COMPUTATIONAL MODELS FOR ARTICULATORY LEARNING OF ENGLISH DIPHTHONGS: ONE DYNAMIC TARGET VS. TWO STATIC TARGETS

Anqi Xu[1], Branislav Gerazov[2], Daniel van Niekerk[3], Paul Konstantin Krug[4], Santitham Prom-on[5], Peter Birkholz[4], Yi Xu[3]
[1]School of Humanities and Social Sciences, Harbin Institute of Technology, Shenzhen, China
[2]Faculty of Electrical Engineering and Information Technologies, UCMS, Skopje, RN Macedonia
[3]Department of Speech Hearing and Phonetic Sciences, University College London, UK
[4]Institute of Acoustics and Speech Communication, Technische Universität Dresden, Germany
[5]Computer Engineering Department, King Mongkut's University of Technology Thonburi, Thailand
a.xu.17@ucl.ac.uk

ABSTRACT

The nature of English diphthongs has been much disputed. By now, the most influential account argues that diphthongs are phoneme entities rather than vowel combinations. However, mixed results have been reported regarding whether the rate of formant transition is the most reliable attribute in the perception and production of diphthongs. Here, we used computational simulation of vocal learning to explore the nature of diphthongs. We tested whether diphthongs have a single dynamic target or two static targets by training an articulatory synthesiser with a three-dimensional (3D) vocal tract model to learn English words. An automatic phoneme recogniser was constructed to guide the learning of the diphthongs. Listening results by native listeners showed that diphthongs learned with dynamic targets were more intelligible and generalisable across variable duration than those learned with static targets. The learning-oriented modelling approach also paves a new way for validating hypotheses of speech perception and production.

Keywords: diphthongs, computational modelling, articulatory synthesis, vocal learning, American English

1. INTRODUCTION

Diphthongs, a special class of vowels, are characterised by transitional formant movements along a path between spectral spaces belonging to two different vowels [1], [2]. Early accounts have treated diphthongs as combinations of two vowels, or sometimes as vowel-semivowel sequences [3]. However, empirical evidence from a comprehensive study by Gay [4] suggests that English diphthongs are more likely to be distinct phonetic units, based on the observation that listeners were more sensitive to the second formant (F2) movement of the synthetic diphthongs than the formant onset and offset. These results are consistent with more recent findings that the most salient perceptual cue of synthetic diphthongs in noise or reverberation is the intensity of F2 transitions [5]. On the other hand, some studies suggest that the crucial cue in the identification of manipulated diphthongs is the endpoint rather than the transitional trajectories [6]. Another line of studies sought to use a classifier to investigate perceptual cues of diphthongs in a speech corpus, and found that instead of F1-F2 onsets and slopes, classification accuracy was the highest when both F1-F2 onsets and offsets were included [7]. A similar approach was adopted in [8], which reported that incorporating F1–F3 onset, offset and transition rates led to the best classification results.

Not only does the debate about the auditorily relevant formant cues of diphthongs continue, contradictory observations have been made regarding the production of diphthongs. Gay [9] investigated the acoustic properties of five American English diphthongs spoken in three different speech rates from slow to fast. The beginning and terminating vowel formants as well as the rate of F2 movement remained the same across different speaking rates. Further, the final portion of the vowel could be eliminated in fast speech. The unfluctuating formant slopes also accords with more recent acoustic evidence from careful and conversational speech [10], as well as loud speech [11]. As far as articulation is concerned, the tongue body exhibits invariant velocity during the production of diphthongs [12]. Recent X-ray data also show that the tongue flesh points undergo minimal changes in different speaking rates [10]. More importantly, the tongue movements and formant transitions of diphthongs are highly correlated, despite some exceptions [13]. In contrast, some researchers found that spectral changes of diphthongs were lowered in clear speech with prosodic prominence [14].

If gliding movements rather than onsets and offsets are a most reliable feature of diphthongs, then diphthongs can be considered distinct phonemes. However, to date no evidence from either perception or production studies has been conclusive. Here, we used computational simulation of vocal learning to explore whether diphthongs have either unitary dynamic targets or consecutive static targets. We trained an articulatory synthesiser with a 3D vocal tract model to learn American English words containing diphthongs, following the approach in [15]–[17]. The learning is guided by a phoneme recogniser, pre-trained by a deep learning model to encode a speaker-normalised perceptual space. After the training, we synthesised speech using the learned articulatory targets with varying duration to further verify their generalisability over different speaking rates. The performance of the two types of articulatory targets is evaluated by the intelligibility of the learned speech in a listening experiment and the plausibility of the learned articulatory kinematics.

2. METHOD

2.1. Speech material

Five diphthongs, /aɪ, eɪ, əʊ, aʊ, ɔɪ/, were embedded in English words with bilabial stops, as follows: ‘buy’, ‘bay’, ‘boy’, ‘bow (and arrows)’ and ‘(to) bow’. Because the two target words ‘bow’ are homographs, we added hints to distinguish the two words, as indicated in brackets. The use of these minimal pairs is to ensure that perception experiments can be carried out naturally by native speakers. The same hints were also given to the participants during the listening experiment.

2.2. Model overview

We trained a vocal tract model to learn the speech material as illustrated in Fig. 1. The learning model consists of a production and a perception system. The model begins with exploration of a set of articulatory targets (Fig. 1A). The kinematic trajectories that approach the articulatory targets are based on the assumption of static or dynamic targets (Fig. 1B). The time-varying vocal tract shapes are then converted to cross-sectional area functions for acoustic simulation (Fig. 1C). The model explores the articulatory parameters iteratively, guided by the perception system. (Fig. 1D).

