ISSP 2024 - 13th International Seminar on Speech Production
13-17 May 2024, Autrans, France

Articulatory speech synthesis without phones and gestures?

Konstantin Sering[1], R. Harald Baayen[1]

[1]*University of Tübingen*
konstantin.sering@uni-tuebingen.de, harald.baayen@uni-tuebingen.de

Abstract

With this work we show how speech production can be modelled on the word level without any symbolic units, neither on the acoustic side like phonemes, nor on the semantic side like word types, nor on the motor side like gestures or articulatory targets. We present and discuss a computational model of articulatory speech production, which implements a predictive planning approach, known from hand and arm movements, into the articulatory domain. This computational model is named Predictive Articulatory speech synthesis Utilizing Lexical Embeddings (PAULE). As articulatory speech synthesizer the VocalTractLab speech synthesizer is used, which simulates the human speech system on a geometrical level with 30 different control parameters (channels) and with a time resolution of 401 Hertz. As the synthesis quality of the PAULE shows decent results, we conclude that human speech production can be modelled without the use of any symbolic units like phones and gestures on the word level.

Keywords: speech production, articulatory speech synthesis, predictive planning, motor control, sequence-to-sequence model

1. Introduction

The Predictive Articulatory speech synthesis model Utilizing Lexical Embeddings (PAULE) is a computational model for speech production that does not use any gestures or targets on the motor side nor any phone representation on the acoustical side (Schmidt-Barbo et al. 2022; Sering 2023). Instead it solves the task of finding suitable control parameter trajectories for the 30-dimensional speech simulator VocalTractLab (Birkholz 2013)[1] by optimizing the effect of the control in an acoustic and semantic goal space.

Several models for speech production have been proposed in the literature. Some are computationally implemented (Dell 1984; Levelt, Roelofs, and Meyer 1999), others provide more programmatic blueprints of what the production architecture might look like Fromkin (1984). What all these theories have in common is that they take sublexical units such as phonemes (the contrastive sounds of a language) and morphemes (taken to be the minimal meaning bearing units) as given, the assumption being that they provide an undisputable ground truth for theory development and computational modeling.

Another conviction shared by all these models is that production and comprehension are largely separated processes. Although, for instance, the model of Levelt, Roelofs, and Meyer (1999) takes into account that speakers are their own listeners, any systematic interaction and integration between comprehension and production is not on the horizon. In fact, the very nature of the cognitive systems underlying production and comprehension were argued by Levelt to be fundamentally different, with comprehension involving statistical inferencing from sound to phoneme sequences, but production involving a cascaded and largely interference-free sequence of selection mechanisms for lemmas, lexemes, morphemes, phonemes, and syllables.

Furthermore, the abovementioned models are static models, models that do not learn. The parameters of these models have to be set by hand. The role that experience and practice play in shaping language and language use are out of reach of these models. Finally, the cognitive models of speech production have little to say about articulation itself. The Levelt, Roelofs, and Meyer (1999) model posits that articulation is driven by syllables, which are conceived of as being, or being associated with, learned articulatory motor programs. The model by Dell (1984) likewise stops at the point that phonemes have been selected and assigned to their proper slots in phonological trees.

There are models that address articulation, but these models are found not in cognitive science, but in linguistics and phonetics. In linguistics, articulatory phonology (Browman and Goldstein 1986) posits articulatory scores. Vocal tract models, including the one implemented by VocalTractLab, create scores for control parameters by setting articulatory targets on a phoneme by phoneme basis. Smooth time series of control parameters for the different articulators are then calculated by connecting the sequences of target positions.

The PAULE model is a computational articulatory speech synthesis model that does not make any use of abstract units such as phonemes and morphemes.

