Simulating gestural undershoot in English diphthongs

Anonymous submission to INTERSPEECH 2023

Abstract

Acoustic reduction is widely attested in speech, but articulatory reduction is not as well understood. The aim of this study is to examine how reduction takes place in articulation, taking the example of the English diphthong [AY].

We identify four articulatory mechanisms that could result in similar reductions in acoustic duration: gestural overlap, undershoot, shortening, and increase in stiffness. We use Task Dynamics to simulate articulatory trajectories exhibiting these mechanisms, examined their correlation with acoustic duration, and compared them to recorded data using a variant of Dynamic Time Warping.

Results indicate that acoustic duration is correlated with shorter gestures, but the best-fit simulations did not show consistent effects of stiffness or undershoot. We interpret this as evidence that details of acoustic duration can result from adjusting gestural duration, and that other parameters are more related to trajectory shape than overall duration.

Index Terms: articulation, gestures, simulation, diphthongs

1. Introduction

Phonetic reduction is a common feature of speech, but the articulatory mechanisms involved are not well understood. Reduction in acoustic duration is correlated with several potential changes in articulation. The goal of this paper is to identify which properties of articulatory gestures are associated with differences in acoustic duration, focusing on the American English diphthong /a͡ɪ/. We demonstrate the use of an analysis-by-synthesis approach to studying articulatory reduction.

1.1. Reduction in diphthong articulation

Previous research on the articulation of diphthongs has found evidence for the temporal coordination in diphthongs, but has not used simulation to test these predictions. [1] tracked the pathways of reduction of Spanish hiatus sequences to dipthongs to monophthongs, but the particular articulatory mechanisms by which these reductions take place are not clear. Evidence for a role of gestural overlap comes from [2] and [3], who demonstrated that Romanian diphthongs crucially differ from glide-vowel sequences in gestural timing, specifically in having more gestural overlap. In a typological study of Romance languages, [4] argue that the temporal stability of diphthongs as compared to hiatus sequences plays a role in diachronic change, and [5] showed that ongoing sound change can involve changes in the timing of a diphthong's inflection point.

In addition to changes in gestural timing, it has also been shown that diphthongs can simply vary in the magnitude of articulatory movements. This was quantified for the /a͡ɪ/ diphthongs of English and German by [6] and [7] in terms of the path distance traveled lingual articulators.

On the basis of this research, we predict that reduction in the English diphthong /a͡ɪ/ should primarily take place in terms of articulatory overlap and reduction in the magnitude of articulatory movements.

1.2. Articulatory analysis-by-synthesis

We frame our approach in interrelated theories of Articulatory Phonology ([8],[9], *et seq.*) and Task Dynamics ([10]), henceforth AP/TD. These theories allow for quantitative phonetic predictions to be made on the basis of a defined set of phonological parameters, specifically with the implementation in the Task Dynamics Application (TADA) [11]. Analysis-by-synthesis was conducted using TADA by [11], who adjusted the start and end times of gestures in order to simulate a phonetic target. Unlike [11], we define four mechanisms by which we predict articulatory reduction could possibly take place; these were selected based on the parameters available in TADA and AP/TD. The first,

The first two types of reduction derive from the definition of an articulatory gesture. According to [12], a gesture is defined by three phonological parameters: the constriction location, constriction degree, and duration of the gesture. The first form of reduction we identify is *undershoot*, the failure to reach a target in either location or degree. The second type of reduction, gestural **shortening**, involves adjusting the gestural duration. The third is an increase in the *overlap* of two gestures, which for the /a͡ɪ/ could result from either shifting the [a] gestures later or the [ɪ] gestures earlier. Finally, the fourth type of reduction derives not from gestural representations or their coordination, but from changes in the *stiffness* parameter in Task Dynamics, which reflects resistance to movement. Reduced stiffness would result in gestures that more rapidly change velocity. In TADA, stiffness is held at one constant value for consonants, and different value for vowels ([13]), but as with the other three parameters, we adjust stiffness as a possible source of variation.

In the following sections, we discuss how we adjusted these four parameters in order to test the hypothesis that *undershoot* and *overlap* would be more closely correlated with reduction in duration than *shortening* or *stiffness*. We simulate a range of tokens of /a͡ɪ/ incorporating adjustments in the four parameters, and systematically compare them to data from the X-Ray Microbeam Database (XRMB) [14].

