Interspeech 2020

Finding intelligible consonant-vowel sounds using high-quality articulatory synthesis

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

UCL

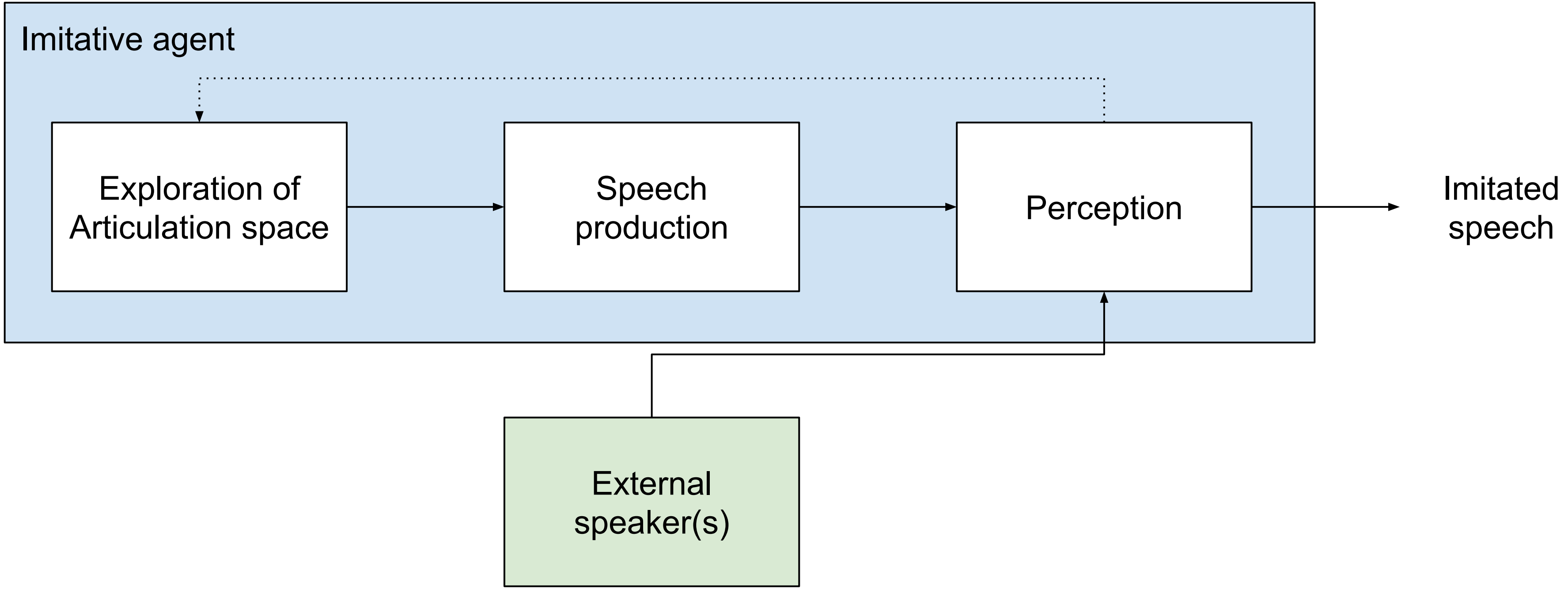
Overview

- ***Derivative-free optimisation*** compared to uniform sampling
- Reduced consonant search-space motivated by ***vowel coarticulation***
- Investigate automatic speech recognition as a means of ***evaluating intelligibility***

Task of obtaining articulatory movements from speech exemplars:

- Model real speech acquisition
  - Understand computational demands
  - Test phonetic assumptions of speech production
- Copy synthesis
  - Reproduce speech with an articulatory synthesiser
  - Speech technology applications

