A data-driven model of acoustic speech intelligibility for optimization-based models of speech production

Benjamin Elie[1], *Juraj Šimko*[2], *Alice Turk*[1]

[1]LEL, PPLS, the University of Edinburgh; Edinburgh, Scotland, United Kingdom
[2]Faculty of Arts; University of Helsinki; Helsinki, Finland

benjamin.elie@ed.ac.uk, juraj.simko@helsinki.fi, a.turk@ed.ac.uk

Abstract

This paper presents a data-driven model of intelligibility which is intended to be used in an optimization-based model of speech production. The BiLSTM-based model is trained as a phoneme classifier and takes a sequence of real articulatory trajectories as input and returns the probability of phonemes over time. The optimization minimizes a cost function which is the weighted sum of the conflicting demands of being intelligible and least articulatory effort. The data-driven intelligibility model presented in this paper is used to compute the intelligibility score. Simulations support Lindblom's hypo- and hyper-articulation theory of speech, as the degree of hyper-articulation of speech can be modified and tuned along a continuum by balancing the importance given to both requirements of intelligibility and least articulatory effort.

Index Terms: Speech production, Articulatory planning, Optimal Control Theory, Intelligibility

1. Introduction

Optimization based models of speech production and articulatory planning assume that purposeful movements, including speech articulation, are planned to satisfy different, potentially conflicting, requirements. These requirements include maximal intelligibility, least articulatory effort, and minimal duration [1–12]. In these models, articulation is optimal with respect to several requirements if it minimizes a multi-objective cost function.

Recent approaches [11–13] have shown that several aspects of speech can be explained by optimization models which estimate articulatory trajectories by minimizing a simplified cost function accounting for two of the possible requirements: maximal intelligibility and least effort, as follows:

$$\mathcal{C}(\theta) = \alpha_{\mathcal{E}}\mathcal{E}(\theta) + \mathcal{P}(\theta), \tag{1}$$

where $\mathcal{C}(\theta)$, $\mathcal{E}(\theta)$, and $\mathcal{P}(\theta)$ are the overall cost, the effort cost, and the parsing cost (related to intelligibility), respectively, all functions of model parameters θ. The conflicting demands of maximal intelligibility and least articulatory effort can be balanced and modulated by adjusting the weight $\alpha_{\mathcal{E}}$ assigned to the effort cost.

The aspects of speech that can be explained by this (simplified) optimization-based approach include – but are not limited to – centralization of vowels in hypo-articulated speech [11,13], lenition of stop consonants [11, 13], and larger lip aperture and shift of F1 of vowels in Lombard speech [12]. Based on the assumption that human perception of phonemes reflects statistical distributions of their characteristics in the acoustic space, the intelligibility component of the parsing cost was modeled in these papers as a function of the probability of phoneme recognition given a vocal tract configuration. This approach requires labeled acoustic and articulatory data, ideally containing a great deal of variability in terms of phonetic context and speech style.

One way of obtaining a parallel corpus of acoustic and articulatory data is to use an articulatory synthesizer. This is the approach followed in [11,13], where the authors used the Maeda model [14] to generate articulatory–acoustic data. Synthesizing a large amount of sufficiently realistic running speech in different contexts and speech styles using the Maeda model is, however, extremely challenging. As a consequence, the authors adopted a solution consisting of training a model on *static* articulatory data with phonemic labels that were derived from the formant frequencies (generated by a physical model) for vowels and degree of constriction for consonants. This approach disregards the contribution of surrounding context to phoneme recognition. This means that this static approach of intelligibility model cannot account for phonemes with sequential targets (e.g., diphthongs, affricates) or for vowel quantity. In addition, using a simplistic articulatory model makes it extremely difficult to make quantitative comparisons of Maeda-based model output to real speech, as the model needs to be adapted to the studied speaker.