2. Method

The procedure used in this study is illustrated in Fig. 1. Details are presented below.

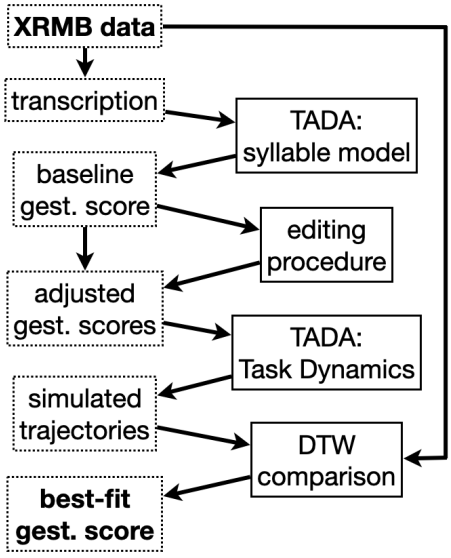


Figure 1: *Simulation and comparison procedure*

2.1. XRMB data

Data for this study comes from the X-Ray Microbeam Database (XRMB) [14], which includes articulatory data from speakers of American English in a range of tasks. Diphthongs allow the study of all four parameters, including overlap. Specifically, we focused on productions of the word <five>, which includes the diphthong /a͡i/ surrounded by only labial consonants, and followed only by a voiced consonant to avoid diphthong raising. A gestural score for <five> is shown in Fig. 2.

In total, we considered 425 tokens from 48 speakers, produced in a variety of conditions: 49 in counting, 253 in number sequences, and 123 produced in the course of reading paragraphs of text. [1]

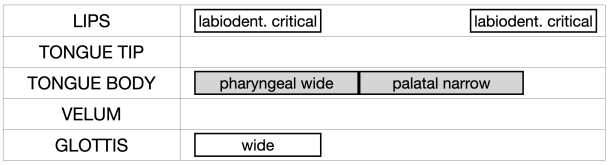


Figure 2: *Gestural score of <five>. Gestures being modified are shaded*

2.2. Simulation procedure

The Task Dynamics Application (TADA) [13] is an implementation of the Task Dynamics model [10]. Given a specification of an utterance as a gestural score as input, it can simulate the trajectories of the vocal tract organs. We created a Python script which successively modifies gestural score files according to a specified set of reduction rules. The "reduced" versions of these utterances produced by our script are then used as input for the TADA trajectory simulation algorithm.

[1]Extracted and synthesized data is available at: `https://osf.io/dr4tk/?view_only=6034ebf4eaf3471c86658728b32919eb`

In TADA's gestural score files, the target for a tongue body constriction gesture is defined using a pair of variables: the constriction location (TBCL, measured in degrees, where 90° is completely palatal and 180° is completely pharyngeal), and constriction degree (TBCD, measured in millimetres from the passive articulator). To simulate the effects of undershoot, we reduce these variables accordingly.

Similarly, the gestural file specifies a value for a damping parameter (denoted β) corresponding to the stiffness of the articulator in the critically-damped mass spring equation. We alter β to increase or reduce stiffness, resulting in a faster approach to the gesture target and faster return to neutral position after the gesture ends. Normally, TADA holds β constant at 6 for consonant gestures, and 3 for vowels, and its various complex effects on speech production are still not well-understood [15]. For this reason, we also produced a version without modifying the stiffness parameter, which produces broadly the same results below.

To shorten a gesture, we move its end point as specified in the gestural score earlier in time. All gestures coupled to the gesture in question will also have their end points moved earlier in time. Increased overlap is modelled analogously, with the end of one gesture moving later in time, and the start of an adjacent gesture moving earlier in time.

As a preliminary study, we simulated <five> with two versions of each gesture: one with the baseline TADA settings and one with the above-described modifications. All possible combinations were simulated. Table 1 summarizes the modifications used in the simulations.

