Journal of Phonetics 105 (2024) 101338

Contents lists available at ScienceDirect

Journal of Phonetics

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

Artificial vocal learning guided by speech recognition: What it may tell us about how children learn to speak

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

[a] School of Humanities and Social Sciences, Harbin Institute of Technology, Shenzhen 518055, China
[b] Department of Speech, Hearing and Phonetic Sciences, University College London, London WC1E 6BT, United Kingdom
[c] Faculty of Electrical Engineering and Information Technologies, Ss Cyril and Methodius University in Skopje, Skopje 1000, RN, Macedonia
[d] Institute of Acoustics and Speech Communication, Technische Universität Dresden, Dresden 01062, Germany
[e] Computer Engineering Department, King Mongkut's University of Technology Thonburi, Bangkok 10140, Thailand
[f] Cognition and Brain Sciences Unit, University of Cambridge, Cambridge CB2 1TN, United Kingdom

ARTICLE INFO

Article history:
Received 29 June 2023
Received in revised form 5 May 2024
Accepted 16 May 2024
Available online xxxx

Keywords:
Computational modelling of vocal learning
Phonological perception
Coarticulation
Speech acquisition
Articulatory synthesis

ABSTRACT

It has long been a mystery how children learn to speak without formal instructions. Previous research has used computational modelling to help solve the mystery by simulating vocal learning with direct imitation or caregiver feedback, but has encountered difficulty in overcoming the speaker normalisation problem, namely, discrepancies between children's vocalisations and that of adults due to age-related anatomical differences. Here we show that vocal learning can be successfully simulated via recognition-guided vocal exploration without explicit speaker normalisation. We trained an articulatory synthesiser with three-dimensional vocal tract models of an adult and two child configurations of different ages to learn monosyllabic English words consisting of CVC syllables, based on coarticulatory dynamics and two kinds of auditory feedback: (i) acoustic features to simulate universal phonetic perception (or direct imitation), and (ii) a deep-learning-based speech recogniser to simulate native-language phonological perception. Native listeners were invited to evaluate the learned synthetic speech with natural speech as baseline reference. Results show that the English words trained with the speech recogniser were more intelligible than those trained with acoustic features, sometimes close to natural speech. The successful simulation of vocal learning in this study suggests that a combination of coarticulatory dynamics and native-language phonological perception may be critical also for real-life vocal production learning.

1. Introduction

Speech is a highly complex cognitive activity that requires sophisticated control of multiple articulators including the tongue, the lips, the jaw and the larynx. The motor skills involved in speaking are intricate and hard to learn, as indicated by the difficulty adults experience when learning a second language. Yet, children learn to speak despite their inability to receive explicit instructions or see many of the relevant speech articulators. They seem to mainly learn from what they hear, but it is still a mystery how they manage to overcome many seemingly insurmountable barriers. One of the most obvious is that their vocal tracts are much shorter and smaller compared to adult's (Fitch & Giedd, 1999). In fact, an infant vocal tract more closely resembles that of a non-human primate (Lieberman et al., 1972). This makes children's formants consistently higher than those of adults (Vorperian & Kent, 2007). This is known as the speaker normalisation problem (Johnson, 2005) for speech perception or the correspondence problem for sensorimotor learning (Nehaniv & Dautenhahn, 2002; Brass & Heyes, 2005). Our prior research, however, has already shown that speaker normalisation may have been an exaggerated problem both for tone and intonation (Xu & Prom-on, 2014; Chen et al., 2022) and for segments (Prom-on, Birkholz, & Xu, 2013). One of the aims of this study is therefore to test if vocal learning can be computationally simulated without explicit speaker normalisation.

* Corresponding author.
E-mail address: yi.xu@ucl.ac.uk (Y. Xu).

https://doi.org/10.1016/j.wocn.2024.101338
0095-4470/

1.1. Sensorimotor learning in humans and other animals

There have been some clues from research on songbirds and certain mammals that are also vocal learners. Their vocal development first undergoes a phase of accumulating auditory experience and then a later phase of vocal practice. During the early sensory phase of perception learning, the learners gain experience of species-specific signals, as observed in songbirds (Konishi, 1965) and humans (Kuhl, 2000). Then there is a phase of vocal practice, during which the vocal systems are calibrated to convert motor commands to sound production, as found in humans (Oller, 1980), songbirds (Thorpe, 1954), marmosets (Elowson et al., 1998b) and bats (Fernandez et al., 2021). For both phases of learning, auditory feedback may play two vital roles: (i) to store auditory experience in long-term memory (Phan et al., 2006), and (ii) to detect production errors during vocal practice (Konishi, 1965). Indeed, there is evidence that for both songbirds (Konishi, 1965) and humans (Oller & Eilers, 1988), a lack of auditory input can lead to severe impairment in their vocal development. It has been further suggested that in humans, production learning can be driven by perception experience via vocal mimicry (Kuhl, 2000). Assuming this is the case, infants probably attempt to match their own vocalisations to the auditory memory of previously heard speech sounds. But it would still be unclear, however, what is the nature of the perceptual representation that guides vocal learning in humans.

The role of perception in vocal learning can be seen in light of learning sensorimotor associations that map articulatory movements to sensory goals. Many studies have demonstrated evidence of sensorimotor coupling (Kuhl et al., 2014; Bruderer et al., 2015; Choi et al., 2021; Fadiga et al., 2002). Several learning mechanisms such as error-based learning and reinforcement learning have been proposed (Wolpert et al., 2011) to account for such coupling. It has also been suggested that associative learning of correlated sensorimotor experience, such as self-observation of actions, synchronous actions and being imitated by social partners, forges the linkage between the motor and sensory systems (Heyes, 2001; Keysers & Perrett, 2004; Cook et al., 2014). Studies on infant sensorimotor learning such as crawling (van Elk et al., 2008) and stepping (de Klerk et al., 2015) have shown support for this assumption, but it has not been widely used to account for speech motor learning, due to the lack of visual cues. Despite some theoretical proposals and behavioural studies, we know remarkably little about the relative contribution of different kinds of sensorimotor experience to vocal learning.

Therefore, although there have been extensive observations and theoretical proposals on vocal learning and sensorimotor learning, the current picture is still blurry, especially in terms of the learning mechanisms involved. What is needed is a way of probing how exactly vocal learning operates rather than merely observing its various characteristics. This can be done through computational simulation of the learning process, which would allow the examination of various hypothetical components and mechanisms by comparing the end products of learning, namely, the learned speech utterances.

1.2. Past models of vocal learning

There have been many computational studies of vocal learning, and a summary chart is shown in Table A1 of Appendix. Several of those studies have probed the neural and cognitive mechanisms at play by modelling the brain network. One of the earliest and the best-known is the DIVA model (Guenther, 1994; Tourville & Guenther, 2011), a neurobiologically motivated framework that simulates the acquisition of sensorimotor interactions based on babbling. It consists of two main components: (a) a feedforward control system that encodes the movement velocities of the articulators, and (b) a feedback control system that encodes the time-varying sensory expectations. The model learns speech by finding appropriate synaptic weights for mapping the phonetic-to-orosensory space and orosensory-to-articulatory space. As a complement to the DIVA model, neurocomputational models (Kröger et al., 2009, 2014) have been proposed to establish a mapping between speech phonetics and sensory signals via self-organising maps (Kohonen, 2001). Similar attempts have been made to simulate the learning of vowels by Hebbian connections, i.e., the correlation of sensory system and motor map through activation of units in both receptive fields. (Westerman & Miranda, 2002, 2004; Heintz et al., 2009). Although these neurobiological models have tried to simulate the neural processes of speech production and perception, they have not demonstrated an ability to generate intelligible speech.

As vocal mimicry has long been regarded as a crucial mechanism for speech acquisition (Kuhl & Meltzoff, 1996), a great deal of research has been carried out to simulate vocal imitation by the distal learning framework. Distal learning describes how a dynamic system can learn actions to generate desired outcomes when supervised by a distal 'teacher' (Jordan & Rumelhart, 1992). HABLAR proposed by Bailly (1997) is perhaps the earliest model that aims to achieve audio-visual-to-articulatory inversion under this framework. Later, Howard and Huckvale (2005) have also built an inverse model between speech acoustics and speech motor control trained by direct mapping, bypassing the utilisation of articulatory data. More recently, Prom-on et al. (2014a, 2014b) successfully trained VocalTractLab, a 3D articulatory synthesiser (Birkholz, 2013) to learn Thai vowels with formant values (F1, F2 and F3) close to natural speech, judged as highly intelligible in a listening experiment by native speakers. Instead of speech acoustics, Philippsen et al. (2016) developed a sensory goal space of vectors derived from acoustic features in order to train the motor space. Nevertheless, the learned speech was limited to vowels and simple CV sequences including /ma/ and /ba/ without intelligibility assessment.

The speaker normalisation problem (i.e., the correspondence problem) has long been considered as a hard or even insurmountable barrier for vocal learners, which has led to a line of research into caregiver feedback as playing a critical role (see Asada, 2016 for a review). Huckvale, Howard and the others have built a virtual infant KLAIR which relies on caregiver's reformulation to reinforce the acquisition of speech (Huckvale et al., 2009; Huckvale, 2011a, 2011b). Following the

same principle, Howard and Messum also proposed an interactive learning model, named Elija (Howard & Messum, 2007; Howard, 2011; Howard & Messum, 2014; Messum & Howard, 2015), which is aimed at establishing a correspondence between his/her own vocal action and adult speech by caregiver's judgement. Other research groups have pursued similar caregiver-infant interaction models (Yoshikawa, Asada, et al., 2003; Yoshikawa, Koga, et al., 2003; Miura et al., 2007; Ishihara et al., 2009; Rasilo et al., 2013; Rasilo & Räsänen, 2017). Judging from the samples of the learned speech generated, however, shifting the burden to the caregiver in these approaches is again not effective in solving the speaker normalisation problem, as far as consonant acquisition is concerned.

Beside the models of infant–caregiver interaction that use reinforcement learning based on extrinsic social rewards, a number of studies have explored the possibility of using intrinsic reinforcement. Warlaumont and her colleagues have proposed a model that produces spontaneous speech with a self-organising map that controls the muscles of a speech synthesiser (Warlaumont et al., 2013; Warlaumont & Finnegan, 2016). The reinforcement signal comes from the auditory salience of these randomly generated sounds. Murakami et al. (2015) has incorporated reinforcement learning with vocal imitation, whereby the agent adjusts motor parameters in an iterative manner to maximise reward signals. The study shows that the supplementary visual reinforcement signal is advantageous in acquiring rounded vowels. Again, however, the scopes of these simulations are restricted to vowel acquisition.

There are also models that focus on how children discover phonological systems. Oudeyer (2005) has proposed a self-organisation model for speech acquisition, which is able to discover vowel inventories based on its own subsystems, without social interactions. A follow-up study compared models of random exploration, random goal reaching and active goal reaching (Moulin-Frier & Oudeyer, 2012). It was found that active learning led to continuing exploration of auditory and acoustic spaces (Moulin-Frier et al., 2014). Recently, the same research group proposed a model for the emergence of phonological systems, referred to as 'Communicating about Objects using Sensory–Motor Operations' (COSMO) (Moulin-Frier et al., 2015; Barnaud et al., 2019). Using a Bayesian modelling approach, motor and auditory systems were linked through linguistic objects to develop a mapping between articulation and acoustics.

Overall, none of the existing simulations has demonstrated successful learning of intelligible words consisting of CVC syllables. Thus, the speaker normalisation problem remains unsolved. As will be reviewed next, the bottleneck may lie in two critical aspects of the speech system: sensory feedback and articulatory dynamics.

1.3. Simulation of speech sensory and motor systems

With regard to auditory feedback[1] (Appendix, Table A1 Sensory control), a vast majority of the studies have used formants (Bailly, 1997; Westermann & Miranda, 2004; Howard & Huckvale, 2005; Miura et al., 2007; Heintz et al., 2009; Ishihara et al., 2009; Rasilo & Räsänen, 2017; Forestier & Oudeyer, 2017; Acevedo-Valle et al., 2020) or normalised formants (de Boer, 2000; Oudeyer, 2005; Kröger et al., 2009; Moulin-Frier & Oudeyer, 2012; Warlaumont, 2012; Warlaumont et al., 2013; Moulin-Frier et al., 2014, 2015; Warlaumont & Finnegan, 2016; Barnaud et al., 2019), while other studies have incorporated more acoustic details in terms of Bark-scaled (Kröger et al., 2014) or gammatone spectrograms (Howard & Messum, 2014; Messum & Howard, 2015). Several studies have attempted to use Mel-frequency cepstral coefficients (MFCCs) (Davis & Mermelstein, 1980) to represent auditory feedback (Kanda et al., 2009; Rasilo et al., 2013; Philippsen et al., 2014; Prom-on et al., 2014a, 2014b; Najnin & Banerjee, 2017), which is the most popular parametric acoustic representation in speech synthesis and recognition (Barry & van Dommelen, 2005). More recently, Plummer et al. (2010) established a mapping between children's vowel production and that of adult productions through manifold learning. In this way, a perceptual space related to language information was simulated.

Only a few computational studies have taken somatosensory information (Guenther, 1994; Kröger et al., 2009; Tourville & Guenther, 2011; Kröger et al., 2014; Acevedo-Valle et al., 2020) into consideration despite its reported influence on speech production in many behaviour studies (Niemi et al., 2006; Ménard et al., 2023). To date, there have been no examinations of how auditory and somatosensory feedback may impact vocal learning through modelling simulations, including the possibility of language-specific perception as a relevant form of auditory feedback.

As for articulatory dynamics (Appendix, Table A1 Motor control & synthesiser), previous models have adopted various articulatory synthesis systems without principled dynamic controls, including the source-filter model (Yoshikawa, Asada, et al., 2003; Yoshikawa, Koga, et al., 2003; Miura et al., 2007), the pipe model (Westerman & Miranda, 2002; Westermann & Miranda, 2004), Praat synthesis (Warlaumont, 2012; Warlaumont et al., 2013; Warlaumont & Finnegan, 2016), the Maeda synthesiser (de Boer, 2000; Kanda et al., 2009) or its modified versions (Guenther et al., 2006; Howard & Messum, 2007; Heintz et al., 2009; Tourville & Guenther, 2011; Moulin-Frier & Oudeyer, 2012; Howard & Messum, 2014; Moulin-Frier et al., 2014; Messum & Howard, 2015; Najnin & Banerjee, 2017; Acevedo-Valle et al., 2017, 2018, 2020; Barnaud et al., 2019). More recent studies have used VocalTractLab2.3 (vocaltractlab.de) (Birkholz, 2013), an articulatory synthesiser with high-dimensional vocal tract parameter control (Prom-on et al., 2014a, 2014b; Philippsen et al., 2014, Murakami et al., 2015; Philippsen et al., 2016). For controlling the dynamics of the synthesiser, the Task Dynamic model (Saltzman & Munhall, 1989; Fowler & Saltzman, 1993) has been adopted (Howard & Messum, 2007, Howard, 2011; Howard & Messum, 2014; Messum & Howard, 2015), which is a second-order dynamical system for generating contextually varying articulatory kinematics. Other researchers have used the Dynamic Movement Primitives (DMPs) framework (Schaal, 2006; Ijspeert et al., 2013) to control a synthesiser (Forestier & Oudeyer, 2017; Philippsen, 2021), which was developed to plan the

[1] Here and throughout the paper, feedback refers only to "offline" feedback which differs from the kind of "online" feedback as seen in perturbations studies such as Houde & Jordan (1998), Tremblay et al. (2003) and Xu et al. (2004), whereby corrective manoeuvres are enacted during the course of an articulatory movement in reaction to continuous sensory feedback.

trajectories for the motor movements of robots with discrete or rhythmic nonlinear dynamic primitives. Again, however, these vocal learning models have been mostly restricted to simulating vowel acquisition. As of now, not a single study prior to the present one has successfully simulated the learning of intelligible words consisting of CVC syllables (Appendix, Table A1 Learning target & Performance). What is missing is the direct simulation of the motor control mechanism of CV coarticulation that generates extensive variations in adjacent consonants and vowels.

1.4. Research questions and major aims

Throughout the animal kingdom, many species show vocal plasticity to a certain extent. Noticeable similarities can be seen in the vocal developmental patterns of songbirds and humans, in which a phase of auditory extraction paves the way for vocal practice (Doupe & Kuhl, 1999). These observations have naturally led to the postulation that auditory experience guides vocal learning in humans and songbirds (Kuhl, 2003). It has been found that child speech perception changes from language-universal to language-specific perception (Werker & Lalonde, 1988; Kuhl, 2004) and production learning follows certain developmental patterns of phoneme acquisition. Unlike birdsong learning, much uncertainty still exists concerning the nature of the auditory guidance for speech acquisition in humans, as the approaches of manipulating auditory feedback in animal studies would be unethical. At the same time, studies on human sensorimotor learning have proposed possible learning strategies such as error-based learning, reinforcement learning, use-dependent learning (Wolpert et al., 2011) and imitative learning (Cook et al., 2014). However, vocal learning is essentially dissimilar to learning other motor movements, because of the lack of visual information. What is yet unclear is whether these sensorimotor learning mechanisms also underlie vocal learning.

Despite extensive observations and theoretical perspectives on vocal learning and sensorimotor learning, the emerging picture is still blurry as questions remain regarding the learning mechanisms. The computational approach is constructive in delineating the underlying cognitive mechanisms because it provides a platform for the verification of different assumptions. If we can recreate the learning process by simulation, then it is possible to probe any component of particular relevance. Nevertheless, so far there has been no clear demonstration of successful learning of intelligible words (see Appendix Table A Performance). In consequence, we are unable to identify which mechanisms are at play, nor can we examine the key aspects of learning quantitatively using listening experiments. Many of these unsolved questions can be investigated if we simulate *end-to-end vocal learning* that starts from audible speech and ends with synthetic words or sentences that can be perceptually evaluated for intelligibility. The developmental stage that we are aiming to simulate is from canonical babbling to first words so that the learned speech can be assessed directly by native listeners. Such simulations would allow quantitative hypothesis testing beyond observational studies.

In this study, we developed a vocal learning model that can learn intelligible speech based on both sensory and motor control mechanisms. A demonstration of the vocal learning model and the learned speech can be found at https://gitlab.com/Anqi_Xu/evoc_learn/-/blob/main/Demo/cvc_cvcv_updated.mp4. The model includes several key innovations. First, unlike previous studies that adopted simplistic vocal tract models that controlled only a few articulators and largely neglected articulatory dynamics, we explicitly modelled the coarticulatory dynamics (as defined in Section 2.1.2) in a high-dimensional vocal tract model (van Niekerk et al., 2023). Second, we systematically examined the impact of speech sensory control on vocal learning using (i) acoustic features to simulate universal phonetic perception that detects any sound differences in any language (Werker & Lalonde, 1988; Kuhl, 2000), (ii) a deep-learning-based speech recogniser to simulate native-language phonological perception that captures key phonetic properties that distinguish words in a native language (Werker & Lalonde, 1988; Kuhl, 2000), and (iii) oral constriction sensing to simulate somatosensory feedback that indicates whether the oral cavity is open or closed (Choi et al., 2021). Thirdly, we confronted the speaker-normalisation problem by training a 3D vocal tract model to learn speech from mismatched speech of different age and sex, i.e., training an adult male and two child vocal tract models to learn from an adult female's speech samples, or to learn from a speech recogniser which is trained by multiple speakers. Finally, unlike most previous studies, we conducted systematic listening experiments (i.e., open-vocabulary dictation and multiple choice) to assess the validity of the model by making direct intelligibility comparisons with natural speech (Krug et al., 2023; van Niekerk et al., 2023). This has set a new benchmark for vocal learning simulations, and potentially for the evaluation of theories and models of vocal learning in general.

The present study is designed to test the following hypotheses based on the successes and inadequacies of previous works as reviewed above.

Hypothesis 1: End-to-end vocal learning can be simulated without explicit speaker normalisation by a combination of auditory guidance and coarticulation dynamics.
Hypothesis 2: Language-specific perception simulated by an automatic speech recogniser provides better auditory guidance than language-universal phonetic imitation simulated by acoustic matching.
Hypothesis 3: Vocal learning can be simulated also with a vocal tract with child anatomical configurations.

2. Methods

2.1. Model structure

The simulation model consists of a motor control component and a sensory component, as illustrated in Fig. 1. The stimuli used in the listening experiments and the code to reproduce the results are available at https://gitlab.com/Anqi_Xu/evoc_learn. The articulatory synthesiser is VocalTractLab (Birkholz, 2013) with a geometrical three-dimensional vocal tract model. Within each learning trial, the motor control model begins with the exploration of a full set of syllabic consonant and vowel targets within the parameter range (Fig. 1A). In other words, the model is trained to learn consonant targets followed by a specific vowel and vowel targets preceded by a

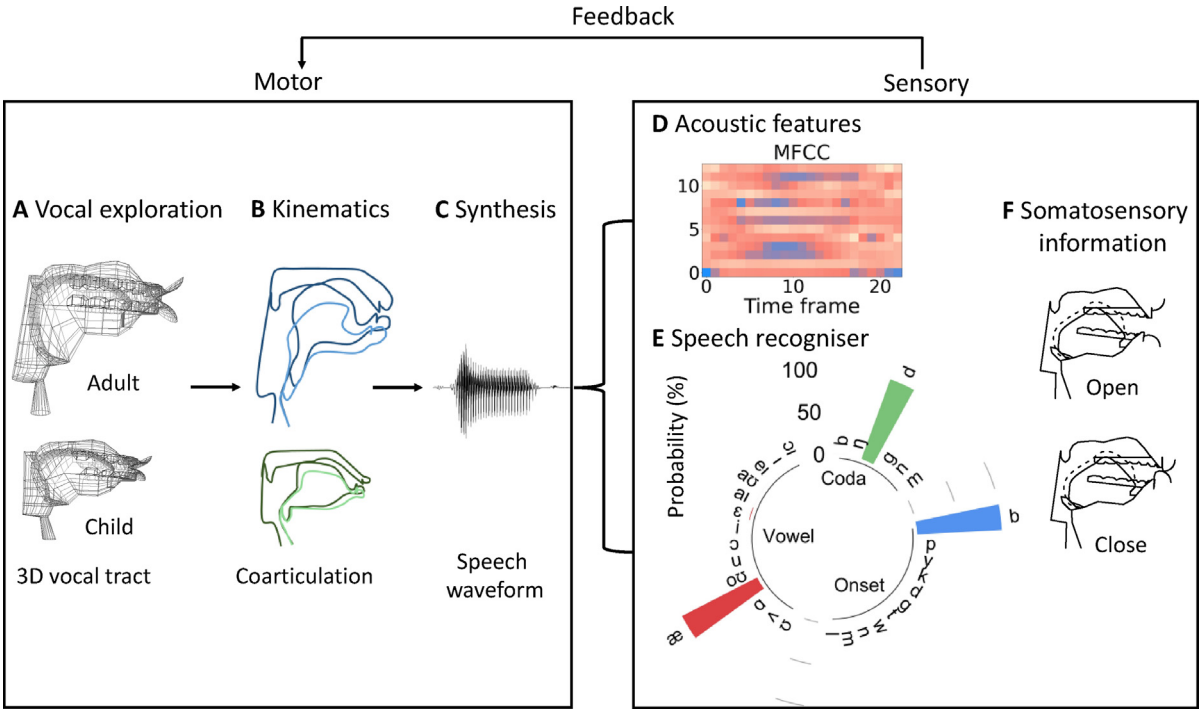


Fig. 1. Schematic overview of the steps involved each time the model tries a new set of articulatory targets (A) The adult vocal tract model is based on the MRI data of an adult male speaker and the child vocal models are scaled versions of the adult model (Birkholz & Kröger, 2007; Davis & Mermelstein, 1980; Goldstein, 1980). (B) Context-sensitive articulator kinematic movements are calculated by a synchronised dimension-specific sequential target approximation model (Liu et al., 2022; Xu, 2020). (C) Aero-acoustic simulation is based on enhanced area functions of time-varying vocal tract shapes to generate speech. (D) Mel-frequency cepstral coefficients (MFCCs) (Davis & Mermelstein, 1980) of natural and synthetic words were extracted. (E) A speech recogniser evaluates the probability of the targeted onset consonant, vowel and coda consonant, as represented by International Phonetic Alphabet (IPA) symbols. (F) Somatosensory information for limiting vocal exploration, provided by the cross-sectional areas to determine whether there is a closure in the vocal tract.

specific consonant (i.e., C_V and V_C hereafter). The kinematic trajectories that approach the dimensional targets and their timing are based on a coarticulation model (Xu, 2020; Liu et al., 2022), which simulates the realisation of consonants and vowels in a syllable frame (Fig. 1B). The time-varying vocal tract shapes are then converted to cross-sectional area functions for the aerodynamic-acoustic simulation (Fig. 1C). The synthetic speech is evaluated either by acoustic features (Fig. 1D), or by a speech recogniser (Fig. 1E), which is a pre-trained deep learning model that maps acoustic features to a contrastive auditory space. In addition, somatosensory feedback is simulated based on the openness of the vocal tract (Fig. 1F). We trained the adult and the child vocal tract systems to learn the production of English words, guided by the sensory feedback options in Fig. 1D–F.

