract parameters. Table 1 details the full set of parameters and corresponding optimisation configuration described above.

3.3. Articulatory objectives

The somatosensory objectives were implemented using proprioceptive or tactile feedback approximated by evaluating the VTL *tube area function* associated with each articulatory target. Three such objectives

[1] Version 2.3 available at https://www.vocaltractlab.de.

Table 1
Target parameters for VTL's "JD2" speaker with geometric glottis model (Birkholz et al., 2019). Neutral (initial) parameters and optimisation ranges are shown; single values in the Range column indicate constants.

Articulatory parameter		Neutral	Range	
Upper vocal tract model				
Hyoid position (horz.)	HX	1.00	$[0.0, 1.0]$	cm
Hyoid position (vert.)	HY	−4.75	$[-6.0, -3.0]$	cm
Jaw position (horz.)	JX	0.00	$[-0.5, 0.0]$	cm
Jaw angle	JA	−2.00	$[-7.0, 0.0]$	deg.
Lip protrusion	LP	−0.07	$[-1.0, 1.0]$	cm
Lip distance	LD	0.95	$[-2.0, 4.0]$	cm
Velum shape	VS	0.00	$[0.0, 1.0]$	
Velic opening	VO	−0.10	−0.10	cm^2
Tongue body (horz.)	TCX	−0.40	$[-3.0, 4.0]$	cm
Tongue body (vert.)	TCY	−1.46	$[-3.0, 1.0]$	cm
Tongue tip (horz.)	TTX	3.50	$[1.5, 5.5]$	cm
Tongue tip (vert.)	TTY	−1.00	$[-3.0, 2.5]$	cm
Tongue blade (horz.)	TBX	2.00	$[-3.0, 4.0]$	cm
Tongue blade (vert.)	TBY	0.50	$[-3.0, 5.0]$	cm
Tongue side elevation 1	$TS1$	0.00	$[0.0, 1.0]$	cm
Tongue side elevation 2	$TS2$	0.00	$[0.0, 1.0]$	cm
Tongue side elevation 3	$TS3$	0.00	$[-1.0, 1.0]$	cm
Glottis model				
Fundamental frequency	$F0_{gl}$	120.00	120.00	Hz
Sub-glottal pressure	SP_{gl}	8000.00	8000.00	dPa
Lower rest displacement	LD_{gl}	0.01	0.01	cm
Upper rest displacement	UD_{gl}	0.01	0.01	cm
Chink area	CA_{gl}	0.00	$[-0.25, 0.25]$	cm^2
Phase lag	PL_{gl}	0.88	0.88	rad.
Relative amplitude	RA_{gl}	1.00	$[-1.00, 1.00]$	
Double pulsing	DP_{gl}	0.00	0.00	
Pulse skewness	PS_{gl}	0.00	0.00	
Flutter	FL_{gl}	25.00	25.00	%
Aspiration strength	AS_{gl}	−40.00	−40.00	dB
Target approximation model				
Vocal tract time-constant	τ_{vt}	0.010	$[0.005, 0.010]$	s
Glottis time-constant	τ_{gl}	0.010	$[0.005, 0.010]$	s

were defined and implemented by sets of simple loss terms included in the overall loss L (Eq. (1)):

- *Vowel objective*: ensures that the vowel target is voiced (L_ω) with an open vocal tract (L_o).
- *Closure objective*: ensures the closure of the vocal tract (L_c) needed for a stop consonant.
- *Visual objective*: ensures that a closure occurs either at the lips or elsewhere while the lips are open (L_v) *depending on the consonant type*.

Similarly, two regularisation objectives were implemented:

- *Precision objective*: prefers consonant targets with precise closures (L_r) at a single place of articulation (L_s).
- *Coarticulation objective*: prefers a smaller articulatory distance between the consonant and vowel targets (L_d).

Fig. 3 shows an example of consonant and vowel targets producing an utterance [dae] with their associated tube areas and the influence of articulatory objectives described here. See Appendix A.2 for precise definitions of the loss terms.

3.4. Optimisation process

We used the Tree-structured Parzen Estimator (TPE) approach (Bergstra et al., 2011) as the optimisation algorithm to drive the articulatory sampling as implemented in the *hyperopt*[2] software package. The loss function was defined as a linear combination of the auditory and articulatory loss functions described earlier (term weights were not optimised but all articulatory loss terms were scaled by a factor 0.2 to ensure that the auditory loss L_p dominates). The optimisation algorithm samples articulatory targets which are synthesised by VTL and each sample is evaluated by the auditory perceptual mapping and tube area function to determine the resulting loss. The optimisation algorithm was configured to sample the first 5% of trials uniformly, including the neutral position. To improve the computational efficiency of the process, synthesis and auditory evaluation is only performed when the somatosensory objectives are satisfied. In the case of failure to achieve these objectives, the loss function is set to an arbitrary large value proportional to the articulatory loss.

Table 2
CVC words from which CV syllables are extracted.

	/b/	/d/	/g/
/iː/	bead	deed	
/ɪ/	bid	did	
/ɛ/	bed	dead	
/ae/	bad	dad	
/ɒ/	bod		god
/uː/	booed		
/ʊ/			good
/ʌ/	bud		

4. Experimental setup

We designed two experiments to address the questions posed in Section 1:

- *Experiment 1*: The simulation is set up to find articulatory gestures that *reproduce specific instances* of CV utterances produced by male and female speakers. The proposed system based on the male vocal tract is compared with a baseline which uses acoustic matching instead of the syllable encoder described in Section 3.1.
- *Experiment 2*: The simulation is configured to find articulatory gestures that produce CV utterances with *specific phonetic identities*. Ideal auditory-perceptual objectives are defined and different sets of articulatory objectives are compared.

On the surface these tasks are distinct in character, however, we consider the success of the outcomes in terms of the production of phonologically relevant utterances. That is, the system should produce or reproduce goal utterances that are equivalent in the particular spoken language context. Consequently, the test utterances are taken from a specific set of words shown in Table 2. These words contain the variation in simple vowels and consonants considered, and all end with the coda consonant /d/ to reduce the complexity of the experiments. These word contexts are used in two ways to enable both tasks to be evaluated in terms of recognition experiments, that is, to obtain a quantitative measure of success related to the function of spoken language (see Section 4.3). Firstly, the goal utterances produced by the male and female speakers (Experiment 1) are extracted from recordings of these CVC words. Secondly, the CV gestures found by vocal exploration (Experiment 1 and 2) are embedded in the same CVC words by appending an articulatory target for the coda /d/ using the predefined set of articulatory parameters from VTL. Thus, each evaluated sample represents an instance of a known word from Table 2.

4.1. Experiment 1

Recordings of the words in Table 2 were made by a British male and female speaker and manually segmented to extract the CV portions of each. Each of these template utterances was used to drive the proposed optimisation process by inputting the audio to the syllable encoder to obtain a goal percept $\boldsymbol{q}$. The exploration algorithm was set to produce trial utterances that were also processed with the syllable encoder, to

[2] https://github.com/hyperopt/hyperopt (v0.2.5).

obtain their encoded percept $\boldsymbol{p}$. The auditory loss L_p was calculated using these two outputs (Eq. (2)). Note, that although the syllable encoder was trained using American English data, it is used here to compare individual utterances relative to each other. That is, it is not important that the goal utterance does not correspond to a specific phonetic category in the perceptual space as long as the language dialects are close enough.[3] For articulatory objectives, only the vowel and closure somatosensory objectives that ensure a basic CV utterance were applied here. Neither the visual nor regularisation objectives were included.