Parameter	Baseline	Edit (steps)
[a] Undershoot	0mm	0-5 (5mm)
[ɪ] Undershoot	0mm	0-5 (5mm)
[a] Overlap	0ms	0-1 (1ms)
[ɪ] Overlap	0ms	0-1 (1ms)
[a] Shortening	0ms	0-1 (1ms)
[ɪ] Shortening	0ms	0-1 (1ms)
[a] Stiffness	4	6 (2)
[ɪ] Stiffness	8	12 (4)

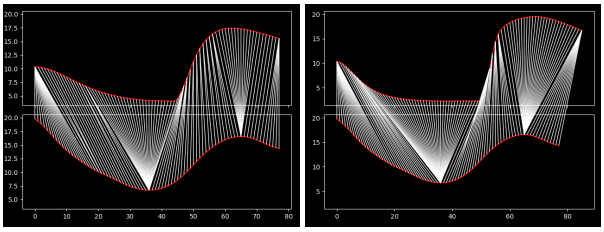
Table 1: *Adjustment of parameters in simulation.*

3. Data analysis

Analysis of the resulting trajectories was performed using Dynamic Time Warping (DTW). DTW is a commonly-used measure of similarity between two sets of time series data, originally designed for use in speech recognition applications [16]. The DTW algorithm finds an alignment between the two time series such that the Euclidean distance is minimized, and allows for arbitrary insertion of repeated points to ensure meaningful comparison of series of differing lengths. Efficient algorithms using dynamic programming have been developed, and the *DTAIDistance* Python package was used here [17].

While DTW is a useful tool for identifying similarity of non-linear time-series data, we found that it does not adequately penalize articulatory trajectories of dissimilar length. In order to identify articulatory trajectories of similar duration, we applied an additional penalty of 100 to the score for each point inserted by the DTW algorithm to penalize simulated results with improbably divergent lengths from the observed values.

An example of the comparison procedure is demonstrated in Fig. 3.



(a) *A particularly close match between simulated and real trajectories of the tongue dorsum*

(b) *A worse match between simulated and real trajectories of the tongue dorsum*

Figure 3: *Example of Dynamic Time Warping comparison. Top images are two simulations; bottom images are the same recorded XRMB trajectory. The better match in 3a was reduced in every dimension except for stiffness (of [a] and [ɪ]) and shortening of [ɪ], while 3b was reduced only in undershoot of [ɪ]*

4. Results

4.1. Best-fit simulations

For each real token in the XRMB corpus, we found the simulated utterance which minimised the DTW distance. Table 2 shows the total number of times each form of reduction (shortening, overlap, undershoot or stiffness) was used in a simulation which produced the closest match to reality, and to which gesture the reduction applied (tongue body constriction degree of [a], and so on.)

	[a] TBCD	[ɪ] TBCD	[a] TBCL	[ɪ] TBCL	sum
sh	352	392	249	388	1381
ov	384	383	369	391	1527
un	211	27	263	19	520
st	352	5	249	57	663
sum	1299	807	1130	855	

Table 2: *Total uses of reduction dimension by articulatory parameter out of the best-fit simulations for the 425 tokens.*

Table 2 indicates that shortening and increased overlap are the most commonly utilized strategies for reduction of gestures.

4.2. Regression analysis

A set of linear mixed-effects models were constructed using the lme4 package in R [18] to investigate the relative contributions of each hypothesized dimension of reduction. Random intercepts were added for each speaker, as well as for each task (the specific utterance being read). Previous mention within each utterance and task type were also included. Possible values for task type were continuous narrative, a random sequence of numbers, or counting up in order.

We constructed a linear mixed-effects model with random effects of speaker and task, and fixed effects of previous mention and task type. The model also included a term for each of the sixteen dimensions of variation appearing in Table 2, each of which had two values (reduced and unreduced).

However, the singular model fit, combined with relatively small dataset and binary choices for each value means that we must interpret these values only with great care. To investigate the effect of different factors, we ran a top-down nested model comparison, constructing a series of models from which we removed the variables corresponding to one of the articulatory dimensions, so that one model was missing the shortening variables, one the overlap variables, and so on. We then performed ANOVA analyses on each of these reduced models to the base model containing every variable. Of these, the base model showed a significant improvement only over the model missing the shortening variables (i.e. the four dimensions of shortening for TBCL and TBCD of [a] and [ɪ]), which supports the interpretation that this is the most important dimension of articulatory reduction in this dataset.

