INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Estimating virtual targets for lingual stop consonants using general Tau theory

Benjamin Elie[1], *Alice Turk*[1]

[1]Linguistics and English Language; the University of Edinburgh; Edinburgh, Scotland, UK

benjamin.elie@ed.ac.ek, a.turk@ed.ac.uk

Abstract

This paper investigates the existence and position of virtual targets during the production of stop consonants. Using the equations from general Tau theory to model the time-course of tongue constriction formation movements, targets were estimated by fitting these equations on observed tongue constriction variables extracted from real EMA data from 2 native speakers of English. Results suggest that targets are virtual for 50 to 60% of movements. For these movements, virtual targets of the tongue tip constriction are predicted to occur around 0.1 cm beyond the palate, and virtual targets for the tongue dorsum constriction are predicted to occur between 0.05 and 0.2 cm beyond the palate. Our results suggest that the time-course of movement is planned so that the onset of closure occurs with relatively high velocity: closure onset is generally located very close in time to the time of peak velocity.

Index Terms: Speech production, Virtual target, Tau Theory

1. Introduction

The production of stop consonants requires contact between an articulator and a rigid part of the vocal tract (or another articulator in the case of bilabial stop consonants) with sufficient force in order to properly seal the vocal tract [1]. Virtual targets have been proposed to account for 1) late velocity peaks for closing movements [1–3] and 2) the compression of articulators during closing phase [1, 4]. In addition, Perrier *et al.* [5] showed the relevance of virtual targets in simulating looping tongue movements for intervocalic velar stops. They used virtual targets located around 1 mm beyond the palate.

To the best of our knowledge, the only attempt to estimate the location of virtual targets from real articulatory was made in [6], where authors fit a sequential target approximation model of speech articulatory trajectories to real EMA data of nasal stop consonants. The authors reported context-dependent virtual targets of lingual stop consonants located between 0.5cm and 3cm beyond the palate. These results differ from those obtained from the simulations by Perrier *et al.* [5] who estimated virtual targets much closer to the palate. One possible explanation for this difference has to do with the nature of the targets used in [6], which are asymptotic. Asymptotic targets are, by definition, located beyond the position actually reached by the articulators at the movement end point.

This paper aims to re-evaluate the location of virtual targets using an interpolation function model which considers articulatory targets as non-asymptotic targets. For that purpose, we use general Tau theory [7], whose equations have been shown by Elie *et al.* [8] to provide a better fit to real articulatory data, as compared to several target approximation models, including the model used by Birkholz *et al.* [6]. This paper focuses on the location of (possible) virtual targets for lingual stop consonants of English using the MOCHA-TIMIT corpus [9]. The estimation is done by curve fitting the equations of Tau theory to tongue constriction signals estimated from EMA data from 2 native speakers of English.

The paper is structured as follows. Section 2 introduces the model of speech articulatory trajectory used in this paper, namely general Tau theory [7]. Section 3 describes the method used to extract the tongue constriction variables from EMA data. The fitting method used to estimate the position of virtual targets is described in Section 4. Finally, Section 5 reports the results of our analyses.

2. General Tau theory

General Tau theory [7] is a theory of control of bodily movements which assumes that purposeful movements close a gap between an end effector state and a target state in a specified time. That is, the target is always reached at the specified time. To accomplish this, general Tau theory introduces the concept of the Tau of any movement X, named τ_X, defined by the following differential equation:

$$\tau_X(t) = \frac{X(t)}{\dot{X}(t)} = \frac{k_X}{2}\left(t - \frac{T^2}{t}\right), \tag{1}$$

where $X(t)$ and $\dot{X}(t)$ are the gap closure function and its time derivative, respectively, and T is the duration of the gap closure. The shape parameter k_X is a constant value that adjusts the velocity profile of the gap closure. Coordination of movements is ensured by coupling Taus of movements so that they are proportional. For instance, given two movements X and Y, $\tau_X = k_{X,Y}\tau_Y$. When there is only one movement, it couples to the canonical, internal Tau-guide τ_G defined by $k_G = 1$, hence $\tau_X = k_{X,G}\tau_G = \tau_X$. Modifying the k-value shapes the velocity profile. For instance, Tau-guided movements with $k = 0.4$ have purely symmetrical velocity profiles, and Tau-guided movements with $k = 1$ accelerate according to the laws of gravity.