For comparison, in the baseline system the auditory loss L_p was replaced by the MSE calculated frame-by-frame over the sequences of acoustic features extracted for the template and trial utterances. The features used were identical to those used in the syllable encoder Appendix A.1 and frame alignment was ensured by using the template segment durations during synthesis of the trial utterances.

4.2. Experiment 2

The goal percepts in this experiment were the one-hot encoded representations of the CVs shown in Table 2. That is, using the chosen perceptual encoding to represent the consonant and vowel identities directly, without the use of template utterances. In this case, the goal utterances are generalised American English pronunciations of the CV syllables as derived from the transcribed speech corpus.

Here we make a comparison between different degrees of articulatory feedback to determine whether this benefits the expected outcomes under the constraint of a finite number of exploratory trials. Independent simulations were initiated to test the following sets of articulatory objectives:

- *Minimal*: includes only the vowel and closure objectives as in Experiment 1 and serves as a baseline condition.
- *Visual*: includes the minimal set of objectives and the visual objective. The fact that L_v is dependent on the consonant type implies a dependence on the auditory perceptual goal.
- *Visual + precise*: adds the precision objective.
- *Visual + precise + coart.*: adds the coarticulation objective.

For the last condition the optimisation process involved two passes. In the first pass both the consonant and vowel targets are optimised using the *Visual + precise* configuration, followed by the second pass optimising only the consonant targets using *Visual + precise + coart.* with the vowel parameters from the first pass fixed. This has the effect that the coarticulation objective only affects the consonant targets given the vowel found in the first pass. Jointly optimising the consonant and vowel using the coarticulation objective can reduce the chances of finding an appropriate vowel target (Van Niekerk et al., 2022). Since the two-pass procedure is not directly comparable with the rest, the *Visual + precise* configuration was also applied in two passes for comparison with the coarticulation condition.

4.3. Evaluation

Two methods were used to recognise the synthesised discovered single-word utterances (Table 2) to determine their intelligibility:

- Automatic speech recognition (ASR) or speech-to-text is an inexpensive and objective mechanism which provides reliable results for VTL speech on this task (Van Niekerk et al., 2020). This allows rapid evaluation of experimental configurations and was used in both experiments.
- An open-ended transcription task that asks listeners to enter a word for each utterance. This is more expensive, with practical limitations, but offers more precise feedback and may be considered more relevant than ASR results. This method was only used in Experiment 2.

Since the exploratory process is non-deterministic, depending on the set of random initial trials, each experimental condition was evaluated through independent repeated experiments. That is, for each configuration described in Sections 4.1 and 4.2 the simulation was run N times for each of the words in Table 2 with different random seeds. The results presented in the next section are in terms of the mean accuracy or recognition rate over all the independent instances of the process and represent the expected intelligibility for the experimental condition over this set of words. For example, given the 13 words and $N = 20$ seeds, the recognition rate is calculated over $13 \times 20 = 260$ utterances. In all cases the best candidate, according to the objective function, was selected after the process sampled 5,000 valid utterances. That is, the process was terminated after synthesising a fixed number of utterances, excluding articulatory targets that did not satisfy the basic somatosensory objectives.

For the ASR evaluation, we used the state-of-the-art *Google Speech-to-Text* service.[4] The service automatically determines the appropriate back-end model to use based on the input, however, we explicitly selected the language and dialect: British English for Experiment 1 and American English for Experiment 2. In addition to the audio samples, the set of 13 words in Table 2 was submitted as "speech contexts". This adapts the language model in favour of this set of words and is considered best-practice for recognising short utterances. A single request was made to the service for each sample and the associated response was in the form of an ordered n-best list of orthographic transcriptions or an empty list (null result) which is interpreted as the rejection of an unintelligible utterance. The system responses were automatically post-processed to obtain a single transcription (or null result) for each sample by applying two operations: (1) only the most likely candidate transcription was retained from the n-best list, and (2) the text was normalised by lowercasing and removal of excess whitespace characters. For all ASR evaluations $N = 20$ instances of each word was evaluated.

The transcription task was set up as an online experiment using the *Gorilla*[5] platform. Each listener was presented with randomised utterances consisting of $N = 10$ instances of each word from the baseline and best conditions (*Minimal* and *Visual + precise + coart.*). The process consisted of basic user and consent forms, a soundcheck to verify the use of headphones (Milne et al., 2020), a short practice transcription task, and the main transcription task. For the main task, listeners were expected to type in the word played back through headphones or indicate if the utterance was unintelligible after listening to it no more than 3 times. Care was taken not to include any implicit or explicit information about the content or quality of the utterances that could introduce bias in the responses. Therefore the practice transcription task contained unrelated 1 or 2 syllable words produced by a female speaker and given the experimental setup it is reasonable to assume that the extent of listeners' prior expectation was to hear short single-word or unintelligible utterances. American English participants were recruited on *Amazon Mechanical Turk*[6] resulting in the following participant funnel: 110 visited the task; 57 passed the soundcheck and 43 submitted a completed task. Three of the completed tasks were excluded after manual inspection revealed anomalies in the response time and/or distributions of answers. The result was a total of 40 participants with valid responses. Responses were post-processed by performing two operations: (1) the text was automatically normalised

[3] In this case, an analogy is of an American speaker perceiving and reproducing a British pronunciation, which only differs marginally in some vowels over the set of words considered here.

[4] https://cloud.google.com/speech-to-text accessed during May–June 2021.

[5] https://gorilla.sc/

[6] https://www.mturk.com/ during June 2021.

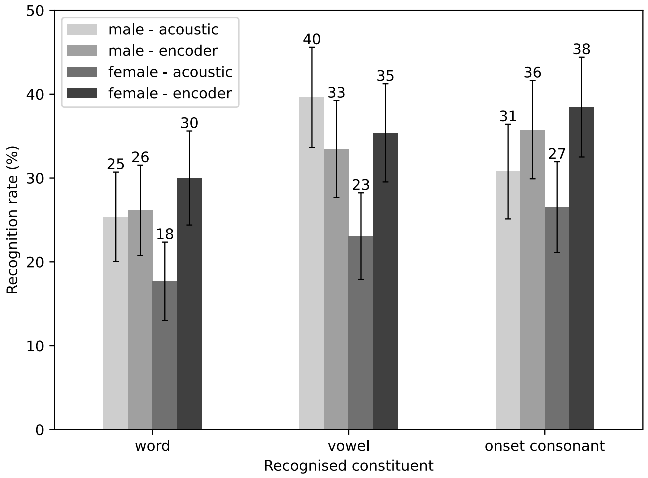


Fig. 4. Experiment 1 – ASR recognition rates comparing the **acoustic matching** and **syllable encoder** auditory objectives for reproducing utterances by a **male** and **female** speaker (error bars indicate the 95% confidence intervals). The recognition rate for the **female-acoustic** utterances is significantly lower than the other conditions.

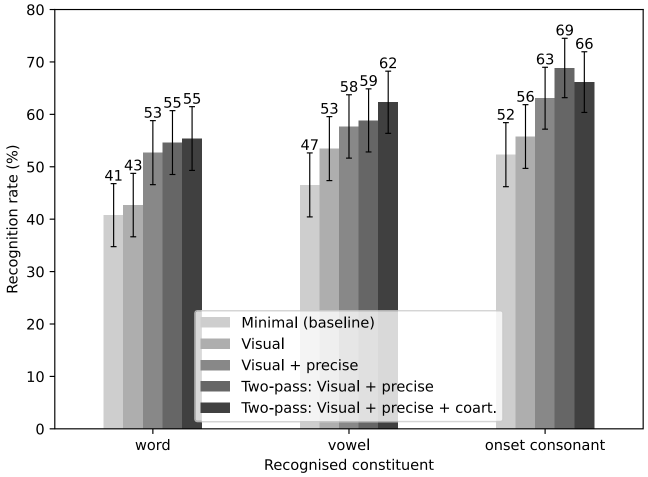


Fig. 5. Experiment 2 – ASR recognition rates comparing conditions with different degrees of articulatory specification (error bars indicate the 95% confidence intervals). The inclusion of the **precision objective** is associated with a significant improvement in recognition rate.