4.3. Correlation analysis

We ran a correlation analysis to determine which phonetic characteristics of the best-fit simulations most closely match the acoustic durations in the corresponding XRMB data. The correlation matrix, created in R [19] using the *corrplot* package [20], is presented in Fig. 5.

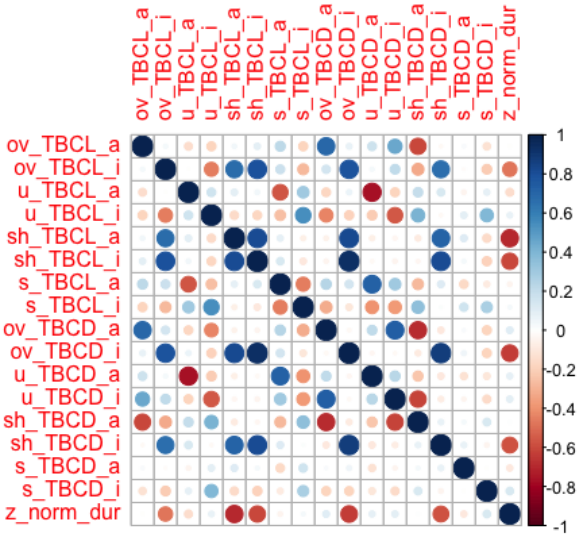


Figure 4: *Correlation plot of best-fit simulated articulatory dimensions and acoustic duration from XRMB. "ov" = overlap, "u" = undershoot, "sh" = shortening, "s" = stiffness; "TBCL" = Tongue Body Constriction Location" and "TBCD" = Tongue Body Constriction Degree" gestures in TADA; "a" and "i" refer to the two targets within the diphthong.*

Of the dimensions tested, the most robust correlations with acoustic duration were from gestural shortening: all shortening dimensions except for the TBCL shortening of [a] were significantly correlated with shorter acoustic duration. The only other significant correlations ($p < .05$) with duration were TBCL and TBCD overlap of [ɪ], indicating an earlier phasing of [ɪ] gestures, TBCL undershoot in [a], TBCD undershoot of [ɪ], and TBCD overlap of [a] (i.e. later phasing of this gesture), and TBCD stiffness in [ɪ]. Given the inconsistent patterning of the latter several effects as well as the very small absolute value of the correlations, we summarize the correlation with duration as

follows: shorter acoustic duration was associated with shortening of [a] and [ɪ] gestures and with overlap (earlier phasing) of [ɪ] gestures.

We also carried out a hierarchical cluster analysis, results of which are shown in Fig. 5. This chart confirms that ov_TBCD_a and sh_TBCD_a are closely associated with each other, suggesting that the interaction of these terms leads to their apparently anomalous behavior. In Fig. 5, it can be seen that shortening of TBCD of [a] was the only shortening dimension with no significant correlation with duration, and TBCD overlap of [a] is the only significant form of reduction correlated with longer acoustic duration. Furthermore, the chart also highlights the close correlation of overlap and shortening in [ɪ] gestures.

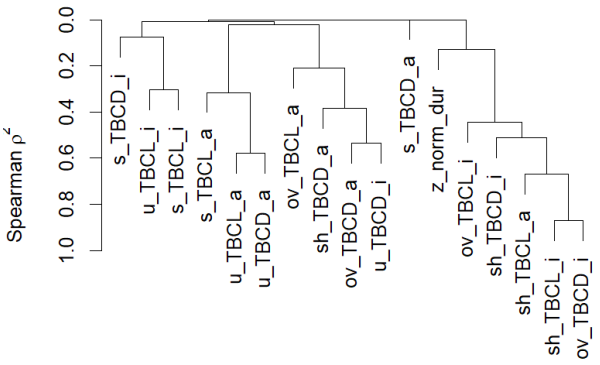


Figure 5: *Hierarchical clustering analysis.* ρ^2 *denotes the square of the Spearman correlation.*

5. Conclusion and future directions

To summarize, our simulations showed that overall gestural duration was correlated with shorter gestures for both [a] and [ɪ] components of the diphthong. Overlap resulting from earlier timing of [ɪ] was also correlated with duration, but was highly correlated with shortening of [ɪ]. The importance of gestural shortening was supported by our modified DTW analysis, which showed that best-matching simulations included not only shortened vowel gestures but also overlap and undershoot. Stiffness neither emerged in the DTW analysis nor was systematically correlated with simulated duration.

