Recurrent Gradient-based Motor Inference for Speech Resynthesis with a Vocal Tract Simulator (ID 67)

ISSP 2020

Konstantin Sering, Paul Schmidt-Barbo, Sebastian Otte, Martin V. Butz, & Harald Baayen

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

Eberhard Karls Universität Tübingen

Overview

- inference principle for speech resynthesis using the **VocalTractLab (VTL)** simulator [1]
- generates smooth and plausible **control parameter (cp-) trajectories** for VTL
- differentiable forward model for imagining acoustic representation as **inner loop**
- physical and geometrical **outer loop** via VTL
- temporal gradient information **minimizes error** between the forward predictor and the target acoustics [4, 2], explicitly incorporating **velocity and jerk constraints**.

Methods

Framework Overview

- **outer loop (slow):** target acoustics ⇒ inverse model ⇒ cp-trajectories ⇒ VTL ⇒ audio ⇒ acoustic representation ⇒ target acoustics
- **inner loop (fast):** cp-trajectories ⇒ predictive model ⇒ acoustic representation ⇒ planning ⇒ cp-trajectories

Action Inference

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

Initial Training

- initial experience for predictive and inverse model (1 hour of speech)
- pairs of cp-trajectories and log mel spectra for German words
- segment based resynthesis of GECO corpus [5, 6]

Loss

- MSE loss: match the acoustics
- jerk loss: as few force changes as possible
- velocity loss: as few position changes as possible

