Interspeech 2022
18-22 September 2022, Incheon, Korea

Evoc-Learn — High quality simulation of early vocal learning

Yi Xu[1], Anqi Xu[1], Daniel R. van Niekerk[1], Branislav Gerazov[2], Peter Birkholz[3], Paul Konstantin Krug[3], Santitham Prom-on[4], Lorna F. Halliday[1,5]

[1]Department of Speech Hearing and Phonetic Sciences, University College London, UK
[2]Faculty of Electrical Engineering and Information Technologies, UCMS, Skopje, RN Macedonia
[3]Institute of Acoustics and Speech Communication, Technische Universität, Dresden, Germany
[4]Computer Engineering Department, King Mongkut's University of Technology Thonburi, Thailand
[5]Medical Research Council (MRC) Cognition and Brain Sciences Unit, University of Cambridge, Cambridge, UK

{yi.xu,a.xu.17, d.vniekerk}@ucl.ac.uk, {paul_konstantin.krug, peter.birkholz}@tu-dresden.de, santitham@cpe.kmutt.ac.th, Lorna.Halliday@mrc-cbu.cam.ac.uk

Abstract

Evoc-Learn is a system for simulating early vocal learning of spoken language in ways that can overcome some of the major bottlenecks in vocal learning. The system consists of VocalTractLab, a geometrical three-dimensional vocal tract model for simulating aeroacoustics and articulatory dynamics, a coarticulation model for controlling the temporal dynamics of articulation, and a sensory feedback system for guiding the learning process. We will demonstrate each component of Evoc-Learn and show how they work together to simulate the learning of highly intelligible speech.

Index Terms: articulatory synthesis, CV coarticulation, speech recognition, speech perception, sensory feedback

1. Introduction

How do children learn to speak without explicit instructions? This question is more than a matter of curiosity, as speech is a lifelong continuous process of acquisition, use and adaptation. One of the best ways to understand vocal learning is to simulate it through computational modeling based on findings of observational studies. Previous modeling attempts have tried to simulate vocal learning as direct imitation [1], caretaker feedback [1,2], reinforcement learning [3], and self-motivation [4,5]. Advances have been restrained, however by difficulties in resolving some of the major bottlenecks, including, in particular, the lack of invariance and speaker normalization. In this Show and Tell we will present a new simulation system with components that can address these difficulties to simulate the learning of words that are intelligible and natural sounding.

2. Framework

Evoc-Learn is a modular simulation system developed to test various ecological assumptions about early vocal learning. As shown in Figure 1, the system consists of the following components.

1. VocalTractLab — A state-of-the-art articulatory synthesizer based on a geometrical 3-D vocal tract model with built-in gestural dynamics [6].
2. A syllable model based on a theory of coarticulation, which controls the temporal dynamics of articulation [7].
3. A learning scheme based on auditory, visual or somatosensory feedback.

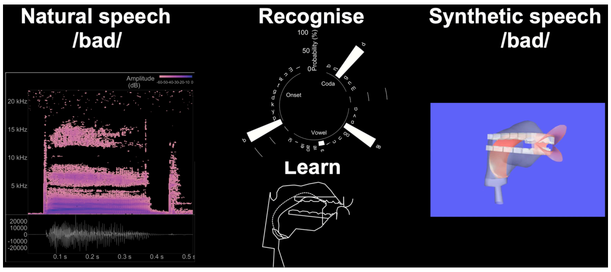

Figure 1: *Schematic of Evoc-Learn.*

2.1. VocalTractLab—An articulatory synthesizer

VocalTractLab is a geometrical three-dimensional vocal tract model with built-in aeroacoustic transformation and articulatory dynamics. The geometrical vocal tract shape and the aeroacoustic transformation allow the generation of spectral properties of consonants, vowels and intonation. The target approximation model simulates articulatory gestures as discrete target approximation movements [6].

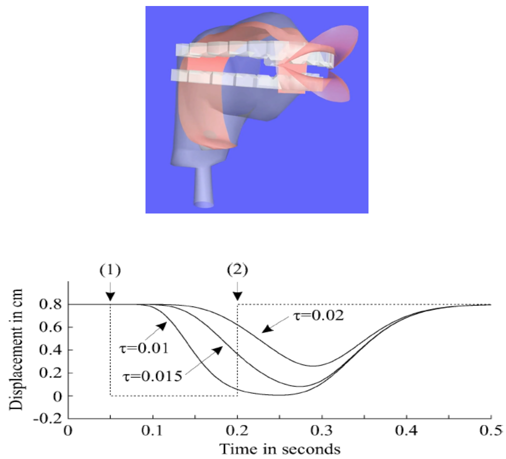

Figure 2: *VocalTractLab. Top: 3D geometric model. Bottom: Target approximation model.*

2.2. Syllable-based coarticulation model

The coarticulation model (Figure 3) is based on a theory that the syllable is a mechanism of reducing temporal degrees of freedom by synchronizing consonantal (C), vocalic (V), and laryngeal (T) gestures at syllable onset to enable neural control of articulation [7]. It also posits that the temporal overlap of consonant and vowels at the syllable onset is realized by strictly sequential target approximation at the level of articulator dimensions, so that different dimensions of the same articulator can be controlled respectively by either the consonant or the vowel [7,8]. We will demonstrate how the coarticulation model helps to resolve a major portion of the variability problem and improves naturalness of synthetic syllables.