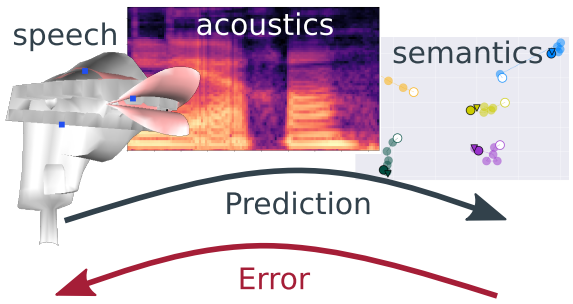

Figure 1: *The predictive principle implemented with PAULE assumes an internal predictive process that predicts the acoustic and semantic effects of an imagined upcoming articulatory motor program.*

[1] https://vocaltractlab.de/index.php?page=vocaltractlab-about

10.21437/issp.2024-2

2. Architecture

PAULE[2] implements a predictive planning approach (see Figure 1) for articulation at the word level. This predictive planning imagines the effect of the control-parameter (cp-)trajectories in terms of perceived acoustics and perceived word semantics. The cp-trajectories are smooth curves over time that define the position of the articulators as well as the parameters for the glottis model in the VTL. PAULE models all 30 control parameters of the VTL with a sampling rate of 401 Hz. For the acoustic representation a log-mel spectrogram is used with a frequency range of 10-12,000 Hz, 60 Mel bins, and a sampling rate of 200.5 Hz. For the semantic representation 300-dimensional fastText (Grave et al. 2018) vectors are used.

The acoustic and semantic representations are used as goal spaces within PAULE. Planning the cp-trajectories is achieved by minimizing the distance of of the predicted effects to given targets in the goal spaces. The minimization in the goal spaces is done along the local gradients of the forward predictions. Figure 2 depicts this process in a simplified form. Through the exploitation of the local gradients PAULE is capable of optimizing those parts of the cp-trajectory which are perceived as most relevant to the predictive forward model.

PAULE connects the different data structures with learned LSTM-based mappings (Hochreiter and Schmidhuber 1997) (Figure 3). These mappings are pre-trained and back-propagate prediction errors from the semantic and acoustic representations. The back-propagated prediction error together with stationarity and constant force constraints are used to plan and optimize the control of the VTL articulatory speech synthesis model.

The LSTM-based mappings are pre-trained on a German corpus containing of 26,271 word tokens distributed over 4,311 word types. The frequency of word types follows a typical language distribution with the most common word /also/ occurring 1,113 times and 2,261 word types only occur once. The duration of the word tokens range from 120 ms to 1,000 ms. A subset of the word types, containing both long and short, and infrequent and frequent words[3], was used to evaluate PAULE.

PAULE is implemented and pre-trained to find suitable cp-trajectories for the 4,311 word types of the German language. These can be synthesised by giving the target label semantic vector and a desired duration. Furthermore, PAULE is capable of re-synthesizing longer chunks of of speech signals even from different languages like English in a copy-synthesis setup.

3. Results

A full implementation of the PAULE model is available for German. When given a word embedding as input, the model produces the sound waves for that word, using the VTL. The quality of the sound waves produced is sufficiently high [4] to provide (1) a strong proof of concept that a shift from mainly reactive feedforward control to predictive goal directed control is feasible and (2) that articulation without intermediate abstract sublexical units such as phonemes and morphemes is possible. Although the PAULE model currently makes use of static word embeddings, nothing prevents the use of dynamic embeddings that are specific to utterance context. Depending on the details

[2] https://github.com/quantling/paule

[3] Beispiel, Freunde, Lehrer, Studium, aber, eigentlich, nämlich, natürlich, praktisch, schwierig, tatsächlich, trotzdem, and zurück.

[4] Examples: https://nc.mlcloud.uni-tuebingen.de/index.php/s/pZPgcCG9MSEHkJT

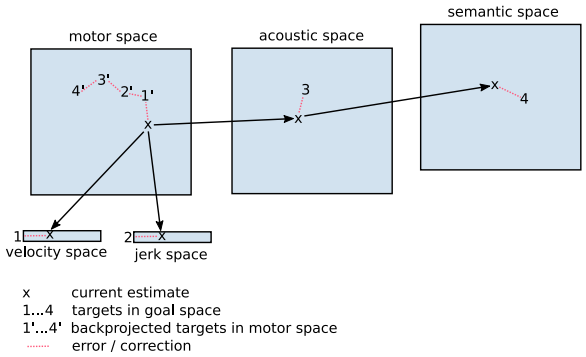

Figure 2: *The predictive principle implemented with PAULE compares the predicted effects in the acoustic and semantic goal space as well as some velocity and jerk constraints. The missmatch (or error) between the predictions and the desired target acoustic respectively semantic representation is used to improve the articulatory motor program along the gradients of the predictions. With this gradient-aware planning (or optimization) only forward models are needed. No explicit model of the error correction is used. Still the error can give locally relevant correctoins. All goal spaces are continuous and therefore no discrete or symbolic representations like phonemes or motor-gestures are used within PAULE.*