In this paper, we propose an alternative approach to this physics-based modeling of intelligibility, which is based on real articulatory data, namely electromagnetic articulography (EMA) data. The optimization procedure uses an acoustic context-dependent intelligibility function trained on real articulatory–acoustic pairings that are associated with phonemic labels. This ensures (1) that all phonemes of the studied language (here English) can be taken into account, (2) a better approximation of acoustic intelligibility which accounts for adjacent short- and long-term acoustic contexts, and (3) predicted articulatory trajectories (forming the optimal solution of the model) can be compared with real articulatory observations. Our data-driven intelligibility model is detailed in Section 2. In Section 3, we illustrate the interest of our data-driven model of speech production by presenting simulations in which we modify real observed trajectories of EMA sensors via optimization with varying demands on intelligibility and articulatory effort.

2. Data-driven modeling of intelligibility

In this paper, we propose a new probabilistic model to compute the cost of not being intelligible, i.e. the parsing cost in Eq. (1), trained on real labeled articulatory data. Inspired by models used for articulatory synthesis [15,16], we used a CNN-BiLSTM architecture. The BiLSTM part of the model ensures that short- and long-term acoustic contexts can be taken into account in our intelligibility model: the model is trained to predict a sequence of phonemes and their time boundaries given a time-sequence of articulatory trajectories. The CNN part consists of 4 1D-convolutional layers that take the sequence of EMA trajectories as input. As proposed in [15, 16], the 4 CNN layers have filter sizes of 30, 60, 90, and 85. The kernel size is 15 and the activation function is tanh for all CNN layers. One BiLSTM layer of size 16 with tanh activation function is used to process the output of the convolutional layers. Finally, one Dense layer with a softmax activation function is connected to the output of the BiLSTM layer. The loss function of the model is the sparse categorical cross-entropy between the predicted and the labeled phonemes. The softmax activation function is used to predict phoneme probability at each time frame.

2.1. Data

In this paper, we used data from the MOCHA-TIMIT corpus [17]. It is a corpus containing synchronized recordings of EMA, speech audio, and laryngograph signals of 2 native speakers of English (one male, labeled *msak0,* and one female, labeled *fsew0*) uttering 460 sentences from the TIMIT corpus [18]. The MOCHA-TIMIT EMA data were recorded at a sampling rate of 500 Hz, and consist of trajectories in the midsagittal plane of sensors glued on the vermilion borders of the lower and upper lips, the lower jaw, a location close to the tongue tip, on two locations posterior to the tongue tip (tongue mid- and back-), and on the velum. One key motive to use this corpus is the presence of a sensor on the velum. This makes it possible to take nasal phonemes into account. In the paper, EMA sensors are denoted as JAW, UL, LL, TT, TM, TB, V, for jaw, upper lip, lower lip, tongue tip, tongue mid, tongue back, and velum, respectively.

Articulatory data were downsampled to 200 Hz and a 11-order LOWESS filter was applied to remove signal noise. We used fundamental frequency, estimated from the laryngograph signals, as an additional feature. In total, we collected 15 features, corresponding to the position of the 7 sensors in the midsagittal plane plus fundamental frequency. We computed DELTA and DELTA-DELTA for all 15 features, yielding an input dimension of 45 for the model. The time step of the BiLSTM layers was the number of frames of the longest utterances in the MOCHA-TIMIT corpus (1283), the shorter utterances being zero-padded. We used the phonemic segmentation provided in the MOCHA-TIMIT corpus to label each articulatory time frame with the corresponding phoneme.

2.2. Training

We used 90% of utterances for training and 10% for validation. Utterances were split at random on condition that all phonemes were present in both training and validation datasets. We trained the model on 50 epochs and kept the epoch which returned the best accuracy on the validation dataset, which was 87.1%.