2.1.1. Articulatory synthesis

The articulatory synthesiser (Fig. 1A) calculates enhanced area functions (Birkholz, 2014) for aerodynamic-acoustic simulations. The adult vocal tract model is adapted from MRI data of a German male speaker and the 1-year-old and 3-year-old boy's vocal tract models are scaled from the adult model based on relative anatomy (Birkholz & Kröger, 2007). Instead of simple linear scaling, the structural modification was based on Goldstein's (1980) cephalometric analysis of the head and the neck from birth to the age of 20. To be compatible with the vocal tract dimension of the articulatory synthesiser, we additionally transformed the craniofacial measurements that were fitted to a growth curve as a function of age and sex (Birkholz & Kröger, 2007).

Table 1
Dimensional vocal tract parameters optimised in the vocal learning model. The location of the tongue body is jointly determined by the tongue body centre position and tongue side positions.

Parameter	Description
HX, HY	Horizontal and vertical hyoid positions
JX, JA	Jaw position and Jaw angle
LP, LD	Lip protrusion and vertical lip distance
VS, VO	Velum shape and velum opening
TTX, TTY	Horizontal and vertical tongue tip positions
TBX, TBY	Horizontal and vertical tongue blade positions
TCX, TCY	Horizontal and vertical tongue body center positions
TS1 – TS3	Tongue side elevation from the posterior to the anterior part of the tongue

The 17 vocal tract parameters define the airway from the glottis to the lips (Table 1). The vocal tract parameters were sampled at 5 ms intervals to ensure precision of articulatory movements. The geometric glottis model accounts for source-filter interaction during synthesis. The vocal folds were set to be fully adducted with moderate longitudinal tension for the $_C$V targets, while the glottis parameters of the C_V target including the distance between vocal cords, glottal gap area and relative amplitude were optimised during the simulated learning. The cross-sectional area function of the vocal tract was converted to a transmission-line model for the aerodynamic-acoustic simulation in the time domain. The pitch contours of the synthetic words were generated using pitch targets learned from the natural speech recordings by PENTAtrainer, an intonation modelling tool (Xu & Prom-on, 2014). The modelling tool automatically optimised the pitch target

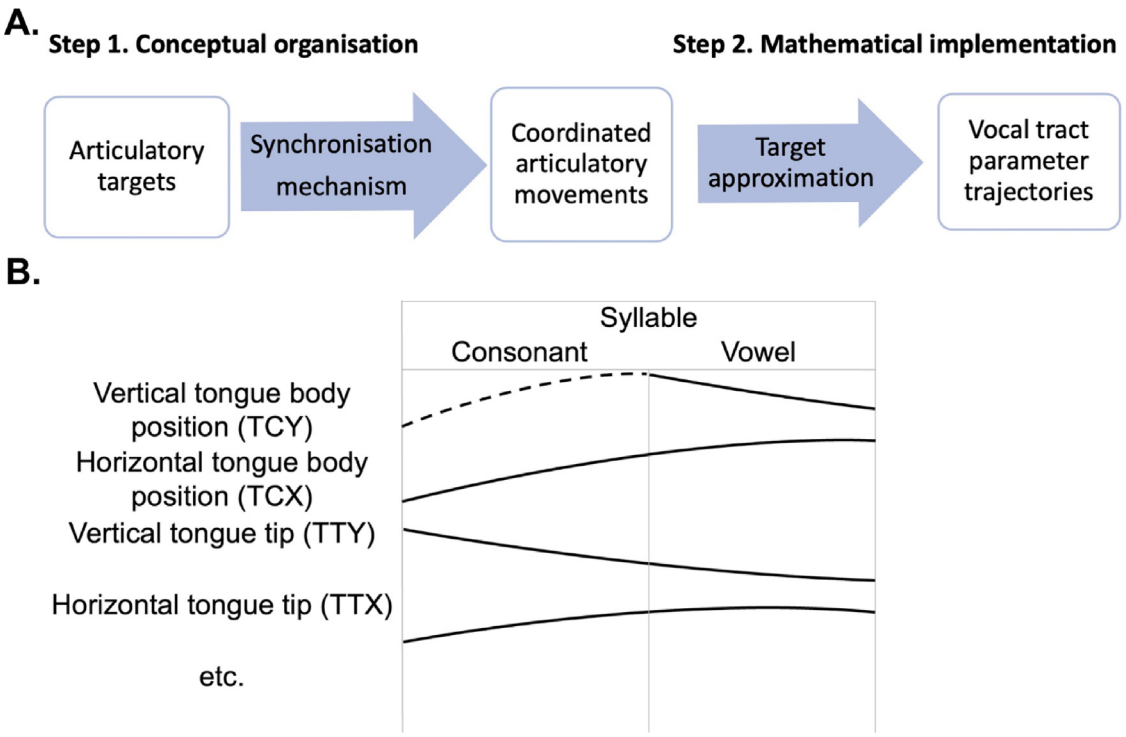


Fig. 2. Illustration of the motor control system. (A) Workflow for controlling the articulatory parameters. (B) Vocal tract parameter trajectories calculated based on the synchronised dimension-specific sequential target approximation (SDSSTA) in the case of velar stop-vowel sequences. The dashed line represents the articulatory trajectory of the C_V target and the solid lines represent the articulatory trajectories of the $_CV$ target.

parameters based on the pitch contours of natural utterances. This was done because monotonous intonation may affect listeners' judgments of segmental quality (Terken & Lemeer, 1988), but we chose to use synthetic f_0 contours over natural ones to make sure that the suprasegmental aspect of the synthetic syllables also came from stochastic-optimisation-based vocal learning algorithms.[2] The audio files were synthesised with a sampling rate of 44.1 kHz and a quantization of 16 bit.

2.1.2. Coarticulatory dynamics of CV syllables

A key assumption of the current simulation is that the learning of articulatory targets starts from the onset of canonical babbling, which we postulate is already driven by the core property of coarticulation dynamics, namely, synchronisation of consonantal, vocalic and laryngeal movements (Xu & Liu, 2006; Xu, 2020). In other words, by the time children start to babble canonically by uttering random sequences like "ba ba ba. . .", they are already synchronising their consonantal and vocalic onsets, and the remaining task is only to learn specific articulatory targets of individual segments. These coarticulatory dynamics are what controls the temporal and spatial movements of the articulators (Fig. 1B), which are simulated by a motor control system that transforms articulatory targets to vocal tract parameter trajectories, as illustrated in Fig. 2A. The relative timing of multiple articulators is controlled by a theoretical coarticulation mechanism—synchronised dimension-specific sequential target approximation (SDSSTA) (Xu, 2020; Liu et al., 2022). SDSSTA posits that the syllable is a mechanism of synchronising the onset of C and V articulation at the beginning of the syllable, and that the ensuing CV co-production is achieved by allowing a specific articulator dimension, e.g., TCY, to approach the C and V targets in succession, while the other dimension of the same articulator, e.g., TCX, approaches the V target from the syllable onset, as illustrated in Fig. 2B. Such separation of articulator dimensions thus avoids the need for gestural blending as a core coarticulation mechanism (Saltzman & Munhall, 1989). After the synchronisation mechanism was applied (step 1 in Fig. 2A), the dynamic trajectories of the 19 vocal tract parameters were then calculated by the target approximation model (step 2 in Fig. 2A). The dynamic trajectories were then passed to the articulatory synthesis to generate synthetic sounds (Fig. 1B and C).

After the articulatory movements toward the syllable-initial consonant and vowel are terminated, all the articulator dimensions begin to approach the next set of articulatory targets in the same manner. The final coda consonant was implemented as another hypothetical CV syllable in CVC words (Xu, 2020). To be more specific, the C_V target ensured a closure in the oral cavity and the $_CV$ target allowed the consonant to be released. The temporal domain of the motor control system is based on the time alignment of the articulatory targets in the natural speech. The articulatory dimensions governed by C_V targets are as follows: LD, JX, JA for bilabial stops; TTY, TBY, TS3, JX, JA for alveolar stops; and TCY, TS2, JX, JA for velar stops.

Next, after the coarticulation model was applied, the dynamic trajectories of the 19 vocal tract and glottal parameters were calculated by the target approximation model. Quantitatively, each articulatory target is defined by its geometrical position, slope (set to zero in the simulation) and strength (i.e., the time constant). The movement of the vocal tract

[2] The PENTAtrainer model was also a learning model, as it extracts the pitch targets from raw f_0 contours via stochastic optimisation (simulated annealing), much like what is done in the current study. Its development was partly what inspired the current study. Because pitch target learning is not what is being tested here, no further details are presented here.

parameters is modelled by a cascade of several identical first-order linear systems with the following transfer function:

$$H(s) = \frac{Y(s)}{X(s)} = \frac{1}{(1+s\tau)^N}$$

where s and N denote the complex frequency and the order of the system respectively. τ denotes the time constant, which determines how quickly the target is approached, hence the (inverse of the) strength of target approximation. Here, N equals 5, that is, we use a fifth-order system that reproduces s-shaped asymptotic movement towards articulatory targets with bell-shaped velocity profiles. The time-domain representation of the aforementioned equation can be derived using inverse Laplace Transform, which results in

$$y(t) = \left(c_0 + c_1 t + \cdots + c_{N-1} t^{N-1}\right) e^{\frac{-t}{\tau}} + x(t)$$

where $x(t) = b$ is the position of the articulator target (neglecting here the slope of the target) and t is the time from the beginning of the target interval. The coefficients are calculated based on the initial state of y and its derivatives of the articulator at the onset of the interval (which is equal to the final state of the previous target approximation movement), as shown in the following equation (Birkholz et al., 2011):

$$c_i = \begin{cases} y(0) - bn = 0 \\ \dfrac{y^{(n)}(0) - \sum_{i=0}^{n-1} c_i a^{n-1} \binom{n}{i} i!}{n!} \quad 0 < n < N \end{cases}$$

2.1.3. Sensory feedback

The learning of the articulatory targets was supervised by both auditory feedback and somatosensory feedback. In this section, we first introduce the two kinds of auditory feedback (Sections 1.2.3.1–1.2.3.2), followed by a description of the somatosensory feedback (Section 1.2.3.3).

To examine our second hypothesis regarding which types of auditory guidance is more advantageous, we included (i) acoustic features (Fig. 1D) to simulate universal phonetic perception, and (ii) a speech recogniser (Fig. 1E) built for American English to simulate native-language phonological perception. For (i), we extracted two types of acoustic features from natural speech as sensory feedback, that is, MFCCs and Log Mel spectrograms (Davis & Mermelstein, 1980). For (ii), we trained a speech recogniser using a deep neural network with convolutional and recurrent layers based on clean speech of multiple speakers from the LibriSpeech corpus (Panayotov et al., 2015). The model was trained to learn a mapping from speech sounds to CVC syllables, which encompasses contextual information of time-series speech signals by combining spectrotemporal feature processing, temporal feature processing and classification.

Moreover, we implemented somatosensory feedback by applying two kinds of constraints on the vocal tract parameters during vocal exploration (Fig. 1F). The vowel constraint ensured that, for each candidate target set, the opening of the vocal tract is larger than a minimal cross-sectional area. The consonant constraint assured a closure over a limited portion of the oral cavity. We implemented the consonant constraint according to the uneven distribution of the sensory receptors on the tongue. Given that the tongue tip is more densely innervated than the tongue dorsum (Marlow et al., 1965), the closure tube length was set to be shorter in the anterior tongue section than in the posterior tongue section.

2.1.3.1. Acoustic features. As a major goal of the model was to address the speaker normalisation problem (Hypothesis 1), we deliberately trained the adult male vocal tract model to learn from female speech. We recorded the natural speech of a female native speaker of American English (age: 27) in a sound-attenuated acoustic laboratory. The sound files were recorded with a studio-grade microphone and audio interface at a sampling frequency of 44.1 kHz with 16-bit quantization. We then extracted MFCCs from the recordings. The Mel-scale approximates human perception of frequency, which is more sensitive to low frequencies than high frequencies (Stevens et al., 1937). We applied high-frequency emphasis through pre-emphasis (coefficient = 0.97). Frames were then extracted using 25 ms Hamming windows with 5 ms overlap, to be consistent with the sampling rate of the vocal tract parameters during synthesis. We applied 26 Mel filters with a maximum frequency of 10 kHz and calculated the log-power of their output to obtain Log Mel spectrograms. We calculated the DCT of the Mel log power to obtain 22-dimensional MFCCs (including energy) with sinusoidal cepstral liftering (coefficient = 2 × number of MFCCs). The acoustic error (E) was calculated using the Euclidean distance between the 22-dimensional MFCCs of the target and the synthetic utterances.

We have tested two types of acoustic features in terms of their proficiency in guiding vocal learning: MFCCs and Log Mel spectrograms. We trained the model with the two features to learn and built a word recogniser to evaluate them. The word recogniser was trained using the Kaldi Speech Recognition Toolkit and the annotated LibriSpeech corpus (Panayotov et al., 2015). The corpus contains speech data extracted from audiobooks recorded by adult male and female speakers of varied ages. The model is based on Weighted Finite State Transducers (WFSTs) that use Gaussian mixture models (GMMs) to model the speech acoustics. The MFCC features were transformed with Linear Discriminant Analysis (LDA) and Maximum Likelihood Linear Transform (MLLT). The model was trained using Speaker Adaptive Training on the 960-hour LibriSpeech mixed training data. The small pretrained trigram language model was used in the decoding.

The two types of acoustic features resulted in very similar word error rates. i.e., no distinction in the intelligibility. We further examined the quality of consonants and vowels separately. The consonant quality was better when trained by Log Mel spectrograms, while MFCCs were more advantageous in training the vowels. Detailed analysis of the synthetic speech trained by MFCCs and Log Mel spectrograms is provided in Appendix Figs. A1–A3. The similarity in the performance of the two acoustic features is consistent with a previous study which suggests that various feature-metric combinations impact minimally on the performance of dealing with the speaker normalisation problem (Gerazov et al., 2020).

2.1.3.2. Speech recogniser. A speech recogniser was trained to simulate a phonological perceptual space that discriminate the phonemes of CVC syllables in American English. To cover the phonemes in the target word list, speech segments were

Table 2
Speech data for training the speech recogniser.

	Number of utterances	Size	Duration
Training	2,711,615	21 G	116.7 h
Validation	337,109	2.6 G	14.4 h
Test	345,263	2.7 G	15 h

extracted with 11 onset consonants (/b/, /d/, /g/, /p/, /t/, /k/, /y/, /w/, /n/, /m/, /l/), 12 vowels and 5 diphthongs (stressed /aɪ/, /aʊ/, /eɪ/, /i/, /oʊ/, /u/, /æ/, /ɑ/, /ɔ/, /ɔɪ/, /ɛ/, /ɪ/, /ʊ/, /ʌ/ and unstressed /i/, /oʊ/, /ʌ/) and 6 coda consonants (/b/, /d/, /g/, /n/, /m/, /ŋ/). The speech data varied in terms of syllable type, including 17 vowels, 187 CV syllables and 1122 CVC words. The details of the speech data used for the training, validation and testing are shown in Table 2. 26-dimensional Log Mel spectrograms of the recordings were computed based on the same settings for extracting acoustic features and pre-padded to a length of 200 frames (spanning 1 s) to be used as the input for the training. The deep neural network is comprised of convolutional layers that extract spectrotemporal features, long short-term memory (LSTM) layers that capture temporal features and finally dense layers for classification. More details about the architecture are provided in Appendix Fig. A4.

The trained speech recogniser had a phoneme accuracy of 94% in the onset position, 88% in the vowel position and 98% in the coda position. We also tested a model trained with 22-dimensional MFCCs and the accuracy was 93%, 87% and 98% respectively. The Log Mel spectrogram had better overall performance (93%) than the 22 MFCCs (92%) and thus it was adopted in the current simulations. The 34-dimensional output vector of the speech recogniser represents a one-hot encoding of the onset consonant, the vowel, and the coda consonant of CVC syllables. The recognition loss of the CVC words is the Euclidean distance between the target vector and the recognised vector of the synthetic speech, by the following equation:

$$L = (p_i - q_i)^2, \quad i = 1, \cdots, N$$

where p_i represents the target phoneme vector and q_i represents the recognised phoneme vector. The p_i and q_i are values between 0 and 1. The output vector is a 34-dimensional perceptual space for a CVC syllable, i.e., N = 34 in the equation.

2.1.3.3. Somatosensory feedback. Somatosensory feedback refers to different neural signals generated by the sense of touch that could inform the learner about the state of their articulators. We implemented it as a type of "offline" feedback similar to the auditory feedback. Given the robustness of the touch sensation of the tongue and the wall of the oral cavity developed for eating, it is conceivable that such somatosensory information could be used to guide vocal learning. We implemented somatosensory feedback by applying two kinds of constraints on the vocal tract parameters during vocal exploration. The somatosensory signal indicates whether the oral cavity is open or closed. The vowel constraint is to ensure that the opening of the vocal tract is larger than a minimal cross-sectional area and the consonant constraint is to ensure a closure over a limited portion of the oral cavity. We implemented the two constraints by checking the tube area during the dynamic articulator movements. According to the constriction settings of VocalTractLab, tube area in the oral cavity larger than 0.25 cm^2 for the adult vocal tract model and 0.15 cm^2 for the child vocal tract models is considered as an open vocal tract. All the $_C$V targets that did not pass the check were filtered out. With regard to C_V target, the number of closed tube areas varied with the place of articulation of the target consonant. The total number of tube area sections is 40. A tube area less than 0.0001 cm^2 in VocalTractLab indicates a closed vocal tract. We allowed up to 4 closed tube sections to ensure closed lips for bilabial stops. Due to the built-in interdependency between lip protrusion parameter and lip distance parameter in the articulatory synthesiser, the threshold of closed tube area was 0.15 cm^2 for bilabial stops preceding rounded vowels in /booed/. Moreover, we implemented the consonant constraint according to the uneven distribution of the sensory receptors on the tongue. Because the tongue tip is highly innervated compared to the tongue dorsum (Marlow et al., 1965; Moayedi et al., 2021), the closure tube length was set to be shorter in the anterior tongue section and longer in the posterior tongue section. Specifically, the number of closed tube area sections were set to be less than 3 for alveolar stops and less than 9 for velar stops, except for alveolar stops before high vowels. In English, alveolar stops preceding high front vowels are likely to be palatalized (Bateman, 2007), which suggests a larger area of contact during the consonant articulation. The number of closed tube area sections was therefore set to be less than 9 for /deed/ and 6 for /did/.

2.1.4. Learning algorithms

Based on the hierarchical mechanisms of associative learning, we simulated the learning process in two stages i.e., exploration and refinement (Makino et al., 2016). In the first stage, the model explores the learning space and selects motor behaviour with preferred outcome based on the sensory information. In the second stage, the model exploits the selected motor behaviour to find the final optimal solution.

As vocal learning is a non-convex optimisation problem where the goal is to find optimal high-dimensional articulatory parameters (i.e., targets), metaheuristic gradient-free optimisation algorithms are appropriate (Larson et al., 2019).[3] We chose simulated annealing (Kirkpatrick et al., 1983) as the optimisation algorithm because it has been shown to be effective in our previous work on modelling tone and intonation (Xu & Prom-on, 2014), which was also one of the best gradient-free algorithms that we have tested, with small loss values as well as low computation time (Krug et al., 2023). It is a stochastic algorithm that seeks an optimal solution through a coarse-to-fine criterion. This algorithm can heuristically optimise models with many degrees of freedom, such as the speech production system. The learning process started with a neutral position (schwa) followed by adjustments of the vocal tract parameters to minimise the sensory errors.

During the optimisation, the articulatory targets were iteratively adjusted and tested, and whether they get accepted is determined by a probability p.

$$p = \begin{cases} 1 & \text{if } \Delta E < 0 \\ e^{-\Delta E/T} & \text{otherwise} \end{cases}$$

[3] We have previously also tried derivative-based algorithms such as gradient descent but found them to be unviable.

where ΔE is the change in the error of the objective function between the current and the previous attempt. T is the temperature that controls the annealing process, which starts at a high value and decreases at each step based on the following equation.

$$T = 1 - k/k_{max}$$

where k is the current iteration and k_{max} is the total number of iterations. The gradually reducing temperature was additionally constrained to be higher than 0.1 to ensure a minimal amount of parameter change in the later optimisation process. A uniformly distributed random number r between 0 and 1 is generated as a criterion for deciding whether the current trial is accepted. If the error is lower than the current error, the current adjustment is accepted. However, the algorithm also keeps some changes that are not ideal. If the probability of acceptance $p > r$, the new attempt is still accepted. This allows a balance between exploration and exploitation of optimal parameters. The gradual decrease of control temperature T throughout the process therefore means that any new target in the earlier stages is likely to be accepted but only targets with low errors are accepted in the later trials.

In the first learning stage, we initiated 10 processes in parallel, each with 2k iterations. Each process started with a neutral position (schwa) followed by random adjustments of the vocal tract parameters and gradually converged to a solution, as displayed in Appendix Fig. A5. Next, we selected the best candidate of each of the 10 processes for a more localised optimisation. In the second stage, the 10 processes randomly walked around these selected sets of articulatory targets for 200 iterations. More specifically, the model generated a neighbour solution based on the previous trial as follows:

$$x_i' = x_i + RW_i, \quad i = 1, \cdots, N$$

in which x_i is the 18-dimensional articulatory target, consisting of 17 vocal tract parameters and 1 time constant ($N = 18$). W_i is added to adjust the relative step of the random walk, based on the range of the vocal tract parameters and the time constant. R is a uniformly sampled random number between -1 and 1. x_i' is further constrained by the range of the parameters.

2.2. Hypothesis testing by listening experiments

To examine the three hypotheses, we conducted listening experiments to assess the learned synthetic speech under different conditions.

Hypothesis 1: End-to-end vocal learning can be simulated without explicit speaker normalisation by a combination of auditory guidance and coarticulatory dynamics.

