RECURRENT GRADIENT-BASED MOTOR INFERENCE FOR SPEECH RESYNTHESIS WITH A VOCAL TRACT SIMULATOR

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

Eberhard Karls Universität Tübingen
konstantin.sering@uni-tuebingen.de, paul.schmidt-barbo@student.uni-tuebingen.de,
sebastian.otte@uni-tuebingen.de, martin.butz@uni-tuebingen.de, harald.baayen@uni-tuebingen.de

This study is part of an ongoing project addressing the challenge of learning an inverse mapping between acoustic features and control parameter trajectories (cp-trajectories) of a vocal tract simulator. We apply a recurrent gradient-based motor inference, which is similar to the active-inference-model-predictive principle (cf. [1], [2], [3]) to the dynamics of the vocal tract simulator developed by [4], the VocalTractLab (VTL). Implementing this principle, We implemented a predictive forward model that starts with the cp-trajectories, which constitute the inputs to the simulator, and learns to predict the corresponding acoustic representations (log-mel-banks spectra). If the predictions of the forward model are of sufficient quality, the gradients of the forward model can be used to plan the cp-trajectories for a given acoustic target.

At first sight, using the gradients of a predictive forward model to infer and plan the trajectories of an inverse mapping might seem to be counter-intuitive. Why not to learn the inverse mapping directly? However, an important advantage is that it becomes easy to adjust the parameters of cp-trajectories dynamically over time. We can first execute pertinent cp-trajectories up to a specific point in time, which places the vocal tract simulator in a specific state, and then plan the optimal follow-up cp-trajectories required to reach a specific acoustic target.

VTL is a 3-dimensional geometrical vocal tract simulator linked with a quasi-1-dimensional acoustic synthesis model. VTL makes it possible to synthesise speech from cp-trajectories that define the geometrical shape of the vocal tract, the properties of the glottis model, and lung pressure, for each subsequent step in discretized time. For each 10 millisecond time step, 33 parameters control the shape of the articulatory system. Until now, cp-trajectories have been derived mostly by means of a dominance model that takes gestural scores as input. Unfortunately, creating and combining gestural scores involves a considerable amount of handcrafting.

Figure 1 lays out the structure of our implementation of the active-inference-model-predictive principle. The explicit inverse model provides a first guess for initialization as to how to move the articulators in order to approximate the intended target acoustics. The active inference tunes the control parameters by projecting forward what acoustics the currently planned cp-trajectories would produce, and changing the control parameters such that the speech output will be more more similar to the targeted acoustics.

Our implementation calculates the log-mel-bank spectrum for an acoustic target. This spectrum is then used as input for the direct inverse model in order to initialise the pertinent cp-trajectories. Subsequently, the initial cp-trajectories are forwarded through the predictive forward model and the predicted log-mel-bank spectrum is compared to the targeted spectrum. On the basis of the MSE-loss between the two spectra, the local gradients for the input cp-trajectories are calculated and the initial cp-trajectories are adjusted by 0.1 times the local gradients. The resulting adjusted cp-trajectories are again feed into the predictive forward model and are again adjusted. This procedure is repeated ten times in total.

Figure 2 shows the wave forms of a recorded utterance of the German word *Wissenschaft* ('science'), which is used as target, along with the wave form of an optimized resynthesised spectrum created with the active-inference-predictive-model principle and the VTL simulator. The overall intonation pattern and the general sound pattern are good enough to afford recognition by the human ear,

but as the resynthesis does not have finer details of articulation available to it, the synthesized speech is approximate only.

A first next step for future research is to model between-word coarticulation patterns. A long-term goal is to integrate the present learning-driven model of articulation into the Linear Discriminative Learning model of the mental lexicon [5].

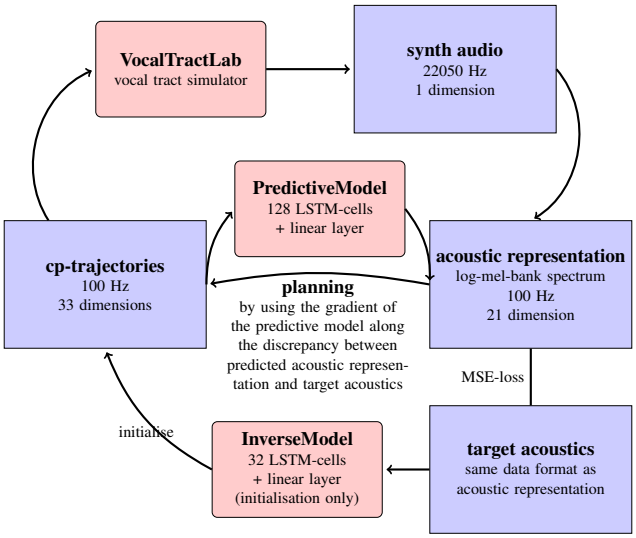

Figure 1 – The active-inference-model-predictive principle applied to a vocal tract simulator. The main idea is using the discrepancy between the predicted acoustics and the target acoustics to fine tune or plan the cp-trajectories, which are the inputs to the vocal tract simulator. The direct inverse model is only used for initialisation.

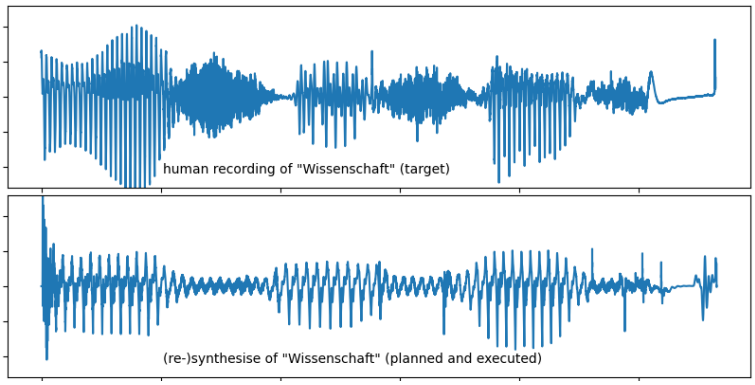

Figure 2 – The top panel shows the original human recording used as the target for the active-inference-model-predictive principle. After planning the cp-trajectories the vocal tract simulator produces the wave form in the bottom panel. The overall pattern is recovered.

[1] FRISTON, K., S. SAMOTHRAKIS, and R. MONTAGUE: *Active inference and agency: optimal control without cost functions*. *Biological cybernetics*, 106(8-9), pp. 523–541, 2012.

[2] OTTE, S., T. SCHMITT, K. FRISTON, and M. V. BUTZ: *Inferring adaptive goal-directed behavior within recurrent neural networks*. In *International Conference on Artificial Neural Networks*, pp. 227–235. Springer, 2017.

[3] BUTZ, M. V., D. BILKEY, D. HUMAIDAN, A. KNOTT, and S. OTTE: *Learning, planning, and control in a monolithic neural event inference architecture*. *Neural Networks*, 117, pp. 135–144, 2019.

[4] BIRKHOLZ, P.: 2018. URL `http://www.vocaltractlab.de/index.php?page=vocaltractlab-about`.

[5] CHUANG, Y., 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*, p. in press, 2020.