SOMATOSENSORY FEEDBACK IN PAULE

Konstantin Sering, Paul Schmidt-Barbo

Eberhard Karls Universität Tübingen
konstantin.sering@uni-tuebingen.de

Abstract: A somatosensory pathway is added to the Predictive Articulatory speech synthesis Utilizing Lexical Embeddings (PAULE) model. The different choices that lead to the specific somatosensory representation and the pathway are discussed. PAULE is a continuously improved control model for the articulatory speech synthesizer VocalTractLab (VTL) that directly facilitates a meaning representation to find suitable motor trajectories and does not use any symbolic units neither for the motor representation nor for the acoustic or semantic representation. The somatosensory representation consists of the minimal cross-sectional area in each of the most frontal 1-centimeter intervals of the oral cavity of the VTL plus the incisor position, the tongue tip elevation, and the velum opening. In the somatosensory pathway the 10-dimensional somatosensory representation is used as an intermediate representation before predictions in the acoustic and semantic goal space are compared against targets. The semantic and acoustic sources of error along the somatosensory and along the acoustic pathway are added together with an effort minimization term on the control parameter (cp-)trajectories of the VTL to form an additive loss. This additive loss is minimized to plan optimal cp-trajectories that result in a copy-synthesis of a target acoustics with the VTL.

1 Introduction

This contribution is part of an ongoing research project that implements articulatory speech production on the word and short phrase level without any symbolic acoustic representation like phones nor any symbolic motor representation like gestural scores. Furthermore, speech production is started from and directly driven by the goal of transferring meaning in form of a semantic vector embedding through the articulatory and acoustic channel. The underlying assumption is that human speech production is a way to transfer meaning from the speaker to the listener. By optimizing the articulatory movement trajectories in the semantic goal space, the word meaning has a direct influence on the articulatory movement trajectories. The model that implements these ideas is the Predictive Articulatory speech synthesis Utilizing Lexical Embeddings (PAULE) model [1].

PAULE is a control model for the articulatory speech synthesizer VocalTractLab (VTL) [2, 3]. PAULE can find suitable control parameter (cp-)trajectories for the VTL in three tasks. The cp-trajectories define the position of the articulators like the jaw and the tongue as well as parameters of the glottis and the sub-glottal lounge pressure. With a given set of cp-trajectories the VTL synthesizes a 44,100 Hz mono audio signal. The three tasks in which PAULE finds suitable cp-trajectories are the acoustic-only (copy-synthesis) task, the semantic-only task, and the semantic-acoustic task. In all tasks PAULE optimizes the cp-trajectories in a semantic and acoustic goal space. All tasks can be solved on the word or short phrase level, which equals to 100 to 1,000 time-steps for the cp-trajectories.

To solve these tasks, PAULE uses an internal forward model to predict the effect of an upcoming articulation. The effect of the articulation is predicted in the acoustic and semantic domain. Both predictions are compared against a target acoustics and a target semantics. From the comparison an error is derived, which is used to improve the upcoming articulation. Through the prediction step and the error driven correction PAULE optimizes cp-trajectories in a goal space and therefore PAULE has no need for predefined motor targets or gestural scores that categorize the motor space. The problem of finding suitable movements in the motor-space is solved indirectly by minimizing the acoustic and semantic distance to an acoustic and semantic target and having soft low-effort constraints in the motor space.

Through the goal directed error driven behavior the planning in PAULE is adaptive to situational changes like conditioning on different past cp-trajectories. The mappings used for the forward prediction and the error driven optimization are learned mappings that encode the experience of articulations produced by the VTL. Therefore, PAULE incorporates learning biases and depends on the knowledge that is already incorporated into the predictive forward model. Furthermore, the better the knowledge in the internal models for the forward prediction are and the better they approximate the VTL, the better the resulting articulations are. This experience based improvements can be interpreted as a training effect of the articulators.

