ESSV 2021

Anticipatory coarticulation in predictive articulatory speech modeling

Konstantin Sering, Fabian Tomaschek, & Motoki Saito

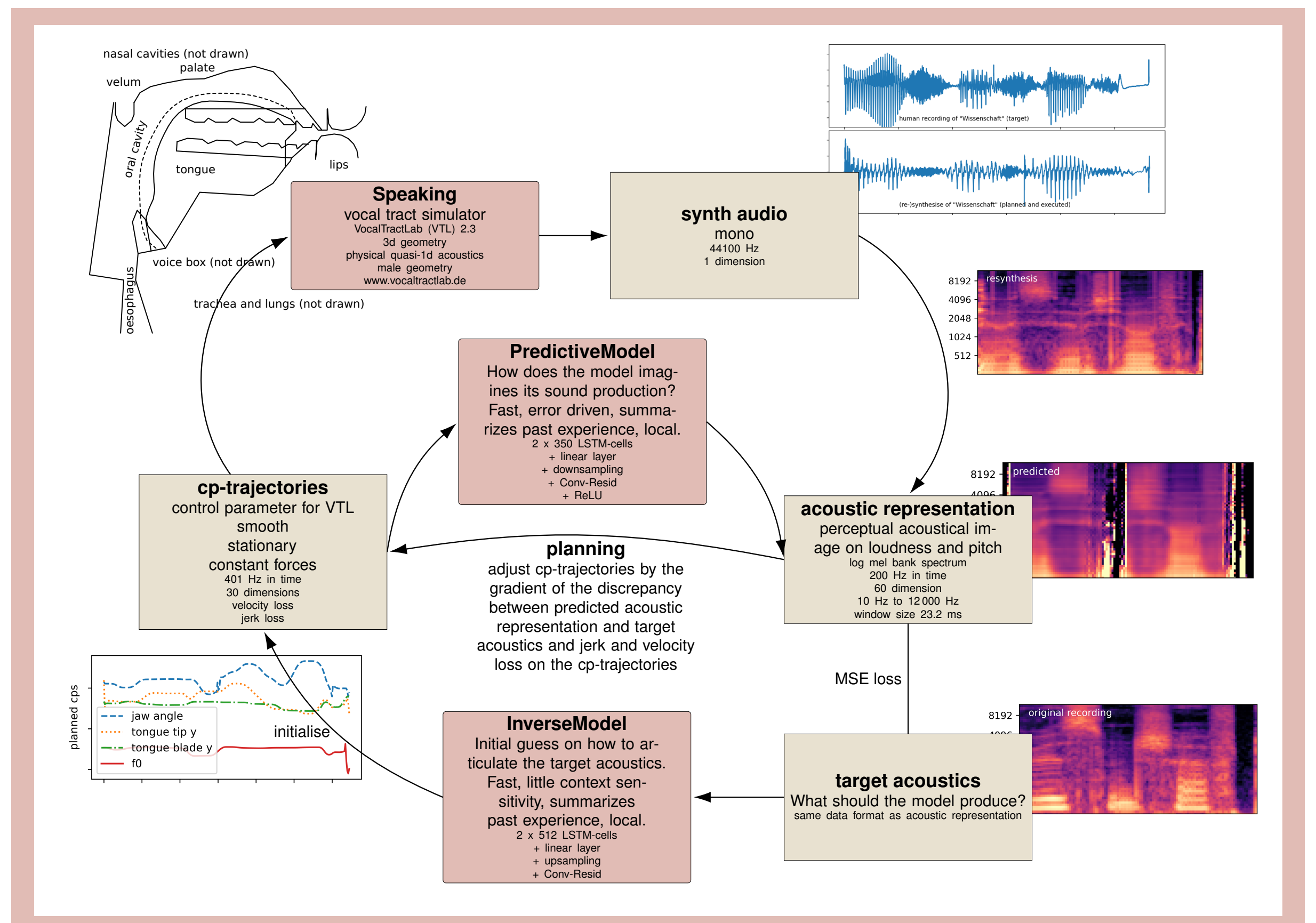
Quantitative Linguistics, University of Tübingen, konstantin.sering@uni-tuebingen.de

