Uncover articulatory correlates of acoustic duration with analysis-by-synthesis: the case of diphthongs

Eoin O'Reilly[1], Christopher Geissler[1] and Kevin Tang[1,2])
[1]*Heinrich Heine University Düsseldorf,* [2]*University of Florida*

Acoustic reduction is widely attested in speech, but how this takes place in articulation is not as well known. The aim of this study is to examine how reduction takes place in articulation, taking the example of the English PRICE /a͡ɪ/ diphthong. We identify four articulatory mechanisms that could potentially result in similar reductions in acoustic duration: increased gestural **overlap**, **undershoot**, **shortening** of gestures, and increase in **stiffness** (resulting in faster movement).

Previous research has found evidence suggesting the nature of temporal coordination in diphthongs, but has not directly tested the predictions of such coordination patterns using simulation. Acoustic reduction of Spanish diphthongs across task conditions was studied by [1], who constructed a continuum of reduction as hiatus→diphthong→monophthong, but the articulatory manifestation of this process is not clear. Differences in articulation of the /a͡ɪ/ diphthong (in English and German) have been studied in terms of the Euclidian distance travelled between the targets [2][3]. Gestural timing has been identified as a key difference between diphthongs and hiatus sequences in Romanian [4][5]. Changes in the timing of gestures play an important role in diachronic sound change, as has been shown in Romance [6] and English, including the PRICE vowel which is the focus of this study [7]. In sum, previous work suggests that diphthong reduction could involve changes in the timing of gestures as well as their targets.

The simulation procedure used the Task Dynamics Application (TADA) [8], with an analysis-by-synthesis approach similar to [9]. Unlike [9], the present study focuses solely on articulation rather than matching acoustics, and adds undershoot and changes to stiffness. After initially generating a gestural score based on the Coupled Oscillator Model of Syllable Structure [10], TADA uses this gestural score as input and simulates the trajectories of the vocal tract organs. We successively modify 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.

This procedure was used to generate 65,536 variations (combinations of 16 parameters each with two values) of the English word *five*, which were compared with 425 examples from 48 speakers in the X-Ray Microbeam Database [11]. Analysis was performed using Dynamic Time Warping (DTW), [12], as implemented by the DTAIDistance Python package [13], with an additional penalty for durations that differ substantially from the XRMB data. Examples of good and bad fits between simulated and actual gestures are shown in Figures 1 and 2. To understand the reduction strategies, regression analyses were used to predict acoustic duration with the best fit articulatory parameters, controlling for speaker, text, task type and giveness.

Results show that the best-fit simulations tended to use gestural shortening and overlap of both [a] and [ɪ] components of the diphthong, but all four strategies were observed. However, a correlation analysis between each dimension of reduction and acoustic duration showed the strongest correlation with gestural shortening. Overlap resulting from earlier phasing of [ɪ] was also correlated with duration, but was even more strongly correlated with shortening of [ɪ]. We interpret this as evidence that gestural shortening is the most important articulatory correlate of overall acoustic duration, but that the other forms of reduction contribute to the shape of an articulatory trajectory, and likely to other aspects of acoustic detail.

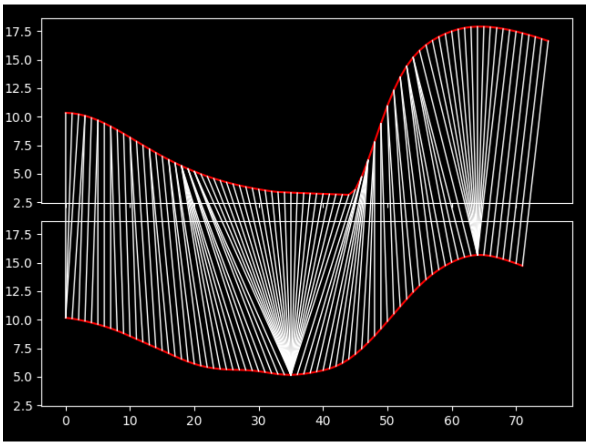

Figure 1. *A good match between real and simulated tongue dorsum trajectories* (*the lower half is the real utterance).*

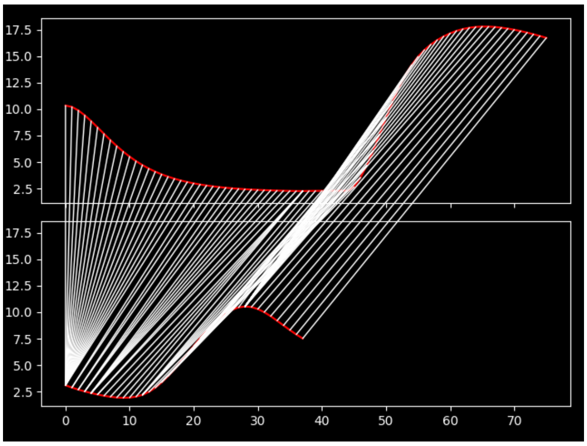

Figure 2. *A poor match between real and simulated trajectories.*

References

[1] Aguilar, L. 1999. Hiatus and diphthong: Acoustic cues and speech situation dierences. *Speech Communication* 18.

[2] Simpson, A. 2002. Gender-specific articulatory–acoustic relations in vowel sequences. *Journal of Phonetics* 30(3). 417–435.

[3] Weirich, M. & Simpson, A. 2018. Individual differences in acoustic and articulatory undershoot in a German diphthong – Variation between male and female speakers. *Journal of Phonetics* 71. 35–50.

[4] Marin, S. & Goldstein, L. 2012. A gestural model of the temporal organization of vowel clusters in Romanian. In Philip Hoole, Lasse Bombien, Marianne Pouplier, Christine Mooshammer & Barbara Kühnert (eds.), *Consonant Clusters and Structural Complexity*, 177–204. De Gruyter.

[5] Marin, S. 2014. Romanian diphthongs /ea/ y /oa/: an articulatory comparison with /ja/-/wa/ and with hiatus sequences. *Revista de Filología Románica* 31(1). 83–97.

[6] Chitoran, I. & Hualde, J.I. 2007. From hiatus to diphthong: the evolution of vowel sequences in Romance. *Phonology*. Cambridge University Press 24(1). 37–75.

[7] Sóskuthy, M., Hay, J. & Brand, J. 2019. Horizontal diphthong shift in New Zealand English. In *Proceedings of the 19th International Congress of Phonetic Sciences*.

[8] Nam, H., Goldstein, L., Saltzman, E. & Byrd, D. 2004. TADA: An enhanced, portable Task Dynamics model in MATLAB. *The Journal of the Acoustical Society of America*. Acoustical Society of America 115(5). 2430–2430.

[9] Nam, H., Mitra, V., Tiede, M., Saltzman, E., Goldstein, L., Espy-Wilson, C. & Hasegawa-Johnson, M.. 2010. A Procedure for Estimating Gestural Scores from Natural Speech. *Interspeech 2010*.

[10] Nam, H. & Saltzman, E. 2003. A competitive, coupled oscillator model of syllable structure. In *Proceedings of the 15th International Congress of the Phonetic Sciences*.

[11] Westbury, J.R., Turner, G. & Dembowski, J. 1994. X-ray microbeam speech production database user's handbook. *University of Wisconsin*.

[12] Sakoe, H. & Chiba, S. 1978. Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing*. IEEE 26(1). 43–49.

[13] Meert, W., Hendrickx, K., Van Craenendonck, T, Robberechts, P. 2020. wannesm/dtaidistance v2.0.0. Zenodo.