Limitations of the PAULE model are that it neither has visual input of the articulation nor has any tactile feedback on the position of the articulators in the VTL. In this contribution, we address the missing somatosensory feedback by introducing and integrating a somatosensory pathway into PAULE. The somatosensory pathway implements a feedback loop that differs from the acoustic feedback in the fact that no somatosensory target is defined. Instead additional mappings between the somatosensory and the acoustic domain and the somatosensory and semantic domain are learned. This can be interpreted the following: PAULE gets an understanding on how a specific somatosensory impression sounds like and which word meanings are directly connected to it. Figure 3 shows the added somatosensory representation and the added mappings, which are highlighted with a yellow background. It can be seen that no somatosensory target is defined, but the planning error flows from the acoustic and semantic representation back through the somatosensory pathway to the cp-trajectories.

The main motivation behind implementing a somatosensory pathway is that this pathway is used by humans for properly controlling the articulators and furthermore because other computational models of human speech production like the DIVA [4] and the FACTS [5] model use a somatosensory error correction mechanism. The somatosensory pathways help the models to be aware of the position of the articulators and hopefully to quickly adapt to external distortions to the articulatory system like a bite block that restricts the jaw movement. In contrast to the PAULE model, both, the DIVA and the FACTS model lack the feature to directly map from and to a semantic representation like semantic embedding vectors of distributional semantics, which is a primary feature of PAULE.

2 Methods

The focus of this section is the somatosensory pathway and not the acoustic pathway or the initialization within PAULE. The interested reader is referred to [1, 6, 7] to read about the different initializations and the acoustic pathway in more detail. Nevertheless, the following explanations can be understood without reading any background material.

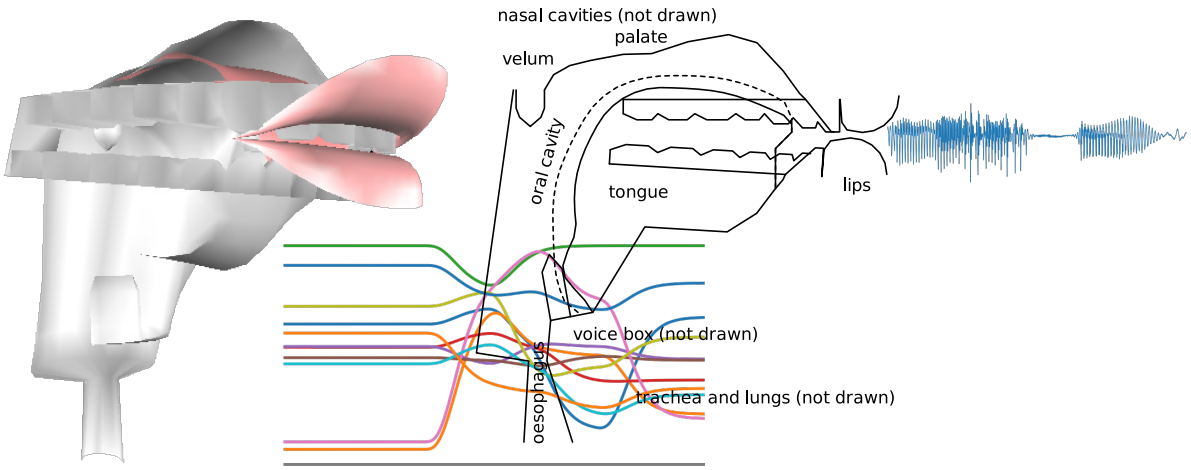

Figure 1 – The VTL uses a 3-dimensional deformable geometrical model of the oral cavity (shown on the left). A mid-saggital slice of this 3-dimensional model shows different articulators. The position of these articulators and additional parameters for the glottis and the lounge pressure are defined through 30 control parameter (cp)-trajectories (11 are depicted as colorful lines in the background). For a given set of cp-trajectories the VTL synthesizes a 44,100 Hz mono audio signal (depicted as blue wave form).