Eberhard Karls Universität Tübingen

Overview

- anticipatory coarticulation in articulatory synthesis and humans
- coarticulation in acoustics (here: **formant shift** [7]) and articulation (here: **tongue raising** [6]) measured with ultrasound
- using **VocalTractLab (VTL)** simulator [1]
- control parameter (cp-) trajectories for VTL derived by fully automatic **segment based** and recurrent gradient based **planning** approach
- human recording shows anticipatory coarticulation in the acoustic and articulation domain
- segment approach fails to recover this coarticulation
- planning approach partially recovers coarticulation

Methods



Speaking
vocal tract simulator
VocalTractLab (VTL) 2.3
3d geometry
physical quasi-1d acoustics
male geometry
www.vocaltractlab.de

synth audio
mono
44100 Hz
1 dimension

PredictiveModel
How does the model imagines its sound production? Fast, error driven, summarizes past experience, local.
2 x 350 LSTM-cells
+ linear layer
+ downsampling
+ Conv-Resid
+ ReLU

cp-trajectories
control parameter for VTL
smooth
stationary
constant forces
401 Hz in time
30 dimensions
velocity loss
jerk loss

planning
adjust cp-trajectories by the gradient of the discrepancy between predicted acoustic representation and target acoustics and jerk and velocity loss on the cp-trajectories

acoustic representation
perceptual acoustical image on loudness and pitch
log mel bank spectrum
200 Hz in time
60 dimension
10 Hz to 12 000 Hz
window size 23.2 ms

InverseModel
Initial guess on how to articulate the target acoustics. Fast, little context sensitivity, summarizes past experience, local.
2 x 512 LSTM-cells
+ linear layer
+ upsampling
+ Conv-Resid

target acoustics
What should the model produce?
same data format as acoustic representation

Figure 1: Recurrent gradient based planning framework [5].

Recurrent gradient based resynthesis [5]

- define acoustic target as log mel spectrogram
- initialize cp-trajectories with inverse model
- plan along equally weighted MSE loss, jerk loss and half weighted velocity loss of predictive model
- adjust cp-trajectories 0.05 times its local gradient (no ADAM; inner loop)
- 40×200 iterations inner loop (planning), 40 iterations outer loop (experience)
- continue training of predictive model with synthesized audio plus 10 initial training samples (outer loop)

Segment based resynthesis [2]

- define phone segment sequence with corresponding durations
- call VTL `vtlSegmentSequenceToGesturalScore` to generate gestural score file
- use gestural score file to synthesize audio and export midsagittal ultrasound pictures

Experimental Setup

- natural articulations of **/baba/**, **/babi/** and **/babu/**
- five speaking rate conditions
- blocked production
- focus on second fastest speaking rate condition: utter the pseudo word five times within 3 seconds, e. g. */babibabibabibabibabi/*
- ultrasound image of midsagittal plane (81.6 fps; 64 directions with 842 pixels each)
- audio recording (22050 Hz) synchronised to ultrasound recordings
- sound attenuated booth; n = 90; one speaker

