ANTICIPATORY COARTICULATION IN PREDICTIVE ARTICULATORY SPEECH MODELING

Konstantin Sering, Fabian Tomaschek, Motoki Saito

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

Abstract: The aim of the present study is to test whether effects of anticipatory coarticulation emerge from a segment based and a recurrent gradient based planning speech resynthesis in the articulatory speech synthesizer VocalTractLab. For this, natural articulations of /baba/, /babi/ and /babu/ are recorded using ultrasound. While anticipatory coarticulation is observable in the articulatory movements and the phonetic signal of the human recording, these patterns are not observable in the segment based resynthesis, and partially observed in the recurrent gradient based planning resynthesis. The recurrent gradient based planning shows anticipatory coarticulation in formant shifts, but not in tongue raising.

1 Introduction

Anticipatory coarticulation is a well known phenomenon that describes the modification of articulatory patterns to anticipate upcoming, not yet articulated phones [1, 2]. In this work, we investigate how well two speech synthesizers are capable to produce anticipatory coarticulation: a recurrent gradient-based predictive modeling framework for articulatory speech synthesis [3] and a segment based gestural score approach [4]. We compare the human and synthesized articulatory trajectories in terms of tongue raising and the audio in terms of formant shifts. The human speaker's tongue is recorded using ultrasound.

Articulatory speech synthesis uses a physical simulation of the pressure changes and resonances of the vocal and nasal cavities as well as the properties of glottis [4]. In contrast to statistical or deep neural network based speech synthesis [5], articulatory synthesis gives insight into how humans produce speech with their speech organ. One challenge in doing articulatory speech synthesis is how to find one set of control parameter (cp) trajectories that drive the speech synthesizer to produce intelligible, meaningful speech, while simultaneously closely matching a given acoustic target.

There are semi-automatic and fully automatic approaches to infer the cp-trajectories for a given target acoustics, which allows to resynthesis or copy-synthesis a target wave file [6, 7]. These approaches have different qualities in terms of acoustic quality and of plausibility of the cp-trajectory and therefore movement plausibility. These different approaches make different assumptions about how the cp-trajectories are structured. One popular assumption is that cp-trajectories result from a blend of different gestures that are tied to a phone repertoire [4, 8, 8].

Blends of different phone sequences lead to systematically different tongue movements, i.e. coarticulation, which in turn produce different acoustical properties. The two resynthesis methods compared in this work are fully automated and have no information about the actual movements of the tongue. Both methods have to infer the underlying movements. The segment based approach has only access to the phone sequence and the duration of each phone. The recurrent gradient based planning approach has only access to the acoustics in the form of a log-mel spectrum, but has no explicit knowledge of the phone sequence or their durations.

To have strong and easy to explain coarticulation effect a pseudo word phrase was uttered by a single speaker in six different speaking rate conditions. Focus is on the tongue movement as an articulatory coarticulation measure and the vowel formant shifts as an acoustical coarticulation measure. It is expected that both resynthesis pipelines produce some amount of coarticulation, although they will differ in amount and direction of the coarticulation.

2 Methods

2.1 Simulator / Articulatory speech synthesis

For the articulatory speech synthesis the VocalTractLab synthesizer is used (VTL, Version 2.3). The VTL models a three dimensional vocal tract shape, i. e. the tongue, the jaw and the lips, together with properties of the glottis and the sub-glottal pressure. From the geometrical shape the VTL derives pressure differences, i. e. acoustical waves, and from that the emitted audio as a 44100 Hz mono signal. For the acoustic simulation up to two branching cavities are coupled with the main cavity of the mouth: the sinus piriformis directly after the glottis and the nasal cavities.

In the configuration used in this study the VTL has 30 input control parameters (cps) that are defined every 110 audio samples (2.5 Milliseconds) and are linearly interpolated in between. The cp-trajectories define the moving parts of the simulated human speech organ, e. g. the tongue and jaw positions and the lip rounding as well as the properties of the glottis model. The size of the oral cavity, the jaw and the teeth are defined by the JD2.speaker that comes with the VTL software.

