COMPARING KEC RECORDINGS WITH RESYNTHESIZED EMA DATA

Konstantin Sering, Fabian Tomaschek

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

Abstract: Comparing simulated articulatory movements to their real counterparts is critical in establishing some trust in the findings produced by a simulator and the simulator's generalizability. Point-wise and global measures are compared in this study on a dataset of human EMA recordings and resynthesised simulated EMA data. Most measures show comparable sensitivity to the hypothesized structure of dissimilarity along different groups of EMA trajectories but fail to fulfill metrical properties. A measure based on the Jensen-Shannon-distance (JSD) seem to show good sensitivity and fulfills the triangle inequality and is therefore suggested to be used in evaluating resynthesis models.

Introduction

The aim of the present work is to find a reliable measure to compare real human articulatory movements, recorded with electromagnetic articulography, to articulatory movements simulated in a vocal tract simulator. This task, however, poses several problems.

Not only do articulatory trajectories of identical words and phones differ between multiple articulations of one and the same speaker due to noise in the kinematic system [1, 2, 3, 4] and short-term practice effects [5, 6, 7, 8, 9]. Also, inter-speaker differences in articulatory strategies due to different physiology of the oral cavity result in different trajectories for the same phones [10, 11, 12, 13, 14].

In spite of these challenges, comparing real articulatory trajectories to their simulated counterparts is critical in establishing trust in the simulator's results and its generalizability We do not want that the trajectories are fitted directly to their human counterparts but rather, they should be a result of simulator-constraints or optimized input parameters, that might be indirectly influenced by how humans articulate, but are optimized primarily on their acoustic results.

In the next section, we will first present the data which we used for our study, followed by a detailed discussion of the challenges coming with establishing a dissimilarity measure on the comparison of real and simulated articulatory trajectories.

Theory

The goal is to establish a (dis-)similarity measure for trajectories of articulography data with a duration ranging between 200 ms to 1000 ms (40 to 200 samples at 200 Hz).

The resulting measure should preferably fulfill the axioms of a metric. Therefore each two trajectories $a, b, c \in A$ can be compared with the measure $d : A \times A \rightarrow [0, \infty)$, which gives a dissimilarity score between 0 and ∞, fulfilling:

1. non-negative $d(a,b) \geq 0$,
2. identity of indiscernibles $d(x,y) = 0 \Leftrightarrow x = y$,

3. symmetry $d(x,y) = d(y,x)$,

4. triangle inequality / subadditivity $d(x,y) \leq d(x,z) + d(z,y)$.

All proposed measures are constructed in a way that they fulfill axioms 1 and 3, but axiom 4 is tested empirically. Axiom 2 is of minor priority for the data at hand as the trajectories are not directly optimized.

Data

Real human articulatory movements

We used articulatory movements provided by the Karl Eberhards Corpus of Southern German (KEC) [15]. The KEC contains recordings of spontaneous dialogues between two speakers, seated in separated recording booths and hearing each other via headphones. Speakers were friends instructed to talk about a common topic. The KEC contains articulatory recordings for 20 speakers, 30 minutes for each, performed with the NDI Wave Articulograph (sampling frequency = 400 Hz). The sensor movements were recorded at the tongue tip (TT), tongue body (TB, 2 cm further back) and tongue mid (in between of TT and TB). In post-recording procedures, articulatory movements were corrected for head movements and further rotated to the center of a biteplate. Simultaneously, the audio signal was recorded and synchronized with the articulography signal. The KEC contains manual annotations at the word level and automatic annotations at the segment level.

To validate and compare the different dissimilarity measures we extracted all instances of *ja* ($N = 1111$) and *halt* ($N = 217$). For comparison with simulated trajectories, EMA trajectories were downsampled to 200 Hz. To get a common reference point, positions within each speaker where shifted by their 95% quantile so that the 95%quantile is the shared origin of all recording. To keep the different orientations in space comparable and preserve the unit of millimeters we deliberately did not apply any scaling across speakers.

Simulator

To obtain simulated articulatory trajectories, we used simulated markers on the tongue of the VocalTractLab (VTL) geometrical model speech synthesis model [16]. VTL is a 3-dimensional geometrical vocal tract simulator linked with a quasi-1-dimensional acoustic synthesis model.