This theoretical framework has been applied to several kinds of bodily movements [10, 11]. Turk and Shattuck-Hufnagel suggested that it could be applied to speech [12], and a recent study shows that it can fit real articulatory movements more accurately than common dynamic articulatory models [8], such as the critically mass-spring model used in Task-Dynamics [13, 14] or the sequential target approximation

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).

10.21437/Interspeech.2023-953

model [6]. The control parameters of Tau theory are the timings at the onset and at the offset of each movement unit, the onset and offset positions at these time points, and the shape parameter k.

3. Estimating tongue constrictions

Following Task-Dynamics based models [15–17], we assume that articulatory targets are best described in terms of degrees and locations of constrictions in the vocal tract. In this vein, we estimate constriction degree targets by fitting the general Tau theory equation to time-varying constriction degree signals for the tongue tip for alveolar stop consonants, and for the tongue body for velar stop consonants. This section describes the method we used to extract the tongue constriction degree at each point in time from the positions of the tongue sensors.

3.1. Data

In this paper, we used data from the MOCHA-TIMIT corpus [9]. It is a corpus containing synchronized recordings of EMA, speech audio, and laryngograph signals of 2 native speakers of English (one male and one female) uttering 460 sentences from the TIMIT database [18]. One key motive to use this corpus is that it provides phonemic segmentation, which is used to extract the sections corresponding to stop consonants (as explained in Section. 4). For our study, we used only the EMA data and the phonemic segmentation. The MOCHA-TIMIT EMA data were recorded at a sampling rate of 500 Hz, and consist of trajectories in the sagittal plane of sensors glued on the vermilion borders of lower and the upper lips, the lower jaw, on the tip, the middle and the back of the tongue, and on the velum.

3.2. Estimating the tongue constriction signal from EMA

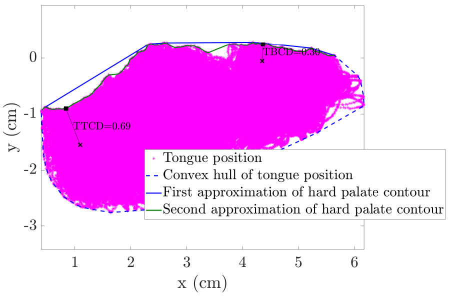

Figure 1: *Estimation of the hard palate contour of the female speaker as the upper bound of the concave hull of the tongue sensors' positions. The values of the tongue constriction variables TTCD and TBCD are taken as the minimal distance from, respectively, the tongue tip and tongue back sensors to the hard palate at each time point. An example of tongue constriction estimation from tongue tip and tongue back positions at a single time point (denoted by cross markers '×') is shown.*

The time-varying tongue constriction variable used in this study was computed separately for the tongue tip and the tongue back sensors, and was defined as the minimal Euclidean distance between each of these tongue sensors and the hard palate. The position of the hard palate was estimated by assuming that the upper limit of tongue sensor positions corresponds to linguopalatal contact. The first step was thus to compute the convex hull of the tongue sensor positions. As shown in Figure 1, these positions do not fully specify the palate contour, as the palate contour may be concave, e.g. around the alveolar ridge. We then computed the concave hull by modifying the convex hull. The first step for this modification consists of oversampling its upper bound, resulting in a line containing a large number L of points (e.g., $L = 10000$ points). Then, for each of these L points, we estimated its nearest neighbor among the tongue sensor positions using a k-NN search with $k = 1$, and then assigned to this point the position of its nearest neighbor. We defined two tongue constriction variables: i) the tongue tip constriction degree (TTCD) for the tongue tip sensor, ii) the tongue body constriction degree (TBCD) for the tongue back sensor. We chose to use the tongue back sensor to compute the tongue body constriction degree because we assume that velar consonants are articulated by the back of the tongue. Figure 1 shows examples of constriction degrees at one time point.

