Interspeech 2022



Exploration strategies for articulatory synthesis of complex syllable onsets

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





LEVERHULME TRUST _____

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



UCL

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**



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



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





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

	C ₁ C ₂ V	$V \to C_1 C_2$	$V {\rightarrow} C_1 {\rightarrow} C_2$	$V \to C_2 \to C_1$
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

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

						V.C	1C2									١	/.C1C2	-coart				
ŀ	HX - 0	.34	0.34	0.34	0.32	0.38	0.36	0.38	0.4	0.35	0.3	HX -	0.27	0.27	0.24	0.28	0.23	0.24	0.34	0.29	0.28	0.27
H	HY - 0	.39	0.39	0.4	0.41	0.38	0.39	0.37	0.46	0.41	0.44	HY -	0.27	0.34	0.32	0.26	0.24	0.28	0.37	0.3	0.35	0.35
	JX - 0	.36	0.31	0.32	0.33	0.35	0.36	0.34	0.38	0.32	0.32	JX -	0.27	0.27	0.23	0.23	0.26	0.24	0.25	0.25	0.22	0.33
	JA - 0	.45	0.28	0.37	0.37	0.36	0.31	0.39	0.35	0.38	0.35	JA -	0.3	0.34	0.29	0.26	0.24	0.27	0.37	0.33	0.22	0.24
~	LP - 0	.34	0.36	0.39	0.34	0.36	0.38	0.34	0.33	0.43	0.28	LP -	0.26	0.31	0.33	0.25	0.3	0.26	0.33	0.3	0.38	0.4
sior	_D - 0.	.53	0.25	0.34	0.22	0.56	0.28	0.24	0.23	0.35	0.34	- DJ -	0.53	0.19	0.34	0.21	0.56	0.24	0.19	0.23	0.27	0.3
١en	/S - 0	.36	0.32	0.31	0.37	0.33	0.36	0.33	0.35	0.33	0.38	ษี VS -		0.26	0.22	0.27	0.19	0.24	0.34	0.28	0.21	0.29
in di	CX - 0.	.33	0.32	0.27	0.36	0.27	0.28	0.39	0.33	0.3	0.38	៏ TCX -	0.21	0.28	0.26	0.29	0.16	0.22	0.31	0.29	0.32	0.31
ъđ	CY - 0	.38	0.43	0.41	0.33	0.39	0.38	0.4	0.36	0.43	0.33	- ۲CY نے	0.2	0.33	0.31	0.27	0.22	0.25	0.41	0.36	0.32	0.24
- El T	TX - 0	.31	0.31	0.35	0.32	0.39	0.35	0.36	0.3	0.27	0.36	- XTT 🗧	0.23	0.24	0.22	0.25	0.23	0.25	0.28	0.24	0.25	0.28
Ĕ⊤	TY - 0	.38	0.38	0.3	0.35	0.29	0.37	0.34	0.39	0.44	0.38	- אדר ב	0.26	0.4	0.3	0.29	0.23	0.28	0.32	0.32	0.51	0.38
TE	3X - 0	.34	0.41	0.32	0.32	0.37	0.35	0.29	0.34	0.27	0.4	⁺ твх -	0.28	0.25	0.25	0.29	0.25	0.24	0.23	0.31	0.32	0.26
Т	BY - 0	.34	0.37	0.39	0.4	0.31	0.38	0.34	0.35	0.37	0.44	TBY -	0.28	0.24	0.29	0.22	0.29	0.25	0.3	0.31	0.37	0.32
T:	51 - 0	.33	0.36	0.32	0.36	0.35	0.36	0.37	0.35	0.38	0.5	TS1 -		0.31			0.24	0.24	0.3	0.29	0.22	0.25
T:	52 - 0	.32	0.34	0.34	0.35	0.35	0.33	0.38	0.39	0.31	0.42	TS2 -	0.27	0.31	0.29	0.29	0.22	0.25	0.39	0.35	0.29	0.37
T	53 - 0	.39	0.33	0.31	0.35	0.36	0.36	0.33	0.34	0.41	0.26	TS3 -	0.27	0.26	0.24	0.24	0.27	0.27	0.27	0.22	0.38	0.39
		b	d	f	k	p	s	ť	g	ŗ	ė		b	d	ŕ	k	р	S	ť	g	ŗ	ė
Initial consonant (C1)												Initi	al cons	onant (C1)							

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

Conclusion



Findings and contribution	Future work
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 results should be extended further to cover the full set of English syllables
The articulatory distance can be used as regularisation objective to discover contextual articulatory coordination if the vowel is fixed	These and future results should be evaluated using formal listening experiments to quantify intelligibility and naturalness
The simulation of articulatory exploration combines auditory and articulatory feedback to discover intelligible CV and CCV syllables	Data generated through articulatory exploration should be used to train forward / inverse models that are able to further improve the precision of gestures