Trajectories predicted by optimal speech motor control using LSTM networks

Tsiky Rakotomalala, Pascal Perrier, Pierre Baraduc

Trajectories predicted by optimal speech motor control using LSTM networks

Tsiky Rakotomalala[1], *Pascal Perrier*[1], *Pierre Baraduc*[1]

[1] Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France

ny-tsiky.rakotomalala@grenoble-inp.fr, pascal.perrier@grenoble-inp.fr, pierre.baraduc@grenoble-inp.fr

Abstract

The question of optimality and its role in trajectory formation is at the core of important debates in motor control research. We present the first speech control model that associates Optimal Feedback Control (OFC) for planning and execution of movements with a biomechanical model of the vocal tract. Simulated trajectories in the VCV sequences are compared with trajectories generated using the GEPPETO model that drives the same 2D biomechanical model; in GEPPETO, the scope of optimality is limited to movement planning and to phoneme-related target motor commands.

In our OFC model commands are estimated via the minimisation of a cost that combines neuromuscular effort, and a penalty on accuracy of the auditory patterns reached for the phonemes. The biomechanics of the plant are implemented by an LSTM trained on simulations of a finite element model of the tongue. The comparison of the OFC model with GEPPETO relies on the time variation of the motor commands, the shape of the articulatory trajectories, and on auditory trajectories in the F1-F2 planes.

Index Terms: speech production, optimal control, internal model, speech motor control, trajectory formation.

1. Introduction

In order to understand speech motor control, that is, how the Central Nervous System (CNS) coordinates the speech articulators in order to generate facial movements and sounds that can be linguistically interpreted by listeners, several models have been presented. An interesting review of the most recent ones has been proposed by Parrell and colleagues [1]. Among them, the DIVA model [2], the Task Dynamics model [3], the FACTS model [4] (implementing principles of the SFC model [5]), and the ACT model [6] use a geometrical representation of the speech articulators; the GEPPETO model [7, 8]is the only model that integrates a realistic biomechanical description of the peripheral system that accounts for the effects on movements of mass, tissue elasticity, force generation mechanisms and some neurophysiological aspects of peripheral motor control with the account of a short delay feedback of the muscle stretch reflex type (see also [9]). However, this model has a strong weakness due to its control strategy: it is unable to react to any kind of perturbation because control is purely feedforward.

On the other hand, optimal control theory gained a wide popularity in motor control modeling because of its ability to account for behavioral observations at several levels of analysis: kinematic (effector trajectories [10]), mechanical (interaction forces [11]), or neurophysiological (electromyography [12], cortical electrophysiology [13]). These predictions emerge from the definition of a goal or set of goals and a cost function. Moreover, optimal feedback control theory [14] provides an important extension by integrating the role of sensory feedback both in planning and during movement execution, the optimal solution integrating an internal estimate of the current state of the effector to determine the future motor commands, allowing to predict how the organism will react to perturbations (e.g. [15]).

Thereby, we wished to present for speech production an optimal feedback control model in which the plant to be controlled is not simply geometrical but takes into account some biomechanical characteristics of the speech production system. For that, we used the same plant as the GEPPETO model. A preliminary version of this work can be found in [16]; while it produced interesting results for vowels, its simplicity made it unable to handle non-linearities in consonants production. Here we tackled this limitation by appling autoencoder networks for dimensionality reduction and recurrent networks (LSTM) for state prediction. Thus, we are now able to compare trajectories from the GEPPETO model and from the OFC model and evaluate the role of optimality in trajectories formation in single vowels, consonants, as well as sequences of phonemes.

2. Method

2.1. The GEPPETO model

In GEPPETO (acronym for "GEstures shaped by the Physics and by a PErceptually-oriented Target Optimization"), a speech sequence is considered as a sequence of phonemes to which is associated a sequence of perceptual targets in the form of the first three formants. GEPPETO integrates optimality at the level of the planning only, by minimizing a cost that combines motor cost and perceptual accuracy. This cost only takes into account the motor commands and the perceptual characteristics of the production at the successive phoneme-related targets, without any consideration for trajectories between the targets. The result of the optimization is a set of target motor commands, from which movement will be generated in a purely feedforward manner, based on principles of the Equilibrium Point Hypothesis (EPH) [17].