4. Estimating virtual targets

We defined targets of an articulatory movement as the value of a tract variable at the movement endpoint. This definition of articulatory targets differs from that used by target approximation models, such as in [6, 16], where the target is defined as an asymptotic value which is never reached by the articulatory movement. Following our definition, articulatory targets are always reached, unless the target is *virtual*. In the case of virtual targets, an obstacle (e.g. the hard palate) prevents the articulatory movement from reaching it. Consequently, the position of virtual targets cannot be extracted from direct observation of articulatory movements, and needs to be estimated via extrapolation using a dynamic articulatory model. Targets (possibly virtual) are estimated by a method based on that described in [6], which involves fitting the tongue constriction signal segments located just before and right after the linguopalatal consonant constriction, using an appropriate dynamic articulatory model. This is done by estimating the control parameters of the dynamic articulatory model such that the modeled trajectories fit the observed tract variables with the smallest residual norm.

4.1. Data preprocessing

The extracted tongue constriction signals were smoothed using a 25-order LOWESS function in order to remove noise. An order of 25 (corresponding to a window of 50 ms) has been found to be a good trade-off between good noise rejection and conservation of the underlying trajectory. Each movement unit was segmented following a zero-crossing method: a movement unit corresponds to a segment of the constriction variable between two successive extrema. Figure 2 shows an example of the tongue tip and tongue body constriction degree signals for the sentence *"Jane may earn more money by working hard"* uttered by the female speaker. As one would expect, the tongue tip constriction degree TTCD exhibits minima for alveolar stop consonants /d, n/ and the affricate consonant /d͡ʒ/. The tongue body constriction degree TBCD exhibits minima for the velar stop consonants /k/ and /ŋ/.

4.2. Extracting consonantal movements

Using the phonetic segmentation provided with the MOCHA-TIMIT database, we first located the movements corresponding to the relevant stop consonants, namely /t, d, n/ for the tongue tip constriction variable, and /k, ɡ, ŋ/ for the tongue body constriction variable. Using the zero-crossing method for the segmentation into movement units, we then extract two succes-

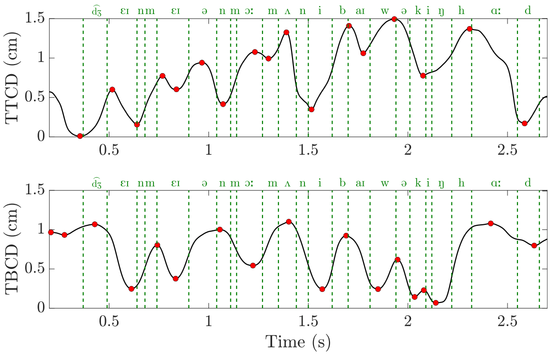

Figure 2: *An example of tongue constriction variables extracted for the sentence* "Jane may earn more money by working hard" *uttered by the female speaker. The top and bottom plots shows the tongue tip and tongue body constriction degree signals, respectively. The phonetic segmentation is shown as vertical dashed lines. The boundaries of the movement units (at velocity zero crossings) are displayed as red circles.*

sive movement units for each stop consonant: i) the descending movement starting before and ending after the start of the stop consonant closure onset and ii) the ascending movement starting before and ending after the stop consonant release onset.

4.3. Curve fitting

Curve fitting consists of estimating the parameters of the dynamic articulatory model which generates a movement which best fits the free segment of the tongue constriction signal surrounding the stop consonant occlusion interval. The free segment corresponds to the segment for which there is no contact between the tongue and palate. The estimation of the parameters is done by minimizing a velocity-weighted normalized root mean square error (NRMSE) between the generated movement and the observed one. We used a similar objective function to the one proposed in [6], namely:

$$C(\theta) = \frac{\sqrt{\left[\sum_n w_n (x_n - \tilde{x}_n(\theta))^2\right] / \sum_n w_n}}{x_{\max} - x_{\min}}, \tag{2}$$

where x_n and $\tilde{x}_n$ are the nth sample of the observed and modeled movement sample of the tongue constriction variable to be fitted, respectively, and $x_{\max}$ and $x_{\min}$ are the maximal and minimal constriction degree values of the analyzed segment. The vector θ contains the model parameters to optimize. Following the method described in [6], we apply local weights w_n based on velocity to penalize discrepancies between the model and the observation in high-velocity regions. We used the same weights as in [6]:

$$w_n = 1 + a\frac{v_n^2}{v_{\max}^2}, \tag{3}$$

where $a = 5$ is a weight coefficient to adjust the relative importance assigned to the fit in the high-velocity region. v_n is the value of the velocity at the nth sample, and $v_{\max}$ is the maximal velocity of the tongue constriction signal in the analyzed segment. Parameters included in θ are estimated using the Nelder-Mead [19] minimization method.