2.1 VocalTractLab (VTL)

The articulatory speech synthesizer VocalTractLab (VTL) uses a 3-dimensional geometrical model of the vocal tract (see Figure 1) and a quasi 1-dimensional acoustic model. The VTL is the simulator that is controlled by PAULE. The movements of the articulators in this 3-dimensional model are defined by control parameter (cp)-trajectories, which are 30 smooth lines over time. The movements of the articulators result in different sizes of the oral cavity. The different sizes are characterized in an intermediate step by an area function (see Figure 2) and finally a 44,100 Hz mono audio signal is synthesized from the time-series of the area function. The area function characterizes the cross-sectional area along the midline of the oral cavity with 40 variably long segments. Therefore, the area function is fully defined by 80 values (40 values for the length and 40 values for the cross-sectional area). Three additional values are important for the the acoustic simulation of the VTL: the incisor position, the tongue tip side elevation, and the velum opening. From all these values and some additional information on quality of the boundaries of the segments in the area function and the nasal cavity the audio signal is derived.

2.2 Somatosensory representation

The somatosensory representation is derived from the frontal part of the area function. As the area function defines the cross-sectional area of the oral cavity, it characterizes the relative proximity of the articulators. We assume that articulators that are close to each other feel each other, even if there is still some opening for the air to travel through. As we aim for a low dimensional representation and assume that the perception within the mouth is not perfectly precise, we use the lowest area for each of the seven 1-centimeter intervals in the frontal part of the mouth. Therefore, the PAULE model knows how close the articulators are in the frontal part of the mouth without giving the precise position. Figure 2 shows the area function as well as the somatosensory representation for the vowels /a/ and /i/. For the full somatosensory representation the seven values, which are derived from the area function, are complemented by the incisor position, the tongue tip elevation, and the velum opening. This results in 10 values per time step. The time step size is the same as in the cp-trajectories, namely every 110 sample of the final 44,100 Hz signal or roughly every 2.5 milliseconds.

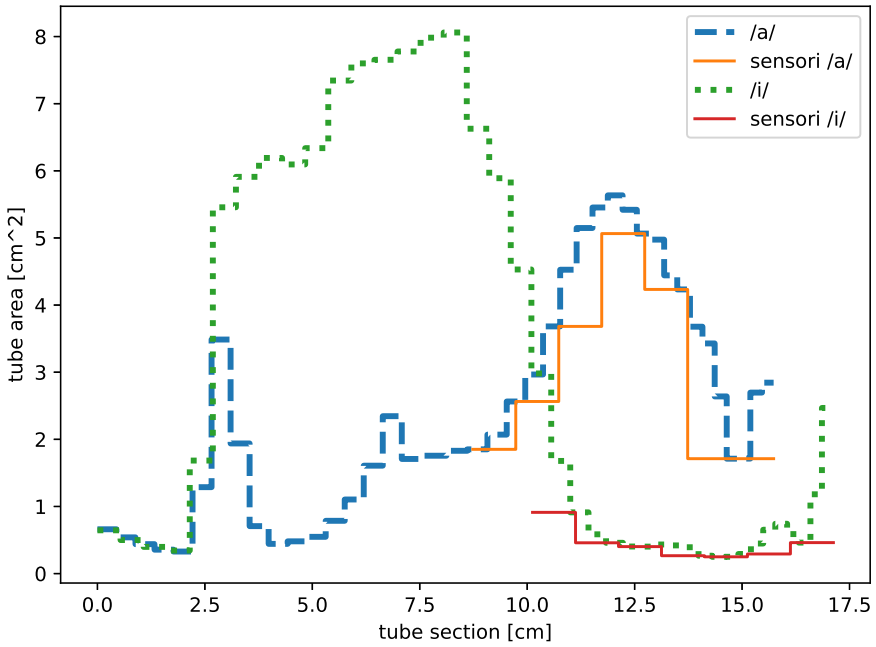

Figure 2 – The somatosensory representation is operationalized as the minimal cross-sectional area for each of the seven most frontal 1-centimeter intervals plus the incisor position, the tongue tip elevation, and the velum opening. The area function with its 40 tube segments of different length (x-axis) and different cross-sectional area (y-axis) is shown for an /a/ and an /i/ sound as well as the seven values derived from the area function for the somatosensory representation.