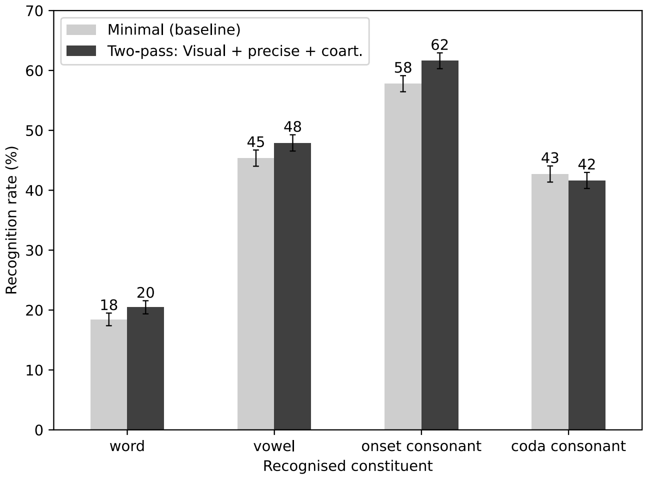
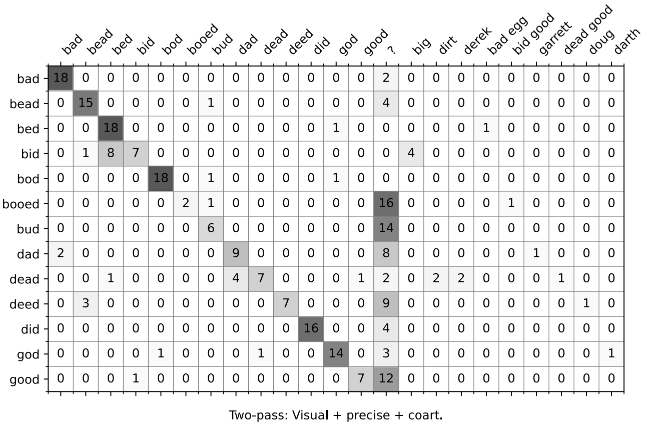


Fig. 6. Experiment 2 – Online transcription recognition comparing the baseline condition with the best configuration identified using the ASR results (error bars indicate the 95% confidence intervals). The inclusion of **regularisation objectives** are associated with a significant improvement in the recognition rate.



	bad	bead	bed	bid	bod	booed	bud	dad	dead	deed	did	god	good	?	big	dirt	derek	bad egg	bid good	garrett	dead good	doug	darth
bad	18	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0
bead	0	15	0	0	0	0	1	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0
bed	0	0	18	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0
bid	0	1	8	7	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0
bod	0	0	0	0	18	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
booed	0	0	0	0	0	2	1	0	0	0	0	0	0	16	0	0	0	0	1	0	0	0	0
bud	0	0	0	0	0	0	6	0	0	0	0	0	0	14	0	0	0	0	0	0	0	0	0
dad	2	0	0	0	0	0	0	9	0	0	0	0	0	8	0	0	0	0	0	1	0	0	0
dead	0	0	1	0	0	0	0	4	7	0	0	0	1	2	0	2	2	0	0	0	1	0	0
deed	0	3	0	0	0	0	0	0	0	7	0	0	0	9	0	0	0	0	0	0	0	1	0
did	0	0	0	0	0	0	0	0	0	0	16	0	0	4	0	0	0	0	0	0	0	0	0
god	0	0	0	0	1	0	0	0	1	0	0	14	0	3	0	0	0	0	0	0	0	0	1
good	0	0	0	1	0	0	0	0	0	0	0	0	7	12	0	0	0	0	0	0	0	0	0

Two-pass: Visual + precise + coart.

Fig. 7. Experiment 2 – The complete ASR confusion matrix for utterances from the best performing configuration. Each row contains the 20 instances of the reference word with outputs labels on each column.

by lowercasing and removal of excess whitespace characters, and (2) cases of unambiguous spelling or typographical errors were manually corrected. The conditions for applying a correction was that the initial response was not a valid word and that the set of possible corrections consisted of only one likely valid word (exceptions were made to refrain from changing responses where the listener may have attempted to transcribe disfluencies), for example, “beed” → “bead”, “gaurd” → “guard”, and so forth, but not “booke” → “book”.

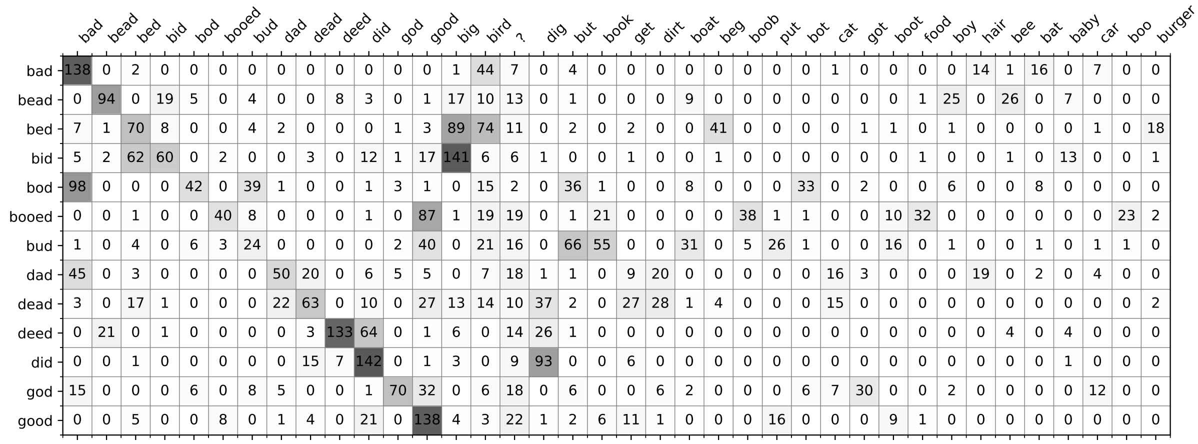
Transcriptions obtained from the ASR system typically consisted of 1–2 words or a null result and most responses from human listeners were single words (compare Figs. 7 and 8 which are discussed later). To determine the recognition rate, transcriptions were either compared directly to the orthographic reference, referred to hereafter as the “word level”, or to phonetic representations of the onset, vowel and coda. The latter was obtained by manually mapping the orthographic forms using the CMU dictionary as reference, or where this was not possible, to a null symbol. For this open-vocabulary transcription task, the set of possible outputs (transcriptions) and its prior probability distribution are unknown – preventing calculation of a chance-level recognition rate directly. However, for our experimental setup the expected recognition rates for a random utterance generator is less than 5% for the large-vocabulary ASR system and 2% for the human transcription task on the word level. Refer to Appendix B for a detailed note on the interpretation of absolute recognition rates. Similarly, vowel and onset recognition rates should be interpreted carefully, not as independent classification tasks, but rather to provide insight into the relative contribution on the word level error rate. The analysis presented in Appendix B establishes that the recognition rates for all the test conditions are consequential. In the following section we therefore focus on the relative performance of different experimental conditions that address the questions posed in Section 1.

To quantify the significance of results throughout, effect sizes are reported in terms of Cohen’s d and two-tailed p values at the 95% level from Welch’s unequal variances t-test.