VTL allows to synthesize speech from *control parameters (cp)-trajectories* which define, for each subsequent step in discretized time, the geometrical shape of the vocal tract, properties of the glottis model, and lung pressure. For each 10 millisecond time step, 33 control parameters need to be adjusted. Until now, cp-trajectories have been derived mostly by means of a dominance model that takes gestural scores as input.

The concatenative synthesis approach used here relies heavily on handcrafted gestural scores (gestural targets) that come with the VTL simulator, defining control parameter targets for the different input dimensions of VTL. These have been derived from fMRI scans and other sources of geometrical knowledge about the sounds produced in the German language, together with some optimization to obtain more realistic acoustics [16].

To resynthesize all instances of *ja* and *halt*, we used the temporal information, i. e. the segment boundaries, provided by the phone annotations which were fed to the dominance model, coordinating the temporal structure of the cp-trajectories. The details of the approach are described [17] and the necessary code can be found under [18].

Figure 1 shows the tongue body sensor positions as a function of time (x-axis) for one random EMA recording of the tokens *ja* (yes) and *halt* (just) in black dots and the resynthesized

VTL positions in red triangles. All three space coordinates are plotted side by side. The different length is due to a padding at the end in the resynthesis.

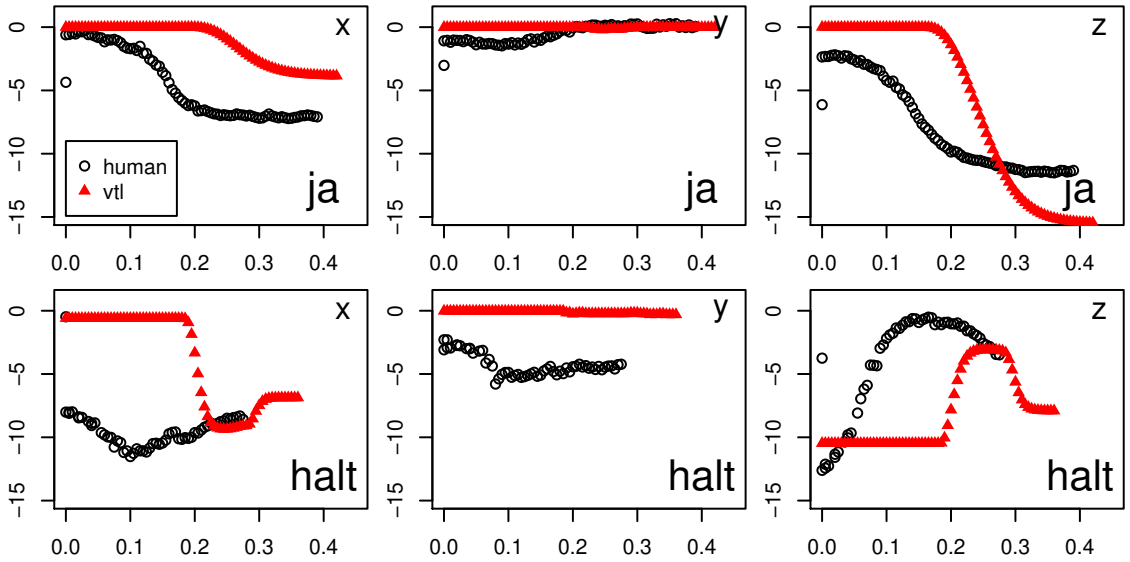

Figure 1 – Trajectories of the tongue body sensor in *back-front*-direction (x, left panels), in *left-right*-direction (y, middle panels) and in *low-high*-direction (z, right panels) for the word *ja* (top panels) and the word *halt* (bottom panels). The black circles show measured EMA data from the first subject and is compared in each plot to the VTL simulated trajectories of a matched sensor (in red triangles). Time is given in seconds, distances in Millimeters. No direct fitting is applied.

Samples

To validate the dissimilarity measures (description below), we created 10,000 random samples. Each sample consists of one trajectory of a random human subject a and a random word v. Depending on this first trajectory $av1$, the following additional trajectories have been sampled for comparison:

- **av2** – one random different trajectory of the same speaker and the same word
- **aw** – one random different trajectory of the same speaker but different word
- **bv** – one random trajectory of a different speaker and same word
- **bw** – one random trajectory of a different speaker but different word
- **vtlav1** – the vtl resynthesized trajectory matching $av1$
- **vtlv** – one random trajectory of the vtl resynthesis of the same word
- **vtlw** – one random trajectory of the vtl resynthesis of a different word