2.3 Somatosensory pathway

Figure 3 shows how the somatosensory representation is integrated into the PAULE model. It is used similarly to the acoustic representation with the important difference that no target somatosensory representation is used. Like the acoustic representation a predictive forward model is used to imagine the upcoming somatosensory representation, instead of comparing the prediction to a target, an acoustic representation and a semantic representation is predicted from predicted somatosensory representation. These predictions can be interpreted the following way: Assuming the model expects these constrictions in the frontal part of the mouth, which acoustics and which semantics is expected from these. The next step is to compare the predicted acoustics and semantics with the target acoustics and target semantics and back-propagate the error between the prediction and target to the cp-trajectories. In the internal loop the cp-trajectories are then improved by combining back-propagated errors from he predicted acoustics, semantics along the somatosensory and the acoustic pathway as well as the low effort constraints of minimal velocity, which results in stationary curves, and minimal jerk, which results in curves that follow from constant forces.

After several iterations of the internal loop, where the cp-trajectories are subsequently refined, the improved cp-trajectories are used as the inputs to the VTL and audio is synthesized. As a byproduct of the synthesis an actual produced somatosensory representation is generated. Together with the cp-trajectories the produced somatosensory representation is used to further improve the predictive somatosensory forward model and the somatosensory-to-acoustic model. In this outer loop, the PAULE model experiences its planned cp-trajectories in terms of produced audio and its corresponding acoustic and semantic representation and its produced somatosensory representation. This can be conceptualized as the model listening to itself and experiencing its somatosensory feedback.

