Analysis of Coarticulation Using EMA Data with a Statistical Shape Space Model of the Tongue

Alexander Hewer[1,2], *Ingmar Steiner*[1,3], *Korin Richmond*[4]
[1]*Multimodal Computing and Interaction, Saarland University, Germany*
[2]*Saarbrücken Graduate School of Computer Science, Germany*
[3]*audEERING GmbH, Gilching, Germany*
[4]*Centre for Speech Technology Research, University of Edinburgh, UK*
hewer@coli.uni-saarland.de

Abstract: This study proposes a novel way to analyze the articulatory features of diphones in electromagnetic articulography (EMA) recordings by using a statistical tongue model to construct a speaker-independent articulatory space. Initial statistical analysis shows that trajectories from different instances of the same diphone are similar to each other in this space, which implies that it is a suitable representation to study coarticulation effects.

1 Introduction

It is well-known that speech is not simply a concatenation of isolated sounds, but is produced by articulatory movements in the vocal tract which transition from the target configuration of one phone to the next in a dynamic and fluent manner. This influence of phonetic context on speech production is referred to as *coarticulation*, and its manifestation can be measured both directly – e.g., through intra-oral motion capture such as electromagnetic articulography (EMA) – and in the resulting acoustic speech signal.

When speech is segmented into a sequence of discrete units, one possible representation is that of *diphones*, which comprise two contiguous half-phones ranging from the (presumed stable) center of one phone to the center of the next. Since diphones capture the immediate phonetic context, they preserve short-range coarticulation, and as such, are an ideal base unit for analyzing the related coarticulation effects.

However, the direct analysis of articulation and coarticulatory effects – even at the local level of diphones – is complicated by the fact that different instances of the same phone sequence have different temporal domains; moreover, comparing the same phone sequences across multiple speakers adds factors such as individual differences in vocal tract anatomy and articulation strategy.

Previous work applying EMA to the study of coarticulation has compared patterns in raw coil motion trajectories or derived statistical measures; Iskarous et al. [1] present a summary of multiple relevant studies and discuss the implications. However, in the past, the analysis of variability between utterances and speakers was typically limited to the EMA data, and a major drawback of this modality is the fact that the tongue surface, or even mid-sagittal contour, cannot be directly observed [2].

In order to model the dynamic aspects of coarticulation in a way that covers the entire tongue surface and allows a direct comparison of different instances of the same phone transitions in a speaker-independent manner, we can turn to a statistical shape space model of the tongue [3] that can separate speaker identity and tongue pose (i.e., phone identity). This model is constructed from magnetic resonance imaging (MRI) speech production data of multiple speakers, and can be fitted to the three-dimensional (3D) movements of sparse point sets, such as EMA fleshpoints on the tongue, as trajectories in a low-dimensional parameter space.

2 Statistical tongue model as articulatory space

This study uses a multilinear statistical tongue model [3] to represent tongue shape configurations with two sets of parameters, $\mathbb{P} \subset \mathbb{R}^m$ and $\mathbb{S} \subset \mathbb{R}^n$. The speaker parameters $\mathbb{S}$ model the anatomical properties of the tongue shape, while the tongue pose parameters $\mathbb{P}$ control the actual pose related features. Both spaces have the dimensions m and n that should be chosen in such a way that the model is not too complex; therefore, low-dimensional spaces are preferable. As the shape representation, a polygon mesh is used that allows to model the whole 3D surface of the tongue. Basically, a polygon mesh is generated from the model by providing the parameters $\mathbf{s} \in \mathbb{S}$ and $\mathbf{p} \in \mathbb{P}$. Furthermore, statistical measures are attached to $\mathbf{s}$ and $\mathbf{p}$ that describe how realistic the generated tongue shape is. In particular, multivariate Gaussian distributions are fitted to the parameter spaces. For $\mathbb{S}$, $\mathcal{N}(\mu_\mathbf{s}, \mathbf{I})$ is used, where $\mu_\mathbf{s}$ corresponds to the mean speaker parameter. Accordingly, $\mathcal{N}(\mu_\mathbf{p}, \mathbf{I})$ is fitted to $\mathbb{P}$ where $\mu_\mathbf{p}$ corresponds to the mean pose parameter. Both means are derived during the model construction process. The actual model is obtained by analyzing volumetric MRI recordings that show speech-related vocal tract configurations.