Limitations of VTL are that it uses a quasi 1D acoustical simulation that only allows for longitudinal waves in the direction along the tube. This leads to good synthesis quality up to 5000 Hz, but might give poor results for higher frequencies. Another limitation is that the VTL has relatively long computation times of roughly three seconds computation time per second synthesised speech. Furthermore, no muscle or motor control is modeled, which is why there is no direct measure on how easy it is to do the deformations that the VTL vocal tract model is doing with a human vocal tract. Therefore, the VTL is not constraint by any muscle groups.

2.2 Segment based resynthesis

For the segment based resynthesis the phone labels and phone durations of a forced alignment are converted into a segment file. Starting with version VTL 2.3 the API can fully automatically synthesize audio from a segment file. This is used here as the first resynthesis framework.

As the segment based synthesis in VTL blends gestures for the different phones, some coarticulation should emerge. But, long-ranging coarticulation along the sequence of phones should be limited. Therefore the segment based resynthesis should be good in modeling coarticulation between phones, but might have difficulties to account for coarticulation that anticipates the two or more steps down the line of the upcoming phone sequence.

2.3 Recurrent gradient based resynthesis

The recurrent gradient based resynthesis framework is depicted in Figure 1. The framework consists of three main data representations and three models that connect these data structures. The three data structures are the control parameter (cp-) trajectories, which are the inputs to the VTL simulator and a log-mel spectrum that represents the acoustics. The cp-trajectories and the acoustic representation are connected in three ways. The physical ground truth (for this