5. Results

5.1. Experiment 1

The results of Experiment 1 are presented in Fig. 4. Experimental conditions relying on the syllable encoder (whether male or female templates) or acoustic matching with the male templates result in



	bad	bead	bed	bid	bod	booed	bud	dad	dead	deed	did	god	good	big	bird	?	dig	but	book	get	dirt	boat	beg	boob	put	bot	cat	got	boot	food	boy	hair	bee	bat	baby	car	boo	burger
bad	138	0	2	0	0	0	0	0	0	0	0	0	0	1	44	7	0	4	0	0	0	0	0	0	0	0	1	0	0	0	0	14	1	16	0	7	0	0
bead	0	94	0	19	5	0	4	0	0	8	3	0	1	17	10	13	0	1	0	0	0	9	0	0	0	0	0	0	0	1	25	0	26	0	7	0	0	0
bed	7	1	70	8	0	0	4	2	0	0	0	1	3	89	74	11	0	2	0	2	0	0	41	0	0	0	0	1	1	0	1	0	0	0	0	1	0	18
bid	5	2	62	60	0	2	0	0	3	0	12	1	17	141	6	6	1	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0	1	0	13	0	0	1
bod	98	0	0	0	42	0	39	1	0	0	1	3	1	0	15	2	0	36	1	0	0	8	0	0	0	33	0	2	0	0	6	0	0	8	0	0	0	0
booed	0	0	1	0	0	40	8	0	0	0	1	0	87	1	19	19	0	1	21	0	0	0	0	38	1	1	0	0	10	32	0	0	0	0	0	0	23	2
bud	1	0	4	0	6	3	24	0	0	0	0	2	40	0	21	16	0	66	55	0	0	31	0	5	26	1	0	0	16	0	1	0	0	1	0	1	1	0
dad	45	0	3	0	0	0	0	50	20	0	6	5	5	0	7	18	1	1	0	9	20	0	0	0	0	0	16	3	0	0	0	19	0	2	0	4	0	0
dead	3	0	17	1	0	0	0	22	63	0	10	0	27	13	14	10	37	2	0	27	28	1	4	0	0	0	15	0	0	0	0	0	0	0	0	0	0	2
deed	0	21	0	1	0	0	0	0	3	133	64	0	1	6	0	14	26	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	4	0	0	0
did	0	0	1	0	0	0	0	0	15	7	142	0	1	3	0	9	93	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
god	15	0	0	0	6	0	8	5	0	0	1	70	32	0	6	18	0	6	0	0	6	2	0	0	0	6	7	30	0	0	2	0	0	0	0	12	0	0
good	0	0	5	0	0	8	0	1	4	0	21	0	138	4	3	22	1	2	6	11	1	0	0	0	16	0	0	0	9	1	0	0	0	0	0	0	0	0

Two-pass: Visual + precise + coart.

Fig. 8. Experiment 2 – The truncated online listener confusion matrix for utterances from the best performing configuration. The full matrix contains 400 instances (10 trials by 40 listeners) of each reference word. The truncated plot shows the 25 most popular responses in addition to the 13 target words comprising 6.7% of unique responses and 70.1% of all instances.

similar recognition rates — no significant differences are found. This indicates that the syllable encoder performs comparatively regardless of the sex of the speaker and that the acoustic matching is also effective when the speaker sex is matched to the male vocal tract. By comparison, the word recognition rate when using acoustic matching against the mismatched templates (female) are significantly lower (d = 0.27, with $t(501.56) = -3.32$, $p = .001$), demonstrating the speaker normalisation problem. Further inspection of the results shows that the syllable encoder sustained or improved recognition rates of all vowels for the female templates and the high vowel /i/ for the male templates. For the male templates, the vowels /ɒ/ and /ɪ/ were significantly more successful when using acoustic matching. However, the types of errors made with acoustic matching were indicative of high variance compared to a consistent bias with the syllable encoder. That is, errors with the syllable encoder involved outputs that were perceptually close, whereas acoustic matching led to less predictable confusions.

5.2. Experiment 2

The results for Experiment 2 based on the ASR and online transcriber tasks are presented in Figs. 5 and 6 respectively. The ASR results show a trend of improvement with the inclusion of additional articulatory objectives with a significant difference in word recognition rate given the precise objective compared to the baseline and visual conditions ($d = 0.24$ with $t(517.87) = -2.74$, $p = .006$ between *Visual + precise* and *Minimal*, and $d = 0.20$ with $t(517.95) = -2.29$, $p = .022$ between *Visual + precise* and *Visual*). There are no significant differences amongst the conditions that include regularisation objectives despite the two-pass results involving double the number of trials of the consonant. It may be noted, however, that informal inspection and subsequent results do suggest that articulatory solutions are more "prototypical" with the regularisation objectives, especially in terms of coarticulation (Van Niekerk et al., 2022). The online transcription task (Fig. 6) also exhibits a significant improvement of outcomes for the two-pass process with precise and coarticulation objectives over the baseline ($t(10381.89) = -2.60$, $p = .009$ for the word recognition rates) albeit with a smaller measurable effect compared to the ASR results ($d = 0.05$). There is no significant difference in recognition rate for the coda consonant, which is expected since all utterances were based on the same preset coda target for /d/. Lastly, a comparison of the word recognition rates for the ASR and online transcription tasks confirms the difference in the nature of the tasks which is illustrated further in Figs. 7 and 8. Fig. 7 shows the full confusion matrix for the ASR results of the best performing condition. It is clear that the inclusion of the 13 words as "speech contexts" adjusts the prior probabilities in favour of these words to the extent that an utterance is most likely to be recognised within this set or judged as unintelligible (the null result indicated with "?" in the figure). By contrast, Fig. 8 reflects an open transcription task without prior knowledge of the set of reference words; 571 unique responses were observed and listeners were less likely to label utterances as unintelligible. Even so, when the results are viewed in terms of their consonant and vowel constituents, the recognition rates on the two tasks are comparable.

To contextualise the recognition rates obtained in this simulation, the intelligibility of natural speech can be evaluated using the ASR system to provide an expected upper bound for this experimental setup. For this purpose approximately 20 instances of each of the target words (Table 2) were extracted from the Librispeech corpus.[7] The resulting word accuracy is 80.6 ± 4.9% compared to the best configuration with 55.4 ± 6.1%. Although the recognition rate for isolated words (not uttered in sentence context) may be higher, the syllable encoder on which the simulation is based has the same intrinsic limitation – i.e., it is trained on continuous speech. The recognition rate is comparable with the test set classification accuracy of 81.7% reported in Appendix A.1.

6. Discussion

The results for Experiment 1 and 2 are not intended to be comparable and should be viewed separately. Firstly, the tasks differ fundamentally due to different definitions of the goal percepts, and secondly, the experimental conditions differ significantly, for example, the two different ASR systems used as evaluators will have distinct performance characteristics due to dialect and construction. It is however notable that there is a significant difference in the absolute recognition rates even for the baseline condition. This could be expected since the agent's auditory discrimination task and the evaluation task are more closely aligned in Experiment 2: (1) the goal percept is an ideal point in American English perceptual space whereas the template utterances are not guaranteed to be optimal British English examples, and (2) the perceptual space of the discriminator and evaluator are matched — both are American English. The remainder of this section focuses on discussing the results for the individual experiments with reference to the research questions posed in Section 1.

[7] The only exceptions due to scarcity were "bod" and "booed" which were substituted with "bon" and "boon" respectively and in the case of the former only 12 instances were found.