In order to allow the fit to generate trajectories that are not modified by linguopalatal contact, the cost function C is computed only for samples that correspond to the free segment of the tongue constriction signal, namely before the closure onset and after the time of release. We estimate the time of closure onset as the time point for which the descending movement has maximal deceleration. This is under the assumption that the tongue-palate contact creates a sudden deceleration of the tongue movement. Following this assumption, the time of release is the first time point for which the ascending movement for which the constriction degree is larger or equal than that of the descending movement at the time of closure onset.

For Tau theory, 4 parameters need to be estimated: the k-values of the two movements (denoted k_1 and k_2, respectively), the virtual target position X_T, and the virtual target time. Since we are dealing with discrete signals, we assume the timing of the virtual target to be an integer, denoted M. Consequently, curve fitting consists of finding the minimal $C_M(k_1, k_2, X_T)$ for different M within the estimated occlusion interval. Each estimation of C_M is done by running 100 optimizations with random initial estimates and keeping the solution that returns the minimal cost C_M. The k-values k_1 and k_2 are randomly initialized between 0.1 and 0.9 following a uniform distribution, and X_T is chosen randomly between 0 and 1 cm lower than the minimal value of the consonantal constriction degree.

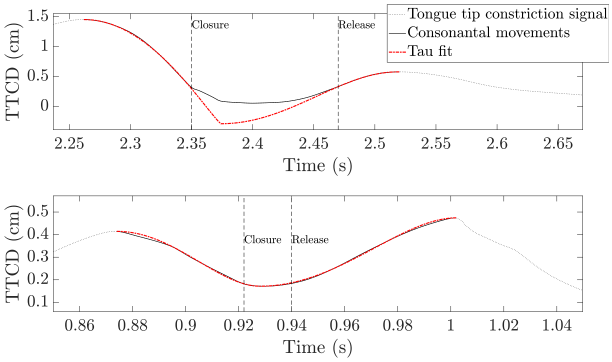

Figure 3: *Example fits for stop consonant movements using general Tau theory (dashed line). The signal is the tongue tip constriction degree signal for a voiceless alveolar consonant /t/ (top plot) and an alveolar nasal consonant /n/ (bottom plot) produced by the female speaker. The estimated closure and release times are denoted by vertical dashed lines.*

Figure 3 shows example fits for stop consonant movements using the Tau equations. The top plot shows a fit that is performed on the tongue tip constriction degree signal for a voiceless alveolar consonant /t/ produced by the female speaker. In this case, the virtual target estimated by Tau theory is located 0.34cm beyond the palate. Figure 3 also shows an example where the estimated target, defined as the constriction value at the movement endpoint, is not located beyond the palate, as it occurs on the observed movement trajectory. Following our definition of virtual targets, this target is not considered virtual.

5. Results

We analyzed the proportion of consonantal movements which used virtual targets. Targets are considered virtual if the movement endpoint returned by the curve fitting is located beyond the minimal constriction degree observed in EMA during the consonantal segment. Table 1 shows the result grouped by speaker and type of consonants. Our results suggest that virtual targets are not ubiquitous, but do occur in the majority (50 to 60%

Table 1: *Number and percentage (in parenthesis) of virtual targets for each consonant and each speaker.*

	Tongue tip				Tongue body			
	Female		Male		Female		Male	
Consonant	# mov.	# virtual (%)	# mov.	# virtual (%)	# mov.	# virtual (%)	# mov.	# virtual(%)
Unvoiced	381	197 (51.7)	387	185 (47.8)	265	168 (63.4)	395	239 (60.5)
Voiced	238	127 (53.4)	245	127 (51.8)	85	53 (62.4)	147	91 (61.9)
Nasal	392	204 (52.0)	478	244 (51.0)	78	41 (52.6)	111	51 (45.9)
All	1011	528 (52.2)	1110	556 (50.1)	428	262 (61.2)	653	381 (58.3)

of cases), confirming intuitions and observations from previous studies [1, 2, 5]. Our results also suggest that velar consonants use virtual targets more often than alveolar consonants. Alveolar consonants use virtual targets for ca. 50% of them, while virtual targets is used for ca. 60% of velar consonants, except for the nasal velar consonant /ŋ/, for which the use of virtual targets occurs less often than other velar consonants. The male speaker used virtual targets in 46% of his produced nasal velar consonant /ŋ/ and the female speaker used virtual targets for 52.5% of her produced /ŋ/. For every type of consonants, the female speaker used virtual targets slightly more often than the male speaker.

