Interspeech 2022

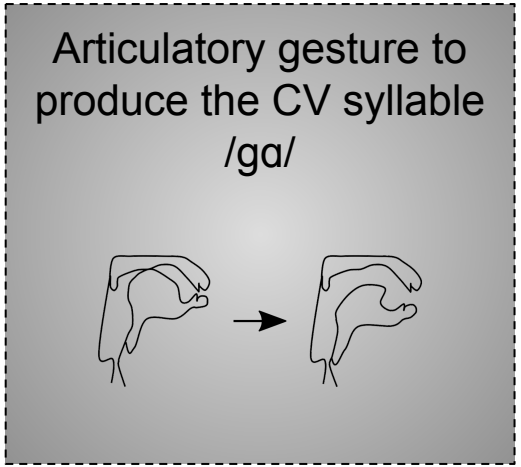
Exploration strategies for articulatory synthesis of complex syllable onsets

Daniel van Niekerk, Anqi Xu, Branislav Gerazov,
Paul Krug, Peter Birkholz, Yi Xu

Introduction

High-quality articulatory synthesis is promising for speech science and technology, but defining **meaningful articulatory gestures** is a bottleneck:

- Requires phonetic expertise
- Time consuming
- Difficult to evaluate
- Language / dialect specific
- Speaker specific



Introduction

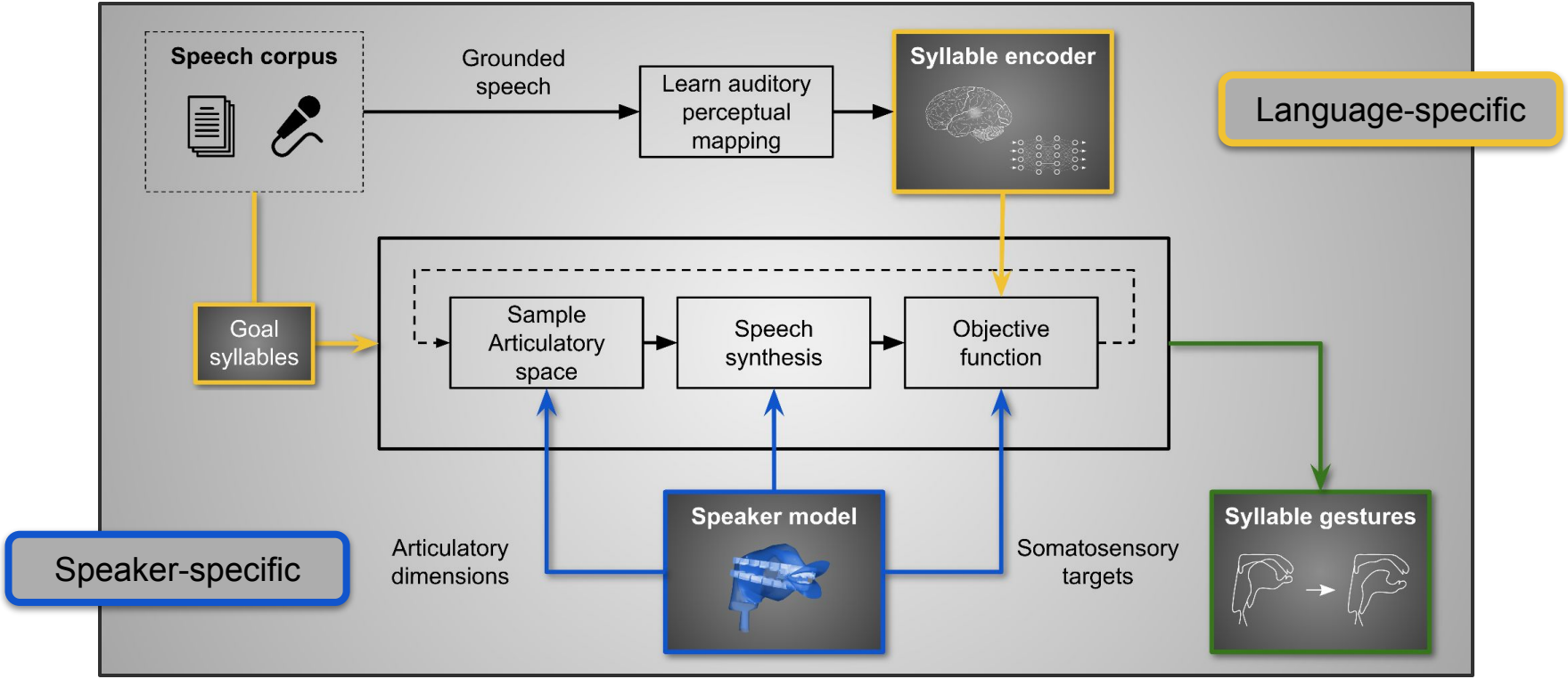
We are interested in a process to self-learn a set of **meaningful gestures** (syllables) given a target **language** and **speaker** geometry with minimal specification / intervention.

