Overview

- inference principle for speech resynthesis using the Vocal-TractLab (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

Results

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

Loss

- 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]

Initial Training

- predict 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
- an initial test data, i.e. segment based resynthesis
- reduction in MSE (produced): 54.2% ± 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% ± 17.8%
- reduction in MSE (predicted): 66.8% ± 5.6%
- MSE produced vs original: 0.0313 ± 0.0101
- 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

- parallel to test data in recovery
- optimized against human audio recording
- female recording vs. male vocal tract geometry
- parallel to test data in recovery
- reduction in MSE (produced): 42.9% ± 17.8%
- reduction in MSE (predicted): 66.8% ± 5.6%
- MSE produced vs original: 0.0313 ± 0.0101
- MSE segment-based vs original: 0.0772 ± 0.0246

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

[6] Konstantin Sering, Eric Schuman, Yaqi Peng, Martin V. Butz, and Harald Bayer. 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. Acknowledgements: This research was supported by an ERC advanced Grant (no. 742545).

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