of a dynamic embedding, the details of the articulated sound waves will change. This illustrates a more general property of the PAULE approach, namely, a shift away from what would be a 'correct' articulation to sufficiently good realizations that balance comprehensibility and minimization of articulatory effort.

Even the question shifts away from "what is the correct articulation for a given word" to which articulatory patterns are sufficient to satisfy the acoustical and semantic target in mind while complying to some articulatory laziness constraints. The PAULE framework therefore proposes that there is not necessarily a single optimal articulatory control, but a multitude of good controls, which satisfy different goals to different degrees and which is inherently dependent on the perceptive experience of the speaker, her knowledge of the target language and her experience with articulating similar words.

4. Discussion and Conclusion

Doing articulatory speech synthesis without any gestures or phones might be seen as a bold claim. But, PAULE is a computational model that does produce control-parameter trajectories for the 30-dimensional articulatory speech synthesiser in VTL on the word-level. PAULE achieves this without the use of any symbolic units in its pipeline.

The current implementation of PAULE has several limitations. *First*, the initialization process builds on approximate cp-trajectories synthesized from a phone-driven gesture-based approach (Sering et al. 2019). This is not a matter of principle, but a matter of convenience. Ideally, the model would be informed by either articulatory measures obtained with electromagnetic articulography or ultrasound or trained from "zero-knowledge" in a goal-babbling approach. At present, however, such empirical data are not available for the task of modeling the articulation of a non-trivial number of words. As a consequence, part of the input to the PAULE model is likely to be too systematic and rule-governed, compared to data from actual