> To test hypothesis 1, we compared the synthetic speech learned by the adult model with natural speech as the baseline condition. If synthetic speech can reach the intelligibility of natural speech, then vocal learning guided by auditory guidance would be deemed possible.

Hypothesis 2: Language-specific perception simulated by an automatic speech recogniser provides better auditory guidance than language-universal phonetic imitation simulated by acoustic matching.

> To test Hypothesis 2, we compared the speech learning performance guided by acoustic features and by the speech recogniser. The auditory feedback that generated more intelligible speech would be deemed as more advantageous.

Hypothesis 3: Vocal learning can be simulated also with a vocal tract with child anatomical configurations.

> To test hypothesis 3, we assessed the speech learned by a 1-year-old and a 3-year-old model. If the children's models also learned intelligible speech, then anatomical structure would be deemed as not constituting a barrier for vocal learning.

In addition, we used two types of listening experiments to evaluate the intelligibility of the synthetic speech, i.e., open-vocabulary dictation and multiple choice. In the open-vocabulary dictation, listeners were free to write down what they heard without any prompt. In the multiple choice, listeners were asked to choose from a fixed set of words. We assured that the online participants were not allowed to participate in the listening tasks more than once to avoid practice effects (Salthouse, 2012).

2.2.1. Speech materials

We trained the adult and the child vocal tract models to learn the target words guided by the sensory feedback options in Fig. 1D–F. The learning targets were minimal pairs of real English words with CVC syllable structures, containing bilabial, alveolar, and velar stops, as follows: "bead", "bid", "bed", "bad", "bod", "booed", "bud", "deed", "did", "dead", "dad", "god", "good", "body", "buddy", "Debbie" and "daddy". The learned articulatory parameters were then also used to synthesise novel CVCV words to verify their generalisability.

2.2.2. Listeners

173 monolingual American English native speakers between 18 and 50 years old participated in the online listening experiment. The participants were born and raised in the US, without any self-reported speech or hearing disorders. Among them, 47 did not pass the headphone screening; 5 were excluded from the experiment because of apparently atypical American accents; and 1 was excluded because of noise in the submitted recordings that suggested a noisy listening environment. We recruited 30 participants separately for four experimental conditions (120 listeners in total). The procedure has been approved by the Department of Speech, Hearing and Phonetic Sciences, University College London and the experiments complied with all relevant ethical regulations. Informed consent from all the participants was obtained online on Gorilla.

2.2.3. Procedure

We conducted a between-subject listening experiment with four conditions to evaluate the acoustic-feature-trained and recogniser-trained models in a set of open-vocabulary dictation tasks and a set of multiple-choice tasks. Participants were recruited and screened on Prolific (prolific.co) and then directed to Gorilla (gorilla.sc) for the online experiment. The participants first filled in a brief questionnaire for demographic and language background information. To verify their accents, participants were asked to read the first two sentences of the story "The North Wind and the Sun", a well-established text recommended by the IPA for eliciting English phonetic contrast. Par-

Table 3
Phonemes are labelled using CMU pronunciation dictionary (Carnegie Mellon University, 2022).

Target	B	AA	D
Insertion	B Correct	L AA Incorrect	D Correct
Deletion	 Incorrect	AA Correct	D Correct

ticipants were asked to undertake the tasks on a computer in a quiet environment without noise or other distractions. A headphone screening was conducted to ensure that the participants were wearing headphones. The listeners were asked to choose the quietest sounds out of three pure tones with one of the tones presented 180° out of phase across the stereo channels. The listeners who were wearing headphones were more likely to discriminate the sounds because a loudspeaker would have resulted in phase cancellation (Woods et al., 2017).

The participants who passed the screening were given five practice trials to get familiarised with the experiment. They were then randomly presented with the words produced by the female speaker, the synthetic sounds learned by the adult male, by the 1-year-old, and by the 3-year-old vocal tract models. 3 unique tokens of the 17 target words (see speech materials) were included in each condition. For the open-vocabulary dictation, the participants were instructed to listen to the audio carefully and freely write down the word they had heard. For the multiple-choice task, the participants were asked to choose the word from a list of 17 words.

2.3. Analysis

After the listening experiments, we analysed the response from the participants. The response was annotated with phone labels using the CMU pronunciation dictionary (Carnegie Mellon University, 2022) by pronouncing package (Parrish, 2022). We then manually added the phone labels for those responses without automatic annotation. In the case of phoneme insertion and deletion, we aligned the recognised phonemes maximally as shown in Table 3. Responses with reaction time shorter than the length of the stimuli were excluded in the analysis. The recognition rate was calculated in terms of how many segments were correctly identified. Due to the skewed distribution of the data and the small sample size, non-parametric statistical tests, including Kruskal-Wallis test, Wilcoxon Signed Rank test and Spearman correlation, were conducted to evaluate reaction time and phoneme accuracy. We report the W statistics and V statistics for unpaired and paired Wilcoxon Signed Rank tests respectively. Post-hoc comparisons were conducted by Wilcoxon Signed Rank test with Bonferroni correction.

3. Results

In this section, we will first report the results of the simulated adult vocal learning guided by a speech recogniser in comparison with natural female speech as the baseline condition to verify whether the perception-guided learning mechanism is plausible without explicit speaker normalisation (Hypothesis 1). We will subsequently examine how sensory feedback may impact on the performance of the model. The synthetic utterances trained by the speech recogniser (simulating language-specific perception) and acoustic features (simulating universal perception) will be compared (Hypothesis 2). Finally, we will report the intelligibility of the children's vocal learning along with its comparison with the adult model for the sake of evaluating the effect of age-related anatomical difference in the vocal tracts (Hypothesis 3).

3.1. Viability of recognition-guided vocal learning (Hypothesis 1)

To assess whether the model is able to accomplish vocal learning, we evaluated the learned synthetic speech in terms of both acoustic matching and perceptual intelligibility. The vocal tract model started with a broad search in the motor space and gradually converged to an optimal solution (Fig. 3A). The learned synthetic male speech trained by the speech recogniser shows a momentary burst of the onset consonant followed by clear vowel formants and high energy aspiration of the coda consonant, similar to the natural speech of a female speaker (Fig. 3B). In the open-vocabulary dictation task, the mean accuracy was 76% for the adult synthetic speech and 95% for the natural speech. Wilcoxon signed-rank tests showed that the natural female speech was more intelligible than the synthetic male speech in the open-vocabulary dictation task (W = 877, p < .001) and the multiple-choice task (W = 893, p < .001). CVC words learned by the adult vocal tract model were highly intelligible, with a median accuracy of 87%, 60%, 82% for the onset, vowel and coda, respectively, although this was short of the near perfect recognition of the reference natural speech (Onset: 95%, Vowel: 97%, Coda: 100%). It is worth noting that some listeners were able to identify the synthetic speech with accuracy close to natural speech (Fig. 3C).

The recogniser-guided adult vocal tract model learned intelligible consonants in both the onset and coda positions, although the vowels in the synthetic speech were not always correctly identified (Fig. 4A). Bilabial stops in 'bed', 'bid' and 'bod' and alveolar stops in 'deed' and 'did' were perfectly identified with identification rates being greater than or equal to the natural female speech in the open-vocabulary task (Fig. 4B). Wilcoxon Signed Rank tests showed that the natural speech had higher phoneme accuracies in all the syllable positions in the open-vocabulary dictation task (Onset: V = 368.5, p < .001, Vowel: V = 465, p < .001, Coda: V = 426.5, p < .001) and the multiple-choice task (Onset: V = 431, p < .001, Vowel: V = 465, p < .001). Wilcoxon Signed Rank tests showed that the phoneme accuracy of CV syllables did not differ significantly between the two types of task conditions (W = 1548, p = .186). The identification rate of phoneme accuracy in each target word is provided in Appendix Additional Analysis: Adult model.

We further analysed the learned articulatory targets of the model. The results show that consonant configurations varied depending on the vowel context (Fig. 3E). The bilabial stops /b/, for example, were articulated with closed lips in both instances but the tongue shape at the consonant closure is

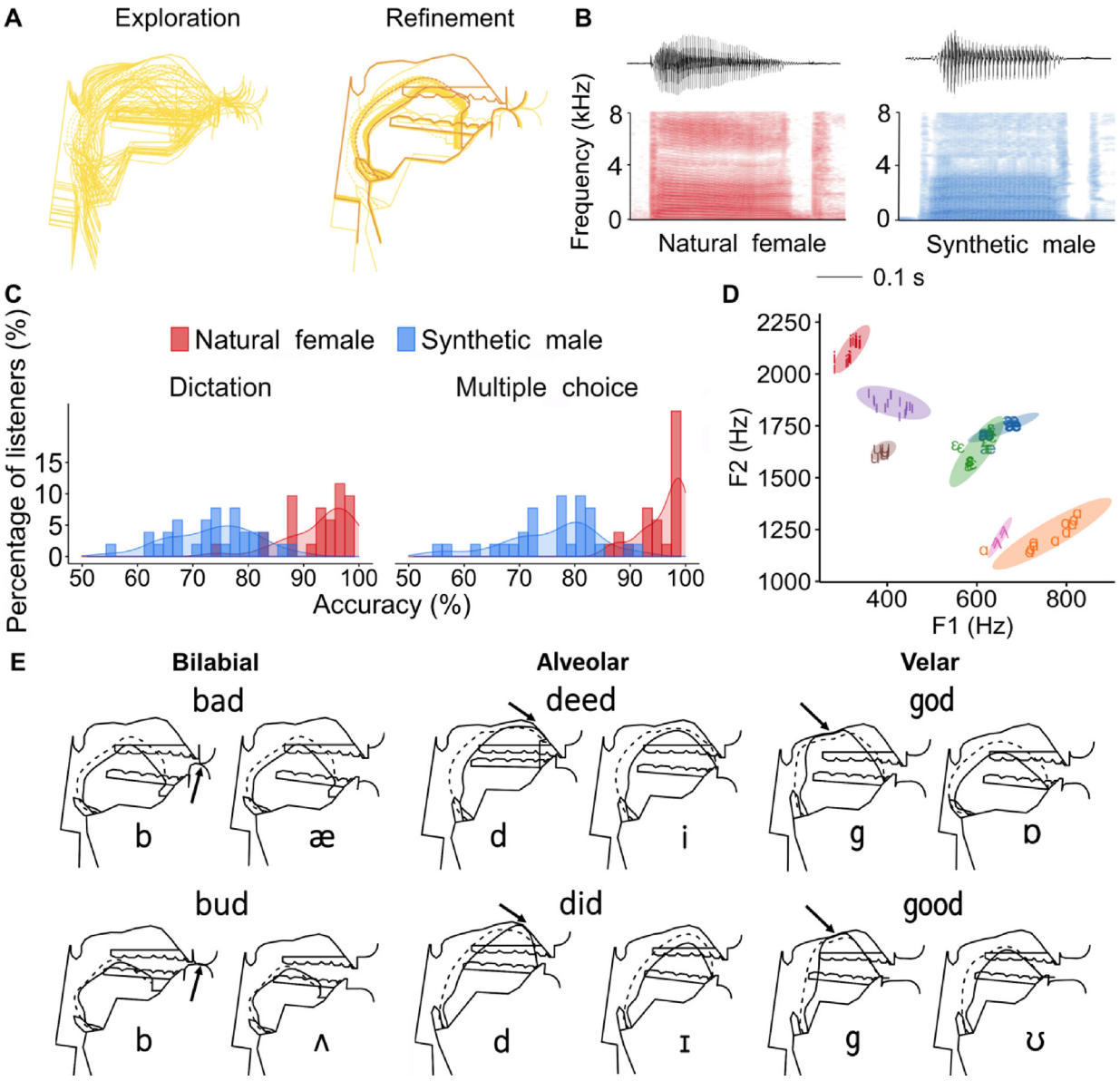


Fig. 3. Performance of vocal learning model. (A) Illustration of vocal learning progress. In the exploration stage, the model was allowed to randomly search around the vocal tract space, while in the refinement stage, solutions around the previous stage were exploited despite a small amount of exploration. Yellow: accepted trials; Orange: converged solution. (B) Waveform and Mel-spectrograms of 'bad' produced by a native speaker and an adult male vocal tract model. (C) Histograms and Kernel density estimates of mean phoneme accuracy of CVC words produced by a female native speaker and learned by an adult male vocal tract model in listening experiments. (D) First formant (F1) and Second formant (F2) of vowels in CVC words with bilabial stops learned by the adult vocal tract model (represented by International Phonetic Alphabet). The best 20 instances per target word were selected based on recognition error. (E) Midsagittal sections of the learned vocal tract shapes. The solid and dashed lines represent the tongue side positions in the front and back respectively. Arrows point at the constrictions formed by the C_V targets. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

already partially controlled by the vowel. In the case of /d/, the tongue body during the closure was higher in 'deed' than in 'did'. For /g/, the tongue body was more advanced in 'good' than in 'god' because the vowel in 'good' is a front vowel. Despite the context-sensitive consonant articulation, acoustic characteristics of the learned vowels in the same category form consistent clusters (Fig. 3D). Furthermore, to verify whether the acquired motor repertoire generalises to novel words, we reused the learned vocal tract parameters to synthesise CVCV syllables. The synthetic words achieved a similar identification rate as the natural female speech in the multiple-choice task (Appendix Fig. A7). The mean phoneme accuracy was 88% for the synthetic male speech and 95% for the natural female speech in the open-vocabulary task. The natural speech had significantly higher identification accuracies than the synthetic speech (Wilcoxon signed-rank, $p < .001$). With respect to the multiple-choice task, the mean identification rate was 96% for the synthetic words and 97% for the natural words. The synthetic speech and natural speech did not differ significantly (Wilcoxon signed-rank, $p = .442$). There were some confusions between 'body' and 'buddy', while 'Debbie' and 'daddy' were almost perfectly identified.

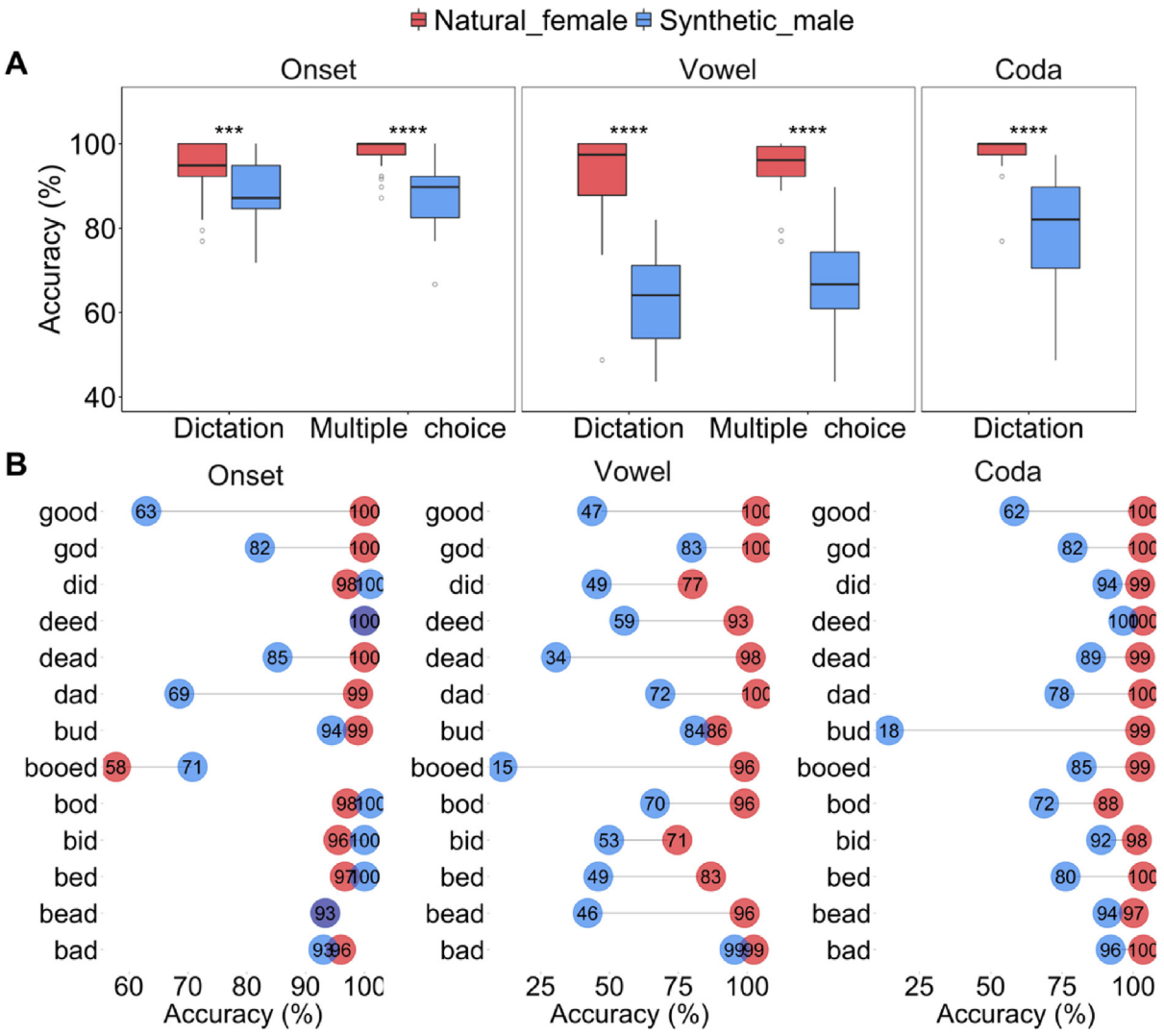


Fig. 4. Comparison between natural female speech and synthetic male speech in the listening experiment. (A) Phoneme accuracy of natural and synthetic speech in different syllable positions, evaluated by open-vocabulary dictation and multiple-choice tasks. Coda accuracy was not evaluated in the multiple-choice task because the coda consonant remains constant in the word list. **** $p \leq 10^{-4}$. (B) Mean phoneme accuracy of CVC words produced by the female native speaker and the male vocal tract model in different syllable positions, measured by open-vocabulary dictation.

The high intelligibility shows that the learned vocal tract parameters can generalise to novel multisyllabic words that the model was not trained on.

3.2. Language-specific perception vs. universal phonetic perception as learning guide (Hypothesis 2)

Having successfully simulated the learning of intelligible English words, we can now evaluate the associated influencing factors. We simulated language-specific perception by a phonological speech recogniser and universal phonetic perception by acoustic features that captures all the details. The results indicated that words trained by the speech recogniser were more intelligible than those trained by acoustic features in both the open-vocabulary dictation task (Fig. 5A: W = 52.5, p < .001) and multiple-choice task (Fig. 5B: W = 790, p < .001, Wilcoxon signed-rank). Post-hoc comparisons showed that the tendency was the same for vocal tract models of adult and children in both the open-vocabulary (1y: p < .001, 3y: p < .001, Adult: p < .001) and multiple-choice tasks (1y: p < .001, 3y: p < .001, Adult: p < .001). In terms of the property of the phoneme, the recogniser-trained words had higher accuracies in the onset (W = 241, p < .001), the vowel (W = 1512.5, p < .001) and the coda (W = 2462, p < .001, Wilcoxon signed-rank) positions than those trained by MFCCs (Fig. 5C). These results suggest that native-language phonological perception simulated by the speech recogniser was more successful than universal phonetic perception simulated by acoustic features.

The benefit of perceptual guidance is most clearly seen in the relationship between the recognition rates by human listeners and the feedback type. Spearman's correlation shows no correlation between human accuracy and acoustic error (Fig. 5D), whereas there is a significant negative correlation between human accuracy and recognition error (Fig. 5E), suggesting a commonality between the recogniser evaluation and human perception. We further compared how target words were identified by the speech recogniser, MFCCs and native listeners. The speech recogniser and the listeners had consistent judgement towards almost all the synthetic speech, while

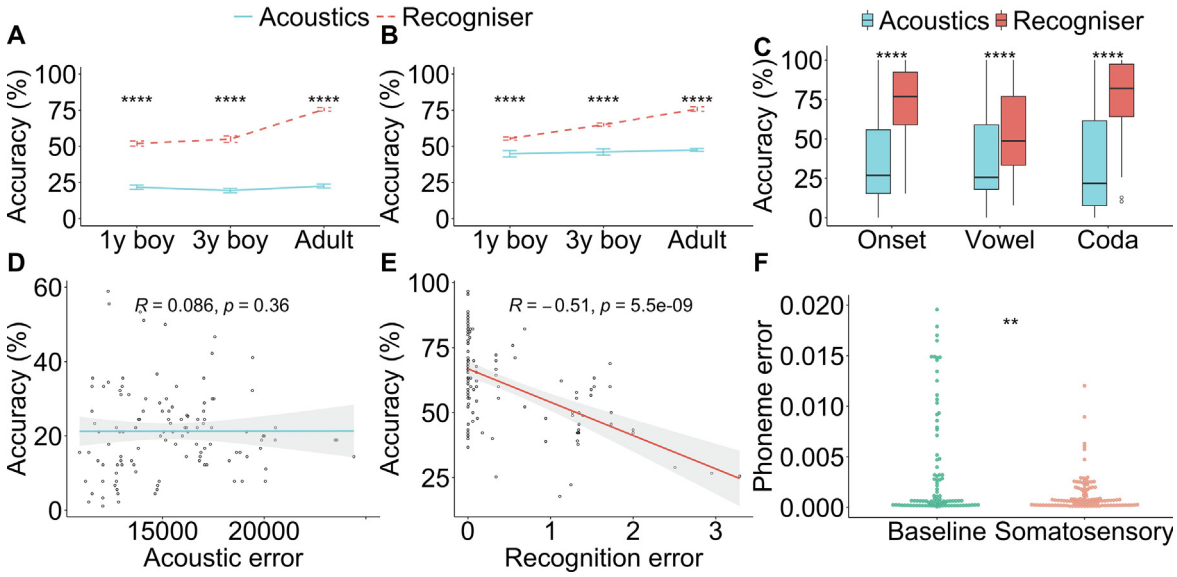


Fig. 5. Effects of sensory feedback on vocal learning. (A, B) Mean phoneme accuracy (39 CVC words) of vocal tract models in the open-vocabulary (A) and multiple-choice (B) tasks. Error bars show standard errors. (C) Phoneme accuracy in different syllable positions. (D, E) Relationship between open-vocabulary dictation accuracy and types of auditory feedback: acoustic features (D) and the recogniser (E). (F) Effect of somatosensory feedback in recognition error of words learned by an adult vocal tract model, evaluated by the recogniser. The best 10 instances per CVC word are included. ** $p \leq 10^{-2}$, **** $p \leq 10^{-4}$.

the normalised identification scores of MFCCs and the listeners were fairly discrepant (Fig. 6). Moreover, listeners spent more time on identifying words trained by acoustic features in both the open-vocabulary dictation and multiple-choice tasks (Fig. 7). Wilcoxon Signed Rank tests showed that reaction time was significantly longer for MFCC-trained speech regardless of the task type (Open-vocabulary: W = 8455108, $p < .001$, Multiple-choice: W = 8294666, $p < .001$).