The biomechanics of the tongue is modeled by the Finite Element method (FE) and seven muscles responsible for the main displacements of the tongue in the sagittal plane are modeled (Fig. 1). The mechanism for generating muscle force is based on the EPH [17], according to which movement is generated from an equilibrium position to the next by a simple shift of the motor commands at a constant rate. Thus trajectories between equilibrium positions are not directly controlled, but result from the interaction between the physical characteristics of the effector and the desired equilibrium positions. In the "λ model" associated with this theory, the force generated by a muscle depends on the difference between its length and an activation threshold λ which is the control variable, this implements a low-level reflex of the "stretch reflex" type.

Muscle force F generated by the model is specified as

$$F = \rho[\exp(cA) - 1] \quad (1)$$

, where c is a form parameter accounting for the gain of the feedback from the muscle to the motoneurons pool and ρ a magnitude parameter directly related to force-generating capability. Muscle activation A is derived from

$$A = l - \lambda + \mu\dot{l} \quad (2)$$

where l and $\dot{l}$ are resp. the actual muscle length and lengthening velocity and μ a damping coefficient [8].

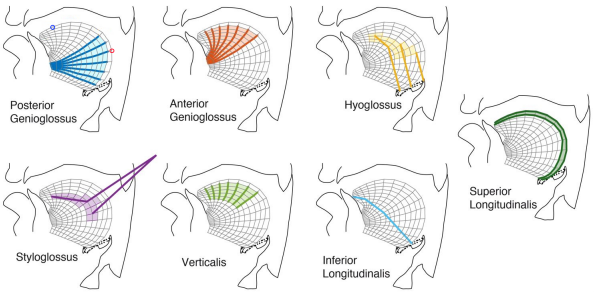

Figure 1: *The seven muscles in the FE model. Bold lines represent the macrofibers, over which overall muscle strength is distributed. Colored quadrilaterals are the elements whose stiffness increases with muscle activation. To visualize tongue kinematics, we chose two points on the upper tongue contour that well characterize the phonetic properties of the articulation: a palatal point (blue dot in the top left panel), and a pharyngeal one (red dot).*

For a given position of the tongue (fixed jaw and labial opening), the area function of the vocal tract is deduced from reference anatomical data [18]. The first three formants are computed with an acoustic harmonic analog of the tract [19].
For motor planning, GEPPETO assumes that the central nervous system stores the association between the values of the control variables (λ) corresponding to the target positions and the corresponding formant frequencies. The model exploits this association to optimize the motor commands carrying out the different phonemes of the sequence, under a perceptual constraint [8].

2.2. The optimal control model

2.2.1. Computation of the optimal motor commands

Like GEPPETO, the optimal control model aims to achieve formant goals using the same biomechanical model. However, while GEPPETO separates the planning phase, in which an optimal choice of target commands is made, from the execution phase (purely feedforward), the model proposed here does not separate planning and execution, and implements optimality during the execution of the movements, once the formant targets and the desired duration T of the movement have been specified.
To determine the motor commands that generate movements of the biomechanical model, a module called *optimal controller* computes the trajectories of the λ commands over time by minimizing a cost function. It contains a term related to neuromuscular effort and a term corresponding to a penalty on accuracy:

$$C = \int_0^T ||\lambda(t)||^2 dt + \alpha||p(T) - p_{goal}||^2 \quad (3)$$

where λ is the motor command vector, p_{goal} is a vector containing the specification of the desired targets, both in terms of auditory goals in formantic space and as a requirement of final stability (null final tongue velocity), $p(T)$ is its actual value at the end of the trajectory, and α is a trade-off parameter between precision and effort. To solve this constrained optimization problem, we used the adjoint method [20].
Since sensory feedbacks are delayed due to physiology, it is important to have an estimator of the state of the system to avoid instability. An *optimal estimator* estimates the state of the system at each time step based on the delayed sensory feedback and a copy of the motor commands (efference copy, see Fig. 2).

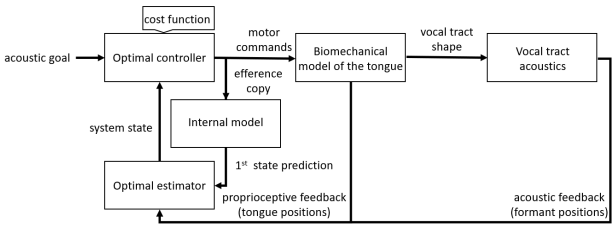

Figure 2: *Architecture of the optimal control model.*

2.2.2. The reduced model

To (among other things) speed up computation, this predictive model is a simplified model of the FE model, called the *reduced* biomechanical model. The upper contour of the tongue, described by the position of 16 nodes in the 2D sagittal plane, is reduced via an autoencoder to a 4D vector. Compared to a preliminary study [16] where we used principal component analysis for dimensionality reduction and a single-layer perceptron for prediction, this model is now able to handle the sharp nonlinearities resulting from the contacts of the tongue with the palate during consonant production. This is achieved thanks to an LSTM network model of the plant dynamics, trained on thousands of simulations of the FE model (for model structure, see Fig. 3 below).