We consider that this accuracy is sufficient for our model given the uncertainty of the phonemic segmentation used for labeling the articulatory data, *i.e.* the phonemic segmentation of the MOCHA-TIMIT corpus, which is based on forced-alignment. This is illustrated, for instance, by the probability matrix for the sentence *Jane may earn more money by working hard* uttered by the male speaker, as shown in Figure 1. In this example, most of the phonemes are correctly estimated and the regions where they are estimated roughly match the original phonemic segmentation based on forced alignment (shown as vertical dashed lines in the figure). It also shows that most prediction errors are related to uncertainty about phoneme boundaries rather than falsely predicted phonemes. For instance the first and second 'ei' phonemes are predicted to begin slightly before their onsets provided in the phonemic segmentation. We believe that this should affect the optimization procedure only marginally. Additionally, one can notice other errors of phoneme prediction which can be attributed to either mispronunciation or errors in the phonemic segmentation: the final /d/ of *hard* has been estimated as /t/, and the model added a schwa /ə/ at the end. Note that better accuracy can be obtained with larger models, but we chose to use only one BiLSTM layer of relatively small size (16) because it provides a good trade-off between accuracy and low complexity (for the sake of computational time). For instance, a model with 4 BiLSTM layers of size 256 returned a slightly better accuracy score of 88.6%, but at the expense of being 9 times slower than our smaller model.

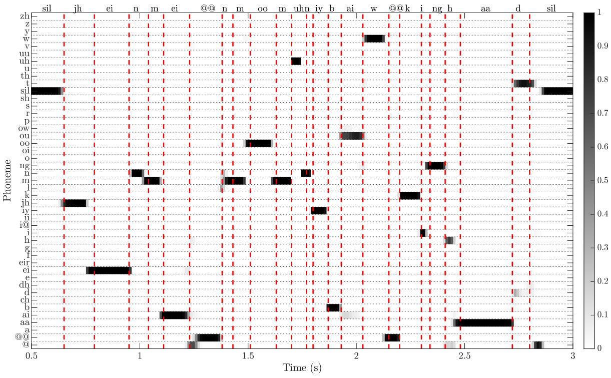

Figure 1: *Probability matrix of the utterance "Jane may earn more money by working hard" uttered by the male speaker. The segmentation of the utterance based on forced alignment of a phonemic transcription is shown with vertical dashed lines.*

3. Simulations

For our simulations, we modified segments of preexisting articulatory trajectories from the MOCHA-TIMIT corpus. These segments correspond to syllables taken from the validation dataset. For each speaker (male and female), we chose 25 closed syllables, either corresponding to a whole word (e.g., *than*) or to one syllable within a polysyllabic word (e.g., *sel* in *seldom*), resulting in a total of 50 syllables. Then, for each of the 50 syllables, we ran the optimization process on the articulatory movements that are included in the target syllable, with increasing values of the weight $\alpha_{\mathcal{E}}$ assigned to the least effort requirement. The aim of these simulations is to show the assumption that our model predicts Lindblom's H&H theory [4]. As such we make the following hypothesis that there is a continuum of hypo- and hyper-articulation of speech, which can be tuned by adjusting the weight $\alpha_{\mathcal{E}}$ assigned to the effort cost.

3.1. Optimization-based articulatory planning of speech

This proof-of-concept paper presents a minimal working version of our model. For that purpose, we model articulatory movements that satisfy only two tasks: 1) maximal intelligibility and 2) least articulatory effort: we use the same objective function as in Eq. (1), as proposed in [11,12]. Similarly to [11], we define the effort cost based on acceleration, as:

$$\mathcal{E}_n = m_n^2 \int_{T_n} |\ddot{x}_n(t)|^2 dt, \qquad (2)$$

where m is the mass of the articulator, and $x_n(t)$ is the time course of the nth articulator during the analyzed segment of duration T_n. For the sake of simplicity, we arbitrarily set the articulator mass m to 100g for all articulators. Tuning the values of m for different articulators is a complex task which is beyond the scope of this paper: this is left for future work. The total articulatory effort $\mathcal{E}$ is the sum of articulatory effort of each individual articulator.

The parsing cost is defined as $\mathcal{P}_p = 1 - \mathcal{I}_p$ where the Intelligibility $\mathcal{I}_p$ of the target phoneme p is:

$$\mathcal{I}_p = \int_{t_1}^{t_2} P(p|\mathbf{X}, t)dt, \tag{3}$$

where $P(p|\mathbf{X}, t)$ is the conditional probability of the target phoneme p at time t, given the time-course of the articulators stored in matrix $\mathbf{X}$, and where t_1 and t_2 are the onset and offset times, respectively, of the segment corresponding to the target phoneme p, as predicted by the forced-alignment.