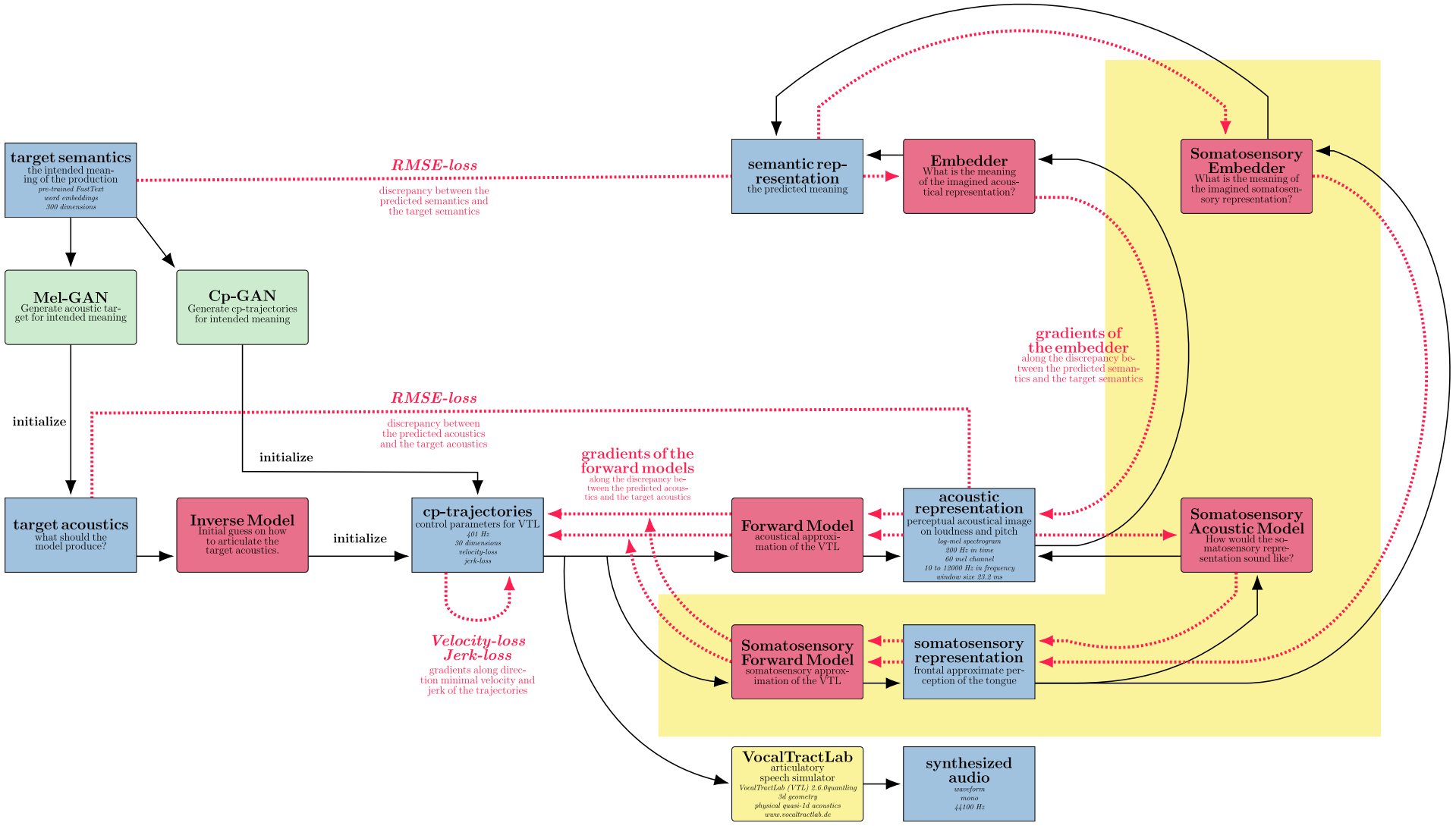

Figure 3 – The somatosensory pathway is added to PAULE (highlighted with the yellow background). It consists of the somatosensory representation as a data structure and three models: the somatosensory forward model, the somatosensory-to-acoustic model, and the somatosensory embedder. In this overview, the forward prediction is shown as black, solid arrows and the backwards error flow is shown as red dotted arrows. This is the internal loop of the PAULE model, which is used to iteratively plan cp-trajectories for the VTL along the acoustic and somatosensory pathway.

2.4 Internal model definitions

The *somatosensory forward model* maps the time-series of the 30-dimensional cp-trajectories to the 10-dimensional somatosensory representation. This model therefore shortcuts the VTL synthesis. The forward model is implemented with a single Long-Short-Term-Memory (LSTM) layer [8] with 360 cells. The *somatosensory-to-acoustic model* maps the time-series of the 10-dimensional somatosensory representation to the 60-dimensional log-mel spectrogram acoustic representation. This model is implemented with a single LSTM-layer with 360 cells followed by an average pooling that combines two consecutive time steps and with that halves the time resolution. The *somatosensory embedder* maps the time-series of the 10-dimensional somatosensory representation onto a fixed 300-dimensional semantic vector. This is achieved with two layers of LSTM-layer with 720 cells each and a dropout during training of 0.7. [1]

2.5 Training

All models are trained on a resynthesized data set derived from a subset of the German Mozilla Common voice corpus. It is the same data set as described and used in [1] with the additionally derived somatosensory representation. The training data is resynthesized with segment-based approach [9]. All models use the ADAM optimizer [10] during training and an Root Mean Squared Error (RMSE) loss. The somatosensory forward model and the somatosensory-to-acoustic model are trained for 100 epochs total. For the first 50 epochs the initial learning rate is set to 0.001 and for the second 50 epochs the initial learning rate was reduced to 0.0001. The somatosensory embedder is trained for 200 epochs with an initial learning rate of 0.0001. To foster a better gradient landscape in the embedder in each epoch random noise was added to the semantic vectors (the output), which was sampled from a Gaussian with a mean of 0 and a variance of 6×10^5. The training of all three models consumed roughly 12 kWh of electric energy.