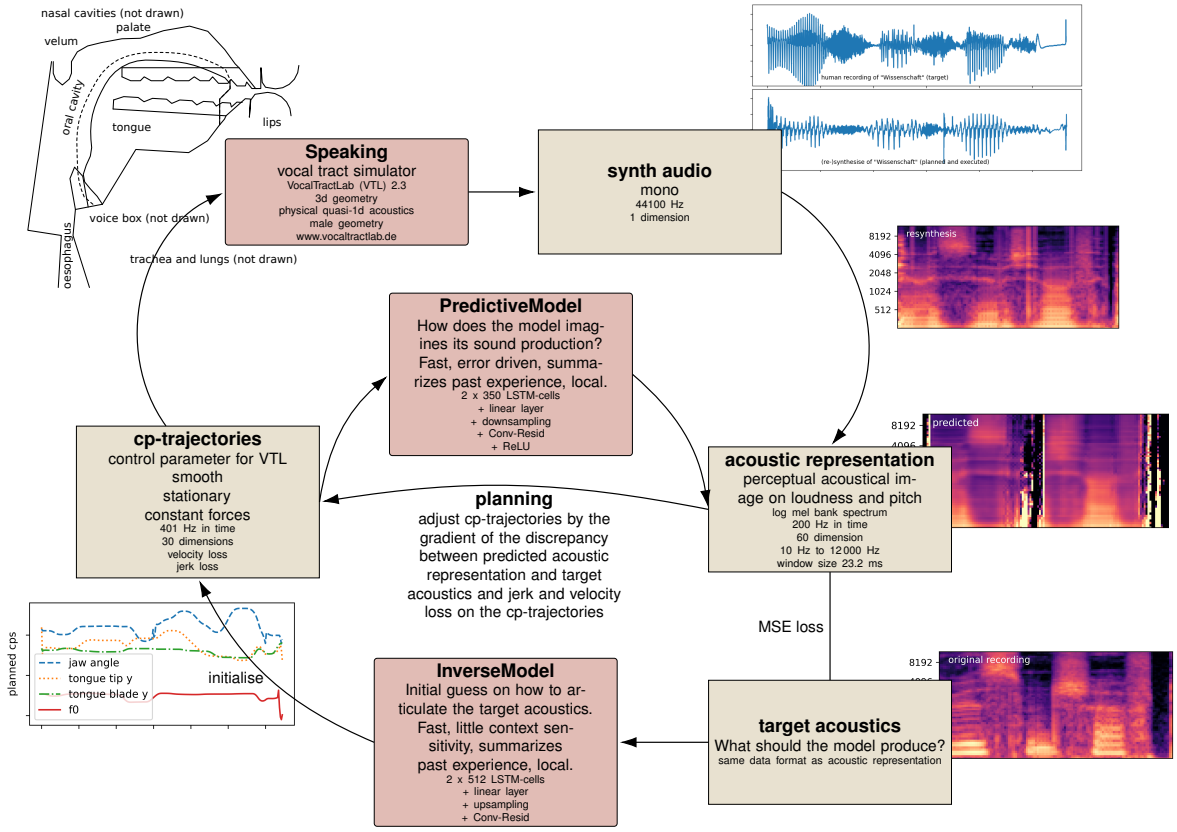

Figure 1 – Implementation of the recurrent gradient-based motor inference principle with LSTM based networks. The predictive model imagines the acoustic representation and allows for adjustment prior to execution. The inverse model is only used for initialisation. [3]

framework) is the forward synthesis by the VTL simulator. The simulator processes the cp-trajectory and synthesises corresponding mono audio from which the acoustic representation is calculated in deterministically. This forward path is emulated by the forward model, while short cutting the actual audio and accumulating learning history or experience in its model. As the forward model is later used for planning but cannot be easily inverted, an explicit inverse model is learned to initialize the planning. The forward model is the most important part of the framework.

To resynthesize a given target acoustics, first the cp-trajectories are initialized by consuming the target acoustics in the inverse model. Next the forward model is used to imagine the initial cp-trajectories and create a predicted acoustical representation. Now some initial cp-trajectory, the target acoustics and the imagined predicted acoustics are present. For the planning, two losses are minimized jointly. The first loss is the jerk and velocity of the cp-trajectory. The jerk and velocity loss favors constant position and constant velocity trajectories. Constant velocity trajectories correspond to trajectories that underlie a constant force. The second loss is the MSE loss between the predicted acoustics and the target acoustics. The two losses are connected by the predictive model. The gradients of the predictive model are used to combine the losses.

After minimizing the joint loss in an iterative procedure and by that planning a new cp-trajectory, the planned cp-trajectory is executed through the VTL synthesizer. This results in a new synthesized audio and therefore a new training ground truth data sample for the forward model is created. To keep the forward model synchronized with the VTL synthesizer the learning of the forward model is continued with this new training data together with ten old training samples. The old training samples are mixed in to prevent total forgetting.

Depending on the desired final synthesis quality the planning and the continue learning of the forward model can be interleaved several times. In this work we interleave the planning and the continue learning 40 times, which gives a good compromise between model accuracy and computation time. The final cp-trajectories are then used to synthesize some final audio on which the formant analysis is performed. To analyze the tongue shape the final cp-trajectories are exported into a graphics of the mid-saggital slice of the VTL vocal tract.

2.4 Experimental Setup

To evaluate the amount of coarticulation patterns in the acoustical and the articulatory domain, one speaker (the first author) is recorded saying the three pseudo words /baba/, /babi/ and /babu/ repeatedly within a three second recording window. The speaking rate is manipulated by instructing the speaker to produce the pseudo word within the three second recording window once to six times. This results in six different speaking rate conditions.

Speaking rate conditions are blocked and pseudo words are interleaved with three repetitions. The speaking rate was increased within one session from the very slow speaking rate to the fastest speaking rate condition. In the very slow speaking rate condition the pseudo word should be uttered once whereas in the fastest speaking rate condition the pseudo word had to be uttered six times within the same 3 second recording window. Three sessions are recorded.

While speaking, an ultrasound image of the tongue is recorded in the mid-saggital plane. The ultra sound image is sampled with 81.6 frames per second and each image consists of 64 vectors, i. e. directions, with 842 pixels each, i. e. depth.. The ultrasound transducer is pressed to the soft part of the jaw between the mandible bone and hold in place by a head mount. The orientation stays relatively fixed to the lower jaw and therefore to the lower teeth, but not to the hard palate. This means that the ultrasound image of the tongue movement has to be interpreted in reference to the lower teeth and not to the hard palate.