This resembles the conditions faced by infants during **vocal learning** of which the following is known:

- Various stages of **vocal babbling** eventually lead to the production of syllables
- **Auditory perception** develops earlier and becomes increasingly **language-specific**
- Vocal learning has to contend with the **normalisation** and **correspondence** problems
- **Auditory feedback** is critical to successful vocal learning
- **Somatosensory and visual feedback** is useful during vocal learning
- Later stages of articulatory exploration is usually considered to be **goal-oriented**

Approach

Goal-directed articulatory exploration implemented as an **optimisation task**:



Empirical questions

Previous work resulted in synthesis of **intelligible** CV syllables and **plausible** articulatory gestures by including regularising articulatory objectives.

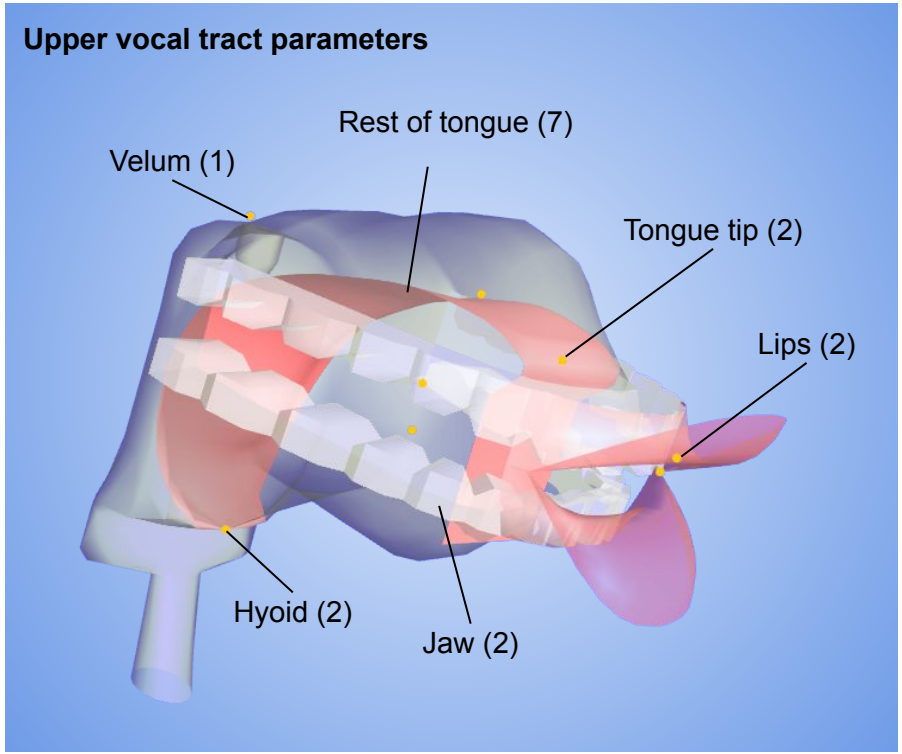
We extend the investigation to **complex syllable onsets** with the following questions:

- **Exploration strategies:** Can / should segment targets be optimised separately / jointly and in which order?
- **Coarticulation:** Can we use the articulatory distance as regularisation objective for complex syllable onsets?
- **Sufficiency:** What is the relative difficulty of discovering different American English CCVs in this framework?