We tested which measures fulfill our hypothesis about the data and which ones seem to be not sensitive enough to capture the systematic structure in the dissimilarities. We hypothesized that the dissimilarities within the same speaker and the same word $d(\mathbf{av1}, \mathbf{av2})$ should be smallest together with the dissimilarities of the same word within the resynthesis $d(\mathbf{vtlav1}, \mathbf{vtlv})$ as theses share the same geometry and the same overall execution pattern. This should be followed by the resynthesis of the same word $d(\mathbf{av1}, \mathbf{vtlav1})$ as these follow the same execution

measure	category	violations sub additivity 1D	3D
mean abs diff	point-wise	872	1042
sqrt mean sq diff	point-wise	569	1042
1 - cor	point-wise	2391	1075
abs t-value paired t-test	point-wise	3026	2518
abs mean diff	global	221	234
abs std diff	global	141	64
JSD	global	3	3
abs t-value unpaired t-test	global	3516	2714

Table 1 – The different measures tested here can be categorised into *point-wise* measures and *global* measures. Unfortunately, all measures except the *JSD* measure seem to violate the sub additivity (triangle inequality) substantially. Violations are counts from the total 10,000 samples.

timing in different geometries or the dissimilarities of a different speaker speaking the same word $d(\mathbf{av1}, \mathbf{bv})$. We should observe highest dissimilarities for different speaker and different word $d(\mathbf{av1}, \mathbf{bw})$, and simulator and a different word $d(\mathbf{av1}, \mathbf{vtlw})$, respectively.

$$\begin{aligned} d(\mathbf{av1}, \mathbf{av2}), d(\mathbf{vtlav1}, \mathbf{vtlv}) &< d(\mathbf{av1}, \mathbf{vtlav1}), d(\mathbf{av1}, \mathbf{bv}), d(\mathbf{vtlav1}, \mathbf{vtlw}) \\ &< d(\mathbf{av1}, \mathbf{bw}), d(\mathbf{av1}, \mathbf{vtlw}) \end{aligned}$$

The sub additivity (triangle inequality) is tested with the following inequality: $d(\mathbf{av1}, \mathbf{vtlav1}) \leq d(\mathbf{av1}, \mathbf{bv}) + d(\mathbf{bv}, \mathbf{vtlav1})$.

Measures

Table 1 shows two groups of measures. In the first group, *point-wise* measures, first some dissimilarity is calculated at each time point and then this point-wise dissimilarities are aggregated over time. In the second group, *global* measures, first each sequence is collapsed over time and then the difference measure is applied onto this time independent measure. All measures have been applied only on the low-high axis (1D-condition) and on all three space axis jointly (3D-condition) of the tongue body sensor.

Into the first, point-wise group would fall a paired t-test as well as calculating a regression coefficient. These measures share that they are sensitive to miss-alignments and that they need to have the same number of (time aligned) data points available. Therefore, for sequences of different length the longer sequence was clipped. This approach was followed here.

Into the second, global, group would fall a t-test for independent samples, differences in energy or frequency component. As these measures operate on sequence descriptions that do not contain time explicitly anymore, it is strongly advised to use some velocity or frequency (phase and energy) measures to disambiguate time reversed or time shifted patterns.

In the following paragraphs, we describe the measures used in the current study.

The **mean abs diff** is the mean of the point-wise euclidean differences between the trajectories at each point in time. In 1D this is equivalent to the absolute differences. In 3D the eucleadian distance at each point is calculated and the mean over all time points is computed.

$$\text{mean abs diff} = \frac{1}{T_{min}} \sum_t \sqrt{\sum_i (a_{it} - b_{it})^2}$$

The **sqrt mean sq diff** is the mean of the point-wise euclidean differences between the

trajectories at each point in time. In 1D this is equivalent to the absolute differences. In 3D, the eucleadian distance at each point is calculated and the mean over all time points is computed.

$$\text{sqrt mean sq diff} = \sqrt{\frac{1}{T_{min}} \sum_t \sum_i (a_{it} - b_{it})^2}$$

For the **one minus corr** measure the Pearson correlation coefficient between the two trajectories is calculated. To obtain data for the 3D computation, a vector containing position data from all three dimensions is created by concatenation. As r from -1 to $+1$ (both including) and the identity is expressed by a correlation coefficient of $+1$, the measure is constructed by

$$\text{one minus corr} = 1 - \text{corr}(a, b).$$

The **tvalue paired** measure is constructed out of a paired t-test applied to the trajectories. The absolute value of the resulting t-value is used as a measure. To obtain data for the 3D computation, all three dimensions are concatenated for both trajectories into one long series per trajectory.