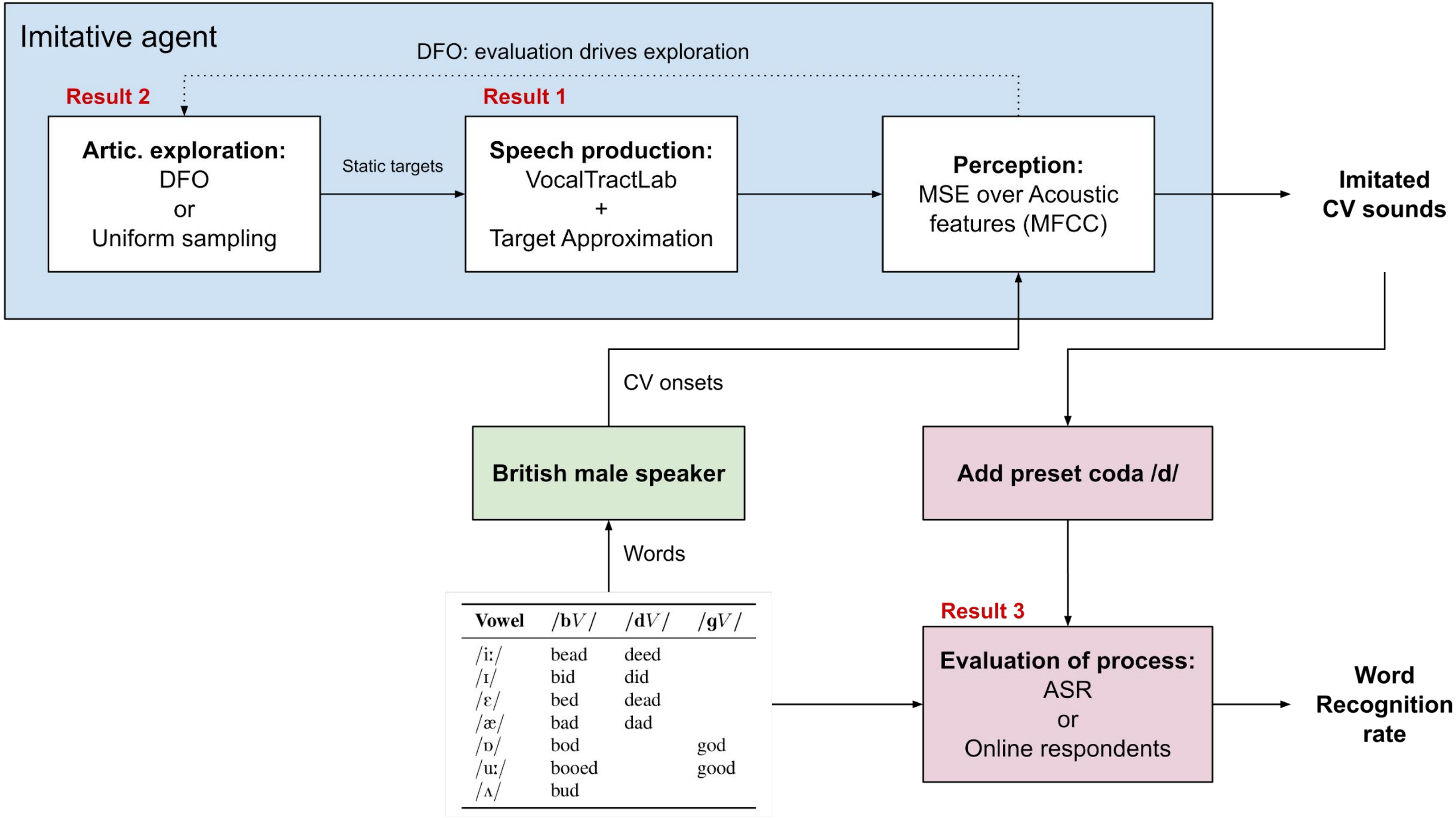
Background: Minimal imitative articulatory exploration process



Using a 3-dimensional articulatory synthesiser, we tested assumptions of a practical and theoretical nature:

- Exploration of articulation space
  - Using **Derivative-free optimisation** (DFO) compared to **Uniform sampling**
- Speech production
  - **Consonant-vowel coarticulation**
  - Articulatory trajectories generated by a **simple kinematic model** from static targets
- Evaluation
  - Use of a standard **automatic speech recognition** (ASR) system during evaluation

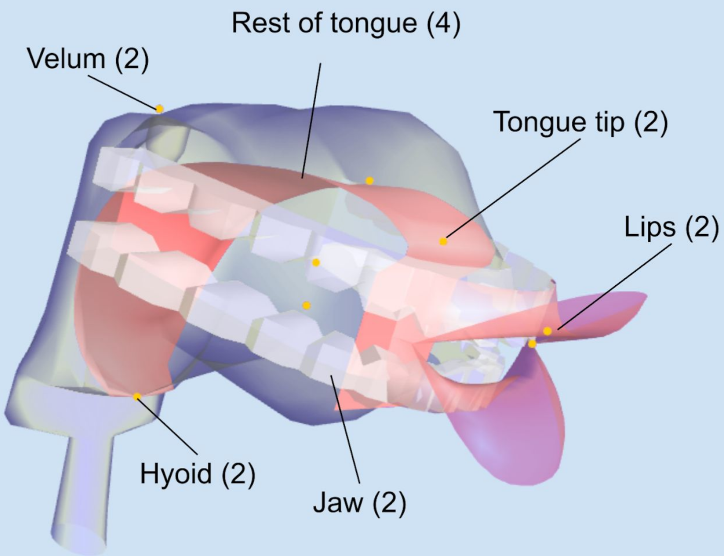
Experimental setup: Overview



Vowel	/bV/	/dV/	/gV/
/iː/	bead	deed	
/ɪ/	bid	did	
/ɛ/	bed	dead	
/æ/	bad	dad	
/ɒ/	bod		god
/uː/	booed		good
/ʌ/	bud		