Experiment 1 demonstrates that articulatory exploration using language oriented perception (Fig. 1) is more successful than acoustic matching at reproducing phonological utterances when a vocal tract mismatch exists. This suggests that the auditory-perceptual objective can be used in an interactive setting where the learner imitates an arbitrary caregiver's stimuli. A secondary observation is that, despite an output representation and training data based on the American English vowel space, the syllable encoder supports the reproduction of British English utterances with comparable success to acoustic matching when considering the male, matched vocal tract, condition. This confirms that it maps to a continuous perceptual space with the ability to represent (interpolate) vowels that are not characteristic of American English (see Section 3.1).

Experiment 2 demonstrates that low-dimensional auditory percepts can be used to produce utterances that reflect aggregated auditory experience. This may be useful for vocal learning in an autonomous setting or, when the mapping is known, to enumerate phonological units. Furthermore, although the perceptual mapping was only trained to discriminate three voiced consonants, it was possible to obtain reasonable recognition rates through the inclusion of basic articulatory objectives and glottal constraints. This suggests that an incomplete discriminative model can still be useful at early stages of development. Experiment 2 also shows that the inclusion of a regularisation objective that prefers more precise articulation of closures results in significantly better recognition rates. The reason for this should be investigated formally in future work, however, inspection of articulatory solutions suggest that imprecise or double-articulations are perceptually ambiguous or sensitive to the articulatory effort controlled by the time constant parameter. Lastly, we have included a condition that prefers solutions where the onset consonant is maximally coarticulated with the vowel (Xu, 2020; Liu et al., 2022). The fact that this configuration is among the best performing conditions is further computational evidence for intra-syllable synchronisation (Xu et al., 2019; Van Niekerk et al., 2020, 2022).

6.1. Relationship to other work

The present work is related to other goal-directed simulations of babbling that produce spoken language utterances such as vowels and CVs (Bailly, 1997; Howard and Huckvale, 2005; Howard and Messum, 2007; Philippsen et al., 2014; Philippsen, 2021; Rasilo and Räsänen, 2017). However, a fundamental distinction of this study is the inclusion of language oriented perception that may affect how ambient language, including multiple speakers, influences vocal exploration (Fig. 1): (1) The ambient language may influence auditory perception early on, before the onset of late-stage babbling (Kuhl, 2004). (2) During the development of auditory perception, the learner may rely on multisensory signals to partially resolve some acoustic ambiguity (Frank et al., 2014). This clarifies the notion of a language oriented goal-space which allows quantitative evaluation in terms of recognition-based experiments which had not yet been applied in this context.

An interactive process and the role of caregiver feedback during vocal learning has been proposed as mechanisms that may alleviate the speaker normalisation (Rasilo and Räsänen, 2017; Plummer, 2012) and correspondence problems (Messum and Howard, 2015). The present work does not preclude the integration of these information sources but asserts that earlier perceptual development should also be accounted for. Under the current conception, the benefits of interactive feedback could be described in terms of continuous development of the perception model with inputs and feedback from the caregiver and exploration process (see Fig. 1). It is also possible to interpret the auditory perception function in our simulation more abstractly. That is, as representing the joint auditory experience of the overall system (i.e., the learner–caregiver combination). In this case, an independent model that is updated during the learning process could represent the learner's own auditory experience which should eventually approximate the joint experience to become independent.

The present study, and Fig. 1 in particular, explicitly draws links between simulations of babbling and computational work on speech perception (Frank et al., 2014; Schatz et al., 2021). While the focus in this article is on vocal exploration, our simulation presents a well-defined task and methodology based on quantitative evaluation that may be useful for testing assumptions about speech perception. The simplifications in perceptual modelling described in Section 2.2 may be relaxed to investigate the simultaneous development of perception and production or the model can be constructed in different ways to investigate different instants during development (Dupoux, 2018).

Our implementation of articulation which relies on syllable synchronisation and the target-approximation model is based on the idea that the syllable is central to simplifying the biomechanical and cognitive demands of articulatory coordination (Xu, 2020). This is supported by observations of intra-syllable coarticulation (Liu et al., 2022) which in turn is the primary reason for selecting the syllable as the temporal domain of perception, see Section 2.2. Furthermore, the alignment of articulatory and perceptual units has the advantage of allowing for a context-free mapping between their respective representations. This significantly simplifies the structure of mappings in either direction, that is, both forward and inverse models could be implemented using a simple feedforward network which maps between percepts and articulatory targets. This interface, forged during vocal learning, may be directly observable (Casile et al., 2011).

Lastly, three important questions fall outside the scope of the current framework: (1) *Adaptive control* plays an important role in the speech motor system (Houde and Jordan, 1998) and has been the basis of successful simulations of articulatory control (Parrell et al., 2019). The importance of somatosensory signals are acknowledged in our work, however, there is no risk of external perturbations in our simulation. This means that it is possible to rely on the empirically derived kinematics of the target-approximation model (Xu and Wang, 2001; Birkholz, 2007) to determine articulatory dynamics. Moreover, to model adaptive control would require a more complete physical simulation of the vocal tract plant, including properties such as mass and elasticity. (2) *Intrinsic motivation* may be responsible for forming the developmental stages of speech acquisition (Moulin-Frier et al., 2014). That is, it could be viewed as a possible mechanism that initiates or controls instances of the current simulation. To be specific, it may determine how to enumerate the optimisation goals and/or replace the general optimisation algorithm used here. (3) *Segmentation* of continuous speech or the mapping to a sequence of percepts representing syllables is not considered here but assumed possible (Jusczyk, 1997; Räsänen et al., 2018).

6.2. Limitations and future work

Our experimental setup relied on synthesis of utterances with predefined segmental durations. This means that the system was constrained to evaluating spectral properties and only basic temporal features affected by articulatory effort. Duration and additional aspects of the glottal model need to be incorporated into the set of optimised parameters to cover a larger set of phonological units in English and other languages. Furthermore, duration and articulatory effort should be allowed to vary to allow for variations in speaking rate and prosody (Birkholz et al., 2011). The simulation could benefit from allowing for tolerances instead of finding articulatory targets that are dependent on a specific value of duration or articulatory effort.

During the course of our experiments, we inspected plots of the articulatory solutions to compare the qualitative impact of different sets of regularisation objectives. As could be expected, including the objectives for precision and coarticulation leads to a greater proportion of "prototypical" solutions that correspond to articulatory phonetic

Table A.3
Network architecture of the syllable encoder.

Architecture	Layer	Output shape	Parameters
Input	Input	200×12	N/A
Recurrent	Bidirectional LSTM	200×512	550912
	Dropout (50%)	200×512	N/A
	Bidirectional LSTM	512	1574912
Feedforward	Dropout (50%)	512	N/A
	Dense (ReLU)	128	65664
	Dense (ReLU)	128	16512
	Dense (ReLU)	128	16512
	Dense (ReLU)	64	8256
	Dense (ReLU)	32	2080
	Dense (Sigmoid)	18	594

descriptions. However, questions of *articulatory correspondence* are beyond the scope of the current work. Two questions could be addressed in future work: (1) what are the conditions – stimuli, constraints or processes – that could lead to establishing sets of articulatory objectives similar to those implemented here, and (2) how do the solutions found in the simulation compare to prototypical articulatory descriptions. For the latter, future work could attempt to quantify this objectively by selecting an appropriate articulatory reference dataset and implementing a procedure for comparing vocal tract configurations.

Lastly, it may be necessary to investigate additional articulatory feedback mechanisms and exploration strategies that facilitate learning of more complex syllable types towards complete language coverage (Van Niekerk et al., 2022).