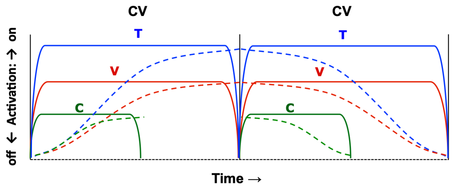

Figure 3: *Coarticulation model of syllable synchronization based on articulatory synchrony.*

2.3. Sensory-guided vocal learning

Because pre-lingual children cannot benefit from caretaker instructions, they have to rely on their own sensory input in vocal learning. In Evoc-Learn, four types of feedback mechanisms are simulated: auditory matching, perceptual recognition, somatosensory constraint, and visual observation.

2.3.1. Auditory matching (imitation)

Direct auditory matching is what is explored in most simulation works. During learning, learner-generated speech sounds are directly compared to the target sounds based on MFCC or Log Mel spectrogram, and the differences are used as a guide to improve further vocal exploration. This learning mechanism resembles direct imitation, whereby the learner tries to match the acoustics of their own practice articulation to that of a target utterance. Direct auditory matching has not yet led to high quality simulation [1-5], but it is included in Evoc-Learn as an option (Figure 1 left) to further explore its full potential and serve as a baseline for other sensory feedback mechanisms.

2.3.2. Perceptual recognition

Perceptual recognition as sensory feedback for guiding vocal learning is a novel mechanism developed in Evoc-Learn (top central in Figure 1). Unlike direct auditory matching, learner-generated speech sounds are assessed by a recognizer without reference to specific natural utterances. The perceptual distance is then used to guide the selection of further candidate articulatory targets. We have found that vocal learning trained by speech recognition can result in synthesis of English words with high intelligibility, sometimes close to natural speech. Importantly, recognizer-based training outperformed acoustic training. Thus, the multi-speaker trained recognizer can help to solve both the contextual variation and speaker normalization problems, two major bottlenecks in vocal learning.

2.3.3. Somatosensory constraint

Somatosensory constraint imposes a limit on the degree of oral opening for each generated vocal tract configuration during vocal exploration. The constraint ensures an open vocal tract for vowels and a narrow vocal tract for consonants. We will show how these constraints can effectively restrict the search space for the articulatory targets.

2.3.4. Visual observation

Visual observation plays a critical role in the learning of sounds with overt facial movements, such as lip rounding or lip spreading [9]. We will demonstrate how a set of articulatory objectives motivated by visually-available signals can have a positive effect on the intelligibility of CV utterances produced by VocalTractLab.

3. Software design

Evoc-Learn is a modular system implemented in Python as a number of standalone packages under the GNU General Public Licence. The system is designed as sets of composable functional components which can be used to construct flexible processing pipelines for experiments either in Python directly or on the UNIX command line. Each package contains implementations for the application of models and processes as well as tools for the construction of models such as neural networks. The system leverages well-known Python infrastructure packages such as Pandas, H5PY, Tensorflow, etc. as far as possible to reduce the learning curve for new users and provides streaming data serialization to human-readable and efficient data formats to ease the implementation of large experiments.

4. Acknowledgements

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

5. References

[1] H. Rasilo, O. Räsänen, An online model for vowel imitation learning. *Speech Communication*. **86**, pp. 1–23, 2017.

[2] P. Messum, Ian. S. Howard, 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**, pp. 125–140, 2015.

[3] A. S. Warlaumont, M. K. Finnegan, Learning to produce syllabic speech sounds via reward-modulated neural plasticity. *PLoS ONE*. **11**, e0145096, 2016.

[4] A. Philippsen, Goal-Directed Exploration for Learning Vowels and Syllables: A Computational Model of Speech Acquisition. *KI - Kunstliche Intelligenz*. **35**, pp. 53–70, 2021.

[5] C. Moulin-Frier, S. M. Nguyen, P. Y. Oudeyer, Self-organization of early vocal development in infants and machines: The role of intrinsic motivation. *Frontiers in Psychology*. **4**, 1006 (2014).

[6] Birkholz, P., "Modeling Consonant-Vowel Coarticulation for Articulatory Speech Synthesis," *PLoS ONE* vol. 8, no. 4, pp. e60603, 2013.

[7] Xu, Y., "Syllable is a synchronization mechanism that makes human speech possible," *PsyArXiv* vol. doi:10.31234/osf.io/9v4hr, 2020.

[8] Xu, A., Birkholz, P., and Xu, Y., "Coarticulation as synchronized dimension-specific sequential target approximation: An articulatory synthesis simulation," in *Proceedings of The 19th International Congress of Phonetic Sciences*, Melbourne, Australia, 2019.

[9] Menard, L., Toupin, C., Baum, S. R. *et al.*, "Acoustic and articulatory analysis of French vowels produced by congenitally blind adults and sighted adults," *The Journal of the Acoustical Society of America* vol. 134, no. 4, pp. 2975-2987, 2013.