Statistical models have been used successfully in literature for registering EMA data [e.g., 4, 5, 6], i.e., the positional data provided by the articulograph was used to derive the 3D surface of the tongue. More importantly, the above description implies that the model treats the tongue pose features independently from the anatomical features. This means that $\mathbb{P}$ can be seen as an articulatory space, that is, a space where tongue shapes associated with two different points in the space only differ with respect to their articulatory features. This property makes the model ideally suited for studying articulatory phenomena, like for example, coarticulation effects during speech production.

In contrast to our previous paper [3] which used the Ultrax dataset [7] to derive the model, in the present study we use the MorphDB dataset [8] to construct the model. This dataset provides access to MRI recordings of 17 native speakers of American English. The phonetic inventory provides a balance between 13 vowels and 14 consonants: [ə, eɪ, æ, iː, ɛ, ɜ˞, ɪ, oʊ, uː, ɔː, ʌ, ɑː, ʊ, f, ʒ, h, l, m, n, ŋ, ɹ, s, ʃ, θ, ð, v, z]. This is different from the Ultrax dataset that mainly focused on vowels and the two sibilants [s] and [ʃ]. The present study also uses the measurements of the pilot speaker that were kindly provided by Sorensen et al., which results in data collected from 18 speakers.

The data was processed and evaluated in a way similar to the approach described by Hewer et al. [3]. The corresponding evaluation of the resulting model led to the decision to select $m = 6$ and $n = 8$. In this regard, it is important to note that a comparison between the new model and the one derived from the Ultrax data revealed that the new model was superior with respect to all investigated evaluation metrics.

3 Mapping EMA to the articulatory space

EMA is a well-established modality for investigating the dynamics of speech production by capturing the intra-oral motion of the tongue [9, 10, 11]. Although the temporal resolution of modern articulograph devices is very high, the spatial coverage of this modality is very limited: often, only 3 to 5 points on the tongue can be tracked in order to avoid impairing the recorded subject's articulation too much, which makes interpretation of such data very difficult. Moreover, rotational measurements of EMA coils are often unreliable, so the tongue is modeled based only on positional data of the selected fleshpoints. Here, the question arises how such sparse positional data may be mapped into the articulatory space described by the tongue pose parameters of the aforementioned tongue model.

The presented model can be used to register EMA data, i.e., to obtain parameters $\mathbf{s}$ and $\mathbf{p}$ at each time frame of the EMA recording such that the generated tongue mesh of the model is near the positional data of the associated tongue coils. To this end, first a correspondence between tongue coils and tongue mesh surface points has to be defined that remains fixed for the entire registration process. These correspondences may be provided manually, or they can be derived in a semi-supervised way. Afterwards, an energy optimization strategy is applied to find good values for $[\mathbf{s}]_t$ and $[\mathbf{p}]_t$ where $[\cdot]_t$ denotes the corresponding parameters at time step t of the recording. This study uses the following energy to perform the registration:

$$E([\mathbf{s}]_t, [\mathbf{p}]_t) = \alpha\, E_{\text{data}}([\mathbf{s}]_t, [\mathbf{p}]_t) + \beta\, E_{\text{bias}}([\mathbf{s}]_t, [\mathbf{p}]_t) + \gamma\, E_{\text{coherence}}([\mathbf{s}]_t, [\mathbf{p}]_t). \qquad (1)$$