In the **abs mean diff** measure first the mean of each trajectory is calculated and then the euclidean distance between both mean vectors is calculated.

$$\text{mean abs diff} = \sqrt{\sum_i \left(\frac{1}{T_a} \sum_t a_t - \frac{1}{T_b} \sum_t b_t \right)_i^2}$$

The **tvalue unpaired** measure is constructed out of a unpaired, independent t-test applied to the trajectories. The absolute value of the resulting t-value is used as a measure. To obtain data for the 3D computation, a vector containing position data from all three dimensions is created by concatenation.

For the **JSD** measure a histogram along the space dimension with 20 evenly spaced bins is created over both trajectories for each dimension. This histogram contains the counts of how many points per trajectory fall into this region in space. These counts per bins then are turned into empirical probability distributions by calculating the relative frequencies. The Jensen-Shannon distance (JSD) is calculated between the two probability distributions per space axis. In the 3D computation the JSD is calculated separately for each dimension and then added up into a single number.

$$\text{JSD} = \sqrt{\sum_i \left(\frac{1}{2} D(p_{ai} || p_{mi}) + \frac{1}{2} D(p_{bi} || p_{mi}) \right)}$$

Here $p_{mi} = \frac{1}{2}(p_{ai} + p_{bi})$ summed probability distributions of the empirical probability distributions that are constructed out of the two trajectories a and b. $D(X||Y)$ denotes the Kullback-Leibler divergence between X and Y.

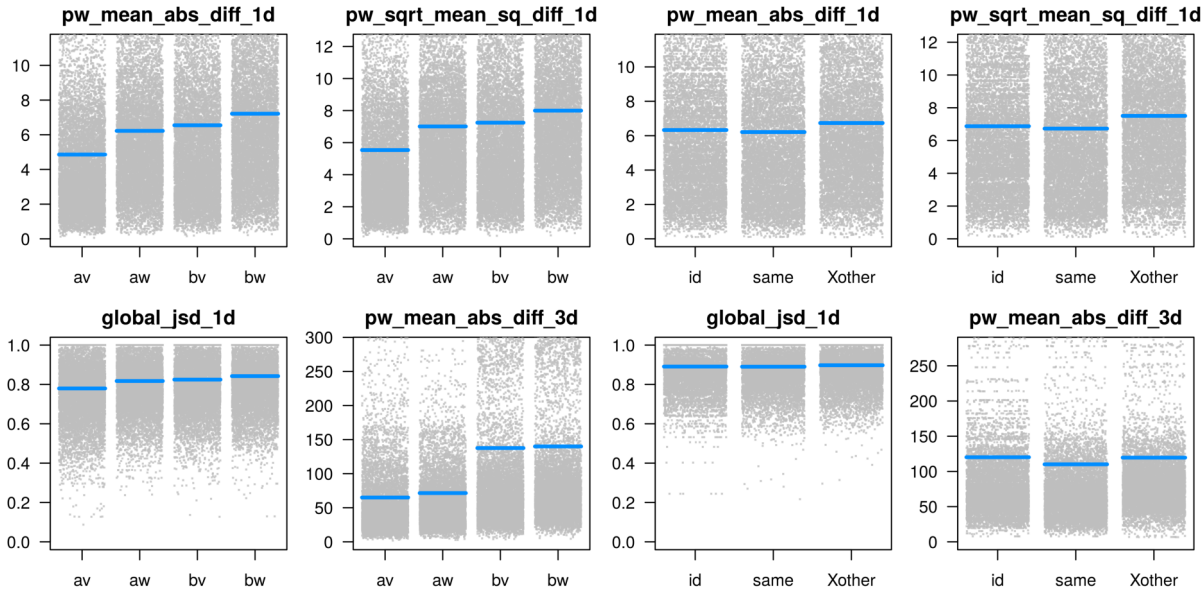

Figure 2 – The two by two plots on the left show mean dissimilarities between the reference and the same speaker and same word (av), same speaker and different word (aw), different speaker but same word (bv) and different speaker and different word(bw). The average dissimilarity should increase from left to right within each plot. The right two by two plot panel show the mean dissimilarities between the reference and the resynthesis in vtl (id), another resynthesis of the same word (same) and the resynthesis of the other word (Xother). The average dissimilarity should increase form left to right within each plot again.