In addition, we simulated somatosensory feedback by a constraint on the degree of oral opening for each generated vocal tract configuration during vocal exploration. The constraint ensured an open vocal tract for vowels and a narrow

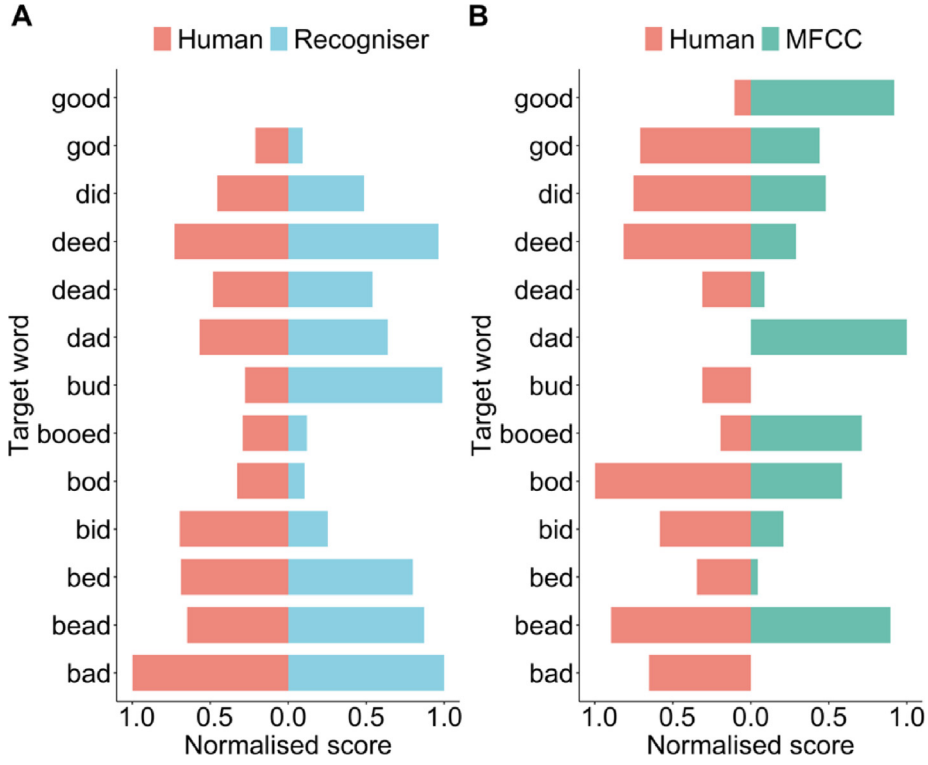


Fig. 6. Comparison of human identification with the speech recogniser and MFCCs by target words. (A, B) Normalised identification scores of the target words judged by the native listeners and the speech recogniser (A) or MFCCs (B). Phoneme accuracy judged by native listeners were normalised to values between 0 and 1. Recognition errors and MFCC errors were also normalised to have the same range but in a reversed order.

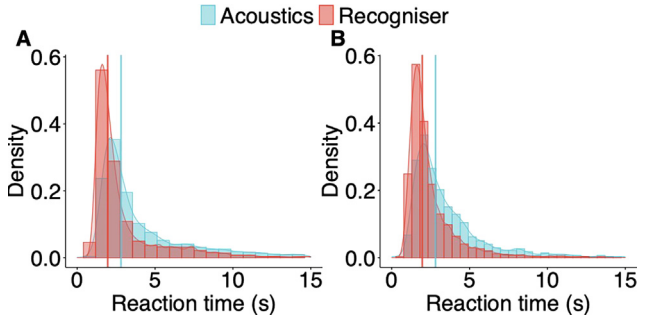


Fig. 7. Effects of type of auditory feedback on simulated vocal learning. (A, B) Reaction time of American English listeners in the open-vocabulary dictation task (A) and in the multiple-choice task (B). The vertical lines represent the median reaction time for the two types of auditory feedback. The listeners were asked to identify CVC and CVCV words learned by vocal tract models of different ages.

vocal tract for consonants. As shown in Fig. 5F, with the same number of iterations, the model with somatosensory feedback learned more intelligible words than the baseline (W = 10429, $p = 0.001$, Wilcoxon signed-rank test). Note, however, that the model yielded speech sounds with low recognition error even without somatosensory feedback.

3.3. Effect of age-related anatomical differences on vocal learning (Hypothesis 3)

We further examined the speech learned by a 1-year-old and a 3-year-old model to see if children's vocal learning can be simulated. We analysed the by-phoneme position accuracy rate of the CVC words learned by the two child vocal tract models. In the open-vocabulary dictation task, the 1-year-old model had a median of 56% phoneme accuracy rate in the onset position, compared with 63% for the 3-year-old model. Listeners correctly transcribed 32% of the vowels learned by the 1-year-old model and 38% with the 3-year-old model. For the coda position, the median identification was 68% for the 1-year-old model and 69% for the 3-year-old model. Both models had higher intelligibility in the consonant positions than the vowel position.

The two models had similar accuracies for bilabial stops and alveolar stops. However, the 1-year-old model learned poorer velar stops in 'god' and 'good', when compared with the 3-year-old model. With respect to the learning of vowels, the two models had comparable performance for most of the vowels. The 3-year-old model yet again showed better results in the case of /ʊ/ in 'good' and /ɒ/ in 'god' than the 1-year-old model. The vowel groups learned by the 3-year-old model distributed more concentratedly than the 1-year-old model (Appendix Fig. A6). Both child models failed to learn intelligible /ɒ/ in 'bod'.

Fig. 8A and B compare the identification accuracy of the phonemes in CVC words learned by the two child models. Each connected line represents the average phoneme accuracy of one listener. Solid lines indicate that the 3-year-old model has higher phoneme accuracies than the 1-year-old model and vice versa for the dashed lines. Native listeners

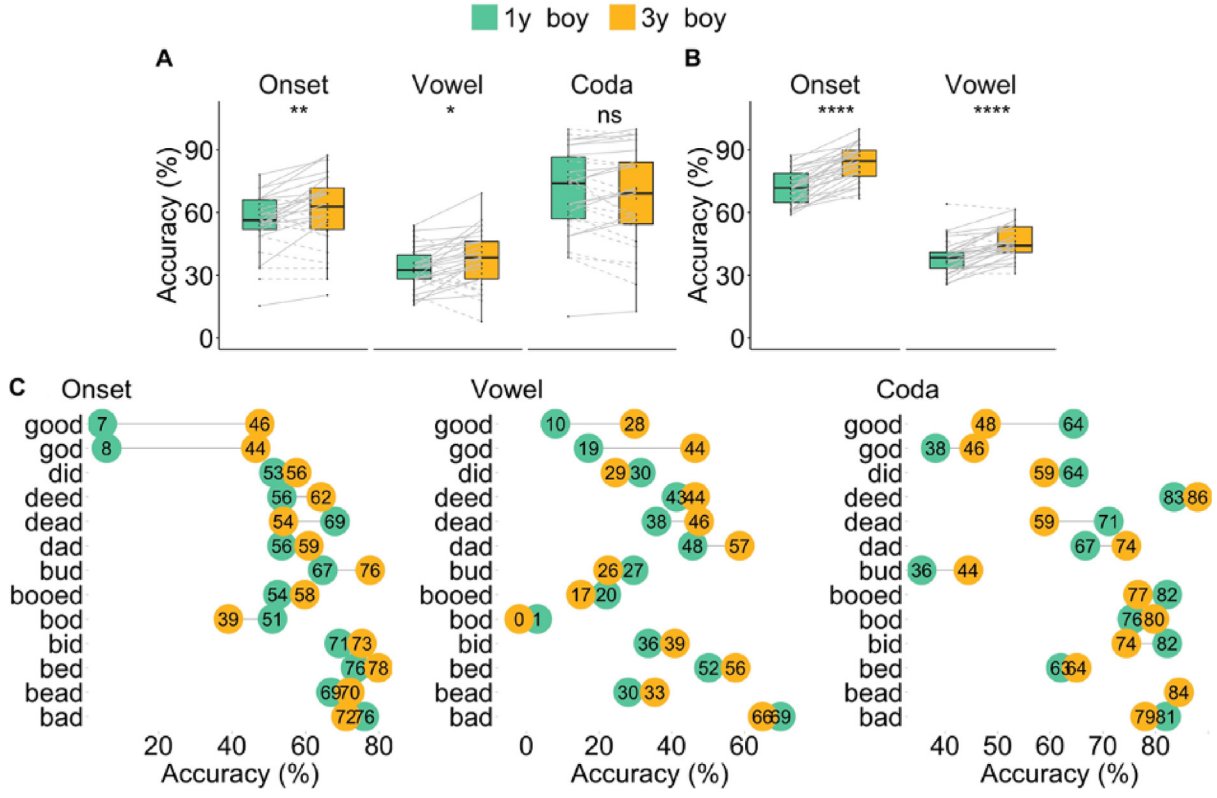


Fig. 8. Consonant and vowel accuracy of child models. (A, B) Boxplots of phoneme accuracy in different syllable positions of CVC words learned by the 1-year-old boy and 3-year-old boy's vocal tract models, measured in an open-vocabulary dictation task (A) and a multiple-choice task (B). Given that the coda consonant remains the same in the word list, what was evaluated in the multiple-choice task was the intelligibility of the initial CV portion of the words. Each connected line represents the average phoneme accuracy of one listener. The solid lines indicate that the 3-year-old model had higher phoneme accuracy than the 1-year-old model and vice versa for the dashed lines. ns $p > 0.05$, *$p < 0.05$, **$p \leq 10^{-2}$, ****$p \leq 10^{-4}$. (C). Mean phoneme accuracy of utterances learned by the 1-year-old boy and a 3-year-old boy's vocal tract models in the onset position, the vowel position, and the coda position of the CVC words, measured in an open-vocabulary dictation task. 30 American English listeners freely transcribed the utterances.

sometimes had higher identification rates when judging speech learned by the 1-year-old model than by the 3-year-old model in the open-vocabulary dictation task (dashed lines). In contrast, we can rarely see such cases in the multiple-choice task, that is, there were only a few cases where the words learned by the 1-year-old model were more intelligible than the 3-year-old model (Detailed intelligibility analysis of each target word learned by the two child models is provide in Appendix Additional Analysis: Child models). Wilcoxon signed-rank tests showed that the 3-year-old model learned more intelligible speech in the onset (p = .019) and the coda position (p = .019), but not in the vowel position (p = .063) in the open-vocabulary task. In the multiple-choice task, similarly, the 3-year-old model had higher accuracies than the 1-year-old model in both the onset and the vowel position (Wilcoxon signed-rank: p < .001). The increase in the perceptual accuracy suggests that the growing child vocal tract has enhanced capability to learn articulatory targets that yield intelligible speech.

We also compared the performance of the speech learned by adult and child vocal tract models. The Kruskal-Wallis tests showed that the age of the vocal tract model had a significant effect on the intelligibility of the learned CVC words, as measured by the open-vocabulary dictation (χ^2 = 50.381, df = 2, p < .001) and multiple-choice (χ^2 = 52.678, df = 2, p < .001) tasks (Fig. 9A). Post-hoc comparisons showed that the adult model had higher mean phoneme accuracy than the two child models (Open-vocabulary: 1y: p < .001 and 3y: p < .001; Multiple-choice: 1y: p < .001 and 3y: p < .001). Consistent with the listening results, the models of different ages also yielded divergent recognition errors with the same number of iterations (Fig. 9B, χ^2 = 62.189, df = 2, p < .001, Kruskal-Wallis test). Post-hoc comparisons showed that the adult male vocal tract model learned CVC words with lower recognition errors than the child models (1y: p < .001 and 3y: p < .001). However, the 3-year-old model had comparable recognition errors to the 1-year-old model (p = .067). Again, we tested the generalisability of the learned articulatory targets by resynthesising CVCV syllables. The identification of the resynthesised CVCV words was easier than the CVC words in the open-vocabulary dictation task (Fig. 9C, W = 1626, p < .001, Wilcoxon signed-rank test). Post-hoc comparisons showed that the accuracy was higher for CVCV words than for CVC words regardless of the age of the model (1y: p < .001, 3y: p < .001, Adult: p < .001).

4. Discussion

In this study, we have built a computational model that can simulate vocal learning guided by either phonetic imitation or phonological perception. Results of open-vocabulary dictation and multiple-choice evaluations show that highly intelligible English words can be learned under perceptual guidance, while only moderate intelligibility can be achieved when guided by phonetic imitation. The performance under both guidance strategies, however, well exceeds those of previous simulation works. Even the vocal tract model configured with two child vocal tract dimensions were able to learn fairly intelligible words, although the quality was not as high as that of the adult model. These results, therefore, show that the speaker normalisation or the correspondence problem previously believed to be insurmountable is no longer a barrier under the learning model developed here. In the following we will discuss the key aspects of the model that have contributed to its learning efficiency.

4.1. Coarticulatory dynamics

Our results have also shown that words trained with a speech recogniser was intelligible (76% phoneme accuracy) and even those trained with acoustic features achieved 48% phoneme accuracy in the multiple-choice task with 17 choices. One of the key aspects absent in previous simulation works on vocal learning is explicit modelling of coarticulation dynamics. The current modelling work adopted SDSSTA (Xu, 2020; Liu et al., 2022) as the coarticulation model (Section 2.1.2). The target approximation (TA) part of SDSSTA assumes that every articulatory movement is an act of asymptotically approaching an underlying target (Xu & Emily Wang, 2001; Birkholz et al., 2011).

This means that the learning process only needs to explore TA generated movement trajectories, which not only massively reduces the total number of vocal explorations needed, but also makes sure that the synthetic utterances generated at the end of learning are also always articulatorily plausible. The power of TA was already seen in our previous studies. Prom-on et al. (2014a, 2014b) found that using VocalTractLab

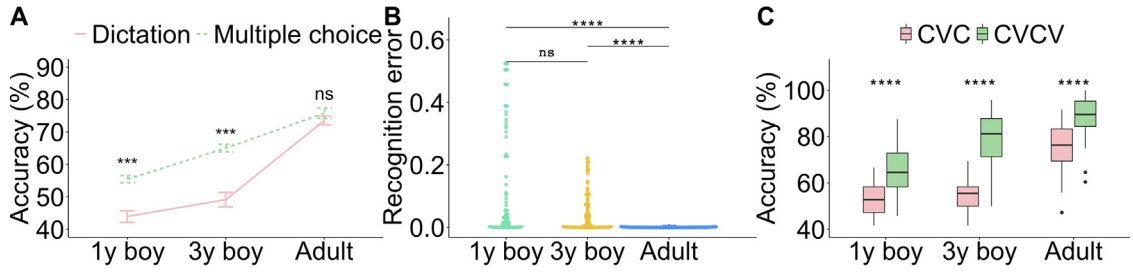


Fig. 9. Performance of vocal learning models of different ages. (A) Mean phoneme accuracy of CVC words learned by the 1-year-old boy, 3-year-old boy and adult vocal tract models, evaluated by open-vocabulary and multiple-choice tasks. Error bars show standard errors. (B) Recognition error distributions of 10-best CVC words evaluated by the recogniser. Outliers outside 1.5 times the interquartile range are not shown. (C) Overall phoneme accuracy of CVC words and CVCV words learned by the adult and child vocal tract models in the open-vocabulary dictation task. ns $p > 0.05$, ***$p \leq 10^{-3}$, ****$p \leq 10^{-4}$.

with built-in TA model and MFCC matching as feedback, Thai vowels could be learned from natural two-vowel Thai words, with synthetic intelligibility equal to the original natural speech. Prom-on et al. (2013) found that syllables with only glide consonants could be learned with high naturalness by informal listening. But that line of modelling work encountered difficulty when trying to simulate the learning of syllables with obstruent consonants. It was not until the expansion of TA into SDSSTA that we started to see initial success in learning some non-glide consonants.

SDSSTA is motivated by not only empirical and modelling studies of tone and segment, but also consideration of the problem of too many degrees of freedom in motor control (Bernstein, 1967). The CV synchrony assumed in SDSSTA removes the need to explore the relative timing of C and V while learning their targets. In other words, the effective exploration of specific articulatory targets starts only from when children are able to synchronise CV onsets, i.e., when they have started canonical babbling. Whether CV synchrony is indeed in place in canonical babbling, however, awaits confirmation from future investigations. Furthermore, additional research is needed to better understand how the learned C and V targets can be deployed across various speaking rates. Finally, SDSSTA solves the conflict of CV co-execution by allowing different articulators involved in a segment and even different dimensions of the same articulator to be sequentially controlled (i.e., without overlap or blending) by either consonant or vowel. This allows the learning process to find solutions that echo existing finding, e.g., variable contact locations in velar consonants when coproduced with different vowels (Fig. 3E).

There are various alternative models to the target approximation component of SDSSTA, the general Tau theory (Elie et al., 2023), the equilibrium point hypothesis (Perrier et al., 1996), the FACTS mode (Parrell et al., 2019), and the Task Dynamic model (TD) (Saltzman & Munhall, 1989), but only TD explicitly models coarticulation, like in SDSSTA. TD nevertheless differs from SDSSTA in some non-trivial ways. First, instead of stipulated CV synchrony in SDSSTA, TD requires settling the relative timing of adjacent segments prior to the articulation of each syllable. Second, instead of strict sequential articulation at the level of articulatory dimensions, TD allows articulatory blending for temporally overlapped gestures. Third, in the implementation of TD, the intergestural timing needed additional models such as Coupled Oscillation model in TADA system as an extension to the original model (Nam et al., 2004). Fourth, in the modelling practice of TD, targets are widely assumed to be virtually reached, because (a) the magnitude of articulatory trajectories is treated as the actual size of the gestural scores in a simplified model (Browman & Goldstein, 1990: 303), and (b) undershoot is interpreted to be generally avoided (Tilsen, 2019: 2). This differs from the simulation work here that assumes that targets are often not reached (Xu & Prom-on, 2019), but can still be learned regardless of the undershoot in real production. The impacts of these differences are unknown, however, and it will require a systematic comparison of simulated vocal learning to find out.

The application of SDSSTA in the current modelling work has also led to an unexpected finding, i.e., the learning process has generated variability along articulatory dimensions that are unrestrained by the C_V target. It is consistent with the notion of uncontrolled manifolds in the motor control literature, i.e., low variability along task-relevant dimensions, but high variability along task-irrelevant dimensions (Scholz & Schöner, 1999; Todorov & Jordan, 2002). As shown in Fig. 10, the variability is low in the tongue positions for the $_C$V target, while the C_V target shows less variance in the articulator dimensions that are critical for forming constrictions. Thus, desirable task-specific variability patterns (Scholz & Schöner, 1999; Todorov & Jordan, 2002) may emerge as a result of vocal learning when two goals (C and V) are executed at the same time. As the functional objectives of speech are more clearly defined and more multifaceted than the artificial tasks often examined in typical motor control studies (Wolpert et al., 2011), the current vocal learning model may contribute significantly to theories of motor control.

A limitation of this study is that the learning model in our current work was only trained to articulate specific syllables. It could be argued that this approach is inconsistent with the well-accepted notion of segments as being independent of each other. However, there is no research that can indicate which is harder: having to learn more syllable-specific targets or having to work out invariant targets that are applicable to all syllabic contexts. Given that children may have greater coarticulation than adults (Zharkova et al., 2011; Zharkova, 2018), i.e., with more vowel-specific consonant articulation, it is conceivable that adults are just further on their way toward fully invariant targets. Since there is no existing simulation of vocal learning that has demonstrated the advantage of invariant over syllable-specific segmental targets, we can only leave this issue to future modelling works.

4.2. Auditory feedback in vocal learning

Our initial modelling based on acoustic matching soon encountered difficulty to improve the synthetic quality beyond a rather moderate level for certain syllables, and to learn some other syllables at all. The adoption of speech recognition as an alternative to acoustic matching brought immediate improvement to the learning performance, which led to the systematic comparison of the two kinds of error signals as auditory feedback that guides vocal learning, as reported in this paper. Acoustic errors are similar to general auditory perception or universal phonetic perception at the early developmental stage, whilst the speech recogniser is comparable to native-language phonological perception at the later stage (Werker & Lalonde, 1988; Kuhl, 2000). The comparison shows that speech-recogniser-simulated language specific phonological perception is far more beneficial than acoustic-matching-simulated universal phonetic imitation.

This finding is in line with previous suggestions that auditory experience gained during perception acquisition may guide vocal learning (Kuhl, 2000), but the nature of such auditory experience has been unclear. It could be in the form of auditory templates as suggested for songbirds (Phan et al., 2006; Zhao et al., 2019). But such templates would take up a lot of neural resources, especially if they need to represent sufficient cross-speaker variability as suggested by the exemplar theory (Goldinger, 1996). The new finding reported here suggests that a neural network trained for the purpose of speech perception

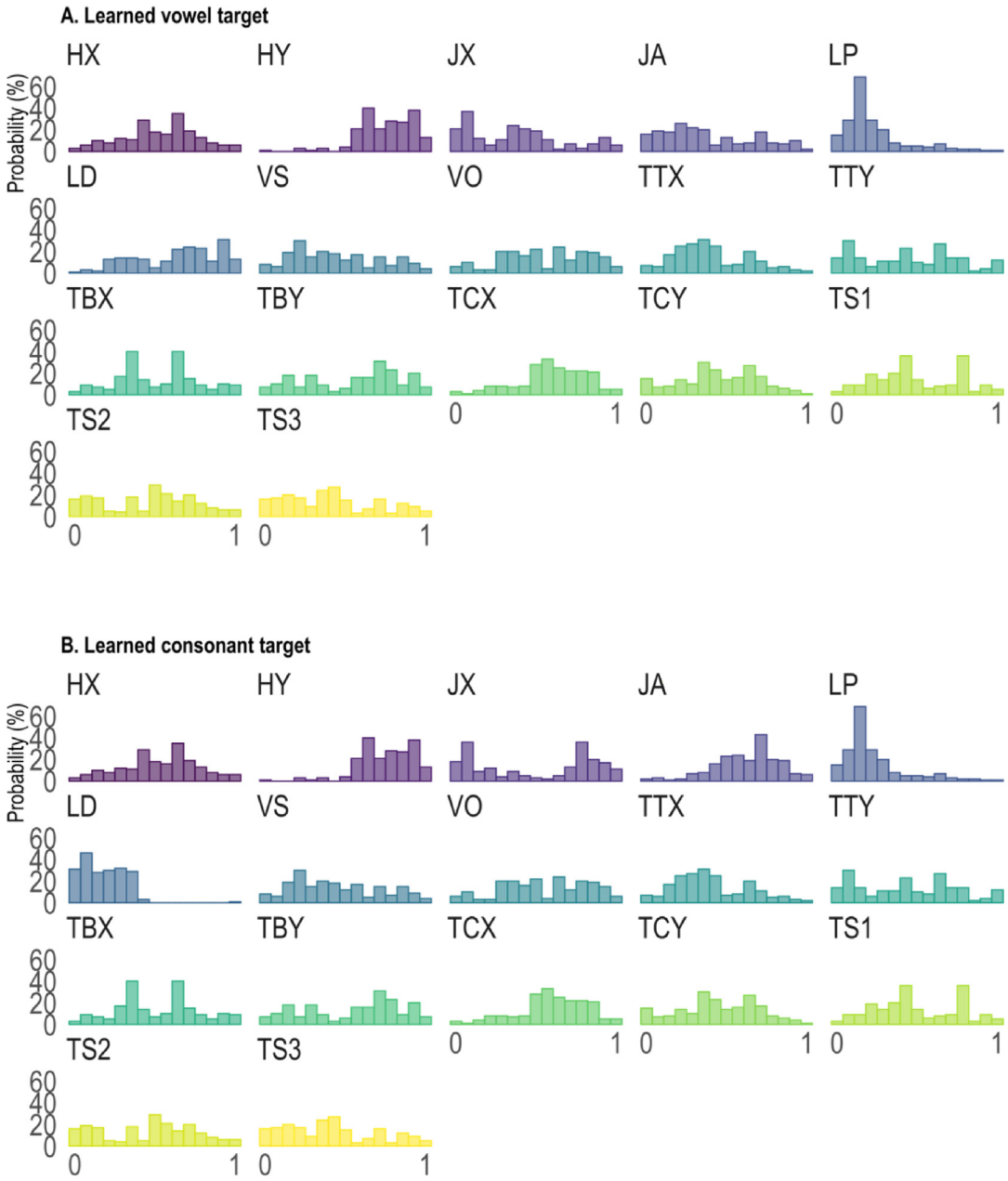


Fig. 10. Learned vocal tract parameters of syllabic $_{C}$V and C$_{V}$ targets after the optimisation of the CV sequence in 'bad'. Histograms show 200 sets of learned targets with lowest recognition errors for the syllabic $_{C}$V and C$_{V}$ target separately. For the $_{C}$V target, tongue tip and tongue body horizontal positions (TCX, TTX), hyoid height (HY), lip protrusion (LP) exhibit concentrated distribution, while jaw angle (JA), jaw horizontal position (JX) lip distance (LD) seem widely distributed. The C$_{V}$ controlled LD, JA and JX during coarticulation and these parameters are narrowly distributed after training.

would be sufficient to guide vocal learning, so that there may be no need to store any auditory templates or exemplars. Furthermore, the recogniser captures the key acoustic details that can sufficiently discriminate between different syllables in a given language, which is precisely what is lacking in the acoustic templates.