Following [11, 13], we use the equations of General Tau Theory [19, 20] to generate the time-course of the articulatory movements. In this framework, individual movement units, namely movements between two successive local extrema, are modeled by 3 parameters: amplitude A, duration T and a shape parameter k which shapes the time-course of the movement, as:

$$x(t) = (x_0 - x_T)\left(1 - \frac{t^2}{T^2}\right)^{\frac{1}{k}} + x_T, \tag{4}$$

where $x(t)$ is the time-course of the movement unit, x_0 and x_T are the positions of the articulator at the onset and offset ($A = x_0 - x_T$), respectively, of the movement unit.

3.2. Articulatory model

Although it is possible to directly optimize the position of the different articulators, we use a specifically-designed articulatory model. The model is based on Principal Component Analysis (PCA) performed on the positions of the sensor for each speaker. It presents the advantages of 1) potentially reducing the number of dimensions, and consequently, the problem complexity, and 2) avoiding unrealistic solutions by constraining positions to be in a certain range of values along the principal components. For that purpose, for each speaker, we performed several PCAs on individual sensors, or groups of sensors (all three tongue sensors, which are interconnected). For the lower lip sensor and the group of tongue sensors, the PCAs have been applied to their position to which the position of the jaw sensor has been subtracted, as these sensors are connected to the jaw. The two articulatory models corresponding to the male and the female speaker both consist of 11 independent components (2 for the jaw, 2 for the upper lip, 2 for the lower lip, 3 for the tongue, and 2 for the velum). We kept only 3 components for the tongue sensors because we assume that they explain sufficient variance (91.8% and 94.2% for the male and female speakers, respectively). Consequently, the number of degrees of freedom of our articulatory model is reduced from 14 to 11. All components have been z-scored: values are expressed in terms of standard deviations around their mean.

In order to ensure that Tau equations can reproduce trajectories which are not too far from those observed in real speech, we performed Tau analysis on the trajectories of each component of our speaker-specific articulatory models. This consists of fitting Tau parameters to the trajectories of our articulatory model parameters, as explained in [21]. Trajectories were computed from all EMA recordings of both *msak0* and *fsew0* speakers included in the MOCHA-TIMIT corpus. The fit error, expressed as a Normalized Root Mean Square Error (NRMSE), has been shown to be low for all components, as it is always between 2 and 3%, whether it is for the male or the female speaker. This is very similar to the Tau fit error applied on EMA trajectories in [21], which was 2.38% overall. Given this relatively low Tau fit error, we assume that it is appropriate to use Tau theory to generate the time-course of the components of our model.

For the sake of simplicity, only the offset positions of the components of the articulatory model are optimized during these simulations. The other parameters, namely the duration and k-values of movements are kept at their original values. To avoid unrealistic articulatory configurations, we bounded these offset positions to be between -4 and +4 (expressed in standard deviations around the mean). We used the same fitting technique as in [21] to estimate Tau parameters for initialization of the optimization procedure.

3.3. Example of a solution

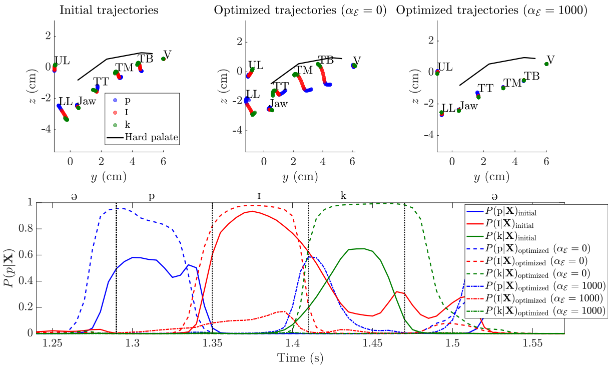

Figure 2: *Example of results returned by the optimization procedure for $\alpha_{\mathcal{E}} = 0$ (no constraint on effort, top center panel) and $\alpha_{\mathcal{E}} = 1000$ (top right panel), for the word "pick" in the sentence "Help Greg to pick a peck of potatoes" uttered by the male speaker. Position of EMA sensors (top panels) and probability functions of /p/, /ɪ/, and /k/ (bottom plot). The top left panel shows the initial position of sensors and the top center and top right panels show the optimized positions of sensors.*