Results

Figure 2 shows four selected measures that show the lowest violations in the sub additivity and confirm mostly our hypothesis along the expected dissimilarities. The 1d point-wise measures have more violations of triangular inequality, but show a good compromise in the mean dissimilarities between word class and speaker. The 3d point-wise measure seems to be sensitive to the geometrical differences and lacks the ability to distinguish the different categories of vtl resynthesis. Overall the JSD measure seems to be most promising in detecting differences between speakers and between word trajectories. The vtl results for the JSD measure could be more sensitive, but does not seem to go in the wrong direction. The JSD measure in 3d (not shown) show very similar results to the 1d case. The JSD achieves this while fulfilling all metric axioms.

Discussion

In the current paper we have evaluated several dissimilarity measures for their suitability to compare real and simulated articulatory trajectories. We identified four measures which seemed to be suitable to gauge dissimilarities between real and simulated articulatory trajectories. The measures based on point-wise distances captured the hypothesized ordering relations in our data, but lacked to fulfil the triangle inequality.

The problematic measures yield unclear results partly because they do local space normalization. In principle, it is possible to go into normalized space, e. g. by speaker, or into normalized time, e. g. by utterance, to calculate all the measures evaluated here. However, space and time normalization which includes local scaling removes meaningful units and therefore makes it difficult to compare different axes or instances. In the present study, local normalization is

applied by the measures derived from the t-test and the correlation. Both yielded unclear results and are therefore rejected.

Surprisingly, however, the most promising measure turned out to be JSD. Although it is not a well established measure of dissimilarity on trajectories, its results not only follow the hypothesized differences between the tested instances, but also it fulfills the definitions of a metric. It should be noted that the current implementation involves arbitrary binning of the data.

The JSD measure now can be extended to multiple sensors simply by adding the distances of each sensor on top of each other. Furthermore the JSD measure should be applied to the velocity curves of the trajectories and added to the distances as well. By adding up all these distances one ensures that the real trajectories are similar in the time domain and not only on the time marginal.

Another important but debatable decision in the current study was focus on rather short signals. Hypothetically, it could have been possible to use very long utterances, even of several minutes. For longer sequences, which consist of repeating similar patterns, a different class of methods might be better suited in comparison to those presented here.

For very long patterns with a lot of re-occurrences that come from a small set of symbols, and that are easily identified, methods that first transform the signal into a string of symbols and then apply a distance measure (or similarity score) to two strings like the Needleman-Wunsch algorithm should work fine. If the re-occurrences should not be identified as being identical to each other – as it is the case in the symbolic approach –, but segment boundaries or other landmarks can be identified in the signal, e.g. by means of phases of rest or silence, an alignment method can be applied first. Some local similarity measures can be applied to the signal between the aligned landmarks. Another approach for longer sequences, it might become a lot more feasable to look at the signal not in the time domain but in the frequency domain. Here, differences between different frequency bands and between phases could be identified and used to construct a difference measure.

Depending on the research question, statistical approaches like generalized additive models (GAM) [19] can could be applied. GAM can fit curved trajectories over time jointly for many sequences. The advantage of such an approaches is that it can find systematic variations between groups even in a nested manner. This would be the right method to ask which systematic differences between the trajectories in *ja* and *halt* exist and how they are different to the ones resynthesized with the simulator.

Yet another approach might lie in the emerging application of recurrent neural networks (RNN) and variational autoencoders (VAE). In principle, these could be used to encode sequences of different durations with an RNN into a time independent latent representation. With the VAE framework, these latent representations are constructed by optimally encoding the structure in the sequence data while fulfilling constrains like following a multidimensional Gaussian distribution and therefore behaving similar to a principle component analysis. On this latent space representation, a difference measure can be constructed and could validated.

In conclusion, finding dissimilarity measures that work on short trajectories of articulatory movements is a difficult endeavour. With the JSD, there seems to be a measure which we can be used to distinguish the quality of resynthesis systems while fulfilling the properties of a metric.

References

[1] HARRIS, C. M. and D. M. WOLPERT: *Signal-dependent noise determines motor planning*. *Nature*, 394(6695), p. 780, 1998.