The data term $E_{\text{data}}(\cdot)$ is minimized at time t if the distance between the tongue EMA coils and the associated surface points of the tongue mesh for the parameters $[\mathbf{p}]_t$ and $[\mathbf{s}]_t$ is small. The mean bias term $E_{\text{bias}}(\cdot)$ produces energy in proportion to how far away the current values of the parameters are from their respective mean. Penalizing the distance to the mean serves the purpose of avoiding unrealistic tongue shapes and providing the registration approach with additional information about the mean shape of the tongue, which may help to deal with the data sparseness of the EMA modality. Finally, the term $E_{\text{coherence}}(\cdot)$ tries to enforce a temporal consistency between consecutive time frames by penalizing differences between the corresponding parameter values. The positive weights α, β, and γ influence the importance of the individual terms. However, the above process optimizes both the speaker and the pose parameters, which implies that the strategy allows anatomy and pose to change over time and thus fails to reduce the data to purely articulatory features represented by the tongue pose parameters. As a remedy, the speaker anatomy can be fixed to appropriate values and the energy in Equation (1) is minimized again to perform a fixed speaker registration. These values can be obtained by averaging the results of the first registration using Equation (1), the full optimization registration. The result of the fixed speaker registration is then the desired mapping of the EMA data into the articulatory space described by the tongue pose parameters. Note only a brief summary of the registration and mapping process is provided here. For full detail, the reader is referred to our previous work in this area [3].

To analyze coarticulatory effects, the results of the fixed speaker registration may be used: they represent the articulatory gestures as trajectories in a low-dimensional space that is speaker-independent. Here, it is of interest to analyze how the overall shape of these trajectories differs across different instances of the same phone transitions. To this end, we can investigate short-range coarticulation by focusing on the transitions from one phone's steady state to the next; the first phone's second half and the second phone's first half comprise a diphone, and we will use this as the basic unit for our analysis of coarticulation.

To transform a given speech corpus into a form suitable for diphone analysis in the articulatory domain of the tongue model, the following operations are performed: first, instances of the diphone of interest are time normalized in such a way that the diphone starts at time 0 and ends at 1. Furthermore, the time 0.5 corresponds to the first time frame of the second half-phone comprising the diphone. This time normalization step is necessary because the durations of the individual instances differ from each other. In the next step, the normalized instances are interpolated by means of natural splines. The interpolation results are used to resample the data at equidistant points, which makes the different instances directly comparable to each other. Finally, the trajectories of each parameter are shifted such that they pass through 0 at time 0.5. This shifting operation serves the purpose of making the overall shape directly comparable. As a result of this processing, both the time and model parameter dimensions are unit-less.

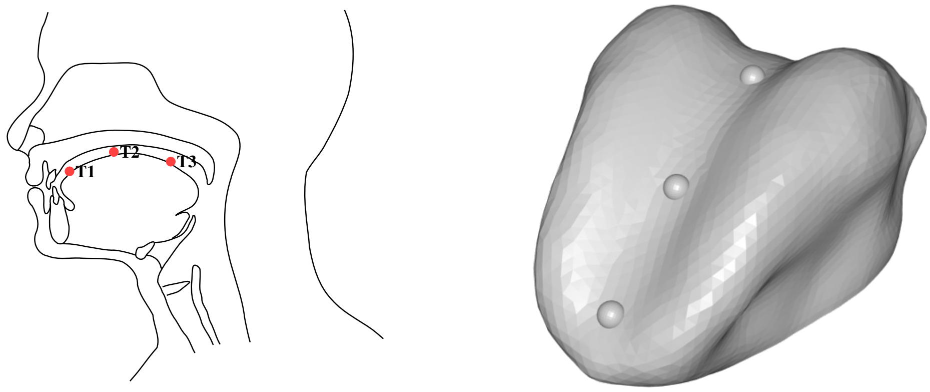


Figure 1 – Illustration of tongue coil layout for the mngu0 dataset (left, adapted from [12]) and corresponding tongue mesh surface points (right). Spheres on mesh highlight the corresponding vertices.