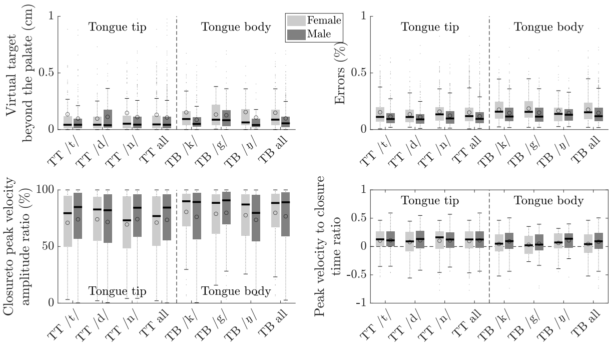

Figure 4: *The location of the estimated virtual targets and fitting errors for different stop consonants and speakers. The top left plot shows the box plots of the estimated locations of virtual targets, expressed in cm beyond the palate. The top right plot shows the box plot of the fit error, expressed in %, corresponding to the optimal value of the cost function C from Eq.* (2). *The bottom left plot shows the box plot of the closure to peak velocity amplitude ratio, expressed in %, and defined as the ratio between the velocity at the closure instant and the peak velocity of the closing movement. The bottom right plot shows the peak velocity to closure time ratio, defined as the time difference between closure time and the time of peak velocity divided by the movement duration. The median value is denoted by a bold horizontal line and the mean value is denoted by a circle.*

Figure 4 shows the estimated location of virtual targets, the fitting error, the closure to peak velocity amplitude ratio, and the peak velocity to closure time ratio. The closure-to-peak amplitude ratio represents the ratio between the velocity at the closure instant and the peak velocity of the modeled Tau-guided movement (possibly not observable when the planned peak velocity occurs after the closure onset). The peak velocity to closure time ratio represents the time difference between closure time and the time of peak velocity divided by the movement duration. The low fitting error, typically less than 0.5%, shows that Tau theory is a good candidate to model speech articulatory trajectories, including tract variables. The estimated locations of virtual targets show a small difference between the alveolar consonants and the velar consonants for the female speaker: most of alveolar virtual targets are located between 0.1 and 0.12mm beyond the palate and it is between 0.4 and 0.16mm for velar consonants. These values for the virtual velar targets are close to the location of the virtual target of velar consonants found by Perrier *et al.* in their simulations [5], which used a virtual target located around 1mm beyond the palate. Within a class of consonants (i.e. alveolar or velar) for a given speaker, the type of consonant (i.e. voiced vs. voiceless vs. nasal) does not have a significant impact on the location of virtual targets, except for the nasal velar consonant /ŋ/ for the female speaker. Additionally, the voiced consonants /d, g/ exhibit more variability in terms of location of their virtual targets than other consonants. There is no such difference between velar and alveolar consonants for the male speaker. The closure to peak velocity amplitude ratio exhibits large variation, but most of the values (75+%) lie between 50 and 100%. This provides support for the hypothesis that virtual targets are used to allow contact to occur at high velocity, as initially proposed in [4]. This is confirmed by the peak-velocity to closure time ratio, which shows that the onset of closure is most often located very close in time to the time of peak velocity. The mean values for the peak velocity to closure time ratio for alveolar consonants is around 0.11 for the male speaker for all type of alveolar consonants, and between 0.08 and 0.11 for the female speaker. For velar consonants, the peak velocity to closure ratio time ratio is more often negative, meaning that the closure onset often occurs before the planned peak velocity. The mean values for the peak velocity to closure time ratio is between 0.03 and 0.06 for the female speaker and 0.05 and 0.1 for the male speaker. For the majority of consonants that use virtual targets, closure onset occurs very close in time to the planned peak velocity.