[2] VAN BEERS, R. J., P. BARADUC, and D. M. WOLPERT: *Role of uncertainty in sensorimotor control.*

Philosophical Transactions of the Royal Society of London B: Biological Sciences, 357(1424), pp. 1137–1145, 2002. doi:10.1098/rstb.2002.1101.

[3] CHURCHLAND, M. M., A. AFSHAR, and K. V. SHENOY: *A central source of movement variability*. *Neuron*, 52(6), pp. 1085 – 1096, 2006. doi:https://doi.org/10.1016/j.neuron.2006.10.034. URL http://www.sciencedirect.com/science/article/pii/S0896627306008713.

[4] BAYS, P. M. and D. M. WOLPERT: *Computational principles of sensorimotor control that minimize uncertainty and variability*. *The Journal of Physiology*, 578(2), pp. 387–396, 2007. doi:10.1113/jphysiol.2006.120121. URL https://physoc.onlinelibrary.wiley.com/doi/abs/10.1113/jphysiol.2006.120121. https://physoc.onlinelibrary.wiley.com/doi/pdf/10.1113/jphysiol.2006.120121.

[5] SOSNIK, R., B. HAUPTMANN, A. KARNI, and T. FLASH: *When practice leads to co-articulation: the evolution of geometrically defined movement primitives*. *Exp Brain Res*, 156, pp. 422–438, 2004.

[6] TIEDE, M., C. MOOSHAMMER, L. GOLDSTEIN, S. SHATTUCK-HUFNAGEL, and J. PERKELL: *Motor learning of articulator trajectories in production of novel utterances*. Hong Kong, 2011.

[7] RAEDER, C., J. FERNANDEZ-FERNANDEZ, and A. FERRAUTI: *Effects of six weeks of medicine ball training on throwing velocity, throwing precision, and isokinetic strength of shoulder rotators in female handball players. J Strength Cond Res.*, 29(7), pp. 1904–14, 2015.

[8] TOMASCHEK, F., D. ARNOLD, F. BROEKER, and R. H. R. BAAYEN: *Lexical frequency co-determines the speed-curvature relation in articulation*. *Journal of Phonetics*, 68, pp. 103–116, 2018.

[9] TOMASCHEK, F., D. ARNOLD, J. VAN RIJ, B. V. TUCKER, and K. SERING: *Proficiency effects on the movement precision during the execution of articulatory gestures*. under revision.

[10] WINKLER, R., S. FUCHS, and P. PERRIER: *The relation between differences in vocal tract geometry and articulatory control strategies in the production of french vowels: Evidence from mri and modeling*. pp. 509–516. 2006.

[11] FUCHS, S., R. WINKLER, and P. PERRIER: *Do speakers' vocal tract geometries shape their articulatory vowel space?* pp. 333–336. 2008.

[12] BRUNNER, J., S. FUCHS, and P. PERRIER: *On the relationship between palate shape and articulatory behavior*. *The Journal of the Acoustical Society of America*, 125(6), pp. 3936–3949, 2009. doi:http://dx.doi.org/10.1121/1.3125313.

[13] WEIRICH, M. and S. FUCHS: *Palatal morphology can influence speaker-specific realizations of phonemic contrasts*. *Journal of Speech, Language, and Hearing Research*, 56, pp. 1894–1908, 2006.

[14] TOMASCHEK, F. and A. LEEMAN: *The size of the tongue movement area affects the temporal coordination of consonants and vowels – a proof of concept on investigating speech rhythm. The Journal of the Acoustical Society of America*, 144(5), pp. EL410–EL416, 2018.

[15] ARNOLD, D. and F. TOMASCHEK: *The karl eberhards corpus of spontaneously spoken southern german in dialogues - audio and articulatory recordings*. In C. DRAXLER and F. KLEBER (eds.), *Tagungsband der 12. Tagung Phonetik und Phonologie im deutschsprachigen Raum.*, pp. 9–11. Ludwig-Maximilians-Universität München, 2016. URL https://ids-pub.bsz-bw.de/frontdoor/index/index/year/2017/docId/5944.

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

[17] 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.

[18] SERING, K., N. STEHWIEN, and Y. GAO: *create_vtl_corpus: Synthesizing a speech corpus with VocalTractLab*. 2019. doi:10.5281/zenodo.2548895. URL https://doi.org/10.5281/zenodo.2548895.

[19] WOOD, S. N., Y. GOUDE, and S. SHAW: *Generalized additive models for large data sets*. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 64(1), pp. 139–155, 2015.