Note that a speech recogniser trained by multiple speakers, like the one used in the present study, may represent an auditory space that is distorted by the language specific phonology, which is reminiscent of the warped auditory map depicted by the magnet effect (Kuhl, 1991; Kuhl et al., 2008). This would be exactly what is needed to guide the learners to find articulatory targets that would satisfy their perception. That this seems to be the case can be best seen in the high correlation between listeners' perception accuracy patterns with the recognition errors by the speech recogniser, which is in sharp contrast to the lack of correlation between acoustic errors and perception accuracy by the native speakers (Fig. 5D, E, Fig. 6).

4.3. Speaker normalisation: A problem no more

A large amount of efforts have made to tackle the problem of child–adult anatomical difference through various means of normalising the acoustic characteristics (Kanda et al., 2009; Moulin-Frier & Oudeyer, 2012; Warlaumont, 2012; Rasilo et al., 2013; Kröger et al., 2014; Philippsen et al.,

2014; Prom-On et al., 2014a, 2014b; Moulin-Frier et al., 2015; Warlaumont & Finnegan, 2016; Najnin & Banerjee, 2017; Barnaud et al., 2019) or associating child acoustic vectors with adult acoustic vectors (Plummer et al., 2010), but these approaches have not been used to train intelligible speech. What the present results have shown is that the problem can be resolved with a combination of coarticulatory dynamics and a language-specific speech recognition. During the iterative vocal practice cycles, the motor search space is first confined by the coarticulatory mechanism, which limits movement trajectories to only those that approach specific targets, and synchronises the onset of consonants and vowels, thus minimising the temporal degrees of freedom. With a deep-learning trained speech recogniser providing feedback, the intelligibility reached over 75% even in open-vocabulary dictation. Note that, because the speech recogniser was trained with the speech of multiple speakers in different linguistic contexts, it has learned an auditory space with categorical phonological distributions (Chang et al., 2010). This space is therefore effectively, though inexplicitly, speaker-normalised (Sjerps et al., 2019). This multi-speaker perception space, which is phonologically warped as mentioned in Section 2.1.5, seems to be also powerful enough to guide the learner's vocalisation gradually toward the language-specific perceptual norm. Assuming that this is indeed the case, it could even explain how each language-specific phonological space came about in the first place.

4.4. Learning mechanisms

The core learning mechanism simulated in this model is vocal learning guided by speech perception, which is in line with previous findings of sensorimotor learning in songbirds and humans (Kuhl, 2000). The model emulated a perceptual learning phase of ambient language using a pre-trained speech recogniser and a production learning phase with biologically-plausible articulatory synthesiser. The two learning phrases share similar attributes with vocal development in other vocal learners including songbirds (Thorpe, 1954) and marmoset monkeys (Elowson et al., 1998a, 1998b). Interestingly, language-related genes such as FoxP2 also show similar expression patterns in the auditory and motor systems in humans and songbirds (Teramitsu et al., 2004). The notable resemblance in behavioural and genetic studies indicates that perception-guided vocal learning can be a shared cognitive mechanism across species.

Yet another critical component of the present simulation model is the learning strategy based on trial and error. This strategy combines vocal exploration, feedback-based selection, and motor refinement, which coincides with the neural mechanisms of sensorimotor learning (Makino et al., 2016). It requires neither directional corrections, nor stored articulatory-acoustic matching data accumulated in the babbling phase as proposed in the DIVA model (Guenther, 1994; Tourville & Guenther, 2011). The feedback in the current model only assesses how good each trial is, without tracking or predicting the direction of improvement. Indeed, from the perspective of child vocal learning, it is inconceivable that the learner would know how to make corrections before they have mastered the articulation of a speech sound. For example, they wouldn't know how to change the tongue shape to make a vowel more front or more back. It is more likely that observed feedback adaptation (Houde & Jordan, 1998) is a property of mature motor skills, as young children are not capable of modifying articulation effectively based on auditory feedback (MacDonald et al., 2012).

In other words, learning novel motor repertoires may not require a fully developed link between articulation and acoustics. Rather, the self-learning process may be the very process that forges the link between speech perception and speech production during development (Makino et al., 2016). It is likely that the correction-based learning (Guenther, 1994; Tourville & Guenther, 2011) modelled by DIVA and offline feedback-based learning simulated in the current study overlap to a certain extent (Makino et al., 2016). We also postulate that the reinforcement learning of motor patterns (Yoshikawa, Asada, et al., 2003; Miura et al., 2012; Warlaumont et al., 2013; Howard & Messum, 2014; Messum & Howard, 2015; Warlaumont & Finnegan, 2016; Rasilo & Räsänen, 2017) may accompany trial-and-error-based learning as extra training signals. There is abundant room for further progress in examining the interface between these learning mechanisms.

4.5. Anatomical structure and vocal learning

The speech learned by the adult model was more intelligible than the child vocal tract models. Furthermore, the 1-year-old model learned speech with lower intelligibility than the 3-year-old model. The vocal tract anatomy between the adult and the children differs greatly (Fitch & Giedd, 1999) and the child's vocal tract undergoes huge changes in the first three years of life (Kent & Murray, 1982; Kent, 1992). While these results may suggest that a mature vocal tract model is more advantageous for vocal production learning, caution is needed in the interpretation of these results because the vocal tract models were all guided by the same mature perceptual system, as the purpose of this experiment is merely to test the effect of the anatomical structure. The speech recogniser was trained without child speech data, whereas in reality, children also hear themselves, and they may hear the speech of other children, both of which would further improve their speech-relevant perceptual space. In addition, the child vocal tract configurations used in this study were based on developmental estimations (Goldstein, 1980; Birkholz & Kröger, 2007), which may contain aspects that differ critically from real-life child vocal tract geometry. Future simulation studies may see improvements with more realistic perceptual representations and vocal tract models based on child MRI data.

Interestingly, the difficult cases of speech sounds for the model seem to correspond well with the ones that are normally acquired later in real life. Children acquire corner vowels before mid vowels (Stoel-Gammon & Pollock, 2008) and our models also had higher accuracies for corner vowels. Mid vowel /ɪ/ learned by the models was frequently mistaken as /i/ and /ɛ/, which are common mistakes in children's production (Vihman, 1996). However, the corner vowel /u/, which is supposed to be easy to acquire, was a difficult case for our models. Also interestingly, the models learned /u/ without lip rounding, which is similar to what is found from the congenitally blind population (Ménard et al., 2014). This further confirms a vital role of visual cues in simulating vocal learning (Murakami et al., 2015).

Regarding consonants, it has been reported that among the voiced stops, the production of bilabials occurs before alveolars and velars, which are also fully acquired the earliest (Crowe & McLeod, 2020). Likewise, both our adult and child models had higher identification rates for bilabials than for alveolars and velars. We also found that the CV combinations with consonants having the place of articulation similar to that of the following vowel were easier to learn, in line with the developmental patterns (MacNeilage & Davis, 2000). For example, CV sequences consisting of alveolar consonants followed by high vowels (e.g., 'deed', 'did') had higher identification rates than the other pairs (e.g., 'dad' and 'dead').

The crucial benefit of native-language phonological perception for vocal learning seen in the simulation in this study seems to be relevant to the question of how children's speech perception space is developed in the first place. A likely way is to learn the phonetic categories through their association with words whose meanings are suggested by the rich context of social interaction (Kuhl, 2007). Because children's initial social interaction is mostly with their caregivers, it is therefore likely that the role of caregivers is to facilitate the child's development of speech perception by providing not only rich auditory input (Yoshikawa, Asada, et al., 2003; Yoshikawa, Koga, et al., 2003; Miura et al., 2012; Messum & Howard, 2015; Rasilo & Räsänen, 2017), but also semantic contexts that necessitate the phonological contrasts. If this speculation is valid, modelling the development of perceptual space through social interaction would be an interesting topic for future studies.

4.6. Broader implications and caveats

The findings of the present study do not rule out that alternative models (e.g., those based on the Task Dynamic model or the DIVA model) may achieve similar or even better results, but they do show the importance of simulating vocal learning *end-to-end,* i.e., from audible input speech to audible output speech, as this is an effective way to reveal hidden weaknesses in theoretical assumptions, and to discover unforeseen alternatives, as happened during the course of this study. In fact, given the increasing availability of computational models, theories of not only vocal learning, but also other aspects of speech should be expected to be scrutinised by modelling simulations in addition to behaviour experiments and theoretical validations.

The current results have also demonstrated the benefit of combining a domain-specific deterministic model (SDSSTA) with data-driven machine learning algorithms (simulated annealing and deep learning) that may be analogous to neural processing in the brain. This finding may be relevant not only for the understanding of the brain, but also for artificial intelligence (AI) such as large language models. In the latter case, despite the impressive successes, a massive amount of data is needed for the training, much more than what a typical human individual receives when acquiring a language. The effectiveness of our modelling shows that language learning can benefit from having deterministic physical laws as a built-in component of the learning system, which would eliminate the need to explore articulatory trajectories that deviate from physical laws, thus saving a huge amount of resources during learning.

The paradigm under which the present study is conducted is still in its early stage of development, however. The present simulation starts from the developmental stage when the SDSSTA mechanism is already in place, presumably as the basic properties of canonical babbly as mentioned earlier. But it takes a child 6–7 months to reach that stage, and so the emergence of canonical babbling from non-canonical babbling (Oller et al., 2019, 2021) needs to be simulated in future studies. Also, continuous speech involving fast syllabic succession may require skills beyond those of isolated syllables (Schiller et al., 1997; Levelt et al., 1999), which may need new modelling strategies to simulate. Additional research is also needed to better understand how children may become proficient in using somatosensory and auditory feedback to perform online (Tremblay et al., 2003; Xu et al., 2004) or offline (Houde & Jordan, 1998) corrections and whether such corrective manoeuvres play any role in vocal learning.

CRediT authorship contribution statement

Anqi Xu: Writing – review & editing, Writing – original draft, Methodology, Conceptualization. **Daniel R. van Niekerk:** Writing – review & editing, Writing – original draft, Methodology, Conceptualization. **Branislav Gerazov:** Writing – review & editing, Writing – original draft, Methodology, Funding acquisition, Conceptualization. **Paul Konstantin Krug:** Writing – original draft, Methodology, Conceptualization. **Peter Birkholz:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Funding acquisition, Conceptualization. **Santitham Prom-on:** Writing – original draft, Methodology, Funding acquisition, Conceptualization. **Lorna F. Halliday:** Funding acquisition, Conceptualization. **Yi Xu:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization.

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.

Acknowledgments

This work was supported in part by the Leverhulme Trust Research Project Grant RPG-2019-241: 'High quality simulation of early vocal learning'. The paper is based on the first author's PhD dissertation.

Appendix

Table A1
Summary table of human vocal learning models.

Model	Motor control	Synthesiser	Sensory control	Learning strategy	Learning target	Performance
HABLAR (Bailly, 1997)	8 articulator parameters (lip, jaw, tongue, apex positions); Consonant goal imposed on the vowel goal to simulate coarticulation	Articulatory-to-acoustic model (Beautemps et al., 1996)	F1, F2, F3, F4 + lip area trajectories	Speech Maps:audio-visual-to-articulatory inversion (Abry et al., 1994)	Single vowel, vowel sequences, VCV	NA
(de Boer, 2000)	3 vocal tract parameters: tongue position, tongue height, and lip rounding	Maeda synthesiser (Maeda, 1990)	Bark-scale F1, F2, F3, F4	Self-organisation, imitation	Vowels	NA
(Westerman & Miranda, 2002; Westermann & Miranda, 2004)	3 glottis parameters + 3 vocal tract parameters	Pipe synthesiser	Auditory map: F1 and F2; Visual information	Sensorimotor integration, Hebbian connections	Vowels	NA
(Heintz et al., 2009)	12 articulatory parameters	VLAM (vocal linear articulatory model): modified Maeda synthesiser (Maeda, 1990)	Vectors derived from F1, F2, F3	Self-organising maps and Hebbian connections	Point vowels / a, i, u/	NA
(Kanda et al., 2009)	7 vocal tract parameters	Maeda synthesiser (Maeda, 1990)	5-dimensional vectors from low-third to low-seventh dimension out of 12-dimensional MFCCs;F0 analysis by STRAIGHT (Kawahara et al., 1999)	Self-organisation, recurrent neural network with parametric bias (RNNPB) (Tani, 2002)	Vowels	NA
(Huckvale & Howard, 2005)	9 articulatory parameters	VTCALCS (Maeda synthesiser (Maeda, 1990))	F1, F2	Distal supervised learning	Sentences containing vowels and consonants	The sound spectrogram shows that the vowel quality is good but the consonant quality is poor
KLAIR (Huckvale, 2011b, 2011a; Huckvale et al., 2009)	6 vocal tract parameters (Huckvale et al., 2009); 8 vocal tract parameters + 4 glottis parameters (Huckvale, 2011a, 2011b)	KLAIR's Synthesiser: infant-sized Maeda synthesiser (Maeda, 1990)	Adult reformulations	Online infant-caregiver interaction (caregiver imitate infant)	Words including vowels and consonants	NA
(Guenther, 1994; Guenther et al., 2006; Tourville & Guenther, 2011)	8 articulators	modified Maeda synthesiser (Maeda, 1990)	Auditory feedback: F1, F2, F3; Somatosensory feedback: 22-dimensional vector	Neurobiological modelling, neural networks	CVC (Guenther, 1994); VV, CV, CVCV (Guenther et al., 2006)	NA

Table A1 (*continued*)

Model	Motor control	Synthesiser	Sensory control	Learning strategy	Learning target	Performance
(Yoshikawa, Asada, et al., 2003; Yoshikawa, Koga, et al., 2003)	5 motor controllers	Source-filter model	Formant vector	Caregiver's imitation of infant speech	Four vowels: /a, i, u, e/	NA
(Ishihara et al., 2009; Miura et al., 2007)	6 vocal tract parameters (Miura et al., 2007); NA (Ishihara et al., 2009);	Source-filter model (Miura et al., 2007); NA (Ishihara et al., 2009)	Social feedback: F1 and F2 of caregiver's imitation of infant speech	Caregiver's imitation of infant speech	Five vowels: /a, i, u, e,o/(Miura et al., 2007); Vowels (Ishihara et al., 2009)	NA
(Miura et al., 2012)	NA	NA	Social feedback: F1 and F2 of caregiver's imitation of infant speech	Auto-mirroring bias (AMB): less imitative caregiver	Five vowels: /a, i, u, e, o/	NA
(Lyon et al., 2012)	NA	eSpeak synthesiser (Aslin et al., 1996)	Auditory feedback: Microsoft SAPI 5.4 (Phoneme recogniser); Social feedback: Teacher's positive/ negative feedback	Human-robot interaction	V, CV, VC and CVC	NA
(Prom-On et al., 2014a, 2014b)	18 vocal tract parameters	VocalTractLab (Birkholz, 2013)	MFCCs	Distal learning, Gradient descent	Thai vowels	Good vowel quality
(Kröger et al., 2009)	270 proto-vocalic states	Articulatory vectors generated by VocalTractLab (Birkholz, 2013)	Auditory feedback: Bark-scale F1, F2, F3; Somatosensory feedback: Vocal tract state	Neurobiological modelling	V, VC and CV	NA
(Kröger et al., 2014)	Motor plan states	Articulatory vectors generated by VocalTractLab (Birkholz, 2013)	Bark-scaled spectrogram Somatosensory feedback: Vocal tract state	Self-organising maps and Hebbian connections	50 CV syllables	78% transcription by one phonetician

(*continued on next page*)

Table A1 (*continued*)

Model	Motor control	Synthesiser	Sensory control	Learning strategy	Learning target	Performance
Elija (Howard & Messum, 2007, 2014; Howard, 2011; Messum & Howard, 2015)	2 vocal tract parameters for young infant, 7 vocal tract parameters for old child + 2 glottis parameters (Howard & Messum, 2007); 7 vocal tract parameters + 2 glottis parameters (Howard & Messum, 2014, 2011; Messum & Howard, 2015); Task dynamics model (Fowler & Saltzman, 1993; Saltzman & Munhall, 1989)	VTCALCS (modified Maeda synthesiser (Maeda, 1990))	Sensory salience: spectral change and low frequency power (Howard & Messum, 2007); Template-based dynamic time warping (Howard & Messum, 2011); Gammatone spectrogram (Howard & Messum, 2014; Messum & Howard, 2015); Social feedback: Caregiver's reformulation of infant speech	Caregiver's imitation of infant speech	NA (Howard & Messum, 2007) CV, VC, or CVV (Howard & Messum, 2011); VV, CV, VC and CVV (Howard & Messum, 2014; Messum & Howard, 2015)	NA (Howard & Messum, 2007, 2011); Synthetic samples are provided. Good vowel quality; Consonants are unintelligible (no trace of consonant burst or frication) (Howard & Messum, 2014; Messum & Howard, 2015);
(Murakami et al., 2015)	16 vocal tract parameters	VocalTractLab (Birkholz, 2013)	Auditory reservoir generated by BRIAN neural network simulator (Fontaine et al., 2011; Lopez-Poveda & Meddis, 2001); Visual input Caregiver's judgment as the reward (Warlaumont, 2012);	Reinforcement learning	Vowels	NA
(Warlaumont et al., 2013; Warlaumont & Finnegan, 2016)	Jaw and lips (Warlaumont, 2012); Lungs, trachea, larynx, pharynx, oral cavity, and nasal cavity (Warlaumont et al., 2013; Warlaumont & Finnegan, 2016);	Praat synthesis of a female vocal tract Muscle activations controlled by a spiking neural network (Maass, 1997);	Mel-scale F0, F1 and F2 (Warlaumont et al., 2013) Estimated auditory salience (Coath et al., 2009) (Warlaumont & Finnegan, 2016);	Reinforcement learning	VCV sequences (Warlaumont, 2012); Vowels (Warlaumont et al., 2013); Single consonant and consonant clusters (Warlaumont & Finnegan, 2016)	NA (Warlaumont, 2012); NA (Warlaumont et al., 2013); Synthetic samples are provided. No trace of consonants in the spectrogram (Warlaumont & Finnegan, 2016)
LeVI (Rasilo & Räsänen, 2017)	9 vocal tract parameters	Rasilo's Articulatory model	Auditory feedback:11 MFCCs without energy (Rasilo et al., 2013); F1, F2 (Rasilo & Räsänen, 2017); Social feedback: Phase1: positive/negative feedback Phase 2: imitation of infants' babbles by caregivers	Caregiver's imitation of infant speech	VCVC sequences containing all 25 Finnish phonemes (Rasilo et al., 2013); CVCV sequences (Rasilo & Räsänen, 2017);	Synthetic samples are provided (unintelligible) LeVI (Rasilo et al., 2013); Synthetic samples are provided (Vowels are not clear; Consonants are unintelligible and no trace of consonant burst/frication) (Rasilo & Räsänen, 2017);

Table A1 (*continued*)

Model	Motor control	Synthesiser	Sensory control	Learning strategy	Learning target	Performance
(Najnin & Banerjee, 2017)	11 vocal tract parameters + 2 glottis parameters	DIVA model (Guenther, 1994; Guenther et al., 2006; Tourville & Guenther, 2011)/modified Maeda synthesiser (Maeda, 1990)	F1, F2, F3 + phonation level + 12 MFCCs	Self-organisation	NN, CN, NC, VN, NV, VV, CV, VC, CC sequences	NA
(Forestier & Oudeyer, 2017)	7 vocal tract parameters	DIVA model (Guenther, 1994; Guenther et al., 2006; Tourville & Guenther, 2011)/modified Maeda synthesiser (Maeda, 1990); Dynamic Movement Primitives (DMPs) (Schaal, 2006) for controlling the articulatory trajectories	Auditory feedback: F1, F2 Social feedback: Simulated caregiver's guidance through objects Sensory feedback: state of the environment including the position of the caregiver, the stick and the toys	Goal-babbling	Vowel sequences including /o, u, i, e, y/	NA
(Cohen & Billard, 2018)	10 words	NA	PerAc (perception/action) architecture (Boucenna et al., 2010; Gaussier & Zrehen, 1995)	Caregiver-infant interaction through objects	CVCV sequences	NA
(Oudeyer, 2005)	3 vocal tract parameters	de Boer's synthesiser (Abstract liner articulatory synthesiser)	Perceptual representations based on Bark-scale F1, F2, F3, F4	Sensory motor interaction	Vowels	NA
(Moulin-Frier & Oudeyer, 2012)	7 vocal tract parameters	VLAM (vocal linear articulatory model): modified Maeda synthesiser (Maeda, 1990)	Bark-scale F1, weighted average of F2 and F3	Goal-babbling Random motor exploration Random goal selection with reaching Curiosity-driven active goal selection with reaching	Five vowels: /a, i, u, e,o/	NA
(Moulin-Frier et al., 2014)	7 parameters based on the PCA of the vocal tract shapes; over-damped spring-mass model for dynamic control	DIVA model (Guenther, 1994; Guenther et al., 2006; Tourville & Guenther, 2011)/modified Maeda synthesiser (Maeda, 1990)	Scaled F1, F2, intensity	Goal-babbling	VV, VC, CV, CC	NA
(Moulin-Frier et al., 2015)	Jaw, tongue body, tongue dorsum, lip protrusion, tongue tip, lip separation, larynx height	VLAM (vocal linear articulatory model): modified Maeda synthesiser (Maeda, 1990)	Bark-scale F1, F2, F3 (Moulin-Frier et al., 2015); Bark-scale F1, F2 (Barnaud et al., 2019)	Bayesian modelling	VV, VC, CV, CC (Moulin-Frier et al., 2015); Vowels (Barnaud et al., 2019)	NA

(*continued on next page*)

Table A1 (*continued*)

Model	Motor control	Synthesiser	Sensory control	Learning strategy	Learning target	Performance
(Acevedo-Valle et al., 2018)	NA	DIVA model (Guenther, 1994; Guenther et al., 2006; Tourville & Guenther, 2011)/modified Maeda synthesiser (Maeda, 1990)	Auditory feedback: F1, F2; Somatosensory feedback: proprioceptive input	Reinforce learning through auditory and somatosensory feedback GMMs (Gaussian mixture models)	NA (Acevedo-Valle et al., 2018)	NA
(Acevedo-Valle et al., 2017, 2020)	7 vocal tract parameters + 2 glottis parameters (Acevedo-Valle et al., 2018); 10 vocal tract parameters + 3 glottis parameters (Acevedo-Valle et al., 2020)	DIVA model (Guenther, 1994; Guenther et al., 2006; Tourville & Guenther, 2011)/modified Maeda synthesiser (Maeda, 1990)	Auditory feedback: F1, F2; Somatosensory feedback: proprioceptive input	Caregiver's imitation of infant speech: GMMs (Gaussian mixture models)	NA (Acevedo-Valle et al., 2018); Vowel sequences containing 17 German vowels (Acevedo-Valle et al., 2020)	NA
(Philippsen et al., 2014)	22 vocal tract parameters + 4 glottis parameters	VocalTractLab (Birkholz, 2013)	39 MFCCs (energy, 12MFCCs, first and second derivatives)	Distal supervised learning by acoustic imitation (Echo State Network)	CV sequences containing 8 vowels and 8 consonants	Perceptual evaluation was conducted by the authors
(Philippsen et al., 2016)	20 vocal tract parameters	VocalTractLab (Birkholz, 2013)	F1, F2, F3 + 39 MFCCs (energy, 12MFCCs, first and second derivatives) projected by Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) to 10-D features	Goal babbling (exploration and adaptation)	6 vowels	NA
(Philippsen, 2021)	18 vocal tract parameters + 3 glottis parameters	VocalTractLab (Birkholz, 2013) Dynamic Movement Primitives (DMPs) (Schaal, 2006) for controlling the articulatory trajectories	Echo State Network (ESN) 10-D vectors were based on F1, F2, F3 + 39 MFCCs (energy, 12MFCCs, first and second derivatives), then PCA and LDA were applied	Goal babbling	6 vowels, /baa/ and /maa/	Good vowel and consonant quality

Fig. A1. Confusion matrices of words learned by adult vocal tract model when guided by MFCCs and Log Mel spectrograms, evaluated by a word recogniser. The score shows the weighted negative log likelihood loss. The two types of acoustic features resulted in very similar word error rates, that is, the intelligibility of the trained speech is indistinguishable.