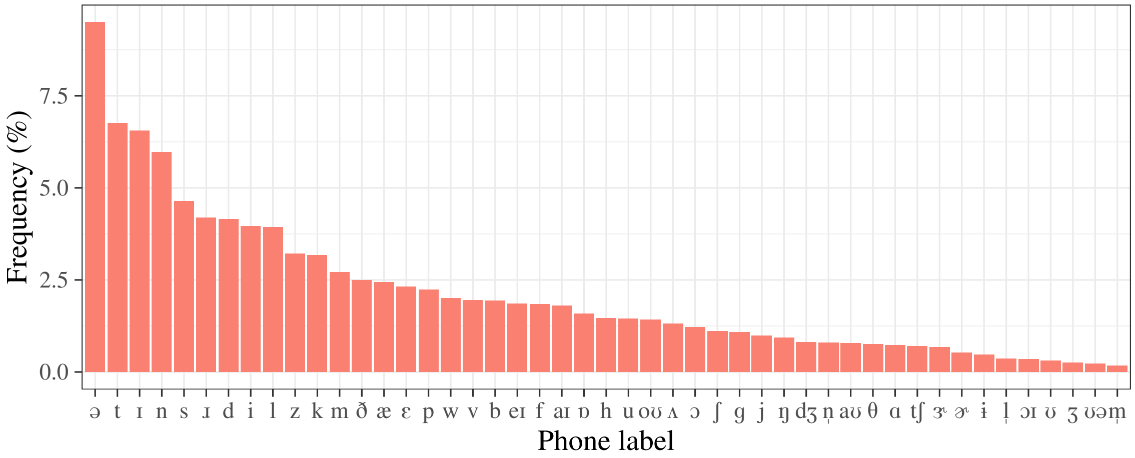


Figure 2 – Histogram of phone occurrence in the mngu0 day 1 dataset. The phone [ɔ̃] is omitted because it only occurs once.

4 Experiments

This section is dedicated to applying the proposed analysis strategy to actual EMA recordings. To this end, the mngu0 dataset [12, 13] is used. This corpus contains EMA recordings and time-aligned phonetic transcriptions of a single male speaker of British English. In particular, over 2000 utterances were recorded that were selected from English newspaper texts with the goal of maximizing the coverage of context sensitive diphones. All data was acquired using a Carstens AG500 articulograph[1] at a sampling rate of 200 Hz. The recordings were performed over two days where a different layout for the EMA coils was used on each day. In this experiment, the recordings of the first day are used, which consist of 1354 utterances corresponding to 67 min of recorded speech.

On this day, the coil layout shown in Figure 1 was applied. This figure also depicts the corresponding surface points on the tongue mesh that were used during the registration operation.

The actual data used in the experiments corresponds to the the following distribution packages that were downloaded from the *mngu0* website:[2]

1. Day1 basic EMA data, head corrected and unnormalized (v1.1.0)
2. Day1 transcriptions, Festival utterances and ESPS label files (v1.1.1)

The label files provide insight into the phonetic coverage of the recorded speech material. A histogram showing the phone distribution is shown in Figure 2. The label files are also critical

[1] Carstens Medizinelektronik GmbH, Bovenden, Germany; `http://www.articulograph.de/`

[2] `http://mngu0.org/`

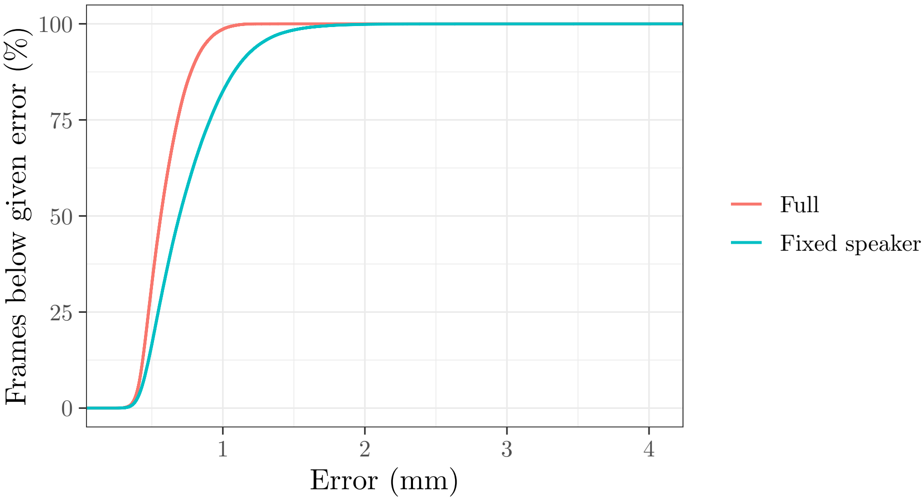


Figure 3 – Cumulative error for the two registrations of the mngu0 data.