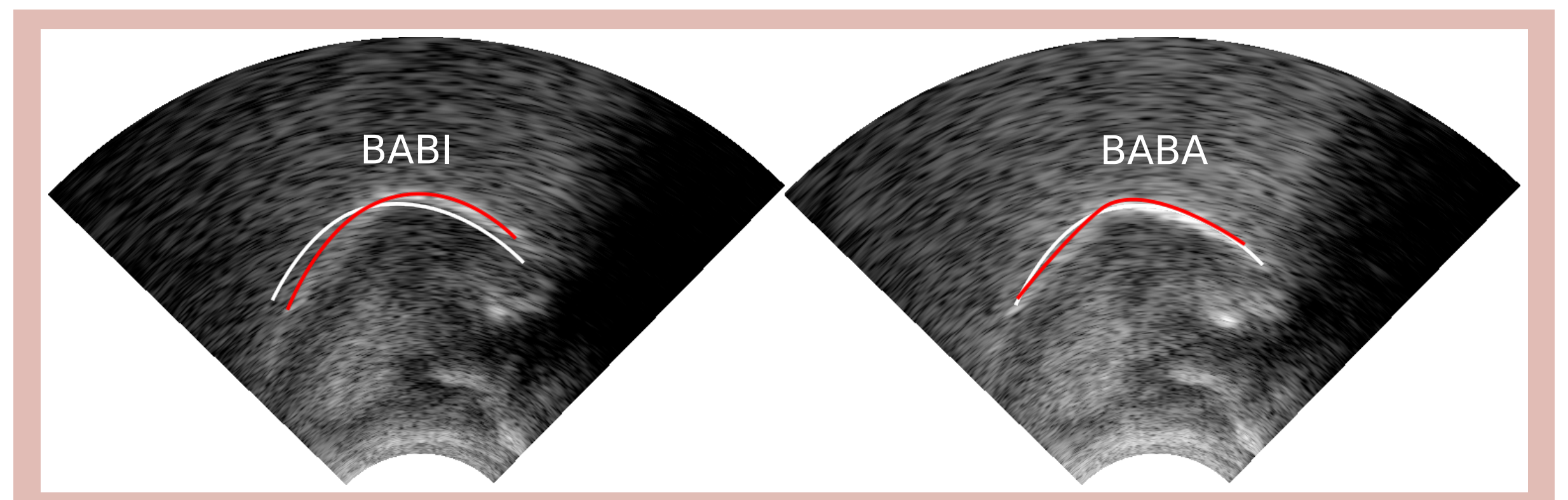


Figure 2: Tongue contours in the midsagittal plane of an ultra sound image at the midpoint and offset in /babi/ and /baba/.