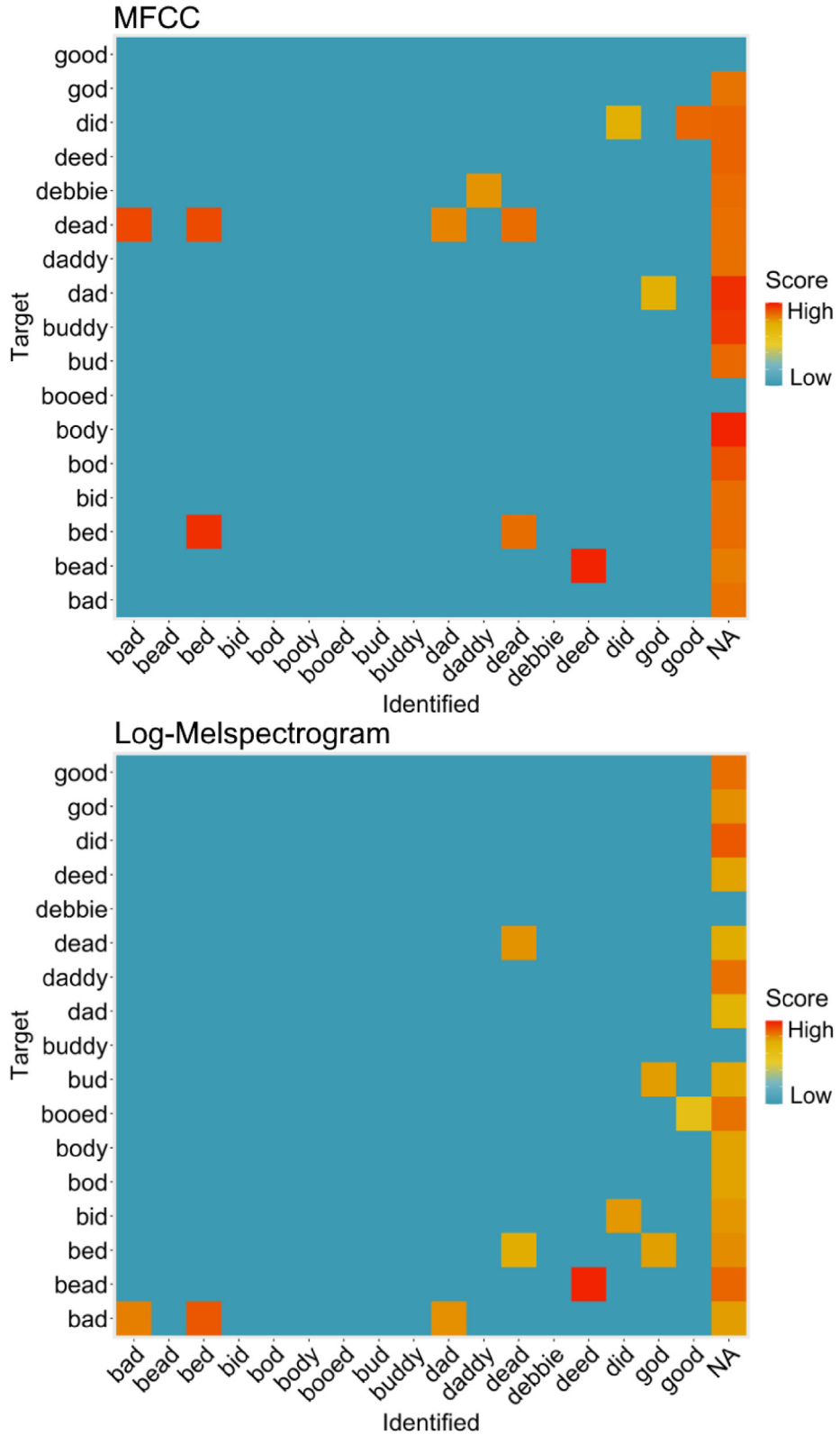


Fig. A2. Confusion matrices of consonants trained by MFCCs and Log Mel spectrograms, evaluated by a word recogniser. The score shows the weighted negative log likelihood loss. Higher scores represent higher phoneme probability. The bilabial stops trained by Log Mel spectrograms had fairly high accuracies, while the alveolar stops trained by MFCCs had higher accuracies. However, the velar stops trained by Log Mel spectrograms were identified as /n/ or /w/. Overall, the speech trained by Log Mel spectrograms were slightly better identified than that trained by MFCCs.

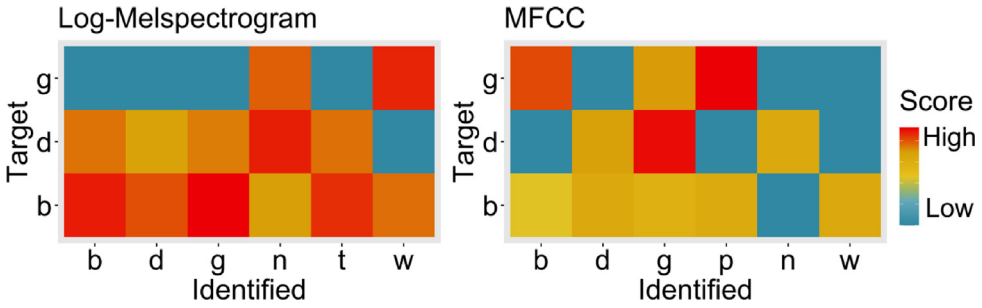


Fig. A3. Confusion matrices of vowels trained by MFCCs and Log Mel spectrograms, evaluated by a word recogniser. The score shows the weighted negative log likelihood loss. Higher scores represent higher phoneme probability. The recognition accuracies were higher for vowels trained by MFCCs than the ones trained by Log Mel spectrograms. Both acoustic features failed to guide the learning of intelligible vowels in 'booed' and 'good'. For the rest of the vowel categories, those trained by MFCCs were identified more correctly than the ones trained by Log Mel spectrograms. Especially, the vowel in 'bud' was less successful when trained by Log Mel spectrograms.

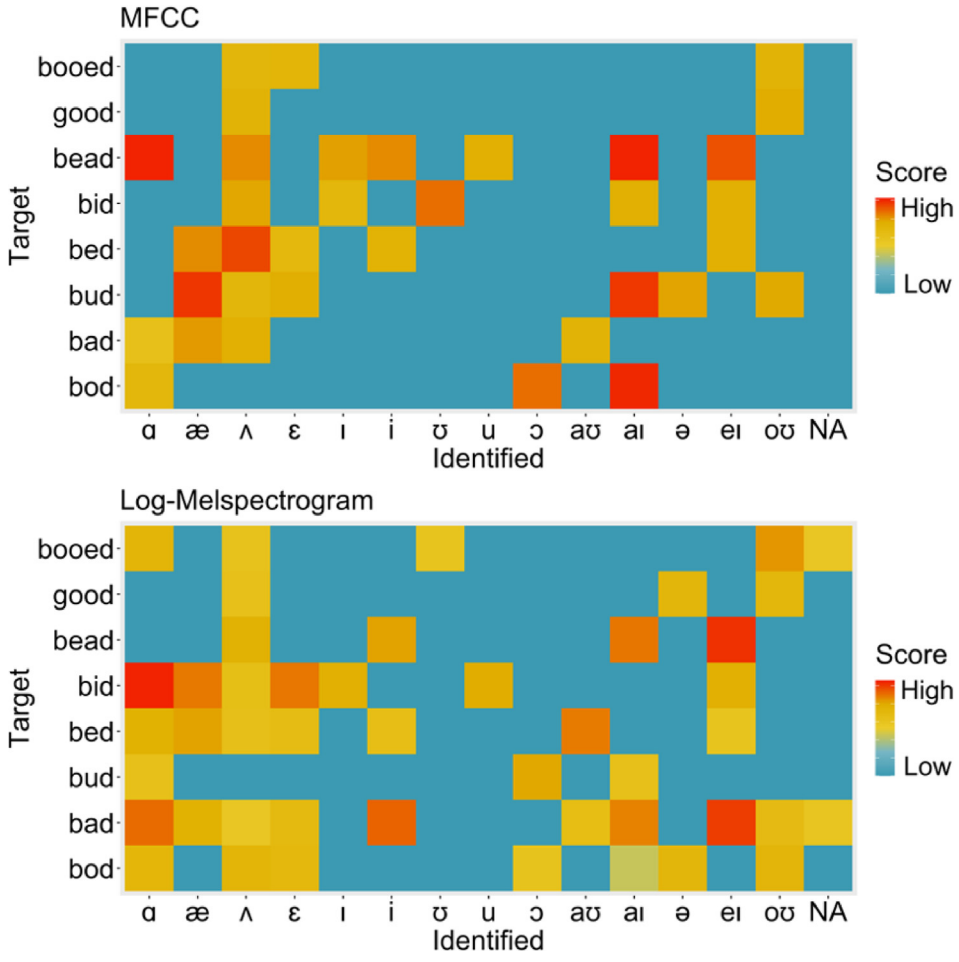


Fig. A4. Schematic diagram of the speech recogniser. The network contains 8 convolutional layers (conv), 6 long short-term memory (LSTM) layers and 3 dense layers (Dense). Batch normalisation layers after each conv, LSTM and Dense layers and dropout layers after each LSTM and Dense layers are not shown in the diagram. The architecture consists of 3 main parts: spectrotemporal feature processing, temporal feature processing and classification. The first convolutional layer module (in blue) was designed to learn the feature representations that may coexist or correlate in the spectral and temporal domains. The temporal feature processing module (in green) was designed to learn the temporal dependency and/or the state-based behaviour of the feature. Lastly, the classification module (in red) was used to learn the relationship between the features for classifying phonemes.

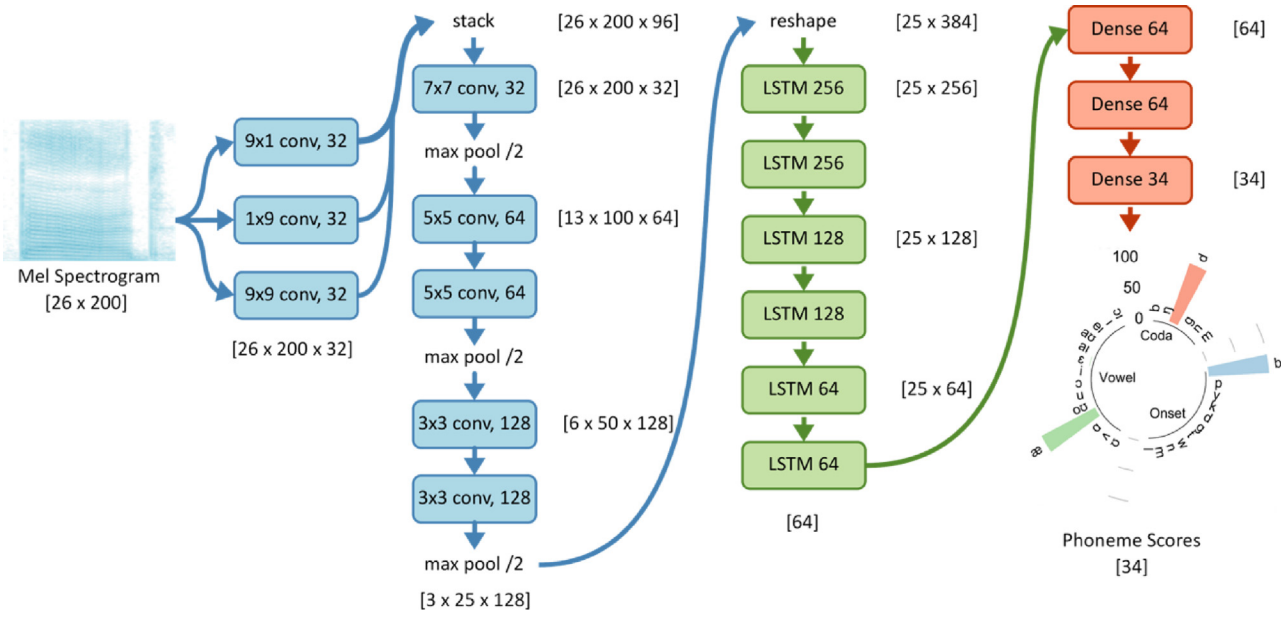


Fig. A5. Learning curves showing iterations with a lowering in the recognition error during the first stage of optimisation (random search).

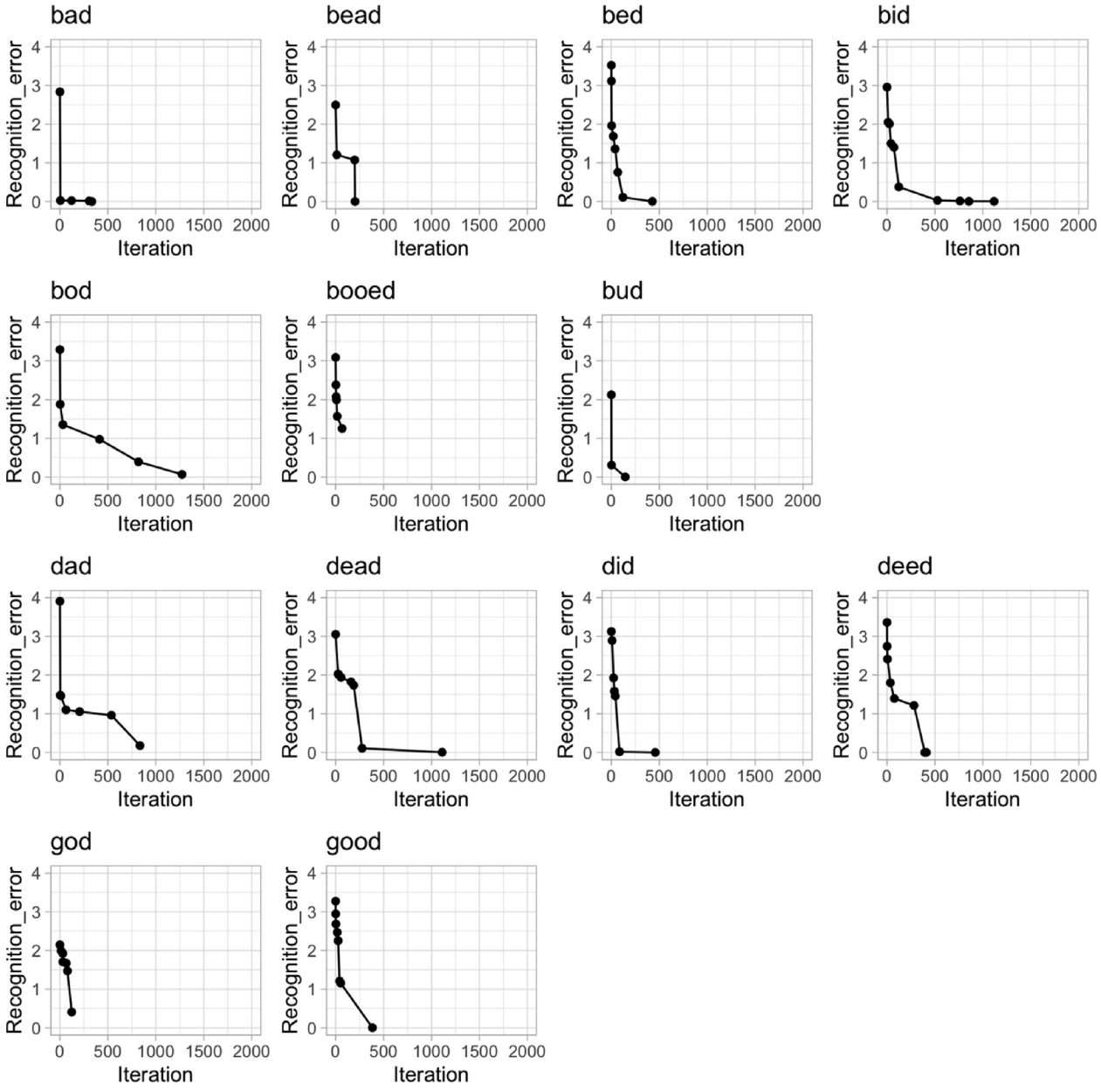
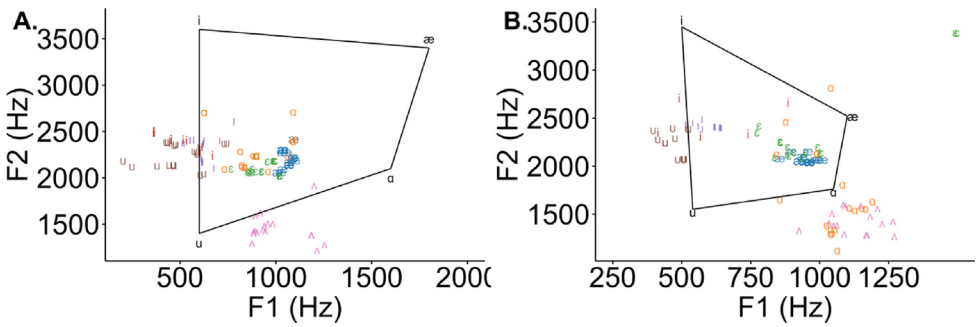


Fig. A6. (A & B) F1 and F2 of twenty vowels with lowest recognition errors learned by 1-year-old model (A) and 3-year-old model (B). The vowel quadrilateral is formed based on the average F1 and F2 across studies (Vorperian & Kent, 2007).

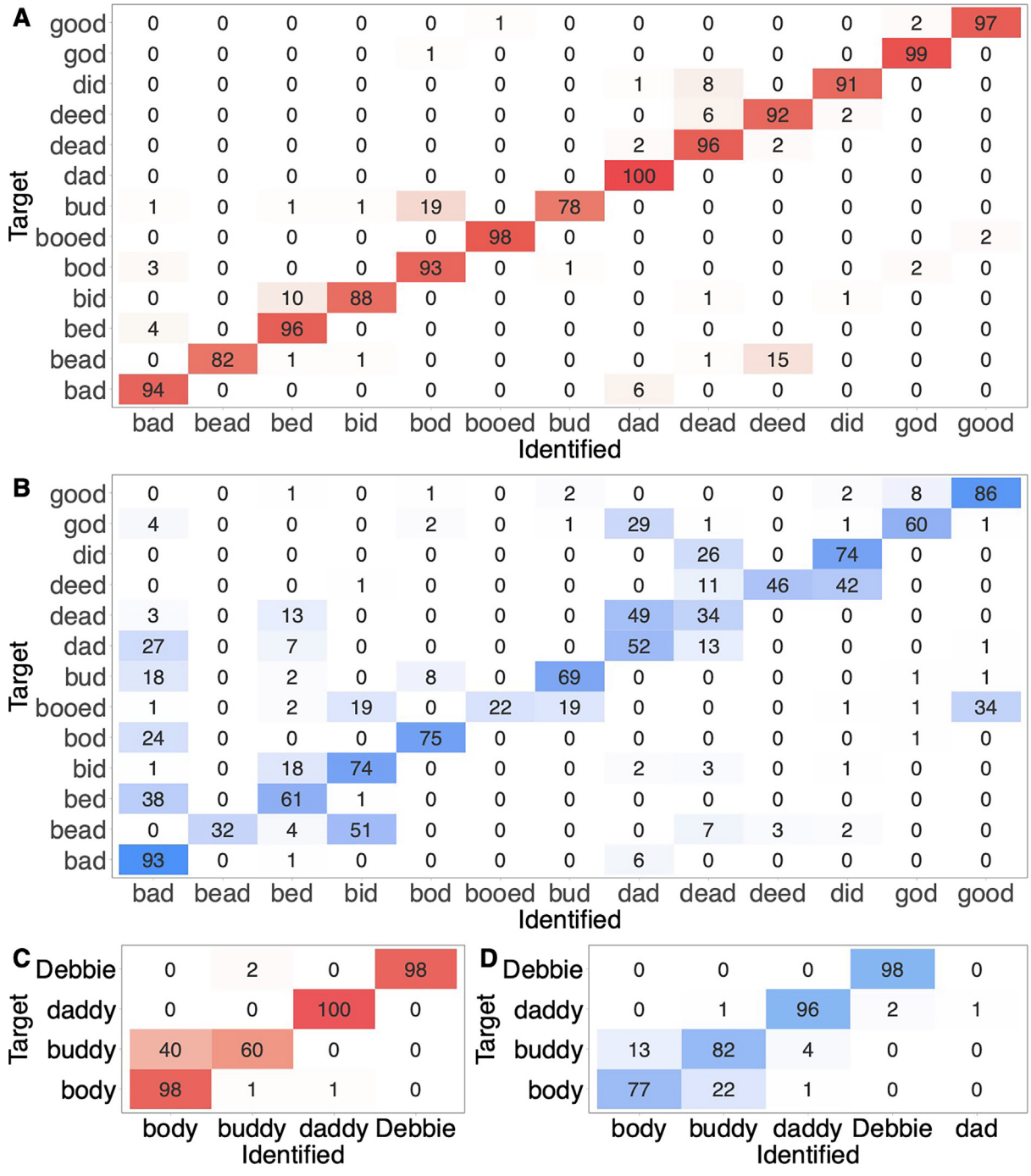


Additional analysis

Adult model

Most of the bilabial stops were correctly recognised but alveolar stops were sometimes identified as bilabial stops in the multiple-choice task (Fig. A7A, B). The velar stops, however, were often identified as alveolar stops. The accuracy of the onset consonant was sometimes lower before certain vowels. Compared with the high identification rate of 'bad' (i.e., 99%), only less than half of the mid vowels in 'bed' and 'bid' were correctly identified. Nevertheless, even for the natural speech some vowels were sometimes misidentified (Fig. A7A, B). In addition to mid vowels, the model had difficulty learning 'booed'. The coda accuracy was stable across vowel contexts with only one exception in the word 'bud', which was frequently heard as 'but'.