Figure 2 shows an example of optimized solutions for the word "pick" in the sentence "Help Greg to pick a peck of potatoes" uttered by the male speaker (msak0). It shows that when there is no constraint on articulatory effort ($\alpha_{\mathcal{E}} = 0$), the optimized trajectories for "pick" exhibit larger amplitude. The tongue mid and tongue back sensors have to go higher toward the hard palate to properly seal the vocal tract in order to produce the velar stop /k/. At the same time, the tongue moves up forward toward the alveolar region to produce /ɪ/. Additionally, note that the optimized lip sensors start closer to each other at the onset of /p/ than the recorded (initial) lip sensors. This results in increased probability functions for /p/, /ɪ/, and /k/ compared to the probability functions associated to the initial trajectories. This shows that our model predicts that increasing hyperarticulation leads to better intelligibility. Conversely, when the constraint on articulatory effort is very high, e.g. $\alpha_{\mathcal{E}} = 1000$, the optimized positions of sensors over time show almost no movement. Lip sensors stay far apart from each other, preventing the formation of the occlusion during /p/, and tongue sensors stay in a low position, preventing the formation of the closure during the production of /k/. As a consequence, our model predicts very low intelligibility for these consonants.

3.4. Effects of the effort penalty

To quantitatively evaluate the degree of hypo- and hyper-articulation in speech, we compute the HH-index, as proposed in [22], which is defined as the ratio between the traveled distance of a sensor and a reference distance. In this paper, the HH-index is the ratio of traveled distances between the one returned by the optimized solution and the one observed in the EMA trajectories recorded in MOCHA-TIMIT (the reference distance). The time segment within which is computed the HH-index is the time segment corresponding to the optimized syllable. High values of the HH-index indicate hyper-articulation, while low values of HH-index indicate hypo-articulation.

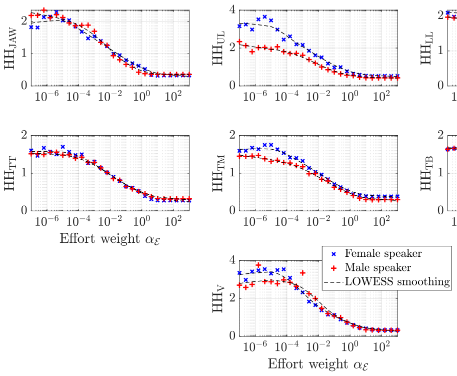

Figure 3: *HH-index of the 7 sensors as a function of the effort weight $\alpha_{\mathcal{E}}$ in a logarithmic scale, for the female ('x') and the male ('+') speakers.*

Figure 3 shows the effect of changing the weight $\alpha_{\mathcal{E}}$ assigned to the effort cost on the hyperarticulation index (HH) for all sensors (JAW, UL, LL, TT, TM, TB, V) and both speakers. Note that, for the sake of clarity, the HH-index values reported in this figure are mean values across utterances (for a given speaker and a given value of $\alpha_{\mathcal{E}}$). For every sensor, the HH-index decreases with increasing $\alpha_{\mathcal{E}}$, showing that more hypo-articulation is predicted when the weight assigned to effort cost is high, which confirms our hypothesis that $\alpha_{\mathcal{E}}$ is used to adjust the degree of hypo- and hyper-articulation of speech.

For a given effort weight $\alpha_{\mathcal{E}}$, the HH-index of the optimized articulatory trajectories for the female speaker is slightly higher than that of the male speaker. This can be explained by the fact that, in the original recordings, the male speaker was found to hyper-articulate slightly more than the female speaker. We observed this difference of hyper-articulation by computing the effort cost and the intelligibility score from the recorded EMA trajectories on the target syllables for both speakers. We found a mean effort cost of 0.36 for the male speaker (averaged over target syllables), but a lower mean effort cost of 0.31 for the female speaker. The mean intelligibility score (averaged over target syllables) was 0.69 for the male speaker and was 0.68 for the female speaker. Since the female speaker was slightly less hyper-articulating than the male speaker in the original recordings, the reference value for computing the HH-index was slightly lower for the female speaker than for the male speaker, hence her higher HH-index.