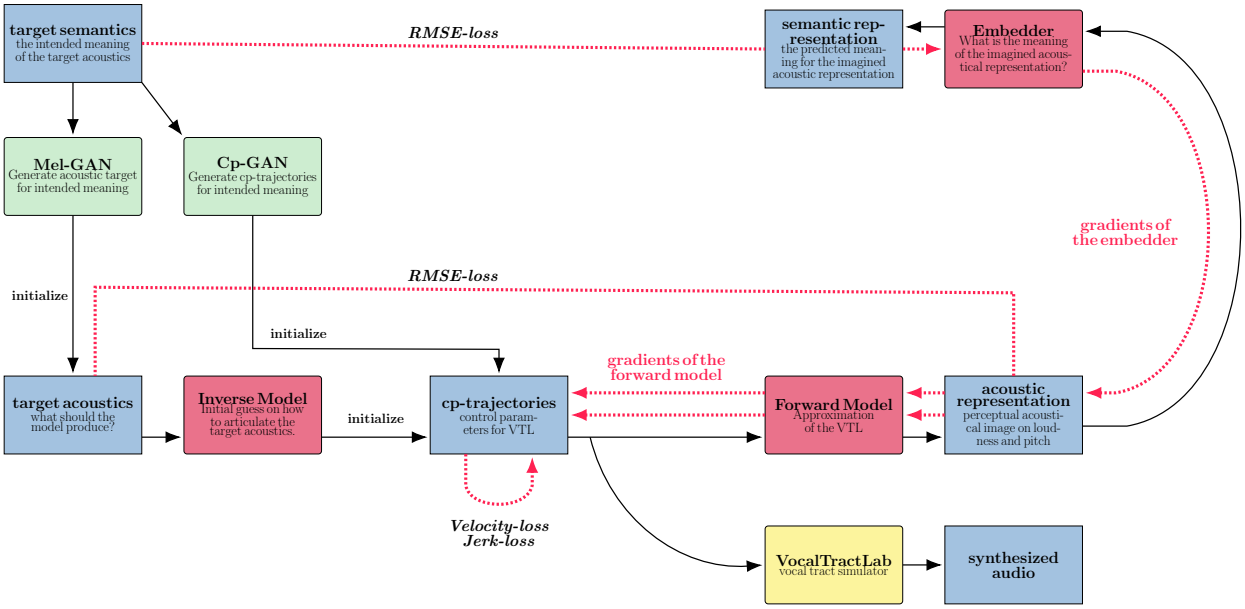

Figure 3: *PAULE plans the articulatory control-parameter(cp-)trajectories of the VocalTractLab geometrical 3-dimensional articulatory synthesizer. The planning uses an internal predictive forward loop and optimizes the upcoming cp-trajectories by minimizing an error in an acoustic and semantic goal space. All data representations uses a fine grained finite time representation as well as a continuous representation for the positions of the articulators as well as for the acoustics features and the semantic lexical embeddings. Therefore, no symbolic representations are needed to synthesize single word tokens with PAULE. The Mel-GAN, Cp-GAN, and the Inverse Model are only used once for initialisation purposes.*

human speech. In future work, we will consider whether it is possible to obtain initialization trajectories using goal babbling learning schemes — although we anticipate that these will be computationally highly demanding. This brings us to a *second issue* we have with our model, namely, that even in its current implementation it is computationally expensive. With a real-time factor of around 3,000, planning one second of speech needs around 50 minutes of computation time. A *third issue* is that the current implementation requires several test-outs of potential articulations using the outer loop. In other words, the model is mumbling to itself before it finalizes on the articulation that it converges to as optimal. This is not how competent language users speak. Although learners need multiple try-outs to master saying a given word, mature learners have automatized what they have learned. PAULE does not utilize its memory efficiently for past experience and routinization. *Nevertheless*, we think that the PAULE model is useful as a proof of concept that considerable progress can be made in learning to articulate words using as input empirical word embeddings and the corresponding audio files within a deep learning architecture.

5. Acknowledgements

Konstantin Sering is funded by the DfG grant 527671319 („Komplexe Wörter im Kontext“). Harald Baayen and Konstantin Sering are members of the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 390727645.

6. References

Birkholz, Peter (2013). “Modeling consonant-vowel coarticulation for articulatory speech synthesis”. In: *PloS one* 8.4, e60603.

Browman, Catherine P and Louis M Goldstein (1986). “Towards an articulatory phonology”. In: *Phonology* 3, pp. 219–252.

Dell, Gary S (1984). “Representation of serial order in speech: evidence from the repeated phoneme effect in speech errors.” In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 10.2, p. 222.

Fromkin, Victoria A (1984). *Speech errors as linguistic evidence*. Vol. 77. Walter de Gruyter.

Grave, Edouard, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov (2018). “Learning Word Vectors for 157 Languages”. In: *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Hochreiter, Sepp and Jürgen Schmidhuber (Dec. 1997). “Long Short-term Memory”. In: *Neural computation* 9, pp. 1735–80.

Levelt, Willem JM, Ardi Roelofs, and Antje S Meyer (1999). “A theory of lexical access in speech production”. In: *Behavioral and brain sciences* 22.1, pp. 1–38.

Schmidt-Barbo, Paul, Sebastian Otte, Martin V. Butz, R. Harald Baayen, and Konstantin Sering (2022). “Using semantic embeddings for initiating and planning articulatory speech synthesis”. In: *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2022*. Ed. by Oliver Niebuhr, Malin Svensson Lundmark, and Heather Weston. TUDpress, Dresden, pp. 32–42.

Sering, Konstantin (2023). *Predictive Articulatory speech synthesis Utilizing Lexical Embeddings (PAULE)*. Tübingen: Universität Tübingen. DOI: 10.15496/publikation-90142.

Sering, Konstantin, Niels Stehwien, Yingming Gao, Martin V Butz, and Harald Baayen (2019). “Resynthesizing the geco speech corpus with vocaltractlab”. In: *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2019*, pp. 95–102.