7. Conclusions

By considering computational work on speech perception and production, we have presented an implementation of vocal exploration which includes semantic, auditory, and articulatory signals. It was suggested that language oriented auditory-perceptual representations can facilitate the inclusion of these information sources to account for the speaker normalisation and phonological correspondence problems associated with imitative vocal learning and the possibility of using such low-dimensional percepts was demonstrated. This approach extends existing work on vocal learning by constructing an appropriate goal-space for vocal learning and adopting a recognition-based methodology for quantitative evaluation. Moreover, the proposed optimisation-based framework was shown to be an effective way of exploring the vocal tract domain which may be useful for self-generating grounded data for developing articulatory synthesisers in new languages (Van Niekerk et al., 2022) or for learning forward and inverse models of articulation (Jordan and Rumelhart, 1992).

CRediT authorship contribution statement

Daniel R. van Niekerk: Software, Investigation, Data curation, Conceptualization, Methodology, Writing – original draft, Writing – review & editing. **Anqi Xu:** Writing – review & editing, Methodology. **Branislav Gerazov:** Writing – review & editing, Software. **Paul K. Krug:** Writing – review & editing. **Peter Birkholz:** Writing – review & editing, Software. **Lorna Halliday:** Conceptualization. **Santitham Prom-on:** Conceptualization. **Yi Xu:** Funding acquisition, Conceptualization, Supervision, Methodology, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

Funding

This work was supported by the Leverhulme Trust with Research Project, United Kingdom Grant RPG-2019-241: “High quality simulation of early vocal learning”.

Appendix A. Implementation details

A.1. Syllable encoder

More than 380,000 CV syllables were extracted from the “train clean” subset of the *Librispeech* corpus using phone-level forced-alignments obtained with the *Kaldi ASR toolkit* (Povey et al., 2011) and partitioned into training, development and test sets as illustrated in Table A.4. From the raw audio, 12-dimensional Mel-frequency cepstral coefficients (including energy) without delta or acceleration coefficients were extracted every 5 ms in a 20 ms Hamming window (zero-padded to 512 samples at 22050 Hz) using *librosa*[8] (McFee et al., 2015) and z-normalised using the statistics of the training data set. These sequences were pre-padded to have a length of 200 samples (spanning 1 s) and used as input for training the encoder. The model thus learns to map a sequence of these acoustic features, spanning a syllable, to a single 18-dimensional vector (as described in Section 3.1). For example, a syllable extracted from the corpus and transcribed as being part of the word “bad” will be assigned the pronunciation /bae/ which has an ideal one-hot vector representation $[1, 0, 0, 1, 0, \dots, 0]$ given that /b/ is the first component for the 3 consonants and /ae/ the first component for the 15 vowels (from Table 2).

Tensorflow[9] was used to train the network consisting of 2 bidirectional long short-term memory (LSTM) recurrent network layers (Hochreiter and Schmidhuber, 1997) followed by 6 dense feedforward layers with dropout regularisation as shown in Table A.3. Since we view the output simply as a point (embedding) in a continuous space, the activations of the output layer are used directly, that is, they are not normalised to represent probabilities over the output dimensions or parts thereof. Training proceeded with early stopping based on the validation set loss with a patience of 6 epochs. When applied as a classifier on the test data, the resulting model obtained an overall accuracy of 79.9% with 96.5% and 81.7% for consonant and vowel identities respectively. This gives an indication of the quality of the data, labels and alignments, and the difficulty of the perception task as well as an explicit upper limit for experimental results described in Section 5.2.

A.2. Articulatory loss functions

Let a CV be represented by a 38-dimensional concatenation vector $\boldsymbol{u} = [\boldsymbol{u}_c, \boldsymbol{u}_v, \boldsymbol{u}_t]$ of the articulatory targets for the consonant $\boldsymbol{u}_c$ and vowel $\boldsymbol{u}_v$ (one 18-dimensional vector each) and the two target-approximation time constants $\boldsymbol{u}_t$ (refer to the free parameters shown in Table 1). Furthermore, let θ represent the set of constant speaker-specific parameters for “JD2” on which all the VTL functions are implicitly dependent. Then the loss terms can be defined as follows (with some examples illustrated in Fig. 3):

[8] https://github.com/librosa/librosa

[9] https://www.tensorflow.org/ (v2.4).

Table A.4
CVs extracted to different dataset partitions from *Librispeech* (Panayotov et al., 2015). Phone labels are in the ARPABET phoneset used in the CMU pronunciation dictionary (Carnegie Mellon University, 2000).

	/aa/	/ae/	/ah/	/ao/	/aw/	/ay/	/eh/	/er/	/ey/	/ih/	/iy/	/ow/	/oy/	/uh/	/uw/	TOTAL
Training set																
/b/	5257	9219	27,333	2476	7367	18,566	10,337	2831	2681	15,865	20,604	3818	2375	2069	569	131,367
/d/	8484	11,333	9964	9384	8506	7094	12,699	621	10,019	18,152	7442	5379	45	1299	11,343	121,764
/g/	7922	4046	2520	1762	226	548	10,704	2396	5244	7791	125	8732	15	5309	250	57,590
TOTAL	21,663	24,598	39,817	13,622	16,099	26,208	33,740	5,848	17,944	41,808	28,171	17,929	2435	8677	12,162	**310,721**
Development																
/b/	626	1107	3294	307	891	2258	1253	333	326	1890	2491	462	310	249	50	15,847
/d/	1020	1361	1200	1145	1025	847	1525	77	1210	2207	906	661	5	131	1364	14,684
/g/	944	487	310	200	20	75	1298	270	615	939	15	1054	18	636	12	6893
TOTAL	2590	2955	4804	1652	1936	3180	4076	680	2151	5036	3412	2177	333	1016	1426	**37,424**
Test																
/b/	544	968	2872	282	777	1944	1080	293	277	1656	2168	400	231	214	48	13,754
/d/	875	1202	1025	996	907	739	1328	61	1053	1915	788	569	7	129	1210	12,804
/g/	840	442	266	200	17	50	1135	245	525	824	7	915	16	538	8	6028
TOTAL	2259	2612	4163	1478	1701	2733	3543	599	1855	4395	2963	1884	254	881	1266	**32,586**

(1) The *voicing loss* tests for a voiced vowel by applying a threshold θ_ω to the magnitude of the volume velocity transfer function $|H_\omega|$[10]

$$L_\omega = \begin{cases} 0, & \text{where } \max |H_\omega(\boldsymbol{u_v})| > \theta_v \\ 1, & \text{otherwise.} \end{cases}$$

(2) The *open tract loss* tests the vowel for a minimum opening of the vocal tract by applying threshold θ_o on the tube area function A_x[11]

$$L_o = \begin{cases} 0, & \text{where } \min A_x(\boldsymbol{u_v}) > \theta_o \\ \frac{\theta_o - \min A_x(\boldsymbol{u_v})}{\theta_o}, & \text{otherwise.} \end{cases}$$

(3) The *closure loss* tests A_x for any complete closure during the consonant

$$L_c = \begin{cases} 0, & \text{where } \min A_x(\boldsymbol{u_c}) < \epsilon, \\ \frac{\min A_x(\boldsymbol{u_c})}{\max A_x}, & \text{otherwise,} \end{cases}$$

where ϵ is the smallest value depending on the numerical resolution of the simulation and $\max A_x$ is the maximum tube area possible given the speaker θ.

(4) The *precise closure loss* applies a threshold θ_r to the tube closure lengths function D_x[12] (as illustrated in Fig. 3) to incentivise consonant closures over a short section of the vocal tract

$$L_r = \begin{cases} 0, & \text{where } \max D_x(\boldsymbol{u_c}) < \theta_r, \\ \frac{\max D_x(\boldsymbol{u_c})}{\max D_x}, & \text{otherwise,} \end{cases}$$

where $\max D_x$ is the maximum vocal tract length of the speaker θ.