While we find the results of the correlation analysis to be suggestive of meaningful effects, we must be cautious in interpreting the correlation results. Even strong correlations only indicate general trends in the relationship between articulation and acoustic duration. That is, an articulatory dimension not correlated with acoustic duration overall may still be important for a small number of tokens. It is entirely possible that the high-level correlations are obscuring substantial variation across speakers, contexts, and tokens. In future work, we intend to more closely investigate this variation by looking more closely at the dimensions along which reduction takes place in each best-fit token (summarized in Table 2).

Nevertheless, the overall results are surprising in light of previous research on the articulation of diphthongs. We had predicted that overlap and undershoot would be the main forms of reduction observed, but instead found a primary role of gestural shortening. We interpret this as evidence that overlap and undershoot are involved less in simple reduction in duration and more in producing the shape of articulatory and/or acoustic trajectories.

The present study demonstrates the value of considering several gesture-based ways to measure phonetic variation. Previous work measuring diphthongs based on articulatory distance was less able to determine which gestures were involved in phonetic variation. By contrast, we found evidence for overlap resulting from shifting [ɪ] earlier rather than shifting [a] later. Moreover, we were able to identify articulatory differences more finely than was possible using the method of [11]. Specifically, we could distinguish shortening from overlap, and also consider undershoot and stiffness.

The lack of observed effect for stiffness is noteworthy. We motivated our choice of stiffness as a parameter able to account for differences in gestural velocity; however, at least one systematic correlate of velocity, gestural movement magnitude, was not systematically varied in this study. Still, a uniform stiffness is consistent with the general approach of Task Dynamics which holds this parameter constant (with one value for consonants and another for vowels) [13]. This finding is also interesting in light of [21], whose approach to articulatory simulation involves an analogous constant that is assumed to vary across gestural targets.

The most robust result from this study is that gestural shortening plays a role in inter-token variation. This is notable for how it differs from the AP/TD approach of emphasizing overlap and relative timing, and in which gesture durations are, by hypothesis, phonologically specified. It is more consistent with approaches that treat phonetic timing independent of phonology, such as that of [22]. Intriguingly, shortening of [ɪ] was found to be closely correlated with overlap resulting from adjusting timing of [ɪ]. This suggests that it is the *endpoint*, not the *beginning* the [ɪ] that is temporally coordinated with [a]. Endpoint-based timing is also consistent with [22] and not with classical AP/TD Though preliminary, we take these results as an intriguing suggestion for future research.

In the future, we intend to expand this research not only to more contexts, but also to a broader range of simulations. As shown in Table 1, the simulations in this study were limited to two steps for each parameter. Though the many combinations of even just this results in a large and time-consuming computational process, we recognize that more steps for each parameter are needed to better understand the range of variation in articulatory data.

6. References

[1] L. Aguilar, "Hiatus and diphthong: Acoustic cues and speech situation differences," *Speech Communication*, vol. 28, no. 1, pp. 57–74, May 1999. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0167639399000035

[2] S. Marin and L. Goldstein, "A gestural model of the temporal organization of vowel clusters in Romanian," in *Consonant Clusters and Structural Complexity*, P. Hoole, L. Bombien, M. Pouplier, C. Mooshammer, and B. Kühnert, Eds. Mouton de Gruyter, Sep. 2012, pp. 177–204. [Online]. Available: https://www.degruyter.com/document/doi/10.1515/9781614510772.177/html

[3] S. Marin, "Romanian diphthongs /ea/ and /oa/: an articulatory comparison with /ja/ - /wa/ and with hiatus sequences," *Revista de Filología Románica*, vol. 31, no. 1, pp. 83–97, 2014. [Online]. Available: http://revistas.ucm.es/index.php/RFRM/article/view/51024

[4] I. Chitoran and J. I. Hualde, "From hiatus to diphthong: the evolution of vowel sequences in Romance," *Phonology*, vol. 24, no. 1, pp. 37–75, May 2007. [Online].