	Full registration				Fixed speaker registration			
	Mean	**Std. dev.**	**Min**	**Max**	**Mean**	**Std. dev.**	**Min**	**Max**
Error (mm)	0.59	0.15	0.23	1.41	0.75	0.27	0.24	4.05

Table 1 – Error statistics for the two registrations of the mngu0 data.

for the diphone analysis: they describe when a specific phone was produced in an utterance and thus allow to extract the speech segments belonging to the diphones of interest. It is important to note that the label information is ignored during the actual registration of the EMA data.

To perform the registration, the following parameters were selected: $\alpha = 1, \beta = \gamma = 5$. These settings were manually tuned to obtain visually satisfying results. Before turning to the diphone-based analysis, it is vital to verify the quality of the registration experiments. To this end, the average Euclidean distance between EMA coils and corresponding mesh surface points was computed at each time frame.

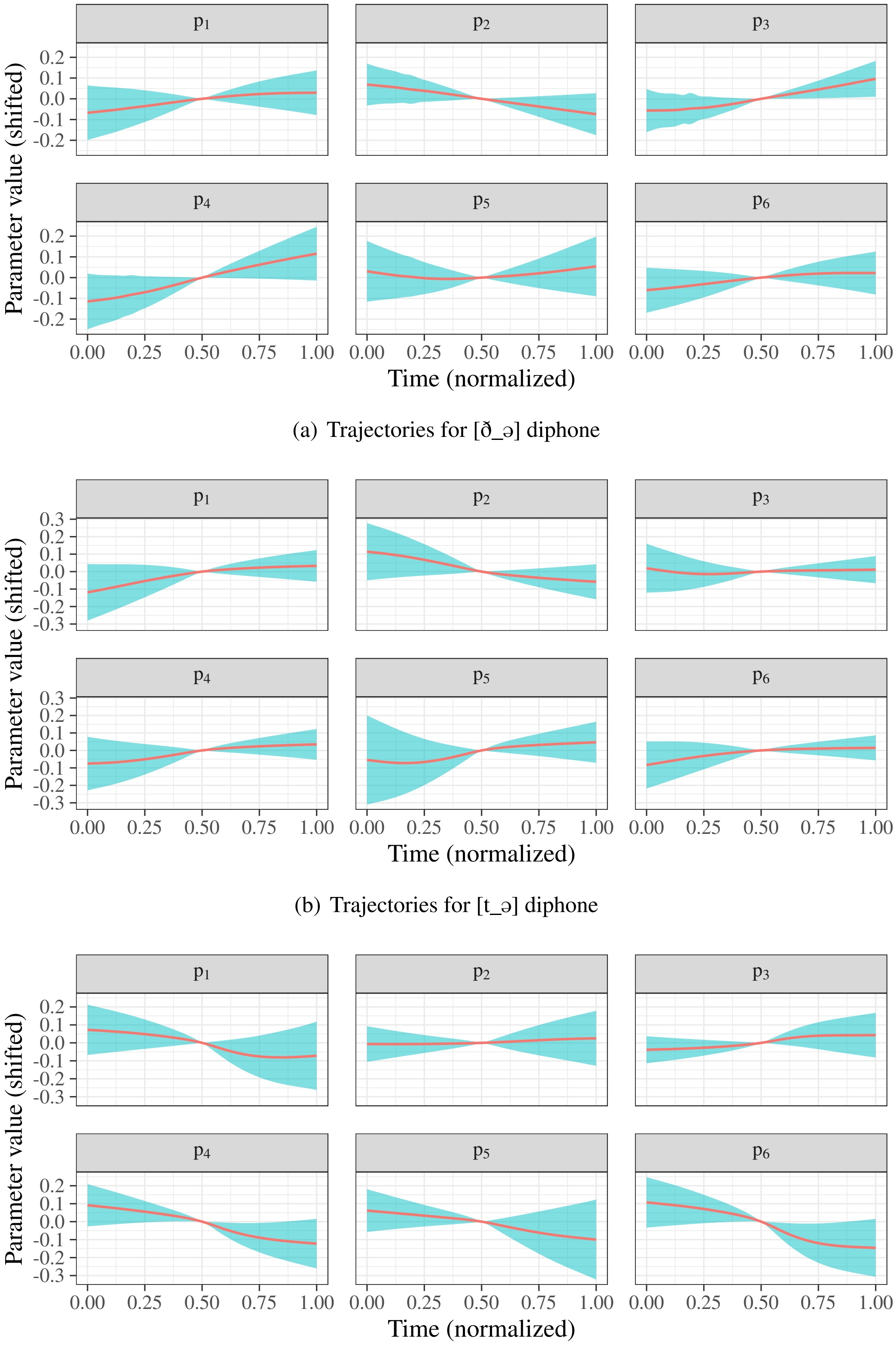
Results of this evaluation are shown in Figure 3 and Table 1. The results reveal that the both registrations provide acceptable results: if all parameters are used during the registration, nearly all errors are below 1 mm. Fixing the speaker parameters to the estimated anatomy and only optimizing the pose parameters leads to larger errors. However, they are mostly below 1.5 mm. Furthermore, the mean error only increases by a small amount. The slightly worse performance of the fixed speaker result can be attributed to the fact that in this case the registration has fewer parameters for adapting the tongue model to the EMA coil positions.