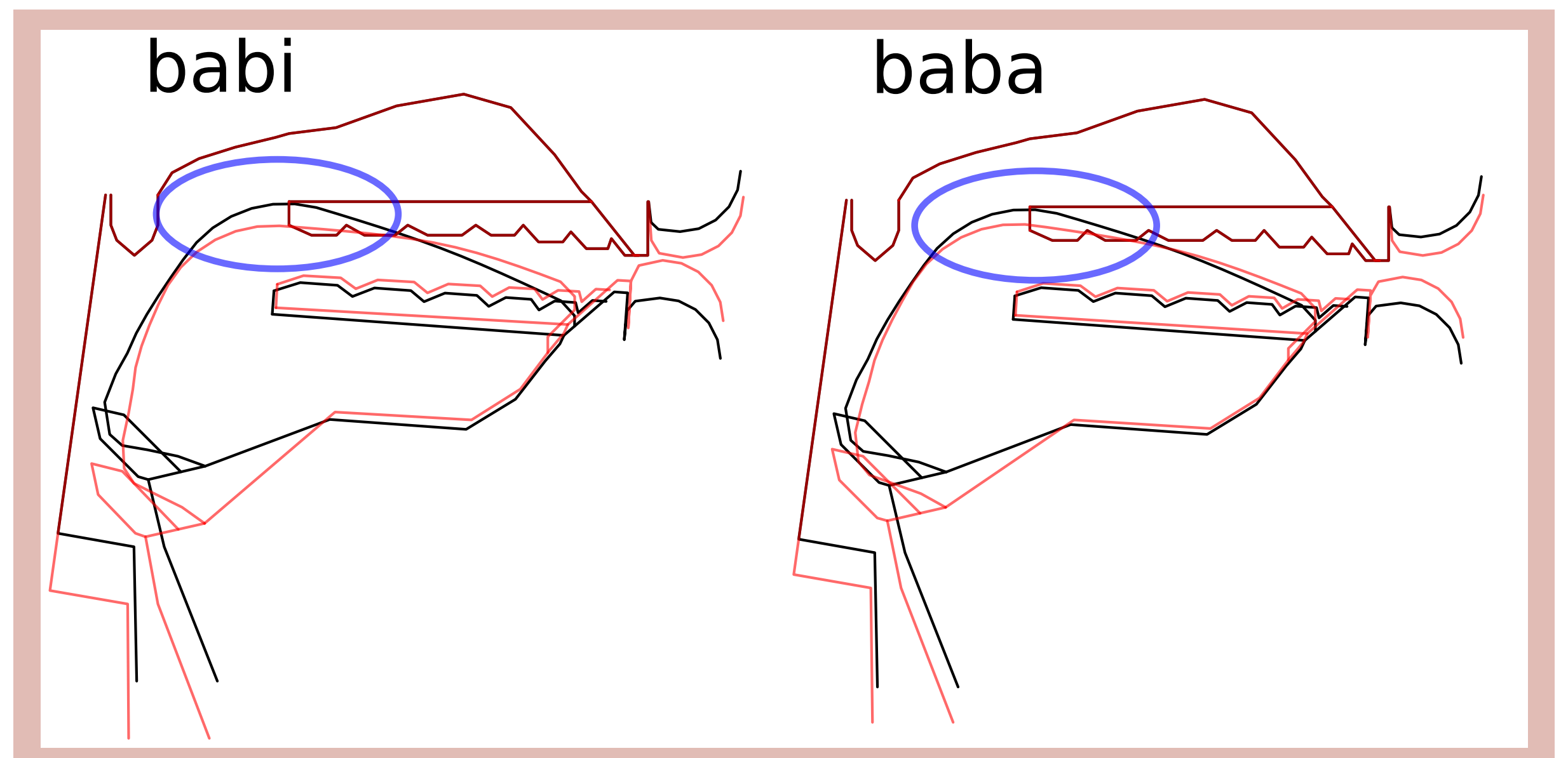


Figure 3: Tongue contours in the midsagittal plane of exported svg image from VTL at the midpoint and offset in /babi/ and /baba/. The virtual tongue raising is not statistically significant.

Data Preparation

- automatically align phone segments with [4]
- extract segments and durations
- time points of interest: midpoint and offset of **first /a/** in each pseudo word
- first two formants in the 20 ms interval before the offset
- tongue height difference between midpoint and offset in relation to ultrasound transducer
- midsagittal picture exported from VTL at midpoint and offset
- picture rotated with respect to lower teeth reference to align with ultrasound transducer
- tongue height difference between midpoint and offset in virtual tongue contour