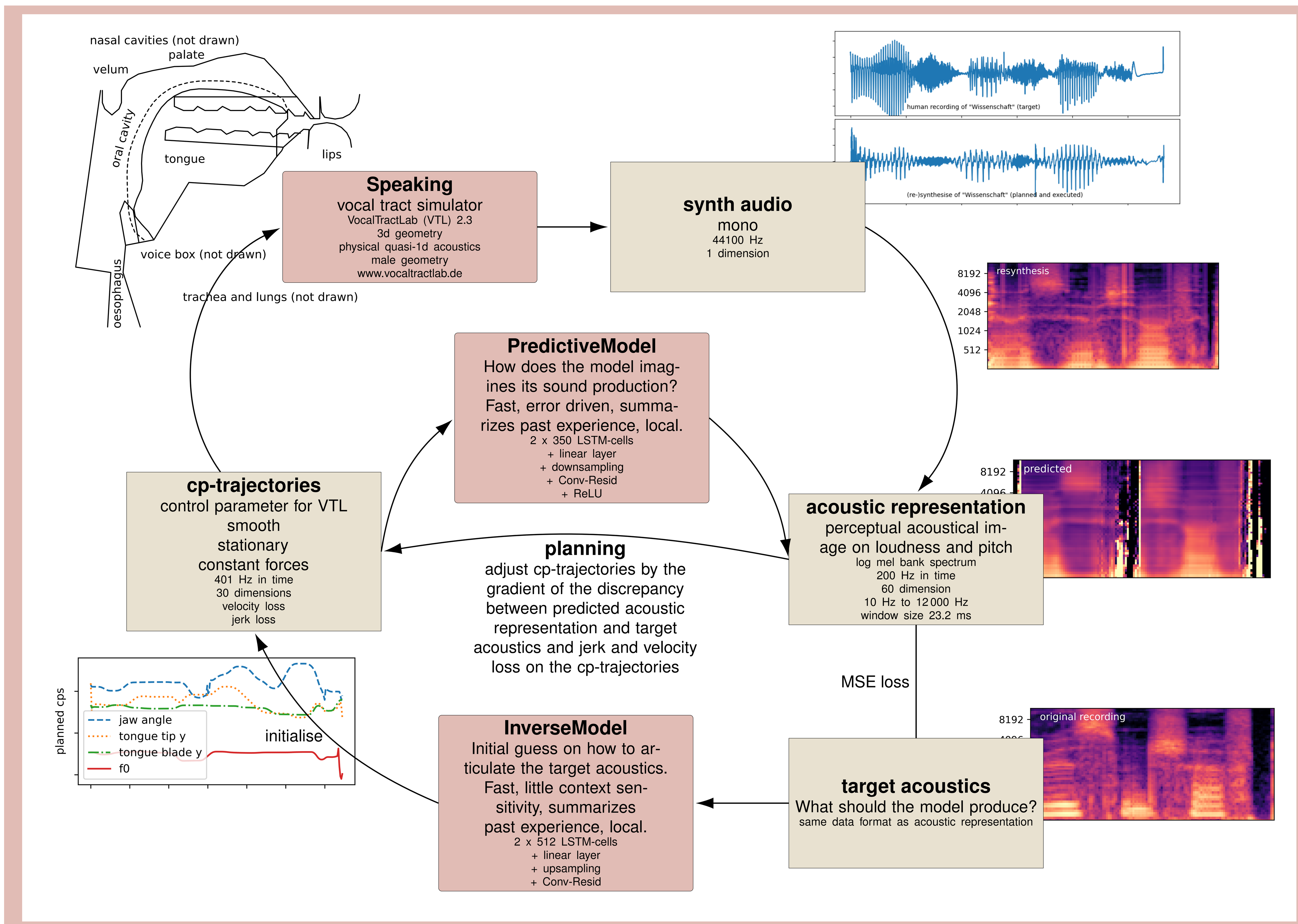

Figure 1: Implementation of the recurrent gradient-based motor inference principle with LSTM based networks. The predictive model imagines the acoustic representation and allows for adjustment prior to execution. The inverse model is only used for initialisation.