2.6 Evaluation

To evaluate the additional models the final validation loss for each model is compared to the validation loss of the corresponding models in the acoustic pathway. The somatosensory forward model has a final validation loss of 0.00231. This is a magnitude smaller than the predictive forward model with a final loss of 0.0188. Note however that the somatosensory forward model maps on the somatosensory representation and not on an acoustic representation and therefore this comparison is only approximated. A better comparison can be made between the somatosensory-to-acoustic model and the predictive forward model. The somatosensory-to-acoustic model has a final RMSE loss of 0.0587, which is roughly three times higher compared to the predictive forward model. As the errors of the somatosensory forward model and the somatosensory-to-acoustic model combine along the somatosensory pathway, the final error is around five times higher compared to the acoustic pathway. The somatosensory embedder has a final RMSE loss of 0.00726, which is 1.5 times smaller compared to the acoustic embedder. Therefore the combined error in the semantic domain is a magnitude smaller along the somatosensory pathway compared to the acoustic pathway on the validation data set.

Evaluating the planning quality along the different pathways in the PAULE model, yields the following preliminary results. The synthesis quality purely along the somatosensory pathway, i. e. without using the acoustic pathway, is less capable of copy-synthesizing the acoustic

[1] The PyTorch definitions of the models can be found in the `paule` Python package `https://github.com/quantling/paule`. Version 0.3.4 is used at the time of writing.

target and of hitting the semantic target, compared to the acoustic pathway. Along the somatosensory pathway only, the production loss is 0.123 in the acoustic domain and 0.0659 in the semantic domain, which is a loss reduction compared to the initial loss of 19% in the acoustic domain and of 22% in the semantic domain. Along the combined acoustic and somatosensory pathway the production loss is 0.106 (loss reduction of 30%) in the acoustic domain and 0.0359 (loss reduction of 57%) in the semantic domain. The acoustic-only pathway shows slightly better results in the acoustic domain with a final production loss of 0.096 (loss reduction of 37%), but shows worse results in the semantic domain with a production loss of 0.0618 (loss reduction of 27%) in the semantic domain. Listening to the synthesized audio gives us the subjective impression that the synthesis, which includes the somatosensory pathway is stronger articulated compared to the acoustic pathway alone. Adding the somatosensory pathway comes with a computational cost though. Computation time is increase by a factor of 1.5, which leads to planning times of PAULE of roughly 5400 seconds (1 hour and 30 minutes) per 1 second of speech that is resynthesized.

3 How to use?

The easiest way to run the PAULE model is to install the Python package named `paule` via `pip` with `pip install paule` and to run the minimal example code [2]. The code for the `paule` Python package is free and open source software and can be found on GitHub [3].

4 Discussion

The somatosensory representation and pathway presented here is the one we found most promising and most natural to implement into the PAULE model. Nevertheless, there are other options available to define a somatosensory representation and with which pathway to integrate it into the PAULE model.

The benefit of the seven minimal cross-sectional areas plus the incisor position, the tongue tip elevation, and the velum opening is that it is lower dimensional compared to the 30-dimensional cp-trajectories and that it is easily derived from data available in the VTL. An alternative representation explored by us is the full 80-dimensional area function plus the three additional variables mentioned earlier, which resulted in a 83-dimensional representation. These 83-values are fully determined by the 30-control parameters, which leads to a numerically unstable mapping, which makes it hard to approximate with a forward model

The somatosensory pathway introduced here is different to the acoustic pathway in the sense that no somatosensory target is defined. Instead the somatosensory representation is directly projected to the acoustic and the semantic domain and there compared to the acoustic and semantic target. This way the somatosensory experience is not explicitly encoded as somatosensory targets, but by the implicit knowledge in the mappings to the acoustic and semantic domain. The better the PAULE model learns which somatosensory representations are connected to which acoustic realisations and to which semantics, the better it can use the error along this mapping to improve the cp-trajectories during planning.