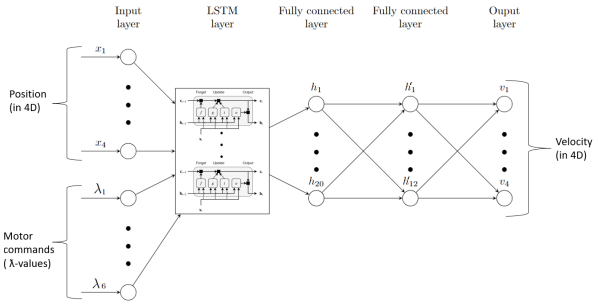

Figure 3: *LSTM Network architecture (LSTM layer scheme taken from https://fr.mathworks.com).*

2.3. Evaluation of the optimal control model

The evaluation of the optimal control model consists of three points: (1) Comparison of the trajectories generated in the articulatory and the auditory spaces with the optimal control model and with the GEPPETO model in Vowel1-Consonant-Vowel2 sequences; (2) Assessment of the capacity of the optimal control model to account for coarticulatory effects as observed in experimental data when Vowel1 or Vowel2 vary; (3) Assess-

ment of the sensitivity of the results obtained with the optimal motor control model when we vary the initial conditions of the optimization process. Regarding this latter point, we tested the following initial conditions: 1) the starting guess for the motor commands was the EPH solution; 2) this starting guess was reduced by 3mm for all muscles (i.e. associated with stronger level of activation); 3) this starting guess was corrupted by Gaussian noise with a 30 dB signal-to-noise ratio.

3. Results

3.1. Comparison of the optimal control model and the GEPPETO model

In this section we present examples of trajectories obtained in three different conditions: (1) Condition 1: the [aki] sequence produced with the optimal control model (Fig. 4); (2) Condition 2: the sequence produced with GEPPETO when the motor commands at the phoneme-related targets are the ones proposed by the optimal control model at the time where targets are reached (Fig. 5); the comparison of the trajectories with those obtained under condition 1 shows the impact of the optimal control model; (3) Condition 3: the sequence obtained with the biomechanical model when the motor commands proposed by the optimal model are used (Fig. 6); the comparison of the trajectories with those obtained under condition 1 will shed light on the influence in the optimal control model of the possible inaccuracy of the reduced LSTM model. The simulation presented in Fig. 4 was obtained under the following specifications: the first target is vowel /a/ as defined by formants F1=581 Hz, F2=1397 Hz and F3=2580 Hz; it should be reached with zero velocity at time 0.12 s; the second target is consonant /k/ as defined by formants F1=359 Hz, F2=1360 Hz and F3=2543 Hz ; it should be reached with zero velocity at time 0.24 s; the third target is vowel /i/ as defined by formants F1=358 Hz, F2=2228 Hz and F3=3073 Hz ; it should be reached with zero velocity at time 0.36 s.

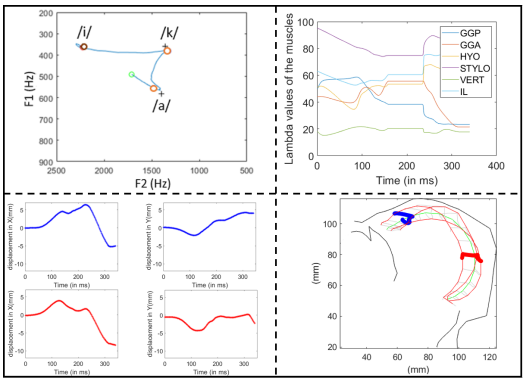

Figure 4: *Results obtained for the sequence [aki], starting from the resting position, with the optimal control model. Top panel, left: auditory trajectories in the F1-F2 plane; the green circle represents the resting position, brown the final position, and orange circles each intermediate phoneme; Top panel, right: time variations of the λ motor commands; Bottom panel, left: time variation of the horizontal and vertical positions of the nodes located on the tongue in the palatal and pharyngeal region (see Fig. 2, top-left panel); Bottom panel, right: articulatory trajectories and tongue shape at targets in the mid-sagittal plane.*

The simulation presented in Fig. 5 was obtained with motor commands varying at a constant rate of shift from the motor commands at a target to the ones of the next target; for each transition the commands vary during 100 ms and the commands are then held fixed during 20 ms.