(5) The *single closure loss* counts the number of distinct closures N_c resulting from the consonant target using the tube area function and open tract threshold θ_o

$$L_s = \begin{cases} 0, & \text{where } N_c(\boldsymbol{u_c}) > 1, \\ 1, & \text{otherwise.} \end{cases}$$

(6) The *visual loss* is conditional on the consonant place of articulation (bilabial or not) and detects a condition where the vocal tract is open except for the lips or the vocal tract is closed except for the lips

$$L_v = \begin{cases} 0 & \text{where } \min A_x(\boldsymbol{u_c}, R)\langle \epsilon, \min A_x(\boldsymbol{u_c}, R')\rangle \theta_o \\ \frac{\min A_x(\boldsymbol{u_c}, R)}{\max A_x} & \text{otherwise} \end{cases}$$

where R is the set of articulators $\{lips, incisors\}$ and R' is its complement when the consonant is a bilabial and vice versa when the consonant is not a bilabial.

(7) Lastly, the *coarticulation loss* is the normalised L_1-distance between the range-normalised upper vocal tract vectors ($\tilde{\boldsymbol{u}}_c$ and $\tilde{\boldsymbol{u}}_v$)

$$L_d = N^{-1} \|\tilde{\boldsymbol{u}}_c - \tilde{\boldsymbol{u}}_v\|_1, \tilde{u}_{c_i} \text{ and } \tilde{u}_{v_i} \in [0, 1] \tag{A.1}$$

with $N = 16$ the length of the vectors $\tilde{\boldsymbol{u}}_c$ and $\tilde{\boldsymbol{u}}_v$ (from the free parameters in common, Table 1).

Appendix B. Interpreting recognition rates

The expected recognition rate for a random utterance generator cannot be determined directly for the free transcription or large vocabulary recognition task because neither the effective number of output classes of the discriminator (whether the proprietary ASR system or human listeners) nor the prior probability distribution over the set of possible outputs (i.e., the language model) is known. This obstacle extends to the recognition rates in terms of the vowel or consonant that are subject to a decision by the discriminator on the word level which involves both the prior probabilities of words as well as the phonotactics of the language (i.e., the context-dependent acoustic model).

However, a reasonable estimate of the upper bound of the expected recognition rate for a random generator would be to take a uniform distribution over the number of output classes observed for the specific experimental condition. For example, we can estimate this for the best performing condition for the ASR and human evaluators from the information presented in Figs. 7 and 8 respectively. For the ASR system, the set of output labels has size 23 resulting in an expected recognition rate of $1/23 \approx 4.4\%$. For the human listeners, if we consider the 38 frequent responses representing approximately 70% of the probability mass, the result is $0.7/38 \approx 1.8\%$.

These estimates should be interpreted as upper bounds since for a process with lower precision, such as random sampling of the articulatory space, the following is expected: (1) Higher variance in the outputs should result in the discriminator producing a larger set of output classes which under the assumption of a uniform distribution will decrease the expected recognition rate. (2) Acknowledging that the prior probabilities modelled by the discriminator are not uniformly distributed, it is expected that more of the probability mass will be distributed outside of the 13 test classes. Concretely, it is expected that a random utterance generator will produce a larger proportion of rejected utterances, at least in the case of the ASR system where the null result is the most probable in the posterior distribution.

Furthermore, since these estimates are derived from the most precise (lowest variance) condition, similar estimates for the other conditions presented in the study are expected to be lower.

[10] Using the C++ function `vtlGetTransferFunction`.

[11] Using the function `vtlTractToTube`.

[12] Also using `vtlTractToTube`.

References

Abramson, A.S., 1977. Laryngeal timing in consonant distinctions. Phonetica 34 (4), 295–303.

Adriaans, F., 2018. Effects of consonantal context on the learnability of vowel categories from infant-directed speech. J. Acoust. Soc. Am. 144 (1), EL20–EL25.

Bailly, G., 1997. Learning to speak. Sensori-motor control of speech movements. Speech Commun. 22 (2), 251–267.

Barnaud, M.-L., Schwartz, J.-L., Bessière, P., Diard, J., 2019. Computer simulations of coupled idiosyncrasies in speech perception and speech production with COSMO, a perceptuo-motor Bayesian model of speech communication. PLOS ONE 14 (1), e0210302.

Bergstra, J., Bardenet, R., Bengio, Y., Kégl, B., 2011. Algorithms for hyper-parameter optimization. In: Proceedings of the 24th International Conference on Neural Information Processing Systems, Red Hook, NY, USA, pp. 2546–2554.

Birkholz, P., 2005. 3D-Artikulatorische Sprachsynthese. Logos Verlag, Berlin.

Birkholz, P., 2007. Control of an articulatory speech synthesizer based on dynamic approximation of spatial articulatory targets. In: Proc. Interspeech, Antwerp, Belgium, pp. 2865–2868.

Birkholz, P., 2013. Modeling consonant-vowel coarticulation for articulatory speech synthesis. PLoS ONE 8 (4).

Birkholz, P., 2014. Enhanced area functions for noise source modeling in the vocal tract. In: International Seminar on Speech Production (ISSP 2014), Cologne, Germany, pp. 37–40.

Birkholz, P., Drechsel, S., Stone, S., 2019. Perceptual optimization of an enhanced geometric vocal fold model for articulatory speech synthesis. In: Proc. Interspeech, Graz, Austria, pp. 3765–3769.

Birkholz, P., Kröger, B.J., Neuschaefer-Rube, C., 2011. Model-based reproduction of articulatory trajectories for consonant–Vowel sequences. IEEE Trans. Audio Speech Lang. Process. 19 (5), 1422–1433.

Birkholz, P., Schmager, P., Xu, Y., 2018. Estimation of pitch targets from speech signals by joint regularized optimization. In: Proc. European Signal Processing Conference (EUSIPCO), Rome, Italy, pp. 2075–2079.

Brass, M., Heyes, C., 2005. Imitation: is cognitive neuroscience solving the correspondence problem? Trends in Cognitive Sciences 9 (10), 489–495.

Carnegie Mellon University, 2000. The CMU pronunciation dictionary. http://www.speech.cs.cmu.edu/.

Casile, A., Caggiano, V., Ferrari, P.F., 2011. The mirror neuron system: A fresh view. The Neuroscientist 17 (5), 524–538.

Davis, B.L., MacNeilage, P.F., 1995. The articulatory basis of babbling. J. Speech Lang. Hearing Res. 38 (6), 1199–1211.

Dupoux, E., 2018. Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. Cognition 173, 43–59, publisher: Elsevier.

Frank, S., Feldman, N., Goldwater, S., 2014. Weak semantic context helps phonetic learning in a model of infant language acquisition. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 1, Baltimore, MD, USA, pp. 1073–1083.

Harnad, S., 1990. The symbol grounding problem. Physica D 42 (1), 335–346.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780.

Houde, J.F., Jordan, M.I., 1998. Sensorimotor adaptation in speech production. Science 279 (5354), 1213–1216.

Howard, I.S., Huckvale, M.A., 2005. Training a vocal tract synthesizer to imitate speech using distal supervised learning. In: International Conference on Speech and Computer (SpeCom), Patras, Greece, pp. 159–162.

Howard, I.S., Messum, P.R., 2007. A computational model of infant speech development. In: XII International Conference Speech and Computer (SPECOM'2007), Moscow, Russia, pp. 756–765.

Jordan, M.I., Rumelhart, D.E., 1992. Forward models: Supervised learning with a distal teacher. Cogn. Sci. 16 (3), 307–354.