Having validated that the mapping to the articulatory space is of acceptable quality, we now turn to the described diphone-based analysis. Here, the three most frequent diphones in the dataset are selected for analysis:

1. [ð_ə] (693 instances)
2. [t_ə] (582 instances)
3. [ə_n] (523 instances)

However, only instances were counted and selected where at least one sample per participating phone is available in the diphone segment. This is due to the fact that the used articulograph might have failed to capture any sample of some phones due to its acquisition rate.

Results of the analysis are shown in Figure 4, where the mean trajectories and the associated standard deviation are visualized. The corresponding plots provide interesting insights: the overall shape of the trajectories is very consistent near the center of the diphone (time 0.5), which is reflected by small standard deviations. This observation implies that around the di-



(a) Trajectories for [ð_ə] diphone

(b) Trajectories for [t_ə] diphone

(c) Trajectories for [ə_n] diphone

Figure 4 – Visualization of the tongue pose trajectories for the three most frequent diphones. Plots show the mean trajectory (line) and the standard deviation (ribbon).

phone center, patterns in the trajectories of the six parameters may be present that could be used to identify the corresponding diphone. As these patterns occur in the articulatory space, they may be considered to be speaker-independent. Moving to the boundaries of the diphone causes the standard deviations to increase, which implies that the trajectories become less consistent. This observed behavior could be related to the context of the diphone: near the boundaries of the diphone segment, the tongue shape configuration is affected by coarticulation effects, i.e., it is influenced by the adjacent phones of the current diphone. Thus, we may conclude that the articulatory space described by the tongue pose parameters $\mathbf{p} \in \mathbb{P}$ is a suitable representation for analyzing and visualizing coarticulation effects.

5 Conclusion

The current study has presented a way of analyzing coarticulation effects by mapping EMA data into an articulatory space. This articulatory space is provided by the tongue pose parameters of a multilinear tongue model that separates anatomical and tongue pose related shape differences. Thus, mapping data into this space leads to a representation that may be regarded as speaker-independent. Initial results from an analysis of diphones show that the chosen representation can be used to observe coarticulation effects during speech production. In particular, the results indicate that there is in fact a stable region around the diphone center where the shape of the trajectories of the pose parameters is consistent.

In the future, the acquired diphone data in the articulatory space could be analyzed to train a model for synthesizing diphone trajectories in this space. Afterwards, these trajectories could be transferred to an arbitrary speaker to create animations of the entire tongue surface by adapting the speaker parameters accordingly. Furthermore, it would be worthwhile to investigate if pose parameter patterns are identifiable that could be used to detect diphones or articulatory gestures. Finally, it would be important to validate the hypothesis that the representation in the articulatory space is speaker-independent by analyzing a multi-speaker EMA dataset and comparing the trajectories across speakers, which could confirm that the used mapping also removes any features from the data that describe the speaker-specific articulation strategy.

References

[1] Iskarous, K., C. Mooshammer, P. Hoole, D. Recasens, C. H. Shadle, E. Saltzman, and D. H. Whalen: *The coarticulation/invariance scale: Mutual information as a measure of coarticulation resistance, motor synergy, and articulatory invariance*. *Journal of the Acoustical Society of America*, 134(2), pp. 1271–1282, 2013. doi:10.1121/1.4812855.