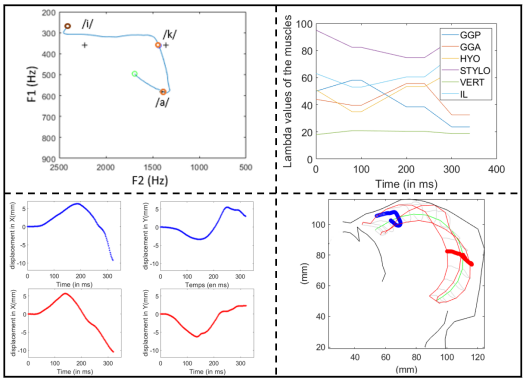

Figure 5: *Results obtained for the sequence [aki], starting from the resting position, with the GEPPETO model, using as targets the λ values of the OFC model at their respective via point time (transitions being made at a constant rate of shift). See Fig. 4 for details.*

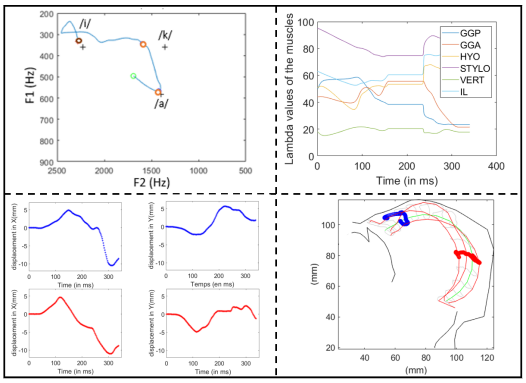

Figure 6: *Results obtained for the sequence [aki], starting from the rest position, when applying the motor commands proposed by the optimal control model to the biomechanical model. See Fig. 4 for details.*

3.2. Assessement of coarticulatory effects

The capacity of the optimal control model to account for the coarticulatory variability observed for velar consonant /k/ was evaluated via the generation of the sequences [iki], [aki] and [aka] and the comparison of the tongue posture reached at the time specified for the production of [k]; the influence of the effort-accuracy trade-off, as implemented with the α parameter in equation (3), was also assessed. Results obtained with α equal to 0.5 and 5 are presented and compared (Fig. 7).

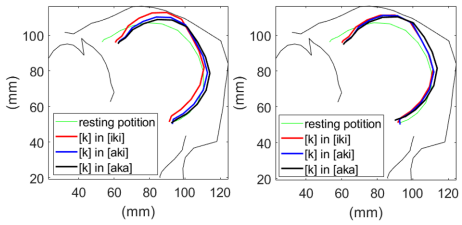

Figure 7: *[k] postures reached in different sequences with the optimal control model with different values of the trade-off parameter α equal to 0.5 (at the left) and 5 (at the right). Green contour: resting position. Red contour: /k/ in /iki/. Blue contour: /k/ in /ika/. Black contour: /k/ in /aka/.*

3.3. Sensitivity of the optimization procedure to its initialization

As a simple assessment of the sensitivity of the optimization algorithm to its initialization, we evaluated the auditory variability of the articulatory patterns generated with the OFC model under the three initialization conditions described above in Section 2.3. We present four estimates of inter-run difference in the F1-F2-F3 auditory space: Euclidian distance between the production of vowel /a/, /k/, /i/ and root mean squared error (RMSE) over the whole sequence /aki/.

Table 1: *Variation of the results obtained in the auditory domain with different initializations of the optimization algorithm. Distances (in Hz) for the separate targets [a], [k], and [i] and RMSE for the complete trajectory.*

Error	Init. 2 vs. Init. 1.	Init. 3 vs. Init. 1.
Max error at [a]	27.9	5.1
Max error at [k]	14.2	1.1
Max error at [i]	7.3	0.7
RMSE in [aki]	32.0	8.1

4. Discussion

Fig. 4 illustrates well that our OFC model is efficient: The intended auditory targets are well reached; at the times that we specified for achieving the targets, the variations in time of the vertical and horizontal positions of the nodes in the palatal and pharyngeal regions either become reversed or slow down, which is consistent with a zero or a very small velocity (the observation of the velocity profiles, non represented here, confirm this statement); the tongue shapes reached at targets are realistic, with a retraction and lowering of the tongue for /a/, a bunching and elevation of the tongue in the velar region for /k/ and a protrusion and elevation of the tongue for /i/; the transitions in the F1-F2 plan (and in the F2-F3 plan non represented here) are consistent with experimental observations from human subjects. We also observe that the articulatory trajectories are curved and globally describe a forward loop in the velar region, in the /aki/ sequence, as observed in several experimental studies (see [21] for a summary). However, the observation of the time variations of the positions in the palatal and pharyngeal regions reveals that the targets are not hold when they are reached. The zero or close to zero velocity does not hold for long. As a consequence the duration of the produced vowels is in general much shorter than the durations usually observed in humans.