All recordings are done in an sound attenuated booth. For the further analysis only the second fastest speaking rate condition with five repetitions is used as this should show solid coarticulation patterns in format shifts and tongue raising.

2.5 Data Preparation

Each audio recording is phonetically aligned with the Montreal Audio Aligner [9] and alignments are stored in a TextGrid. The TextGrids are used to accomplished two goals: First, the segment based resynthesis is done using the phone durations of all the phones in each word. Second, the recording timestamps of the midpoint and the offset of the /a/ in the first syllable of each pseudo word are extracted. The midpoint and offset are then used to extract formant values and tongue positions as measures of coarticulation. The gradient based resynthesis uses only the recorded audio as input. It uses the audio of each spliced out word and therefore is not resynthesizing the five repetitions of the pseudo word at once.

To remove adaptation effects to a new speaking rate condition the first block in each new speaking rate condition is removed. This results in 30 recordings of each pseudo word or 90 recordings in total in the second fastest speaking rate condition.

2.6 Formant Shifts

To evaluate coarticulation in the acoustic domain, shifts in the the first and second formants of the /a/ in the first syllable are inspected. From each audio a KlattGrid file is generated and the first and second formant frequencies are extracted in the 20 Milliseconds before the offset of the first /a/. The following settings were used: pitch floor is set to 50 and pitch ceiling is set to 350

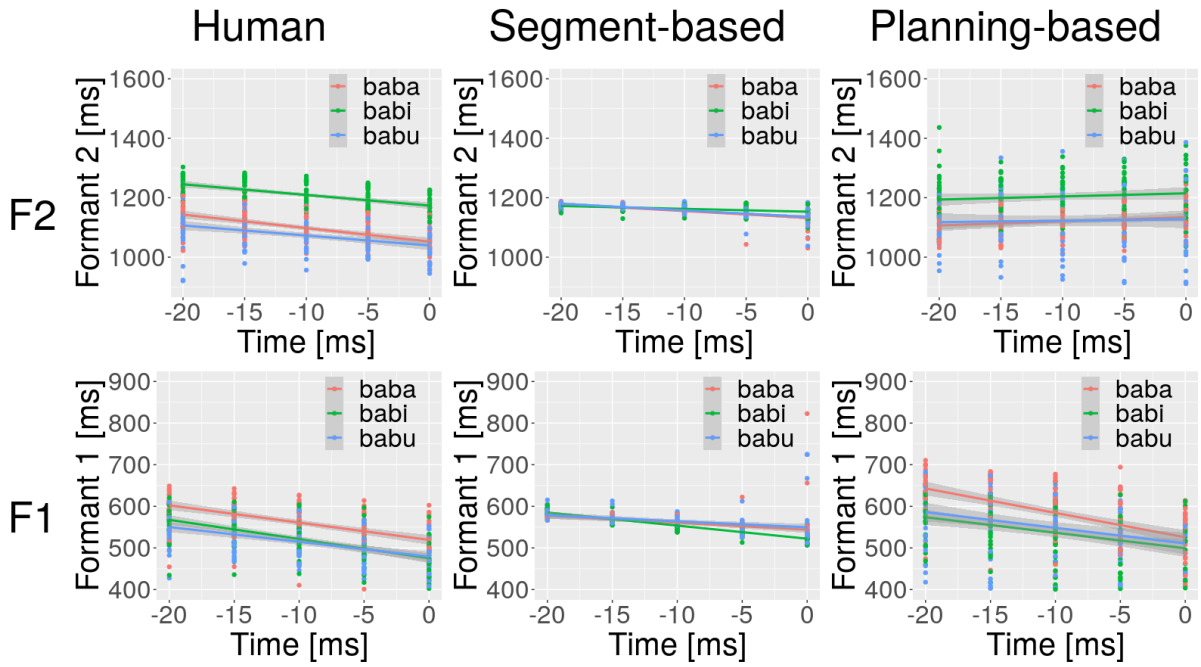

Figure 2 – The six panels show the formant frequency in Hz for the first (bottom panels) and second (top panels) in the 20 Milliseconds before the offset of the first /a/. The human speaker shows anticipatory coarticulation in both formants (left panels). The segment based resynthesis shows no anticipatory coarticulation (middle panels), in contrast the recurrent gradient based planning approach shows anticipatory coarticulation in the formant space but fails to mimic the full richness of the human formant transitions (right panels).