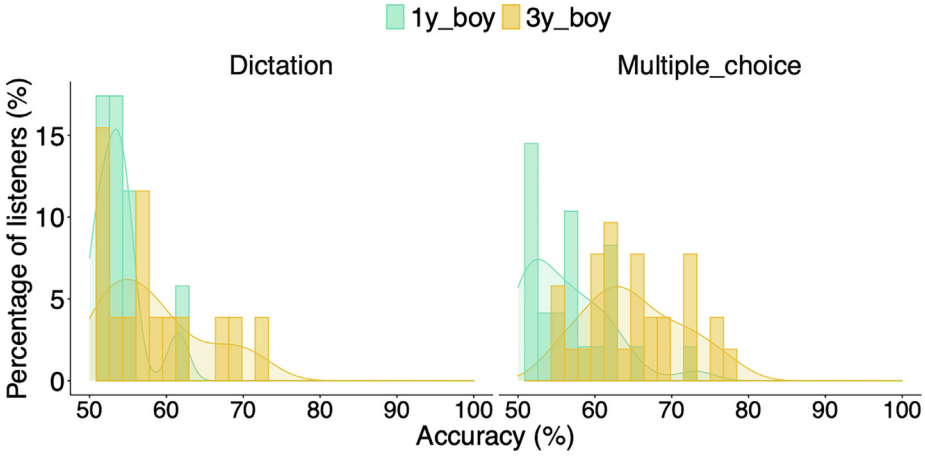
Fig. A7. Confusion matrix of natural female speech and synthetic male speech in the listening experiment. (A, B) Confusion matrices (%) of CVC words produced by a female American English speaker (A) and learned by an adult male vocal tract model (B), measured in the multiple-choice task. 30 American English listeners identified the utterances by selecting words from a fixed set of words. (C, D) Confusion matrices of CVCV words produced by the native speaker (C) and generated with the vocal tract parameters learned by the adult vocal tract model (D).



Child models

We analysed the by-phoneme accuracy of the speech learned by the 1-year-old and 3-year-old vocal tract model in the two types of listening experiments. Fig. A8 shows the mean phoneme accuracy rate of the CV syllables in the CVC words learned by the two child vocal tract models. As can be seen from the plot, the distribution of the two child vocal tract models overlaps greatly in the open-vocabulary dictation. Still, the 3-year-old vocal tract model had an overall higher mean accuracy rate than the 1-year-old model. The 3-year-old model had a mean phoneme accuracy rate of 49% for the target CV syllables in the open-vocabulary task, while the 1-year-old model had a mean accuracy rate of 44%. The mean phoneme accuracies in the multiple-choice task were 65% and 55% for the 3-year-old and 1-year-old models, respectively. Wilcoxon signed-rank tests showed that the 3-year-old vocal tract model had higher phoneme accuracy rate than the 1-year-old model in the multiple-choice task ($p < .001$), but not in the open-vocabulary task ($p = 0.120$). The type of tasks did not influence the phoneme accuracies for either of the child vocal tract models (Wilcoxon signed-rank: $p = 1.000$).

Fig. A8. Distribution of by-listener mean phoneme identification accuracy of CVC words learned by the 1-year-old and 3-year-old vocal tract models, obtained from the listening experiment. Kernel density estimate and histogram show the distribution of the performance of the listeners.

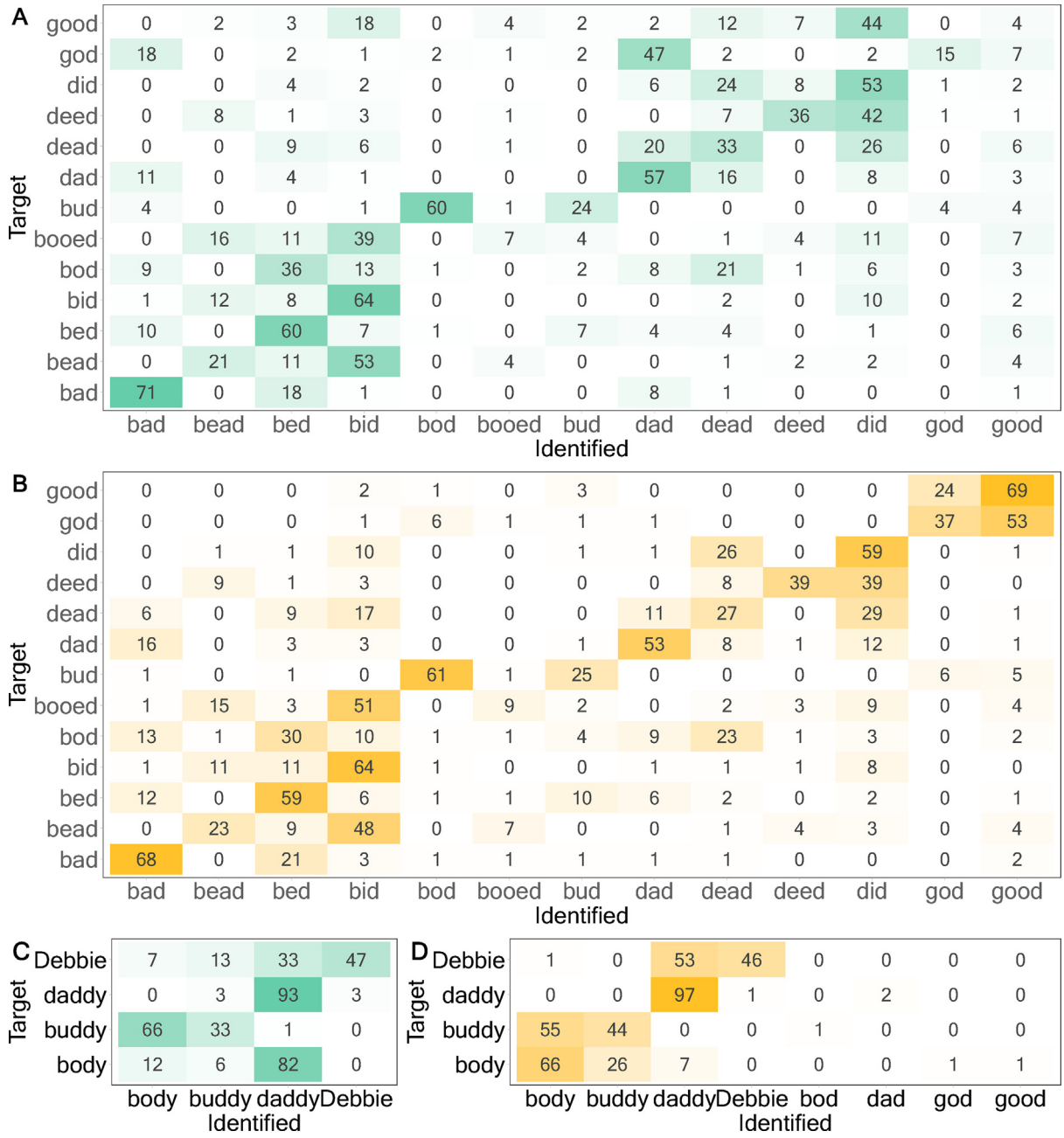


When the word list was given, listeners could identify the words learned by both models more easily. There were less variances in the phoneme accuracy of the multiple-choice task than the open-vocabulary dictation task. The median phoneme accuracies increased in the onset consonant position (Wilcoxon signed-rank: 1y: $p < .001$, 3y: $p < .001$), which was 85% and 72% for the 1-year-old model and 3-year-old model respectively. There was improvement in the vowel accuracy as well (Wilcoxon signed-rank: 1y: $p = 0.017$, 3y: $p = 0.003$). The median vowel accuracy was 44% and 38% for the 3-year-old model and 1-year-old model respectively.

The confusion matrices of the CVC words of the two child vocal tract models in the multiple-choice task are shown in Figs. A9A, B. The child vocal tract model learned relatively intelligible vowels in 'bad' (1y: 71%; 3y: 68%), 'bed' (1y: 60%; 3y: 59%), and 'bid' (1y: 64%; 3y: 64%). The bilabial stops were rarely mistaken as other types of consonants except for the one in 'bod', which was sometimes mistaken as an alveolar stop. The place of articulation of alveolar stops was almost always correctly identified for both child vocal tract models. The learning of velar stops was relatively successful for the 3-year-old vocal tract model but not for the 1-year-old model. The velar stops learned by the 1-year-old model were often identified as alveolar stops and bilabial stops. Only a very small proportion of velar stops was correctly identified (4% in 'good' and 22% in 'god') for the 1-year-old model. In contrast, the velar stops learned by the 3-year-old model had fairly high accuracy, which was 89% in 'god' and 93% in 'good'.

Both models had difficulty in learning vowels with similar openness and tongue height. For instance, 'bead' was often mistaken as 'bid', and 'bud' as 'bod'. The learning of the vowel /ɒ/ in 'bod' was unsuccessful for both child models, which was heard as /ɪ/ in 'bid'. The rounded vowel /u/ was difficult for both child vocal tract models. Compared with the 3-year-old model, the 1-year-old model learned much less intelligible vowels following velar stops in 'god' and 'good'. Only 7% was correctly identified for the 1-year-old model and 9% for the 3-year-old model.

Fig. A9. Comparison between the two child models. (A, B) Confusion matrices (%) of CVC words learned by the 1-year-old child (A) and 3-year-old child's (B) vocal tract models, measured in a multiple-choice task. 30 American English listeners identified utterances by selecting words from a fixed set of words. (C, D) Confusion matrices of CVCV words regenerated by learned vocal tract parameters of the 1-year-old child (C) and 3-year-old boy's (D) vocal tract models, measured in a multiple-choice task.



In the open-vocabulary dictation task, the mean identification accuracy of CVCV words was 64% for the 1-year-old vocal tract model and 78% for the 3-year-old model. The 3-year-old model achieved a significantly higher identification rate than the 1-year-old model in the open-vocabulary dictation task (Wilcoxon signed-rank: $p < .001$). Furthermore, the mean accuracy of CVCV words was 76% for the 1-year-old model and 86% for the 3-year-old model in the multiple-choice task. The learning performance of the two child vocal tract models was significantly different in the multiple-choice task as well (Figs. A9C, D, Wilcoxon signed-rank: $p < .001$).

Task types

We have also explored how task types can influence the intelligibility of the synthetic speech. The identification accuracy was higher in the multiple-choice task than in the open-vocabulary task (W = 2534, $p < .001$, Wilcoxon signed-rank test), as shown in Fig. 9A. Post-hoc comparisons showed that the type of listening task influenced the perception of words learned by the child models but not by the adult model (1y: $p < .001$, 3y: $p = .002$, adult: $p = .570$).

References

Abry, C., Badin, P., & Scully, C. (1994). Sound-to-gesture inversion in speech: The Speech Maps approach. In K. Varghese, S. Pfleger, & J. Lefevre (Eds.), *Advanced speech applications* (pp. 182–196). Germany: Berlin.

Acevedo-Valle, J. M., Angulo, C., & Moulin-Frier, C. (2018). Autonomous discovery of motor constraints in an intrinsically motivated vocal learner. *IEEE Transactions on Cognitive and Developmental Systems, 10*(2), 314–325. https://doi.org/10.1109/TCDS.2017.2699578.

Acevedo-Valle, J. M., Hafner, V. v., & Angulo, C. (2017). Social reinforcement in intrinsically motivated sensorimotor exploration for embodied agents with constraint awareness. In *2017 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, 255–262. https://doi.org/10.1109/DEVLRN.2017.8329815.

Acevedo-Valle, J. M., Hafner, V. V., & Angulo, C. (2020). Social reinforcement in artificial prelinguistic development: A study using intrinsically motivated exploration architectures. *IEEE Transactions on Cognitive and Developmental Systems, 12*(2), 198–208. https://doi.org/10.1109/TCDS.2018.2883249.

Asada, M. (2016). Modeling early vocal development through infant-caregiver interaction: A review. *IEEE Transactions on Cognitive and Developmental Systems, 8*(2), 128–138. https://doi.org/10.1109/TCDS.2016.2552493.

Aslin, R. N., Woodward, J. Z., LaMendola, N. P., & Bever, T. G. (1996). Models of word segmentation in fluent maternal speech to infants. In J. Morgan & K. Demuth (Eds.), *Signal to Syntax*. Lawrence Erlbaum.

Bailly, G. (1997). Learning to speak. Sensori-motor control of speech movements. *Speech Communication, 22*(2–3), 251–267. https://doi.org/10.1016/S0167-6393(97)00025-3.

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.

Barry, W. A., & van Dommelen, W. A. (2005). *The integration of phonetic knowledge in speech technology* (W. J. Barry & W. A. van Dommelen, Eds.; Vol. 25). Springer Netherlands. https://doi.org/10.1007/1-4020-2637-4.

Bateman, N. (2007). *A Crosslinguistic Investigation of Palatalization* [University of California, San Diego]. https://escholarship.org/uc/item/13s331md.

Beautemps, D., Badin, P., Bailly, G., Galvan, A., & Laboissiere, R. (1996). Evaluation of an articulatory-acoustic model based on a reference subject. In *Proceedings of 1st ESCA Tutorial and Research Workshop on Speech Production Modeling* (pp. 45–48).

Bernstein, N. A. (1967). *The co-ordination and regulation of movements*. Pergamon Press.

Birkholz, P. (2013). Modeling consonant-vowel coarticulation for articulatory speech synthesis. *PLoS ONE, 8*(4), e60603.

Birkholz, P. (2014). Enhanced area functions for noise source modeling in the vocal tract. In *Proc. of the 10th International Seminar on Speech Production (ISSP 2014)*, 37–40. https://www.vocaltractlab.de/publications/birkholz-2014-issp.pdf.

Birkholz, P., & Kröger, B. J. (2007). Simulation of vocal tract growth for articulatory speech synthesis. In *Proc. of the 16th International Congress of Phonetic Sciences (ICPhS 2007)*, 377–380. www.icphs2007.de.

Birkholz, P., Kröger, B. J., & Neuschaefer-Rube, C. (2011). Model-based reproduction of articulatory trajectories for consonant-vowel sequences. *IEEE Transactions on Audio, Speech and Language Processing, 19*(5), 1422–1433. https://doi.org/10.1109/TASL.2010.2091632.

Boucenna, S., Gaussier, P., Andry, P., & Hafemeister, L. (2010). Imitation as a communication tool for online facial expression learning and recognition. *IEEE/RSJ International Conference on Intelligent Robots and Systems, 2010*, 5323–5328. https://doi.org/10.1109/IROS.2010.5650357.

Brass, M., & Heyes, C. (2005). Imitation: Is cognitive neuroscience solving the correspondence problem? *Trends in Cognitive Sciences, 9*(10), 489–495. https://doi.org/10.1016/j.tics.2005.08.007.

Browman, C. P., & Goldstein, L. (1990). Gestural specification using dynamically-defined articulatory structures. *Journal of Phonetics, 18*(3), 299–320. https://doi.org/10.1016/S0095-4470(19)30376-6.

Bruderer, A. G., Kyle Danielson, D., Kandhadai, P., & Werker, J. F. (2015). Sensorimotor influences on speech perception in infancy. *Proceedings of the National Academy of Sciences of the United States of America, 112*(44), 13531–13536. https://doi.org/10.1073/pnas.1508631112.

Carnegie Mellon University (2022). *The CMU Pronouncing Dictionary*. http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

Chang, E. F., Rieger, J. W., Johnson, K., Berger, M. S., Barbaro, N. M., & Knight, R. T. (2010). Categorical speech representation in human superior temporal gyrus. *Nature Neuroscience, 13*, 1428–1432. https://doi.org/10.1038/nn.2641.

Chen, Y., Gao, Y., & Xu, Y. (2022). Computational Modelling of Tone Perception Based on Direct Processing of f0 Contours. *Brain Sciences, 12*(3), 337. https://doi.org/10.3390/brainsci12030337.

Choi, D., Dehaene-Lambertz, G., Peña, M., & Werker, J. F. (2021). Neural indicators of articulator-specific sensorimotor influences on infant speech perception. *Proceedings of the National Academy of Sciences, 118*(20). https://doi.org/10.1073/pnas.2025043118 e2025043118.

Coath, M., Denham, S. L., Smith, L. M., Honing, H., Hazan, A., Holonowicz, P., & Purwins, H. (2009). Model cortical responses for the detection of perceptual onsets and beat tracking in singing. *Connection Science, 21*(2–3), 193–205. https://doi.org/10.1080/09540090902733905.

Cohen, L., & Billard, A. (2018). Social babbling: The emergence of symbolic gestures and words. *Neural Networks, 106*, 194–204. https://doi.org/10.1016/j.neunet.2018.06.016.

Cook, R., Bird, G., Catmur, C., Press, C., & Heyes, C. (2014). Mirror neurons: From origin to function. *Behavioral and Brain Sciences, 37*(2), 177–192. https://doi.org/10.1017/S0140525X13000903.

Crowe, K., & McLeod, S. (2020). Children's english consonant acquisition in the united states: A review. In *American Journal of Speech-Language Pathology* (Vol. 29, Issue 4, pp. 2155–2165). American Speech-Language-Hearing Association. https://doi.org/10.1044/2020_AJSLP-19-00168.

Davis, S. B., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing, 28*(4), 357–366. https://doi.org/10.1109/TASSP.1980.1163420.

de Boer, B. (2000). Self-organization in vowel systems. *Journal of Phonetics, 28*, 441–465. https://doi.org/10.006/jpho.2000.0125.

de Klerk, C. C. J. M., Johnson, M. H., Heyes, C. M., & Southgate, V. (2015). Baby steps: Investigating the development of perceptual-motor couplings in infancy. *Developmental Science, 18*(2), 270–280. https://doi.org/10.1111/desc.12226.

Doupe, A. J., & Kuhl, P. K. (1999). Birdsong and human speech: Common themes and mechanisms. *Annual Review of Neuroscience, 22*(1), 567–631. https://doi.org/10.1146/annurev.neuro.22.1.567.

Elie, B., Lee, D. N., & Turk, A. (2023). Modeling trajectories of human speech articulators using general Tau theory. *Speech Communication, 151*, 24–38. https://doi.org/10.1016/j.specom.2023.04.004.

Elowson, A. M., Snowdon, C. T., & Lazaro-Perea, C. (1998a). Infant 'Babbling' in a nonhuman primate: complex vocal sequences with repeated call types. *Behaviour, 135*(5), 643–664 www.jstor.org/stable/4535550.

Elowson, A. M., Snowdon, C. T., & Lazaro-Perea, C. (1998b). 'Babbling' and social context in infant monkeys: parallels to human infants. *Trends in Cognitive Sciences, 2*(1), 31–37. https://doi.org/10.1016/s1364-6613(97)01115-7.

Fadiga, L., Craighero, L., Buccino, G., & Rizzolatti, G. (2002). Speech listening specifically modulates the excitability of tongue muscles: a TMS study. *European Journal of Neuroscience, 15*(2), 399–402. https://doi.org/10.1046/j.0953-816x.2001.01874.x.

Fernandez, A. A., Burchardt, L. S., Nagy, M., & Knörnschild, M. (2021). Babbling in a vocal learning bat resembles human infant babbling. *Science, 373*(6557), 923–926. https://doi.org/10.1126/science.abf9279.

Fitch, W. T., & Giedd, J. (1999). Morphology and development of the human vocal tract: A study using magnetic resonance imaging. *The Journal of the Acoustical Society of America, 106*(3), 1511–1522. https://doi.org/10.1121/1.427148.

Fontaine, B., Goodman, D. F. M., Benichoux, V., & Brette, R. (2011). Brian hears: online auditory processing using vectorization over channels. *Frontiers in Neuroinformatics, 5*(9). https://doi.org/10.3389/fninf.2011.00009.

Forestier, S., & Oudeyer, P.-Y. (2017). A Unified Model of Speech and Tool Use Early Development. In *39th Annual Conference of the Cognitive Science Society (CogSci 2017)*. https://github.com/sebastien-forestier/CogSci2017.

Fowler, C. A., & Saltzman, E. (1993). Coordination and coarticulation in speech production. *Language and Speech, 36*(23), 171–195. https://doi.org/10.1177/002383099303600304.

Gaussier, P., & Zrehen, S. (1995). PerAc: A neural architecture to control artificial animals. *Robotics and Autonomous Systems, 16*(2–4), 291–320. https://doi.org/10.1016/0921-8890(95)00052-6.

Gerazov, B., van Niekerk, D., Xu, A., Krug, P. K., Birkholz, P., & Xu, Y. (2020). *Evaluating features and metrics for high-quality simulation of early vocal learning of vowels*.

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*(5), 1166–1183. https://doi.org/10.1037/0278-7393.22.5.1166.

Goldstein, U. G. (1980). *An articulatory model for the vocal tracts of growing children* [Massachusetts Institute of Technology]. https://dspace.mit.edu/handle/1721.1/22386.

Guenther, F. H. (1994). A neural network model Of speech acquisition and motor equivalent Speech production. *Biological Cybernetics, 72*, 43–53. https://doi.org/10.1007/BF00206237.

Guenther, F. H., Ghosh, S. S., & Tourville, J. A. (2006). Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and Language, 96*(3), 280–301. https://doi.org/10.1016/j.bandl.2005.06.001.

Heintz, I., Beckman, M., Fosler-Lussier, E., & Ménard, L. (2009). Evaluating parameters for mapping adult vowels to imitative babbling. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 688–691. https://doi.org/10.21437/interspeech.2009-238.

Heyes, C. (2001). Causes and consequences of imitation. *Trends in Cognitive Sciences, 5*(6), 253–261. https://doi.org/10.1016/S1364-6613(00)01661-2.

Houde, J. F., & Jordan, M. I. (1998). Sensorimotor adaptation in speech production. *Science, 279*(5354), 1213–1216. https://doi.org/10.1126/science.279.5354.1213.

Howard, I. S., & Huckvale, M. A. (2005). Training a vocal tract synthesizer to imitate speech using distal supervised learning. In *Proceedings of the 10th International Conference on Speech and Computer (SPECOM 2005)*, 159–162.

Howard, I. S., & Messum, P. (2007). A Computational Model of Infant Speech Development. In *Proceedings of XII International Conference 'Speech and Computer' (SPECOM'2007)*, 756–765.

Howard, I. S., & Messum, P. (2014). Learning to Pronounce First Words in Three Languages: An Investigation of Caregiver and Infant Behavior Using a Computational Model of an Infant. *PLoS ONE, 9*(10), e110334.