Fig. 5 shows that the GEPPETO model does not reach all the auditory targets, despite the fact that it was controlled on the basis of the same target commands as the ones proposed by the OFC model. In addition, the positions closest to the desired targets are not reached within the specified time. In general we observe a clear overshoot of the targets, both in the F1 (for /k/ and /i/) and in the F2 direction (for /i/). Qualitatively we do not observe relevant differences of the shape of the trajectories either in the auditory F1-F2 plan or in articulatory sagittal plan. The comparison of the time variation of the motor commands reveals that the respect of time constraints on the achievement of targets is due in the OFC model to a strong decrease, for a short duration, of the λ commands of the muscles that are the agonists of the corresponding movement (Hyoglossus for /a/, Styloglossus for /k/, Posterior Genioglossus for /i/), which induces strong peaks of activation for these muscles. This is clearly different from the hypotheses underlying the EPH on which the motor control model of GEPPETO is based. In addition the very low values suggested by the OFC model for these agonist muscles induce clear overshoot of the articulatory positions in the GEPPETO model as shown in the tongue shapes reached for /k/ (with a larger contact region between the tongue and the palate) and for /i/ (for which we observe palatal contacts).

The comparison of Fig. 6 and Fig. 4 shows that the motor commands inferred by the OFC model do not enable a satisfactory reaching of the targets, either in time or in the auditory and articulatory domains, when they are applied to the biomechanical model. This is due to the inaccuracy of the LSTM model that is used in the optimization process in the OFC model. Surprisingly, while the LSTM provides a very good account of the dynamics of the biomechanical model in our learning and testing sets, its inaccuracy has strong consequences on the target reaching when it is included in the optimization process. Further tests are needed to thoroughly investigate the origins of this phenomenon. In particular it will be important to assess whether imposing a hold duration for the targets and avoiding strong peaks of muscle activation in the OFC model can enable us to reduce the consequences the inaccuracy of the internal model in the optimization process.

Fig. 7 shows that the OFC model is able to nicely reproduce the trends of coarticulation observed in human subjects: the tongue shape reached for /k/ is more anterior in the anterior vocalic context /iki/ than in the intermediate /ika/ and posterior /aka/ contexts. In addition, as expected the variation is larger when the weight of the perceptual accuracy in the optimized cost is smaller. The magnitude of the variation is smaller that the one observed in human data (see [22]). Further investigations are needed to more carefully assess this point, but the trend supports the hypothesis that the OFC model can realistically account for coarticulation phenomena.

Finally, Table 1 demonstrates the efficacy of the optimization procedure, since the global variation of the results is limited and within a range that prevent any perceptual inaccuracy.

5. Conclusion

The goal of this paper was to evaluate the role of optimality in tongue trajectories during speech production, by comparing the predictions of the more complex OFC model to an EPH model. We observed that applying optimal control to a biomechanical model does not significantly affect the main characteristics of the articulatory trajectories, suggesting a main influence of biomechanics in this regard. While the OFC model is able of a better precision, and able to account for an effort-precision trade-off, it is also sensitive to inaccuracy in internal model specification. The use of LSTM networks for the internal model is efficient compared to simple feedforward networks to handle non-linearities in consonant production. It enables us to avoid potential instabilities associated with feedback delays, which is a clear perspective of this work. However, its inaccuracy can generate errors in the articulatory and auditory domains. Future work will further investigate this aspect of our OFC model.

6. Acknowledgements

This work was funded by the Multidisciplinary Institute in Artificial Intelligence of the Université Grenoble Alpes (MIAI@Grenoble Alpes, ANR-19-P3IA-0003) and the French Agence Nationale de la Recherche (ANR), under grant ANR-16-CE19-0005 (project BrainSpeak).