Experimental setup

Use the VocalTractLab synthesizer to find gestures for American English syllables:

- Focus on the **upper vocal tract** (presets for the glottis model)
- The set of **valid CCVs** excluding nasals and diphthongs
- **Auditory objective** based on a CCV encoder trained on Librispeech
- **Articulatory objectives** represent basic somatosensory feedback (tract closure / opening)
- **Coarticulation:** Test the articulatory distance objective
- **Strategy:** Test the order of exploration



Results – Exploration strategies

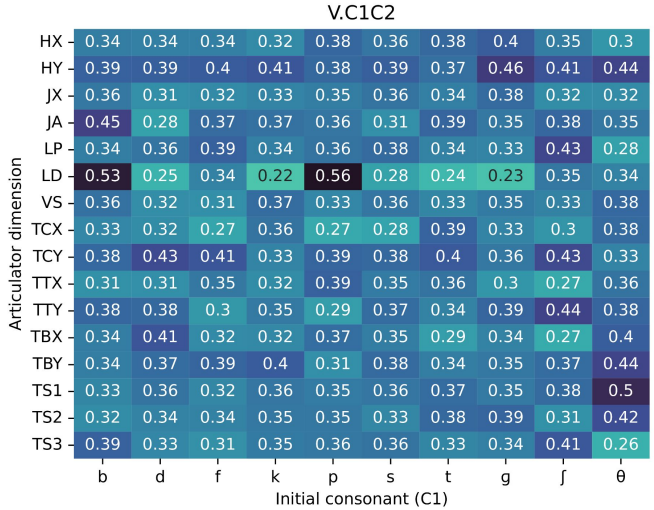
Optimisation success rates (%) for different strategies in terms of the auditory objective:

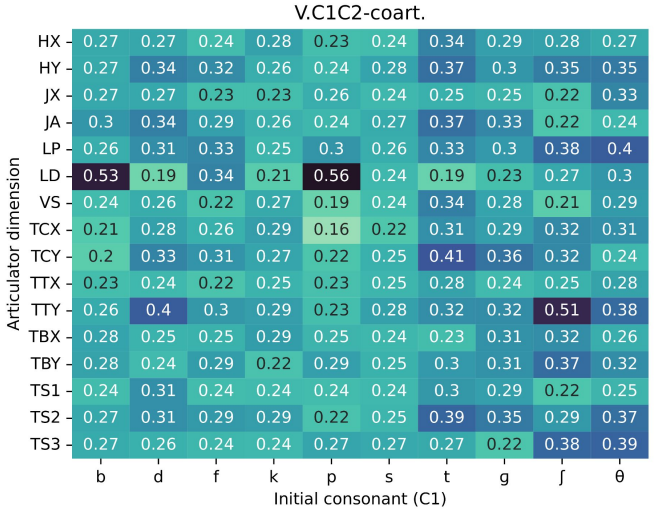
	C_1C_2V	$V \rightarrow C_1C_2$	$V \rightarrow C_1 \rightarrow C_2$	$V \rightarrow C_2 \rightarrow C_1$
syllable	80	**85**	81	**71**
vowel	**89**	94	93	93
onset	90	90	88	**77**

- Joint optimisation of the onset was **most successful**
- Optimising C1 on the basis of C2V was significantly **less successful**
- Joint syllable optimisation with coarticulation objective results in **less successful vowels**

Results – Coarticulation

Requires two-pass optimisation strategy to fix the vowel first:





- Reduces articulatory distance without affecting the success rate
- May be used to discover articulatory coordination in different phonological contexts

Conclusion

Findings and contribution

Joint optimisation of the onset is most successful and if necessary C_2 should be optimised in the context of C_1 rather than the reverse

The articulatory distance can be used as regularisation objective to **discover contextual articulatory coordination** if the vowel is fixed

The simulation of articulatory exploration **combines auditory and articulatory feedback** to discover intelligible CV and CCV syllables

Future work

The results should be extended further to cover the **full set of English syllables**

These and future results should be evaluated using formal listening experiments to **quantify intelligibility and naturalness**

Data generated through articulatory exploration should be used to train forward / inverse models that are able to further **improve the precision of gestures**