Results

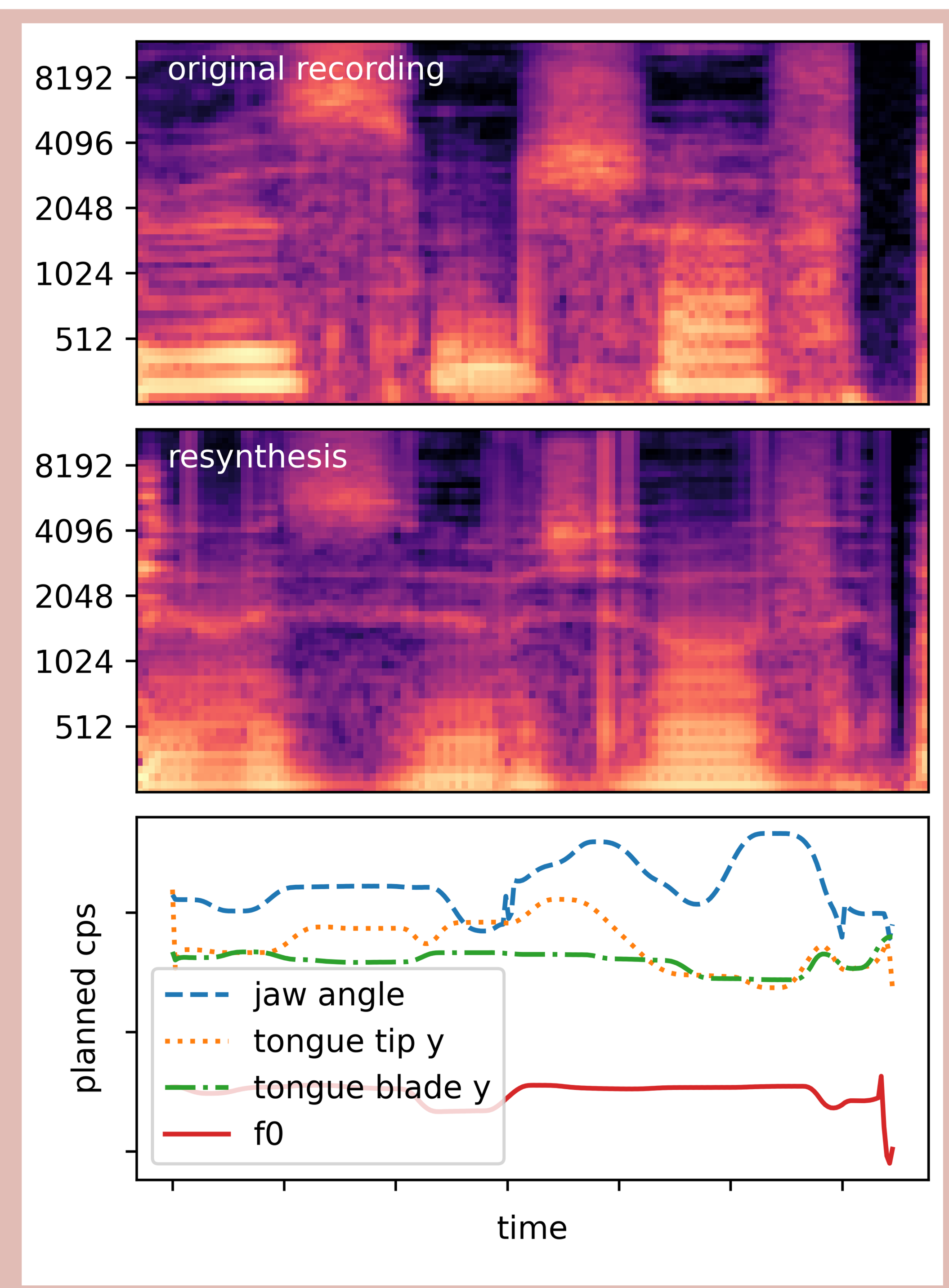

Figure 2: The top panel shows the log mel spectrum of the original human recording *Wissenschaft* (science) used as the target. The middle panel shows the resulting log mel spectrum after the planned trajectories are executed by the VTL. The bottom panel shows four selected cp-trajectories after planning.

- predictive model much faster than VTL synthesis
- good recovery, good generalisation
- optimizes imitation instead of intelligibility
- fails to recover cp-trajectories when initialized with flat neutral gesture
- no global loss-landscape of the VTL
- more evaluations needed, e. g. coarticulation patterns, language transfer

Recovery

- optimize against VTL synthesis
- on initial test data, i. e. segment based resynthesis
- reduction in MSE (produced): $53.2\% \pm 15.8\%$
- final MSE: 0.0706 ± 0.0266
- smoothing of cp-trajectories while keeping MSE error low

Generalisation

- optimized against human audio recording
- female recording vs. male vocal tract geometry
- parallel to test data in recovery
- reduction in MSE (produced): $42.9\% \pm 17.8\%$
- reduction in MSE (predicted): $66.8\% \pm 5.65\%$
- MSE produced vs original: 0.0313 ± 0.0103
- MSE segment-based vs original: 0.0772 ± 0.0246

Limitations

- only longitudinal waves in VTL
- no motor or muscle modeling (pure geometry)
- long computation times
- wave form vs. mel spec vs. mfcc
- imitating on the cost of intelligibility

Future plans

- tool for studying mechanics of human speech generation
- change objective to intelligibility
- evaluate motor dynamics
- compare coarticulation patterns with humans
- from isolated words to words in context
- goal babbling, learning without initial training data
- second language acquisition and dialect
- integrate into the Linear Discriminative Learning model of the mental lexicon [3]

Conclusion

Recurrent gradient-based motor inference for speech resynthesis with a vocal tract simulator successfully generates input control-parameter trajectories for a vocal tract simulator. Initial evaluation runs indicate that the model combines both flexibility and stability, but more stringent testing is required.

Acknowledgments: This research was supported by an ERC advanced Grant (no. 742545).

References

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

[2] Martin V Butz, David Bilkey, Dania Humaidan, Alistair Knott, and Sebastian Otte. Learning, planning, and control in a monolithic neural event inference architecture. *Neural Networks*, 117:135–144, 2019.

[3] Y.Y. Chuang, M-L. Vollmer, E. Shafaei-Bajestan, S. Gahl, P. Hendrix, and R. H. Baayen. The processing of pseudoword form and meaning in production and comprehension: A computational modeling approach using linear discriminative learning. *Behavior Research Methods*, page in press, 2020.

[4] Sebastian Otte, Theresa Schmitt, Karl Friston, and Martin V Butz. Inferring adaptive goal-directed behavior within recurrent neural networks. In *International Conference on Artificial Neural Networks*, pages 227–235. Springer, 2017.

[5] Antje Schweitzer and Natalie Lewandowski. Convergence of articulation rate in spontaneous speech. In *INTERSPEECH*, pages 525–529, 2013.

[6] Konstantin Sering, Niels Stehwien, Yingming Gao, Martin V Butz, and Harald Baayen. Resynthesizing the geco speech corpus with vocaltractlab. *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2019*, pages 95–102, 2019.