Jusczyk, P.W., 1997. The Discovery of Spoken Language. MIT Press, Cambridge, MA, USA.

Kröger, B.J., Kannampuzha, J., Kaufmann, E., 2014. Associative learning and self-organization as basic principles for simulating speech acquisition, speech production, and speech perception. EPJ Nonlinear Biomed. Phys. 2 (1), 1–28.

Kröger, B.J., Kannampuzha, J., Neuschaefer-Rube, C., 2009. Towards a neurocomputational model of speech production and perception. Speech Commun. 51 (9), 793–809.

Krug, P.K., Birkholz, P., Gerazov, B., Van Niekerk, D.R., Xu, A., Xu, Y., 2022. Efficient exploration of articulatory dimensions. Studientexte Zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2022, 51–58.

Kuhl, P.K., 2004. Early language acquisition: Cracking the speech code. Nat. Rev. Neurosci. 5 (11), 831–843.

Liu, Z., Xu, Y., F.-f. Hsieh, 2022. Coarticulation as synchronised CV co-onset – Parallel evidence from articulation and acoustics. J. Phonetics 90, 101116.

McFee, B., Raffel, C., Liang, D., Ellis, D.P.W., McVicar, M., Battenberg, E., Nieto, O., 2015. librosa: Audio and music signal analysis in python. In: Proc. Python in Science Conference (SciPy), Austin, Texas, USA, pp. 18–24.

Messum, P., Howard, I.S., 2015. Creating the cognitive form of phonological units: The speech sound correspondence problem in infancy could be solved by mirrored vocal interactions rather than by imitation. J. Phonetics 53, 125–140.

Mills, A.E., 1988. Visual handicap. In: D. Bishop, K. Mogford (Ed.), Language Development in Exceptional Circumstances. Longman, pp. 150–164, Ch. 9.

Milne, A.E., Bianco, R., Poole, K.C., Zhao, S., Oxenham, A.J., Billig, A.J., Chait, M., 2020. An online headphone screening test based on dichotic pitch. Behav. Res. Methods.

Moulin-Frier, C., Nguyen, S.M., Oudeyer, P.-Y., 2014. Self-organization of early vocal development in infants and machines: the role of intrinsic motivation. Front. Psychol. 4.

Murakami, M., Kröger, B., Birkholz, P., Triesch, J., 2015. Seeing [u] aids vocal learning: Babbling and imitation of vowels using a 3D vocal tract model, reinforcement learning, and reservoir computing. In: International Conference on Development and Learning and on Epigenetic Robotics (ICDL-EpiRob), Providence, Rhode Island, USA, pp. 208–213.

Nam, H., Goldstein, L.M., Giulivi, S., Levitt, A.G., Whalen, D.H., 2013. Computational simulation of CV combination preferences in babbling. J. Phonetics 41 (2), 63–77.

Nasir, S.M., Ostry, D.J., 2006. Somatosensory precision in speech production. Curr. Biol. 16 (19), 1918–1923.

Oller, D.K., Eilers, R.E., 1988. The role of audition in infant babbling. Child Dev. 59 (2), 441–449.

Oohashi, H., Watanabe, H., Taga, G., 2013. Development of a serial order in speech constrained by articulatory coordination. PLOS ONE 8 (11), 1–10.

Panayotov, V., Chen, G., Povey, D., Khudanpur, S., 2015. Librispeech: An ASR corpus based on public domain audio books. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5206–5210.

Parrell, B., Ramanarayanan, V., Nagarajan, S., Houde, J., 2019. The FACTS model of speech motor control: Fusing state estimation and task-based control. PLoS Comput. Biol. 15 (9), e1007321.

Philippsen, A., 2021. Goal-Directed Exploration for Learning Vowels and Syllables: A Computational Model of Speech Acquisition. KI - Künstliche Intelligenz.

Philippsen, A.K., Reinhart, R.F., Wrede, B., 2014. Learning how to speak: Imitation-based refinement of syllable production in an articulatory-acoustic model. In: International Conference on Development and Learning and on Epigenetic Robotics (ICDL-EpiRob), Genoa, Italy, pp. 195–200.

Plummer, A.R., 2012. Aligning manifolds to model the earliest phonological abstraction in infant-caretaker vocal imitation. In: Proc. Interspeech 2012, Portland, OR, USA, pp. 2482–2485.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlíček, P., Qian, Y., Schwarz, P., Silovský, J., Stemmer, G., Veselý, K., 2011. The Kaldi speech recognition toolkit. In: IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, Hawaii, USA.

Räsänen, O., 2012. Computational modeling of phonetic and lexical learning in early language acquisition: Existing models and future directions. Speech Commun. 54 (9), 975–997.

Räsänen, O., Doyle, G., Frank, M.C., 2018. Pre-linguistic segmentation of speech into syllable-like units. Cognition 171, 130–150.

Rasilo, H., Räsänen, O., 2017. An online model for vowel imitation learning. Speech Commun. 86, 1–23.

Rasilo, H., Räsänen, O., Laine, U.K., 2013. Feedback and imitation by a caregiver guides a virtual infant to learn native phonemes and the skill of speech inversion. Speech Commun. 55 (9), 909–931.

Saltzman, E.L., Munhall, K.G., 1989. A dynamical approach to gestural patterning in speech production. Ecol. Psychol. 1 (4), 333–382.

Schatz, T., Feldman, N.H., Goldwater, S., Cao, X.-N., Dupoux, E., 2021. Early phonetic learning without phonetic categories: Insights from large-scale simulations on realistic input. Proc. Natl. Acad. Sci. 118 (7).

Serkhane, J.E., Schwartz, J.L., Boë, L.J., Davis, B.L., Matyear, C.L., 2007. Infants' vocalizations analyzed with an articulatory model: A preliminary report. J. Phonetics 35 (3), 321–340.

Tourville, J.A., Guenther, F.H., 2011. The DIVA model: A neural theory of speech acquisition and production. Lang. Cogn. Process. 26 (7), 952–981.

Turk, A., Shattuck-Hufnagel, S., 2020. Timing evidence for symbolic phonological representations and phonology-extrinsic timing in speech production. Front. Psychol. 10, 2952.

Van Niekerk, D.R., Xu, A., Gerazov, B., Krug, P.K., Birkholz, P., Xu, Y., 2020. Finding intelligible consonant-vowel sounds using high-quality articulatory synthesis. In: Proc. Interspeech 2020, Shanghai, China, pp. 4457–4461.

Van Niekerk, D.R., Xu, A., Gerazov, B., Krug, P.K., Birkholz, P., Xu, Y., 2022. Exploration strategies for articulatory synthesis of complex syllable onsets. In: Proc. Interspeech 2022, Incheon, South Korea, pp. 635–639.

Wakita, H., 1977. Normalization of vowels by vocal-tract length and its application to vowel identification. IEEE Trans. Acoust. Speech Signal Process. 25 (2), 183–192.

Xu, Y., 2020. Syllable is a synchronization mechanism that makes human speech possible. PsyArXiv.

Xu, A., Birkholz, P., Xu, Y., 2019. Coarticulation as synchronized dimension-specific sequential target approximation: An articulatory synthesis simulation. In: Proceedings of the International Congress of Phonetic Sciences (ICPhS), Melbourne, Australia, pp. 205–209.

Xu, Y., Wang, Q.E., 2001. Pitch targets and their realization: Evidence from Mandarin Chinese. Speech Commun. 33 (4), 319–337.

Zhan, P., Waibel, A., 1997. Vocal tract length normalization for large vocabulary continuous speech recognition. Technical Report CMU-LTI-97-150, Carnegie Mellon University, Language Technologies Institute, Pittsburgh, PA, USA.