When looking at the hyper-articulated case (low $\alpha_{\mathcal{E}}$), the HH-index is higher for jaw and lip sensors than for tongue sensors with mean HH-index around 2 for jaw and lip sensors (except for the upper lip sensor of the female speaker which is more than 3), and mean HH-index is around 1.5 for tongue sensors. This observation is similar to the observations made in [22], where lips and jaw sensors have been shown to exhibit higher HH-index than tongue sensors in Lombard speech in the presence of high level of background noise.

4. Conclusion and future work

This paper has presented a data-driven model of intelligibility for use in models of speech production based on multi-objective optimization. The main contribution of this paper is the presentation of a model of intelligibility trained on real articulatory and acoustic data which can take acoustic context into account. Using the EMA MOCHA-TIMIT corpus [17], we trained a BiLSTM-based classifier to predict a sequence and time-boundaries of phonemes from EMA trajectories.

The paper has illustrated the interest of the approach via optimization-based modifications of real existing EMA trajectories from the MOCHA-TIMIT corpus, where optimization consisted of varying the trade-off between minimization of articulatory effort (hypo-articulation) and maximization of intelligibility (hyper-articulation). Our simulations implement the predictions of Lindblom's H&H theory [4], as the degree of hypo- and hyper-articulation of the optimized articulatory trajectories can be tuned along a continuum by balancing the importance given to both constraints (least effort and maximal intelligibility). In addition, our simulations predict different effects of hyper-articulation across articulators: in our simulations, jaw and lip sensors were found to be more hyper-articulated than tongue sensors when the constraint on least articulatory effort was low. Similar effects have been observed in Lombard speech [22].

The approach to intelligibility presented here has the potential to provide a significant improvement for computational implementations of optimization-based models of speech production [1, 11–13] as it takes into account long- and short-term acoustic context in the approximation of intelligibility. Additionally, it is not restricted in terms of phonemes to model. The model is currently at a preliminary stage of development and more work is required to tune the different parameters of the model (e.g., the mass of articulators). This tuning procedure could be done, for instance, via finding model parameter values for which the distance between the optimized solution and the observed trajectories is minimal. The use of real data makes our approach suitable for quantitative comparison with real articulatory observations. In addition, the use of real data could make possible the analysis of recorded articulatory trajectories via inversion methods, such as estimating the weights for which the solution of the optimization would return the observed trajectories. This could provide a new way to investigate speech production with higher level cognitive parameters.

In the future, intelligibility models could be trained on purely acoustic data that cover a wider range of speakers, speech styles and contexts. For that purpose, it will be necessary to use articulatory speech synthesizers to predict acoustics from planned articulatory trajectories [15, 16, 23]. In addition, connecting an articulatory synthesizer to our data-driven model of speech production could offer a way of investigating the acoustic effects of changes in the model parameters (e.g., weights of the cost function), and thus better understand the cognitive mechanisms which are involved in articulatory planning and speech variation.

5. Acknowledgments

This work was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (PlanArt: Planning the Articulation of Spoken Utterances; Grant awarded to A. Turk; Grant agreement No. 101019847).

6. References

[1] J. Simko and F. Cummins, "Embodied task dynamics," *Psychological review*, vol. 117, no. 4, pp. 1229—-1246, 2010.

[2] W. L. Nelson, "Physical principles for economies of skilled movements," *Biological cybernetics*, vol. 46, no. 2, pp. 135–147, 1983.

[3] J. S. Perkell, F. H. Guenther, H. Lane, M. L. Matthies, P. Perrier, J. Vick, R. Wilhelms-Tricarico, and M. Zandipour, "A theory of speech motor control and supporting data from speakers with normal hearing and with profound hearing loss," *Journal of phonetics*, vol. 28, no. 3, pp. 233–272, 2000.