[2] Zharkova, N. and N. Hewlett: *Measuring lingual coarticulation from midsagittal tongue contours: Description and example calculations using English /t/ and /ɑ/*. *Journal of Phonetics*, 37(2), pp. 248–256, 2009. doi:10.1016/j.wocn.2008.10.005.

[3] Hewer, A., S. Wuhrer, I. Steiner, and K. Richmond: *A multilinear tongue model derived from speech related MRI data of the human vocal tract*. *Computer Speech & Language*, 51, pp. 68–92, 2018. doi:10.1016/j.csl.2018.02.001. URL `https://arxiv.org/abs/1612.05005`.

[4] Badin, P., F. Elisei, G. Bailly, and Y. Tarabalka: *An audiovisual talking head for augmented speech generation: Models and animations based on a real speaker's articula-*

tory data. In F. J. PERALES and R. B. FISHER (eds.), *Articulated Motion and Deformable Objects*, pp. 132–143. Springer, 2008. doi:10.1007/978-3-540-70517-8_14.

[5] ENGWALL, O.: *Making the tongue model talk: merging MRI & EMA measurements*. In *Eurospeech*, pp. 261–264. Aalborg, Denmark, 2001. URL `https://www.isca-speech.org/archive/eurospeech_2001/e01_0261.html`.

[6] ENGWALL, O.: *Combining MRI, EMA and EPG measurements in a three-dimensional tongue model*. *Speech Communication*, 41(2–3), pp. 303–329, 2003. doi:10.1016/S0167-6393(02)00132-2.

[7] ESHKY, A., M. S. RIBEIRO, J. CLELAND, K. RICHMOND, Z. ROXBURGH, J. SCOBBIE, and A. WRENCH: *UltraSuite: A repository of ultrasound and acoustic data from child speech therapy sessions*. In *Interspeech*, pp. 1888–1892. Hyderabad, India, 2018. doi:10.21437/Interspeech.2018-1736.

[8] SORENSEN, T., Z. SKORDILIS, A. TOUTIOS, Y.-C. KIM, Y. ZHU, J. KIM, A. LAMMERT, V. RAMANARAYANAN, L. GOLDSTEIN, D. BYRD, K. NAYAK, and S. S. NARAYANAN: *Database of volumetric and real-time vocal tract MRI for speech science*. In *Interspeech*, pp. 645–649. Stockholm, Sweden, 2017. doi:10.21437/Interspeech.2017-608.

[9] SCHÖNLE, P. W., K. GRÄBE, P. WENIG, J. HÖHNE, J. SCHRADER, and B. CONRAD: *Electromagnetic articulography: Use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract*. *Brain and Language*, 31(1), pp. 26–35, 1987. doi:10.1016/0093-934X(87)90058-7.

[10] PERKELL, J. S., M. H. COHEN, M. A. SVIRSKY, M. L. MATTHIES, I. GARABIETA, and M. T. JACKSON: *Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements*. *Journal of the Acoustical Society of America*, 92(6), pp. 3078–3096, 1992. doi:10.1121/1.404204.

[11] HOOLE, P. and A. ZIERDT: *Five-dimensional articulography*. In B. MAASSEN and P. VAN LIESHOUT (eds.), *Speech Motor Control: New Developments in Basic and Applied Research*, chap. 20, pp. 331–349. Oxford University Press, 2010. doi:10.1093/acprof:oso/9780199235797.003.0020.

[12] RICHMOND, K., P. HOOLE, and S. KING: *Announcing the electromagnetic articulography (day 1) subset of the mngu0 articulatory corpus*. In *Interspeech*, pp. 1505–1508. Florence, Italy, 2011. URL `http://www.isca-speech.org/archive/interspeech_2011/i11_1505.html`.

[13] STEINER, I., K. RICHMOND, I. MARSHALL, and C. D. GRAY: *The magnetic resonance imaging subset of the mngu0 articulatory corpus*. *Journal of the Acoustical Society of America*, 131(2), pp. EL106–EL111, 2012. doi:10.1121/1.3675459.