otherwise Praat's default values are used [10]. It is expected that the first formant is lower in /babi/ and /babu/ compared to /baba/ and that the second formant is higher in /babi/ and lower in /babu/ compared to the formant in /baba/.

2.7 Tongue Raising

To evaluate the articulatory coarticulation, the midpoint and the offset of the first */a/* in the aligned TextGrids of the human recordings are used as reference points. For these points in time the highest tongue position is extracted from the ultrasound recordings. The highest point is measured in the coordinate system of the ultrasound device and therefore relative to the jaw or lower teeth.

For the resynthesis the highest point on the tongue contour of the svg pictures of the midsaggital plane in both resynthesize conditions is used. The raising or lowering is defined by the pixel difference of the offset minus the midpoint. To account for the fact that the lower teeth are the reference all svg pictures are rotated and translated to align on the lower teeth. The reference teeth are in the position of the VTL gesture that corresponds to /a/, i.e. in an open mouth position. An anticipatory tongue raising is expected in /babi/ and /babu/ in comparison to the the /baba/ condition.

3 Results

3.1 Formant Shifts

The human recordings show a lowering of the first formant in the 20 Milliseconds before the offset of the first /a/ in /babi/ and /babu/. The second formant shows raising in /babi/ and small lowering in /babu/ compared to /baba/ (Fig. 2 left panels). This is in line with the well-

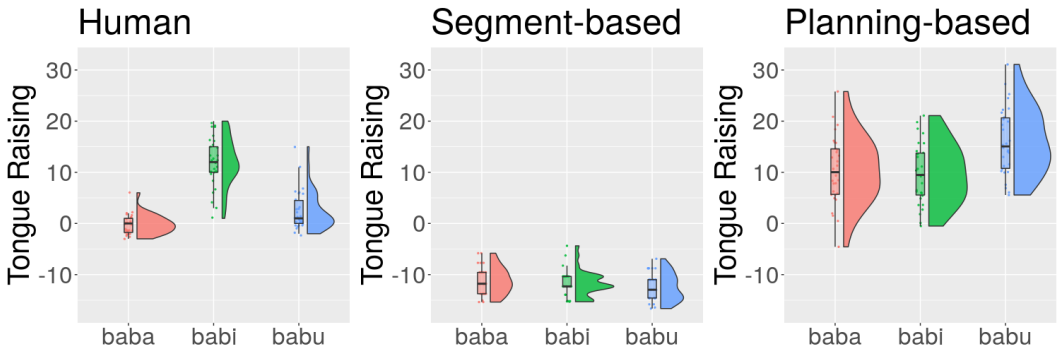

Figure 3 – Tongue raising from the midpoint to the offset of the first /a/ in each pseudo word. For the human recording (left panel) no systematic raising is visible for /baba/, but the tongue is systematically raised strongest in /babi/ and moderate in /babu/. For the resynthesis neither the segment based nor the recurrent gradient based planning approach capture the lowering and raising pattern of the human speaker.

known relationships between the first and second formants and the tongue height [e.g. 11]. Results for the segment based resynthesis show no systematic raising or lowering of the formant frequencies before the offset. Nevertheless it shows a similar decrease in formant frequency over time, which might be seen as some form of coarticulation, but not as anticipatory vowel-to-vowel coarticulation (Fig. 2 middle panels). The recurrent gradient based planning approach shows similar formant frequencies to the human recordings. The overall decreasing trend of the first formant is correctly taken up (Fig. 2 bottom right panels). In addition, the first formant in /baba/ is higher than the other two, which is also in line with the human recordings. The second formant is in line with the human recordings with respect to the higher frequency for /babi/, compared to the other two. However, the overall decreasing trend observed in the human recordings is not picked up correctly in this approach (Fig. 2 right panels).