6. Conclusion

This paper has presented a study about the existence and the location of virtual targets during the production of alveolar and velar consonants. It used Tau theory for extrapolating tongue constriction signals, extracted from EMA recordings of a male and a female speaker, after consonantal closure. Our results suggest that virtual targets are not ubiquitous, as only 50 to 60% of movements use them. These movements use virtual targets which are located very close to the palate. Virtual targets of the tongue tip constriction are located around 0.1 cm beyond the palate, and virtual targets for the tongue body constriction are located between 0.05 and 0.2 cm beyond the palate. The results are in agreement with the simulations by Perrier *et al.* for velar consonants, which used virtual targets located slightly beyond the palate (around 0.1cm). An analysis of the time of closure onset and the planned time of peak velocity (as predicted by the Tau equations) shows that closure occurs at a moment close to the time of peak velocity. This supports the theory that virtual targets are used to enable enough velocity at closure to properly seal the vocal tract [4]. This work provides new evidence for the use of virtual targets and presents a new method to estimate their location. The dynamic articulatory model is able to fit the data very accurately, with errors less than 0.5%. However, the present study has a few limitations, mainly due to the nature of the used data. The sparse nature of EMA data requires an approximation of the hard palate contour, as well as an estimation of the time of closure onset, both of which add uncertainty. One possible way to address these issues would be to apply curve fitting to MRI.

7. References

[1] A. Löfqvist and V. L. Gracco, “Control of oral closure in lingual stop consonant production,” *The Journal of the Acoustical Society of America*, vol. 111, no. 6, pp. 2811–2827, 2002.

[2] J. Brunner, S. Fuchs, and P. Perrier, “Supralaryngeal control in korean velar stops,” *Journal of Phonetics*, vol. 39, no. 2, pp. 178–195, 2011.

[3] S. A. Frisch and S. M. Wodzinski, “Velar–vowel coarticulation in a virtual target model of stop production,” *Journal of phonetics*, vol. 56, pp. 52–65, 2016.

[4] A. Löfqvist and V. L. Gracco, “Lip and jaw kinematics in bilabial stop consonant production,” *Journal of Speech, Language, and Hearing Research*, vol. 40, no. 4, pp. 877–893, 1997.

[5] P. Perrier, Y. Payan, M. Zandipour, and J. 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.

[6] P. Birkholz, B. J. Kröger, and C. Neuschaefer-Rube, “Model-based reproduction of articulatory trajectories for consonant–vowel sequences,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1422–1433, 2010.

[7] D. N. Lee, “Guiding movement by coupling taus,” *Ecological psychology*, vol. 10, no. 3-4, pp. 221–250, 1998.

[8] 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.

[9] 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.

[10] D. Lee, D. Young, P. Reddish, S. Lough, and T. Clayton, “Visual timing in hitting an accelerating ball,” *The Quarterly Journal of Experimental Psychology*, vol. 35, no. 2, pp. 333–346, 1983.

[11] M. W. Rodger, S. O’Modhrain, and C. M. Craig, “Temporal guidance of musicians’ performance movement is an acquired skill,” *Experimental brain research*, vol. 226, no. 2, pp. 221–230, 2013.

[12] A. Turk and S. Shattuck-Hufnagel, *Speech timing: Implications for theories of phonology, speech production, and speech motor control.* Oxford University Press, USA, 2020, vol. 5, ch. How do timing mechanisms work?, pp. 238–263.

[13] E. L. Saltzman and K. G. Munhall, “A dynamical approach to gestural patterning in speech production,” *Ecological psychology*, vol. 1, no. 4, pp. 333–382, 1989.

[14] T. Sorensen and A. Gafos, “The gesture as an autonomous nonlinear dynamical system,” *Ecological Psychology*, vol. 28, no. 4, pp. 188–215, 2016.

[15] C. P. Browman and L. M. Goldstein, “Towards an articulatory phonology,” *Phonology*, vol. 3, pp. 219–252, 1986.

[16] C. P. Browman, L. Goldstein *et al.*, “Dynamics and articulatory phonology,” *Mind as motion: Explorations in the dynamics of cognition*, vol. 175, p. 194, 1995.

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

[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] J. A. Nelder and R. Mead, “A simplex method for function minimization,” *The computer journal*, vol. 7, no. 4, pp. 308–313, 1965.