Results

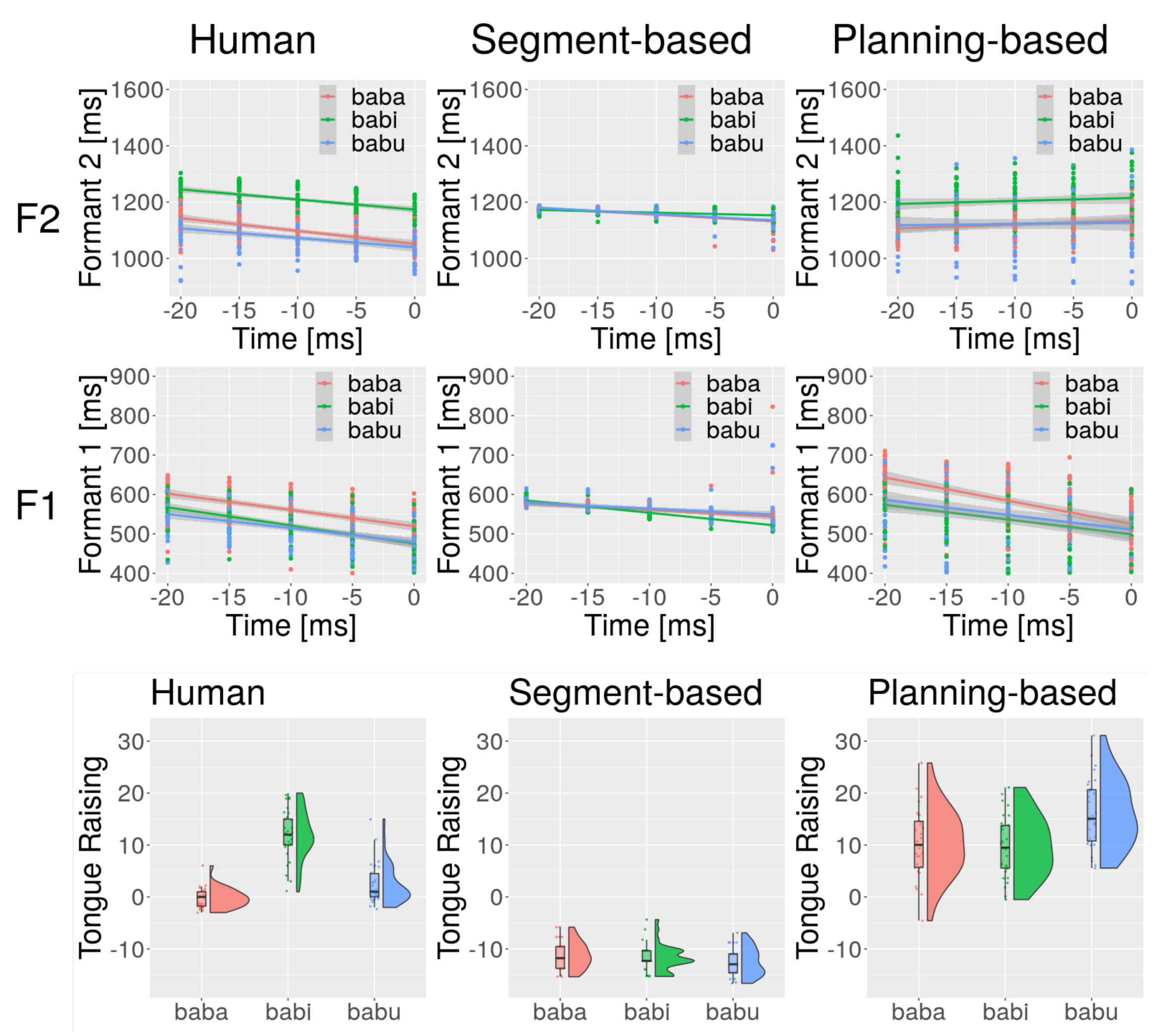


Figure 4: Top panels: Formant shifts of the 20 ms before the offset of the first /a/. Bottom panels: Tongue raising from the midpoint to the offset of the first /a/ in each pseudo word.

Formant Shifts

- human: anticipatory coarticulation in both formants
- segment based resynthesis: no anticipatory coarticulation
- planning based resynthesis: anticipatory coarticulation but fails to mimic the full richness

Tongue Raising

- human: no raising in /baba/, raised in /babi/ and /babu/
- segment based resynthesis: overall lowering
- planning based resynthesis: overall raising

Discussion

- mimicking formant transitions, but with different tongue movements compared to humans
- only one speaker and 90 data points
- planning substantially slower then segment based resynthesis
- no optimisation in the segment based resynthesis as in [3]
- minimal changes in segment based approach might give already very good results
- focus only on first two formants, very specific points in time, and highest point on the tongue, but data is much richer, how to facilitate?
- improve gradient based planning to be informed by semantic embeddings

Conclusion

The artificial utterances /baba/, /babi/, /babu/ repeatedly spoken with a high speaking rate show robust anticipatory coarticulation effect in formant shifts and tongue raising. For fully automatic resynthesis frameworks it is still a challenge to model the full range of human coarticulation. With the recurrent gradient based resynthesis framework anticipatory coarticulation patterns are partially recovered, but it does not seem to achieve this by means of anticipatory tongue raising as humans do.

Acknowledgments: This research was supported by an ERC Advanced Grant (no. 742545) and a collaborative grant from the Deutsche Forschungsgemeinschaft (Spoken Morphology, BA 3080/3-2), awarded to R. Harald Baayen.

References

[1] Peter Birkholz. Modeling consonant-vowel coarticulation for articulatory speech synthesis. *PLOS ONE*, 8(4):1–17, 04 2013.

[2] Peter Birkholz, 2018.

[3] Yingming Gao, Simon Stone, and Peter Birkholz. Articulatory Copy Synthesis Based on a Genetic Algorithm. In *Proc. Interspeech 2019*, pages 3770–3774, 2019.

[4] Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. Montreal Forced Aligner: trainable text-speech alignment using Kaldi. *Proceedings of the 18th Conference of the International Speech Communication Association*, 2017.

[5] Konstantin Sering, Paul Schmidt-Barbo, Sebastian Otte, Martin V Butz, and Harald Baayen. Recurrent gradient-based motor inference for speech resynthesis with a vocal tract simulator. In *12th International Seminar on Speech Production*, 2020.

[6] F. Tomaschek, B. V. Tucker, M. Fasiolo, and R. H. Baayen. Practice makes perfect: the consequences of lexical proficiency for articulation. *Linguistics Vanguard*, 4(s2), 2018.

[7] S.E.G. Öhman. Coarticulation in VCV Utterances: Spectrographic Measurements. *Journal of the Acoustical Society of America*, 39(151):151–168, 1966.