Alternatively, a somatosensory target could be used, which is either given a priory or by deriving it from the target semantics or target acoustics. Deriving it from the target semantics equates to having a clear expectation on how the somatosensory experience should be for uttering a specific phrase or word meaning. Deriving it from the target acoustics equates to having a clear expectation on how the somatosensory experience should be for a specific acoustic image.

[2] `https://github.com/quantling/paule/blob/main/docs/examples/minimal_example.py`
[3] `https://github.com/quantling/paule`

As human cognition tends to form these kinds of expectations, implementing the somatosensory route by defining an explicit somatosensory target would be a valid design choice as well. Nevertheless, with this contribution our goal was to show how a somatosensory feedback loop can be implemented without any explicit target and therefore in an indirect way. The pathway presented here is still only optimizing in the acoustic and semantic domain, but fosters this optimization not only along an acoustic pathway, but additionally along a somatosensory pathway. This is possible as the optimization, also called planning, is conducted along an additive loss, which integrates different sources of error.

4.1 Conclusion

We show how a somatosensory pathway can successfully be integrated into the PAULE model. The PAULE model yield cp-trajectories, the inputs of the VTL articulatory speech synthesizer, that results typically in intelligible speech, without using any symbolic acoustic, motor, or semantic representations. Furthermore, PAULE generalizes well to acoustic targets it has never experienced or synthesized beforehand. Interested readers can use the `paule` Python package to create their own copy-synthesis.

Acknowledgments This research was supported by an ERC Advanced Grant (no. 742545) awarded to R. Harald Baayen. Konstantin Sering and Paul Schmidt-Barbo are members of the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 390727645.

References

[1] SCHMIDT-BARBO, P., S. OTTE, M. V. BUTZ, R. H. BAAYEN, and K. SERING: *Using semantic embeddings for initiating and planning articulatory speech synthesis*. In O. NIEBUHR, M. S. LUNDMARK, and H. WESTON (eds.), *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2022*, pp. 32–42. TUDpress, Dresden, 2022.

[2] BIRKHOLZ, P.: *Modeling consonant-vowel coarticulation for articulatory speech synthesis*. *PLOS ONE*, 8(4), pp. 1–17, 2013. doi:10.1371/journal.pone.0060603. URL `https://doi.org/10.1371/journal.pone.0060603`.

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

[4] TOURVILLE, J. A. and F. H. GUENTHER: *The diva model: A neural theory of speech acquisition and production*. *Language and cognitive processes*, 26(7), pp. 952–981, 2011.

[5] PARRELL, B., V. RAMANARAYANAN, S. NAGARAJAN, and J. HOUDE: *The FACTS model of speech motor control: Fusing state estimation and task-based control*. *PLoS computational biology*, 15(9), p. e1007321, 2019.

[6] SCHMIDT-BARBO, P., E. SHAFAEI-BAJESTAN, and K. SERING: *Predictive articulatory speech synthesis with semantic discrimination*. *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2021*, pp. 177–184, 2021.

[7] SERING, K., P. SCHMIDT-BARBO, S. OTTE, M. V. BUTZ, and H. BAAYEN: *Recurrent gradient-based motor inference for speech resynthesis with a vocal tract simulator*. In *12th International Seminar on Speech Production*. 2020.

[8] HOCHREITER, S. and J. SCHMIDHUBER: *Long short-term memory*. *Neural computation*, 9, pp. 1735–80, 1997. doi:10.1162/neco.1997.9.8.1735.

[9] SERING, K., N. STEHWIEN, Y. GAO, M. V. BUTZ, and H. BAAYEN: *Resynthesizing the geco speech corpus with vocaltractlab*. *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2019*, pp. 95–102, 2019.

[10] KINGMA, D. P. and J. L. BA: *Adam: A method for stochastic optimization*. *3rd International Conference for Learning Representations*, abs/1412.6980, 2015.