Howard, I. S., & Messum, P. (2011). Modeling the Development of Pronunciation in Infant Speech Acquisition. In *Motor Control* (Vol. 15).
Huckvale, M. (2011a). Recording caregiver interactions for machine acquisition of spoken language using the KLAIR virtual infant. In *Proceedings of Interspeech 2011*.
Huckvale, M. (2011b). The KLAIR toolkit for recording interactive dialogues with a virtual infant. In *Proceedings of Interspeech 2011*, 28–31.
Huckvale, M., & Howard, I. (2005). Teaching a vocal tract simulation to imitate stop consonants. In *Proceedings of Interspeech 2005*, 3213–3216.
Huckvale, M., Howard, I. S., & Fagel, S. (2009). KLAIR: a Virtual Infant for Spoken Language Acquisition Research. In *Proceedings of Interspeech 2009*.
Ijspeert, A. J., Nakanishi, J., Hoffmann, H., Pastor, P., & Schaal, S. (2013). Dynamical movement primitives: learning attractor models for motor behaviors. *Neural Computation, 25*(2), 328–373. https://doi.org/10.1162/NECO_a_00393.
Ishihara, H., Yoshikawa, Y., Miura, K., & Asada, M. (2009). How caregiver's anticipation shapes infant's vowel through mutual imitation. *IEEE Transactions on Autonomous Mental Development, 1*(4), 217–225. https://doi.org/10.1109/TAMD.2009.2038988.
Johnson, K. (2005). Speaker normalization in speech perception. In D. Pisoni & R. E. Remez (Eds.), *The handbook of speech perception* (pp. 363–389). Blackwell.
Jordan, M. I., & Rumelhart, D. E. (1992). Forward models: supervised learning with a distal teacher. *Cognitive Science, 16*(3), 307–354. https://doi.org/10.1016/0364-0213(92)90036-T.
Kanda, H., Ogata, T., Takahashi, T., Komatani, K., & Okuno, H. G. (2009). Continuous vocal imitation with self-organized vowel spaces in recurrent neural network. In *Proceedings – IEEE International Conference on Robotics and Automation*, 4438–4443. https://doi.org/10.1109/ROBOT.2009.5152818.
Kawahara, H., Masuda-Katsuse, I., & de Cheveigné, A. (1999). Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. *Speech Communication, 27*(3–4), 187–207. https://doi.org/10.1016/S0167-6393(98)00085-5.
Kent, R. D. (1992). The biology of phonological development. In C. A. Ferguson, L. Menn, & C. Stoel-Gammon (Eds.), *Phonological Development: Models, Research, Implications* (pp. 65–90). York Press.
Kent, R. D., & Murray, A. D. (1982). Acoustic features of infant vocalic utterances at 3, 6, and 9 months. *Journal of the Acoustical Society of America, 72*(2), 353–365. https://doi.org/10.1121/1.388089.
Keysers, C., & Perrett, D. I. (2004). Demystifying social cognition: a Hebbian perspective. *Trends in Cognitive Sciences, 8*(11), 501–507. https://doi.org/10.1016/j.tics.2004.09.005.
Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science, 220*(4598), 671–680. https://doi.org/10.1126/science.220.4598.671.
Kohonen, T. (2001). *Self-organizing maps.* Springer.
Konishi, M. (1965). The role of auditory feedback in the control of vocalization in the white-crowned sparrow. *Zeitschrift Für Tierpsychologie, 22*(7), 770–783. https://doi.org/10.1111/j.1439-0310.1965.tb01688.x.
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 Biomedical Physics, 2*(2) http://www.epjnonlinearbiomedphys.com/content/2/1/2.
Kröger, B. J., Kannampuzha, J., & Neuschaefer-Rube, C. (2009). Towards a neurocomputational model of speech production and perception. *Speech Communication, 51*(9), 793–809. https://doi.org/10.1016/j.specom.2008.08.002.
Krug, P. K., Birkholz, P., Gerazov, B., van Niekerk, D. R., Xu, A., & Xu, Y. (2023). Artificial vocal learning guided by phoneme recognition and visual information. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 31*, 1734–1744. https://doi.org/10.1109/TASLP.2023.3264454.
Kuhl, P. K. (1991). Human adults and human infants show a 'perceptual magnet effect' for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics, 50*(2), 93–107.
Kuhl, P. K. (2000). A new view of language acquisition. *Proceedings of the National Academy of Sciences, 97*(22), 11850–11857. https://doi.org/10.1073/pnas.97.22.11850.
Kuhl, P. K. (2003). Human speech and birdsong: Communication and the social brain. *Proceedings of the National Academy of Sciences, 100*(17), 9645–9646. https://doi.org/10.1073/pnas.1733998100.
Kuhl, P. K. (2004). Early language acquisition: Cracking the speech code. *Nature Reviews Neuroscience, 5*(11), 831–843. https://doi.org/10.1038/nrn1533.
Kuhl, P. K. (2007). Is speech learning 'gated' by the social brain? *Developmental Science, 10*(1), 110–120. https://doi.org/10.1111/j.1467-7687.2007.00572.x.
Kuhl, P. K., Conboy, B. T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., & Nelson, T. (2008). Phonetic learning as a pathway to language: New data and native language magnet theory expanded (NLM-e). *Philosophical Transactions of the Royal Society B: Biological Sciences, 363*(1493), 979–1000. https://doi.org/10.1098/rstb.2007.2154.
Kuhl, P. K., & Meltzoff, A. N. (1996). Infant vocalizations in response to speech: Vocal imitation and developmental change. *The Journal of the Acoustical Society of America, 100*(4 Pt 1), 2425–2438. https://doi.org/10.1121/1.417951.
Kuhl, P. K., Ramírez, R. R., Bosseler, A., Lin, J.-F.-L., & Imada, T. (2014). Infants' brain responses to speech suggest Analysis by Synthesis. *Proceedings of the National Academy of Sciences, 111*(31), 11238–11245. https://doi.org/10.1073/pnas.1410963111.
Larson, J., Menickelly, M., & Wild, S. M. (2019). Derivative-free optimization methods. *Acta Numerica, 28*, 287–404. https://doi.org/10.1017/S0962492919000060.
Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences, 22*(1), 1–75. https://doi.org/10.1017/S0140525X99001776.
Lieberman, P., Crelin, E. S., & Klatt, D. H. (1972). Phonetic ability and related anatomy of the newborn and adult human, neanderthal man, and the chimpanzee. *American Anthropologist, 74*(3), 287–307.
Liu, Z., Xu, Y. i., & Hsieh, F. (2022). Coarticulation as synchronised CV co-onset – Parallel evidence from articulation and acoustics. *Journal of Phonetics, 90*. https://doi.org/10.1016/j.wocn.2021.101116 101116.
Lopez-Poveda, E. A., & Meddis, R. (2001). A human nonlinear cochlear filterbank. *The Journal of the Acoustical Society of America, 110*(6), 3107–3118. https://doi.org/10.1121/1.1416197.
Lyon, C., Nehaniv, C. L., & Saunders, J. (2012). Interactive language learning by robots: The transition from babbling to word forms. *PLoS ONE, 7*(6), e38236.
Maass, W. (1997). Networks of spiking neurons: the third generation of neural network models. *Neural Networks, 10*(9), 1659–1671. https://doi.org/10.1016/S0893-6080(97)00011-7.
MacDonald, E. N., Johnson, E. K., Forsythe, J., Plante, P., & Munhall, K. G. (2012). Children's development of self-regulation in speech production. *Current Biology, 22*(2), 113–117. https://doi.org/10.1016/j.cub.2011.11.052.
MacNeilage, P. F., & Davis, B. L. (2000). On the origin of internal structure of word forms. *Science, 288*(5465), 527–531. https://doi.org/10.1126/science.288.5465.527.
Maeda, S. (1990). Compensatory articulation during speech: evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In *Speech production and speech modelling: Vol. NATO ASI Series, 55* (pp. 131–149). Springer. https://doi.org/10.1007/978-94-009-2037-8_6.
Makino, H., Hwang, E. J., Hedrick, N. G., & Komiyama, T. (2016). Circuit mechanisms of sensorimotor learning. *Neuron, 92*(4), 705–721. https://doi.org/10.1016/j.neuron.2016.10.029.
Marlow, C. D., Winkelmann, R. K., & Gibilisco, J. A. (1965). General sensory innervation of the human tongue. *The Anatomical Record, 152*, 503–511. https://doi.org/10.1002/ar.1091520410.
Ménard, L., Beaudry, L., & Perrier, P. (2023). Effects of somatosensory perturbation on the perception of French /u/. *JASA Express Letters, 3*(5). https://doi.org/10.1121/10.0017933.
Ménard, L., Leclerc, A., & Tiede, M. (2014). Articulatory and acoustic correlates of contrastive focus in congenitally blind adults and sighted adults. *Journal of Speech, Language, and Hearing Research, 57*(3), 793–804. https://doi.org/10.1044/2014_JSLHR-S-12-0395.
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. *Journal of Phonetics, 53*, 125–140. https://doi.org/10.1016/j.wocn.2015.08.005.
Miura, K., Yoshikawa, Y., & Asada, M. (2007). Unconscious anchoring in maternal imitation that helps find the correspondence of a caregiver's vowel categories. *Advanced Robotics, 21*(13), 1583–1600. https://doi.org/10.1163/156855307782148596.
Miura, K., Yoshikawa, Y., & Asada, M. (2012). Vowel acquisition based on an auto-mirroring bias with a less imitative caregiver. *Advanced Robotics, 26*(1–2), 23–44. https://doi.org/10.1163/016918611X607347.
Moayedi, Y., Michlig, S., Park, M., Koch, A., & Lumpkin, E. A. (2021). Somatosensory innervation of healthy human oral tissues. *Journal of Comparative Neurology, 529*(11), 3046–3061. https://doi.org/10.1002/cne.25148.
Moulin-Frier, C., Diard, J., Schwartz, J. L., & Bessière, P. (2015). COSMO ('Communicating about Objects using Sensory-Motor Operations'): A Bayesian modeling framework for studying speech communication and the emergence of phonological systems. *Journal of Phonetics, 53*, 5–41. https://doi.org/10.1016/j.wocn.2015.06.001.
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. *Frontiers in Psychology, 4*(JAN), 1006. https://doi.org/10.3389/fpsyg.2013.01006.
Moulin-Frier, C., & Oudeyer, P.-Y. (2012). Curiosity-driven phonetic learning. In *2012 IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL)*, 1–8. https://doi.org/10.1109/DevLrn.2012.6400583.
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 *Proceedings of 5th International Conference on Development and Learning and on Epigenetic Robotics*, 208–213. https://doi.org/10.0/Linux-x86_64.
Najnin, S., & Banerjee, B. (2017). A predictive coding framework for a developmental agent: Speech motor skill acquisition and speech production. *Speech Communication, 92*, 24–41. https://doi.org/10.1016/j.specom.2017.05.002.
Nam, H., Goldstein, L., Saltzman, E., & Byrd, D. (2004). TADA: An enhanced, portable Task Dynamics model in MATLAB. *The Journal of the Acoustical Society of America, 115*(5_Supplement), 2430. https://doi.org/10.1121/1.4781490.
Nehaniv, C. L., & Dautenhahn, K. (2002). The correspondence problem. In K. Dautenhahn & C. L. Nehaniv (Eds.), *Imitation in animals and artifacts* (pp. 41–61). Boston Review.
Niemi, M., Laaksonen, J.-P., Ojala, S., Aaltonen, O., & Happonen, R.-P. (2006). Effects of transitory lingual nerve impairment on speech: An acoustic study of sibilant sound /s/. *International Journal of Oral and Maxillofacial Surgery, 35*(10), 920–923. https://doi.org/10.1016/j.ijom.2006.06.002.
Oller, D. K. (1980). The emergence of the sounds of speech in infancy. In G. H. Yeni-Komshian, J. F. Kavanagh, & C. A. Ferguson (Eds.), *Child phonology* (pp. 93–112). Elsevier. https://doi.org/10.1016/B978-0-12-770601-6.50011-5.

Oller, D. K., Caskey, M., Yoo, H., Bene, E. R., Jhang, Y., Lee, C. C., Bowman, D. D., Long, H. L., Buder, E. H., & Vohr, B. (2019). Preterm and full term infant vocalization and the origin of language. *Scientific Reports, 9*(1). https://doi.org/10.1038/s41598-019-51352-0.

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

Oller, D. K., Ramsay, G., Bene, E., Long, H. L., & Griebel, U. (2021). Protophones, the precursors to speech, dominate the human infant vocal landscape. *Philosophical Transactions of the Royal Society B: Biological Sciences, 376*(1836). https://doi.org/10.1098/rstb.2020.0255.

Oudeyer, P. Y. (2005). The self-organization of speech sounds. *Journal of Theoretical Biology, 233*(3), 435–449. https://doi.org/10.1016/j.jtbi.2004.10.025.

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 Computational Biology, 15*(9). https://doi.org/10.1371/journal.pcbi.1007321.

Parrish, A. (2022). *Pronouncing* (0.2.0). https://pronouncing.readthedocs.io/en/latest/tutorial.html.

Perrier, P., Ostry, D. J., & Laboissière, R. (1996). The equilibrium point hypothesis and its application to speech motor control. *Journal of Speech, Language, and Hearing Research, 39*(2), 365–378. https://doi.org/10.1044/jshr.3902.365.

Phan, M. L., Pytte, C. L., & Vicario, D. S. (2006). Early auditory experience generates long-lasting memories that may subserve vocal learning in songbirds. *Proceedings of the National Academy of Sciences, 103*(4), 1088–1093. https://doi.org/10.1073/pnas.0510136103.

Philippsen, A. (2021). Goal-directed exploration for learning vowels and syllables: a computational model of speech acquisition. *KI – Künstliche Intelligenz, 35*(1), 53–70. https://doi.org/10.1007/s13218-021-00704-y.

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 *IEEE ICDL-EPIROB 2014 – 4th Joint IEEE International Conference on Development and Learning and on Epigenetic Robotics*, 195–200. https://doi.org/10.1109/DEVLRN.2014.6982981.

Philippsen, A. K., Reinhart, R. F., & Wrede, B. (2016). Goal Babbling of Acoustic-Articulatory Models with Adaptive Exploration Noise. *Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, 72–78. https://doi.org/10.0/Linux-x86_64.

Plummer, A. R., Beckman, M. E., Belkin, M., Fosler-Lussier, E., & Munson, B. (2010). Learning speaker normalization using semisupervised manifold alignment. *Interspeech, 2010*, 2918–2921. https://doi.org/10.21437/Interspeech.2010-758.

Prom-on, S., Birkholz, P., & Xu, Y. (2013). *Training an articulatory synthesizer with continuous acoustic data*. Interspeech, 2013, 349-353. https://doi.org/10.21437/Interspeech.2013-98.

Prom-on, S., Birkholz, P., & Xu, Y. (2014a). Estimating vocal tract shapes of Thai vowels from contextual vowel variation. In *2014 17th Oriental Chapter of the International Committee for the Co-Ordination and Standardization of Speech Databases and Assessment Techniques (COCOSDA)*, 1–6. https://doi.org/10.1109/ICSDA.2014.7051442.

Prom-on, S., Birkholz, P., & Xu, Y. (2014b). Identifying underlying articulatory targets of Thai vowels from acoustic data based on an analysis-by-synthesis approach. *EURASIP Journal on Audio, Speech, and Music Processing, 2014*(1), 23. https://doi.org/10.1186/1687-4722-2014-23.

Rasilo, H., & Räsänen, O. (2017). An online model for vowel imitation learning. *Speech Communication, 86*, 1–23. https://doi.org/10.1016/j.specom.2016.10.010.

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 Communication, 55*(9), 909–931. https://doi.org/10.1016/j.specom.2013.05.002.

Salthouse, T. A. (2012). Robust cognitive change. *Journal of the International Neuropsychological Society, 18*(4), 749–756. https://doi.org/10.1017/S1355617712000380.

Saltzman, E. L., & Munhall, K. G. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology, 1*(4), 333–382. https://doi.org/10.1207/s15326969eco0104_2.

Schaal, S. (2006). Dynamic movement primitives – a framework for motor control in humans and humanoid robotics. In H. Kimura, K. Tsuchiya, A. Ishiguro, & H. Witte (Eds.), *Adaptive motion of animals and machines* (pp. 261–280). Springer. https://doi.org/10.1007/4-431-31381-8_23.

Schiller, N. O., Meyer, A. S., & Levelt, W. J. (1997). The syllabic structure of spoken words: evidence from the syllabification of intervocalic consonants. *Language and Speech, 40*(2), 103–140.

Scholz, J. P., & Schöner, G. (1999). The uncontrolled manifold concept: identifying control variables for a functional task. *Experimental Brain Research, 126*(3), 289–306. https://doi.org/10.1007/s002210050738.

Sjerps, M. J., Fox, N. P., Johnson, K., & Chang, E. F. (2019). Speaker-normalized sound representations in the human auditory cortex. *Nature Communications, 10*(1). https://doi.org/10.1038/s41467-019-10365-z.

Stevens, S. S., Volkmann, J., & Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. *Journal of the Acoustical Society of America, 8*(3), 185–190. https://doi.org/10.1121/1.1915893.

Stoel-Gammon, C., & Pollock, K. (2008). Vowel development and disorders. In M. J. Ball, M. R. Perkins, & M. Nicole (Eds.), *The handbook of clinical linguistics* (pp. 525–548). Blackwell Publishing Ltd.. https://doi.org/10.1002/9781444301007.ch33.

Tani, J. (2002). Self-organization of behavioral primitives as multiple attractor dynamics: a robot experiment. In *Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat. No.02CH37290)*, 489–494. https://doi.org/10.1109/IJCNN.2002.1005521.

Teramitsu, I., Kudo, L. C., London, S. E., Geschwind, D. H., & White, S. A. (2004). Parallel FoxP1 and FoxP2 expression in songbird and human brain predicts functional interaction. *Journal of Neuroscience, 24*(13). https://doi.org/10.1523/JNEUROSCI.5589-03.2004.

Terken, J., & Lemeer, G. (1988). Effects of segmental quality and intonation on quality judgments for texts and utterances. *Journal of Phonetics, 16*(4), 453–457. https://doi.org/10.1016/S0095-4470(19)30521-2.

Thorpe, D. W. H. (1954). The process of song learning in the chaffinch as studied by means of the sound spectrograph. *Nature, 173*, 465–469. https://doi.org/10.1038/173465a0.

Tilsen, S. (2019). Motoric mechanisms for the emergence of non-local phonological patterns. *Frontiers in Psychology, 10*. https://doi.org/10.3389/fpsyg.2019.02143.

Todorov, E., & Jordan, M. I. (2002). Optimal feedback control as a theory of motor coordination. *Nature Neuroscience, 5*(11), 1226–1235. https://doi.org/10.1038/nn963.

Tourville, J. A., & Guenther, F. H. (2011). The DIVA model: a neural theory of speech acquisition and production. *Language and Cognitive Processes, 26*(7), 952–981. https://doi.org/10.1080/01690960903498424.

Tremblay, S., Shiller, D. M., & Ostry, D. J. (2003). Somatosensory basis of speech production. *Nature, 423*(6942), 866–869. https://doi.org/10.1038/nature01710.

van Elk, M., van Schie, H. T., Hunnius, S., Vesper, C., & Bekkering, H. (2008). You'll never crawl alone: neurophysiological evidence for experience-dependent motor resonance in infancy. *NeuroImage, 43*(4), 808–814. https://doi.org/10.1016/j.neuroimage.2008.07.057.

van Niekerk, D. R., Xu, A., Gerazov, B., Krug, P. K., Birkholz, P., Halliday, L., Prom-on, S., & Xu, Y. (2023). Simulating vocal learning of spoken language: Beyond imitation. *Speech Communication, 147*, 51–62. https://doi.org/10.1016/j.specom.2023.01.003.

Vihman, M. M. (1996). *Phonological development: the origins of language in the child [Book]*. Blackwell.

Vorperian, H. K., & Kent, R. D. (2007). Vowel acoustic space development in children: a synthesis of acoustic and anatomic data. *Journal of Speech, Language, and Hearing Research, 50*(6), 1510–1545. https://doi.org/10.1044/1092-4388(2007/104).

Warlaumont, A. S. (2012). A spiking neural network model of canonical babbling development. In *2012 IEEE International Conference on Development and Learning and Epigenetic Robotics, ICDL 2012*, 1–6. https://doi.org/10.1109/DevLrn.2012.6400842.

Warlaumont, A. S., & Finnegan, M. K. (2016). Learning to produce syllabic speech sounds via reward-modulated neural plasticity. *PLoS ONE, 11*(1), e0145096.

Warlaumont, A. S., Westermann, G., Buder, E. H., & Oller, D. K. (2013). Prespeech motor learning in a neural network using reinforcement. *Neural Networks, 38*, 64–75. https://doi.org/10.1016/j.neunet.2012.11.012.

Werker, J. F., & Lalonde, C. E. (1988). Cross-language speech perception: initial capabilities and developmental change. *Developmental Psychology, 24*(5), 672–683. https://doi.org/10.1037/0012-1649.24.5.672.

Westerman, G., & Miranda, E. R. (2002). Modelling the development of mirror neurons for auditory-motor integration. *Journal of New Music Research, 31*(4), 367–375. https://doi.org/10.1076/jnmr.31.4.367.14166.

Westermann, G., & Miranda, E. R. (2004). A new model of sensorimotor coupling in the development of speech. *Brain and Language, 89*(2), 393–400. https://doi.org/10.1016/S0093-934X(03)00345-6.

Wolpert, D. M., Diedrichsen, J., & Flanagan, J. R. (2011). Principles of sensorimotor learning. *Nature Reviews Neuroscience, 12*(12), 739–751. https://doi.org/10.1038/nrn3112.

Woods, K. J. P., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, and Psychophysics, 79*(7), 2064–2072. https://doi.org/10.3758/s13414-017-1361-2.

Xu, Y. (2020). Syllable as a synchronization mechanism that makes human speech possible. *PsyArXiv*. https://doi.org/10.31234/osf.io/9v4hr.

Xu, Y., & Emily Wang, Q. (2001). Pitch targets and their realization: Evidence from Mandarin Chinese. *Speech Communication, 33*(4), 319–337. https://doi.org/10.1016/S0167-6393(00)00063-7.

Xu, Y., Larson, C. R., Bauer, J. J., & Hain, T. C. (2004). Compensation for pitch-shifted auditory feedback during the production of Mandarin tone sequences. *The Journal of the Acoustical Society of America, 116*(2), 1168–1178. https://doi.org/10.1121/1.1763952.

Xu, Y., & Liu, F. (2006). Tonal alignment, syllable structure and coarticulation: Toward an integrated model. *Italian Journal of Linguistics, 18*, 125–159.

Xu, Y., & Prom-on, S. (2014). Toward invariant functional representations of variable surface fundamental frequency contours: synthesizing speech melody via model-based stochastic learning. *Speech Communication, 57*, 181–208. https://doi.org/10.1016/j.specom.2013.09.013.

Xu, Y., & Prom-on, S. (2019). Economy of effort or maximum rate of information? Exploring basic principles of articulatory dynamics. *Frontiers in Psychology, 10*. https://doi.org/10.3389/fpsyg.2019.02469.

Yoshikawa, Y., Asada, M., Hosoda, K., & Koga, J. (2003). A constructivist approach to infants' vowel acquisition through mother-infant interaction. *Connection Science, 15* (4), 245–258. https://doi.org/10.1080/09540090310001655075.

Yoshikawa, Y., Koga, J., Asada, M., & Hosoda, K. (2003). Primary vowel imitation between agents with different articulation parameters by parrot-like teaching. *IEEE International Conference on Intelligent Robots and Systems, 1*, 149–154. https://doi.org/10.1109/iros.2003.1250620.

Zhao, W., Garcia-Oscos, F., Dinh, D., & Roberts, T. F. (2019). Inception of memories that guide vocal learning in the songbird. *Science, 366*(6461), 83–89. https://doi.org/10.1126/science.aaw4226.

Zharkova, N. (2018). An ultrasound study of the development of lingual coarticulation during childhood. *Phonetica, 75*(3), 245–271. https://doi.org/10.1159/000485802.

Zharkova, N., Hewlett, N., & Hardcastle, W. J. (2011). Coarticulation as an indicator of speech motor control development in children: an ultrasound study. *Motor Control, 15*(1), 118–140.