3.2 Tongue Raising

The amount of the tongue raising was measured by subtracting the tongue height at the midpoint from that of the offset of the first /a/ in each pseudo word (y-axis in each panel in Figure 3). Therefore, bigger y-values in Figure 3 correspond to greater distance of the tongue raising.

In the human recordings, the distance of the tongue raising is biggest in /babi/, mild in /babu/, and almost zero (no raising) in /baba/. This means that /babi/ and /babu/ show anticipatory coarticulation, while /baba/ does not. This observation is in line with the formant shifts described in the last subsection. These different degrees of tongue raisings were picked up neither by the segment based resynthesis, where the tongue is lowered (rather than raised), nor by the recurrent gradient based planning framework where the tongue is raised for all pseudo words. Fig. 4 shows one example for the differences in anticipatory coarticulation in the first /a/ of /baba/ and /babi/ for the human recording and the resynthesis. Remember that all the raising and lowering is in reference to the lower teeth.

4 Conclusion and Future Work

The artificial utterances /baba/, /babi/, /babu/ repeatedly spoken with a high speaking rate show robust coarticulation effect in formant shifts and anticipatory tongue raising. For fully automatic resynthesis frameworks it is still a challenge to model the full range of human coarticulation. Still, articulatory synthesis allows to pinpoint differences in the underlying vocal tract geometry and allows to reason about missing coarticulation patterns in a straight forward and transparent way. The segment based resynthesis shows no anticipatory coarticulation and shows lower

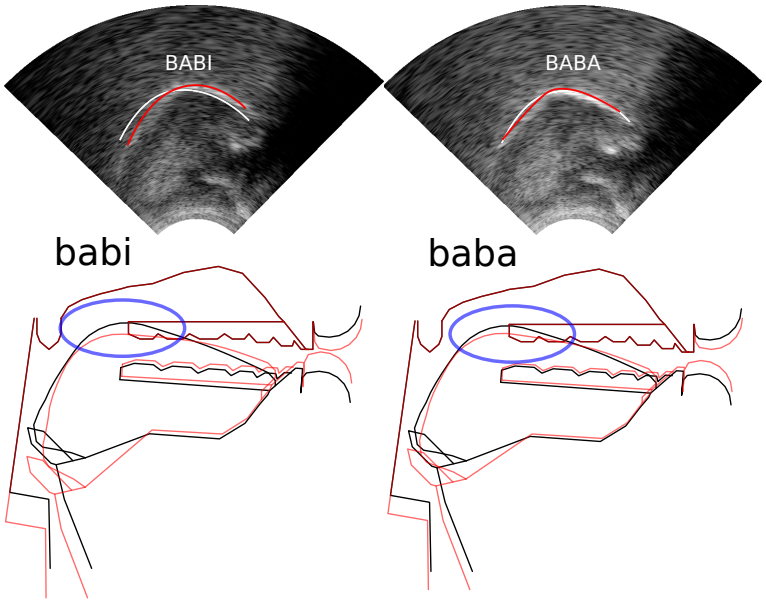

Figure 4 – Top panel: Tongue position of the human tongue in the ultrasound recording in the middle (white) and at the offset (red) of the first /a/ in /babi/ and /baba/. An anticipatory raising of the tongue body is visible in /babi/ compared to /baba/.
Bottom panel: Tongue position of the simulated vtl tongue in the middle (black) and at offset (red) of the first /a/ in the resynthesized /babi/ and /baba/ with the recurrent gradient based synthesis approach. A small anticipatory raising of the tongue body is visible (blue ellipses). The statistical analysis indicates that the articulatory tongue raising in the resynthesis models might not be present.

variability in its formant compared to the human recording. The recurrent gradient based resynthesis shows anticipatory coarticulation in the formants but does not seem to achieve this by means of anticipatory tongue raising. The variability in the recurrent gradient based framework is too high compared to the human recordings. This evaluation gives us valuable insight into optimizing the planning process of the recurrent gradient based planning framework to produce more human like trajectories and thereby softly restrict the flexibility of the VTL simulator.