7. References

[1] B. Parrell, A. C. Lammmert, G. Ciccarelli, and T. F. Quatieri, "Current models of speech motor control: A control-theoretic overview of architectures and properties," *The Journal of the Acoustical Society of America*, vol. 145, no. 3, pp. 1456–1481, 2019.

[2] F. H. Guenther, S. S. Ghosh, and J. A. Tourville, "Neural modeling and imaging of the cortical interactions underlying syllable production," *Brain and Language*, vol. 96, no. 3, pp. 280–301, 2006.

[3] E. Saltzman, "Task dynamic coordination of the speech articulators: A preliminary model," *Experimental Brain Research Series*, vol. 15, pp. 129–144, 1986.

[4] B. Parrell, V. Ramanarayanan, S. Nagarajan, and J. Houde, "The facts model of speech motor control: Fusing state estimation and task-based control," *PLoS Comput. Biol.*, vol. 15, no. 9, p. e1007321, 2019.

[5] J. F. Houde and S. S. Nagarajan, "Speech production as state feedback control," *Frontiers in Human Neuroscience*, vol. 5, p. 82, 2011.

[6] B. J. Kröger, J. Kannampuzha, and C. Neuschaefer-Rube, "Towards a neurocomputational model of speech production and perception," *Speech Communication*, vol. 51, pp. 793–809, 2009.

[7] P. Perrier, L. Ma, and P. Yohan, "Modeling the production of vcv sequences via the inversion of a biomechanical model of the tongue," in *Proceedings of Interspeech2005*. International Speech Communication Association, 2005, p. 1041–1044.

[8] 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, no. 6, pp. 611–626, 2015.

[9] J. Dang and K. Honda, "Construction and control of a physiological articulatory model," *The Journal of the Acoustical Society of America*, vol. 115, no. 2, pp. 853–870, 2004.

[10] T. Flash and N. Hogan, "The coordination of arm movements: an experimentally confirmed mathematical model," *The Journal of Neuroscience*, vol. 5, pp. 1688–1703, 1985.

[11] M. Kawato, Y. Maeda, Y. Uno, and R. Suzuki, "Trajectory formation of arm movement by cascade neural network model based on minimum torque-change criterion," *Biological Cybernetics*, vol. 62, no. 4, pp. 275–288, 1990.

[12] J. Collins, "The redundant nature of locomotor optimization laws," *Journal of Biomechanics*, vol. 28, no. 3, pp. 251–267, 1995.

[13] E. Guigon, P. Baraduc, and M. Desmurget, "Coding of movement- and force-related information in primate primary motor cortex: a computational approach." *The European journal of neuroscience*, vol. 26, no. 1, pp. 250–260, Jul. 2007.

[14] E. Todorov, "Optimality principles in sensorimotor control," *Nature Neuroscience*, vol. 7, no. 9, pp. 907–915, 2004.

[15] J. Diedrichsen, R. Shadmehr, and R. B. Ivry, "The coordination of movement: optimal feedback control and beyond," *Trends in cognitive sciences*, vol. 14, no. 1, pp. 31–39, 2010.

[16] P. Baraduc and P. Perrier, "Motor control of the tongue during speech: Predictions of an optimization policy under sensorimotor noise," in *Society for Neuroscience Annual meeting (Neuroscience 2017)*, 2017, pp. 408–03.

[17] A. G. Feldman, "Once more on the equilibrium-point hypothesis (λ model) for motor control," *Journal of Motor Behavior*, vol. 18, no. 1, pp. 17–54, 1986.

[18] P. Perrier, L. Boë, and R. Sock, "Vocal tract area function estimation from midsagittal dimensions with ct scans and a vocal tract cast: Modeling the transition with two sets of coefficients," *Journal of Speech and Hearing Research*, vol. 35, no. 1, pp. 53–67, 1992.

[19] P. Badin and G. Fant, "Notes on vocal tract computation," *STL QPSR*, vol. 2, no. 3, pp. 53–108, 1984.

[20] A. E. Bryson, *Dynamic Optimization*. Addison Wesley, 1999.

[21] P. Perrier, Y. Payan, M. Zandipour, and J. S. Perkell, "Influences of tongue biomechanics on speech movements during the production of velar stop consonants: A modeling study," *The Journal of the Acoustical Society of America*, vol. 114, no. 3, pp. 1582–1599, 2003.

[22] A. Bothorel, P. Simon, F. Wioland, and J.-P. Zerling, *Cinéradiographie des voyelles et des consonnes du français*. Travaux de l'Institut de Phonétique de Strasbourg, 1986.