Experimental setup: Articulatory space and speech production

Full set of free parameters per segment (15):



Target Approximation time constant (1)

Result 1

Free-onset configuration:

C - 15 parameters

V - 15 parameters

Tied-onset configuration:

/b/ - 4 params. (jaw + lips)

/d/ - 7 params. (jaw + tongue)

/g/ - 4 params. (jaw + tongue)

V - 15 params.

Static targets

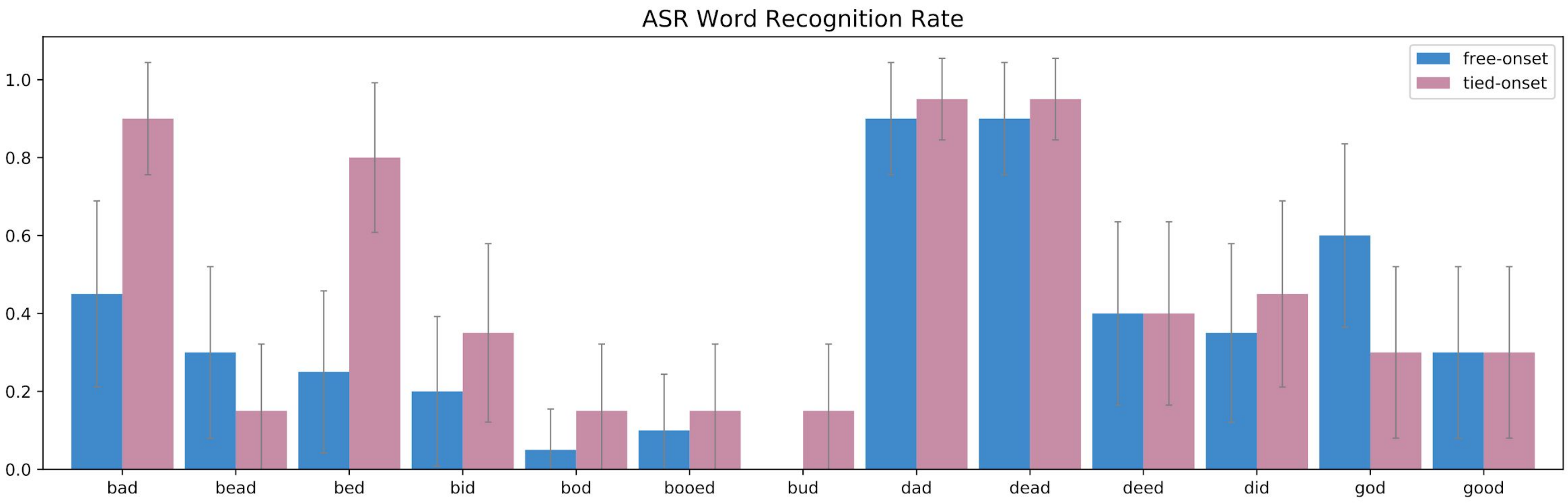
Target Approximation

Trajectories

VocalTractLab

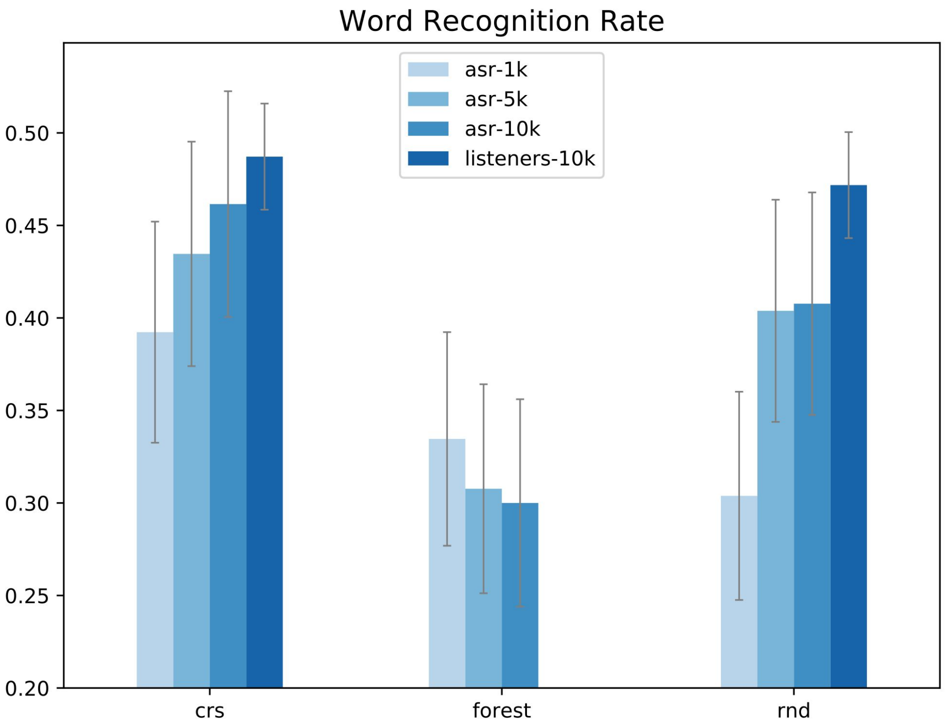
Speech

Results I: Consonant-vowel coarticulation



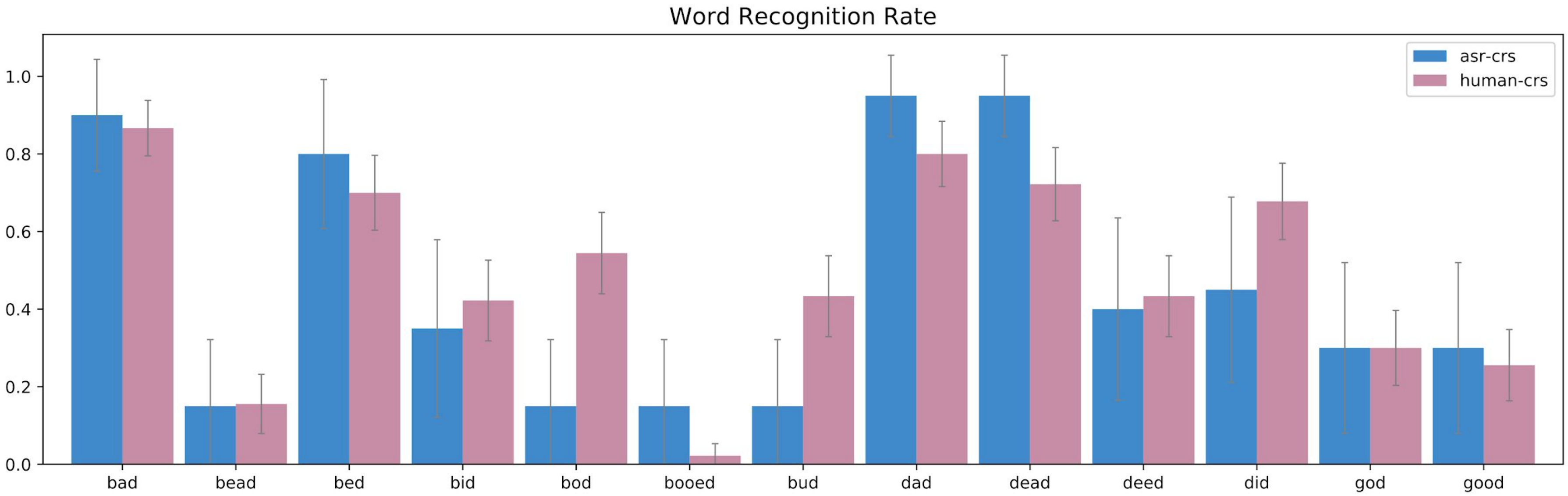
- Using 10k iterations and the best optimisation algorithm
- **Significantly better results overall with tied-onset configuration**
- **Reversed trend in “bead” and “god”** may indicate over-constrained setting (but not significant in our experiment)

Results II & III: Optimisation algorithms and ASR evaluation



- **Controlled-Random-Search** (CRS) performed significantly better than Uniform sampling (RND) with fewer iterations
- **Random Forest** model-based optimisation did not benefit from more iterations
- **Overall recognition rate** using ASR exhibited more variation but not significantly different to listeners in our experiment

Results III: ASR evaluation



- Using 10k iterations and the CRS optimisation algorithm
- ASR recognition rate lower for **bod, bud, did** (presumably language model)
- Human recognition rate lower for **dad, dead** (/b/ vs /d/ confusion)

Conclusions	Future work
Parameter tying motivated by coarticulation has the potential to reduce the search-space significantly	**Control parameters for velars** need further investigation (possibly over-constrained)
Controlled-Random-Search and ASR are **viable tools to speed up exploration and evaluation**	**Automatically adding codas** to form words is difficult (recognition rates affected)
Results baseline and implementation guidelines for future experiments	Future work will involve **learned models** and possibly incorporate intelligibility as simulated objective