Available: https://www.cambridge.org/core/product/identifier/S095267570700111X/type/journal_article

[5] M. Sóskuthy, J. Hay, and J. Brand, "Horizontal diphthong shift in New Zealand English," in *Proceedings of the 19th International Congress of Phonetic Sciences*, S. Calhoun, P. Escudero, M. Tabain, and P. Warren, Eds. Canberra, Australia: Australasian Speech Science and Technology Association Inc., 2019, pp. 597–601.

[6] A. P. Simpson, "Gender-specific articulatory–acoustic relations in vowel sequences," *Journal of Phonetics*, vol. 30, no. 3, pp. 417–435, Jul. 2002. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0095447002901713

[7] M. Weirich and A. P. Simpson, "Individual differences in acoustic and articulatory undershoot in a German diphthong – Variation between male and female speakers," *Journal of Phonetics*, vol. 71, pp. 35–50, Nov. 2018. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0095447017300827

[8] C. P. Browman and L. M. Goldstein, "Towards an articulatory phonology," *Phonology*, vol. 3, pp. 219–252, May 1986. [Online]. Available: https://www.cambridge.org/core/product/identifier/S0952675700000658/type/journal_article

[9] C. P. Browman and L. Goldstein, "Some Notes on Syllable Structure in Articulatory Phonology," *Phonetica*, vol. 45, no. 2-4, pp. 140–155, Mar. 1988. [Online]. Available: https://www.degruyter.com/document/doi/10.1159/000261823/html

[10] E. Saltzman and J. A. Kelso, "Skilled actions: A task-dynamic approach." *Psychological Review*, vol. 94, no. 1, pp. 84–106, 1987. [Online]. Available: http://doi.apa.org/getdoi.cfm?doi=10.1037/0033-295X.94.1.84

[11] H. Nam, V. Mitra, M. Tiede, E. Saltzman, L. Goldstein, C. Espy-Wilson, and M. Hasegawa-Johnson, "A procedure for estimating gestural scores from natural speech," in *Interspeech 2010*. ISCA, Sep. 2010, pp. 30–33. [Online]. Available: https://www.isca-speech.org/archive/interspeech_2010/nam10_interspeech.html

[12] C. P. Browman and L. Goldstein, "Articulatory Phonology: An Overview," *Phonetica*, vol. 49, no. 3-4, pp. 155–180, May 1992. [Online]. Available: https://www.degruyter.com/document/doi/10.1159/000261913/html

[13] H. Nam, L. Goldstein, E. Saltzman, and D. Byrd, "TADA: An enhanced, portable Task Dynamics model in MATLAB," *The Journal of the Acoustical Society of America*, vol. 115, no. 5, pp. 2430–2430, May 2004. [Online]. Available: http://asa.scitation.org/doi/10.1121/1.4781490

[14] J. R. Westbury, G. Turner, and J. Dembowski, *X-ray microbeam speech production database user's handbook, version 1.0*. University of Wisconsin Waisman Center, 1994.

[15] K. D. Roon, A. I. Gafos, P. Hoole, and C. Zeroual, "Influence of articulator and manner on stiffness," in *Proceedings of the 16th International Congress of Phonetic Sciences*, Saarland University, 2007, pp. 409–412.

[16] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 1, pp. 43–49, 1978.

[17] W. Meert, K. Hendrickx, T. Van Craenendonck, P. Robberechts, H. Blockeel, and J. Davis, "DTAIDistance," Aug. 2020. [Online]. Available: https://github.com/wannesm/dtaidistance

[18] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[19] R Core Team, *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2022. [Online]. Available: https://www.R-project.org/

[20] T. Wei and V. Simko, *R package 'corrplot': Visualization of a Correlation Matrix*, 2021. [Online]. Available: https://github.com/taiyun/corrplot

[21] P. Birkholz, B. J. Kroger, 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, Jul. 2011. [Online]. Available: http://ieeexplore.ieee.org/document/5634084/

[22] A. Turk and S. Shattuck-Hufnagel, *Speech Timing: Implications for Theories of Phonology, Phonetics, and Speech Motor Control*, 1st ed. Oxford University Press, Feb. 2020. [Online]. Available: https://academic.oup.com/book/36950