[4] B. Lindblom, "Explaining phonetic variation: A sketch of the H&H theory," in *Speech production and speech modelling*. Springer, 1990, pp. 403–439.

[5] E. Todorov, "Optimal control theory," in *Bayesian brain: probabilistic approaches to neural coding*, D. K, Ed. MIT press Cambridge, MA, 2006, pp. 268–298.

[6] T. Flash and N. Hogan, "The coordination of arm movements: an experimentally confirmed mathematical model," *Journal of neuroscience*, vol. 5, no. 7, pp. 1688–1703, 1985.

[7] G. Ananthakrishnan and O. Engwall, "Mapping between acoustic and articulatory gestures," *Speech Communication*, vol. 53, no. 4, pp. 567–589, 2011.

[8] H. Rasilo, O. Räsänen, and U. K. Laine, "Feedback and imitation by a caregiver guides a virtual infant to learn native phonemes and the skill of speech inversion," *Speech Communication*, vol. 55, no. 9, pp. 909–931, 2013.

[9] N. Hogan, "An organizing principle for a class of voluntary movements," *Journal of neuroscience*, vol. 4, no. 11, pp. 2745–2754, 1984.

[10] J.-F. Patri, J. Diard, and P. Perrier, "Optimal speech motor control and token-to-token variability: a bayesian modeling approach," *Biological cybernetics*, vol. 109, pp. 611–626, 2015.

[11] B. Elie, J. Šimko, and A. Turk, "Optimal control of speech with context-dependent articulatory targets," in *Interspeech 2023, Dublin*, 2023.

[12] B. Elie, J. Šimko, and A. Turk, "Optimization-based modeling of Lombard speech articulation: Supraglottal characteristics," *JASA Express Letters*, vol. 4, no. 1, p. 015204, 01 2024. [Online]. Available: https://doi.org/10.1121/10.0024364

[13] B. Elie, J. Šimko, and A. Turk, "Optimization-based planning of speech articulation using general Tau Theory," *Speech Communication*, vol. 160, p. 103083, 2024.

[14] S. Maeda, "Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model," in *Speech production and speech modelling*. Springer, 1990, pp. 131–149.

[15] Y. Yu, A. H. Shandiz, and L. Tóth, "Reconstructing speech from real-time articulatory MRI using neural vocoders," in *2021 29th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021, pp. 945–949.

[16] P. Wu, T. Li, Y. Lu, Y. Zhang, J. Lian, A. W. Black, L. Goldstein, S. Watanabe, and G. K. Anumanchipalli, "Deep speech synthesis from MRI-based articulatory representations," *arXiv preprint arXiv:2307.02471*, 2023.

[17] A. Wrench, "A multichannel articulatory speech database and its application for automatic speech recognition," in *Proc. 5th seminar on speech production: models and data, 2000*, 2000.

[18] L. F. Lamel, R. H. Kassel, and S. Seneff, "Speech database development: Design and analysis of the acoustic-phonetic corpus," in *Speech Input/Output Assessment and Speech Databases*, 1989, pp. 161–170.

[19] D. N. Lee, "Guiding movement by coupling Taus," *Ecological psychology*, vol. 10, no. 3-4, pp. 221–250, 1998.

[20] ——, "General Tau Theory: Evolution to date," *Perception*, vol. 38, no. 6, pp. 837–850, 2009.

[21] B. Elie, D. N. Lee, and A. Turk, "Modeling trajectories of human speech articulators using general Tau theory," *Speech Communication*, vol. 151, pp. 24–38, 2023.

[22] J. Šimko, Š. Beňuš, M. Vainio, N. Campbell, D. Gibbon, and D. Hirst, "Hyperarticulation in Lombard speech: A preliminary study," in *Proceedings of Speech Prosody*, 2014, pp. 869–873.

[23] Y.-W. Chen, K.-H. Hung, S.-Y. Chuang, J. Sherman, W.-C. Huang, X. Lu, and Y. Tsao, "Ema2s: An end-to-end multimodal articulatory-to-speech system," in *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2021, pp. 1–5.