An alternative to the method presented here is to compare the virtual tongue movements to electromagnetic-articulography (EMA) data by tracking selected vertices of the virtual tongue that correspond to the glued EMA-sensors [12]. In contrast to ultrasound the reference frame is usually the skull and therefore the upper teeth rather than the jaw and the lower teeth.

Next steps in this line of research involve testing to what degree effects of prosody [13] as well as lexical phenomena [2, 14, 15, 16] in natural articulation are mirrored in the resynthesised audio.

Acknowledgments This research was supported by an ERC Advanced Grant (no. 742545) and a collaborative grant from the Deutsche Forschungsgemeinschaft (Spoken Morphology, BA 3080/3-2), awarded to R. Harald Baayen.

References

[1] Öhman, S.: *Coarticulation in VCV Utterances: Spectrographic Measurements. Journal of the Acoustical Society of America*, 39(151), pp. 151–168, 1966.

[2] Tomaschek, F., B. V. Tucker, M. Fasiolo, and R. H. Baayen: *Practice makes perfect: the consequences of lexical proficiency for articulation. Linguistics Vanguard*, 4(s2), 2018. doi:10.1515/lingvan-2017-0018.

[3] 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.

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

[5] WANG, Y., R. SKERRY-RYAN, D. STANTON, Y. WU, R. J. WEISS, N. JAITLY, Z. YANG, Y. XIAO, Z. CHEN, S. BENGIO, Q. LE, Y. AGIOMYRGIANNAKIS, R. CLARK, and R. A. SAUROUS: *Tacotron: Towards end-to-end speech synthesis*. In *Proc. Interspeech 2017*, pp. 4006–4010. 2017. doi:10.21437/Interspeech.2017-1452.

[6] GAO, Y., S. STONE, and P. BIRKHOLZ: *Articulatory Copy Synthesis Based on a Genetic Algorithm*. In *Proc. Interspeech 2019*, pp. 3770–3774. 2019. doi:10.21437/Interspeech.2019-1334.

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

[8] GUENTHER, F. H.: *Neural Control of Speech*. MIT Press, 2016.

[9] MCAULIFFE, M., M. SOCOLOF, S. MIHUC, M. WAGNER, and M. SONDEREGGER: *Montreal Forced Aligner: trainable text-speech alignment using Kaldi*. *Proceedings of the 18th Conference of the International Speech Communication Association*, 2017.

[10] BOERSMA, P. and D. WEENINK: *Praat: doing phonetics by computer (version 6.0.48)*. 2019. URL `http://www.praat.org`.

[11] DELATTRE, P.: *The Physiological Interpretation of Sound Spectrograms*. *PMLA*, 66(5), pp. 864–875, 1951.

[12] SERING, K. and F. TOMASCHEK: *Comparing kec recordings with resynthesized ema data*. In R. BÖCK, I. SIEGERT, and A. WENDEMUTH (eds.), *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2020*, pp. 77–84. TUDpress, Dresden, 2020.

[13] FOUGERON, C. and P. A. KEATING: *Articulatory strengthening at edges of prosodic domains*. *The Journal of the Acoustical Society of America*, 101(6), pp. 3728–3740, 1997. doi:10.1121/1.418332.

[14] TOMASCHEK, F., D. ARNOLD, K. SERING, J. VAN RIJ, B. V. TUCKER, and M. RAMSCAR: *Articulatory variability is reduced by repetition and predictability*. *Language and Speech*, pp. 1–27, 2020.

[15] SAITO, M., F. TOMASCHEK, and H. BAAYEN: *Semantic measures determining coarticulatory movements of the tongue tip*. In *12th International Seminar on Speech Production*. 2020.

[16] SAITO, M., F. TOMASCHEK, and H. BAAYEN: *An ultrasound study of frequency and coarticulation*. In *12th International Seminar on Speech Production*. 2020.