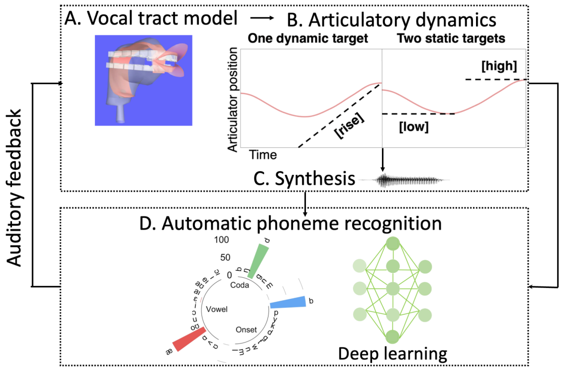

Figure 1: Overview of the learning process.

2.3. Vocal tract model

The articulatory synthesiser used in the study is VocalTractLab 2.3 (www.vocaltractlab.de), with a geometrical 3D vocal tract model (Fig. 1A). The vocal tract model was adapted from MRI data of a German male speaker, involving sixteen free vocal tract parameters (Table 1).

Parameter	Description
HX, HY	Horiz. and vert. hyoid positions
JX, JA	Horiz. jaw position and jaw angle
LP, LD	Lip protrusion and vert. lip distance
TTX, TTY	Horiz. and vert. tongue tip positions
TBX, TBY	Horiz. and vert. tongue blade positions
TCX, TCY	Horiz. and vert. tongue body centre positions

Table 1: Free vocal tract parameters in the simulation.

2.4. Articulatory targets

The temporal and spatial movements of the articulators were simulated by a coarticulation model, synchronised dimension-specific sequential target approximation model [15]–[17], which generates dynamic trajectories of vocal tract parameters. Quantitatively, each articulatory target is represented by height (i.e., positions of the articulators), slope and strength. As illustrated in Fig. 1B, the same asymptotic articulatory curves of diphthongs can be the result of two static targets or one dynamic target. The two static targets have a slope of zero but the target height and the duration ratio need to be optimised. In contrast, the dynamic target requires the optimisation of both height and slope.

2.5. Automatic phoneme recogniser

We trained a deep learning-based phoneme recognition system to guide the optimisation process. We extracted speech with 11 onset consonants, 17 vowels (stressed and unstressed) and 6 coda consonants from the LibriSpeech corpus [23]. The training, validation and test set contains 116.7, 14.4 and 15 hours of speech, respectively. We applied pre-emphasis (coefficient = 0.97) and calculated the log Mel spectrogram (25 ms Hamming window, 5ms overlap) with 26 Mel filters. The log Mel spectrograms were used as the input for the training with a length of 200 frames (spanning 1 s). The model was trained to learn a mapping from the Log Mel spectrograms to a 34-dimensional vector one-hot encoding the phonemes listed in Fig. 1D.

2.6. Optimisation

We use simulated annealing [22] to optimise the vocal tract parameters, which is a stochastic algorithm that seeks an optimal solution through a coarse-to-fine criterion, suitable for problems with many degrees of freedom, such as the speech production system. The learning process started from a neutral position (schwa) followed by adjustments of the vocal tract parameters and gradually converged to a solution. We initiated 10 processes in parallel for each target word, each with 4k iterations.

2.7. Listening experiment and statistical analysis

After optimisation, we selected five items with the lowest recognition errors for both the static and dynamic articulatory targets. In addition to the original duration of 400ms, we synthesised the target words with longer (450ms and 500ms) and shorter (350ms and 300ms) durations to examine the generalisability over speaking rates. For the static targets, the duration ratio of the two targets was kept the same as the learned ratio.

The listeners were 20 American English native speakers (male: 12; mean age: 36), invited and screened via Prolific. The stimuli were randomised and presented to the participants via Gorilla. Before the experiment, the participants filled a brief questionnaire for demographic and language background. Listeners were asked to undertake the experiment on a computer in a quiet environment. A headphone screening was conducted [23] and five practice trials were presented. In the experiment, participants were instructed to listen to the audio carefully up to five times and choose the word that they heard from the word list. The experiment lasted around 20 minutes.

After the experiment, non-parametric statistical analysis, Wilcoxon rank sum tests with Bonferroni corrections were chosen for statistical analysis, due to the skewed distribution of the data.

3. RESULTS

The simulation learned intelligible English words containing diphthongs. A demonstration video and learned synthetic samples can be found in https://gitlab.com/Anqi_Xu/dynamic_diphthongs.

The identification accuracy of the learned diphthongs across target words judged by native listeners is shown in Fig. 2. Error bars show standard errors. The average accuracy was 34.36% and 64.92% for diphthongs synthesised with two static targets and one dynamic target, respectively. The single dynamic target yielded diphthongs that were significantly more intelligible than those synthesised with two static targets (W = 7514, $p < .001$) with only except for /aɪ/. The difference was significant for /aɪ, eɪ , ɔɪ/ ($p < .05$) but not for /aʊ/ and /əʊ/. /eɪ/ and /ɔɪ/ with dynamic targets had fairly high accuracy, both over 90%, whereas the two static targets had the highest accuracy for /aɪ/.

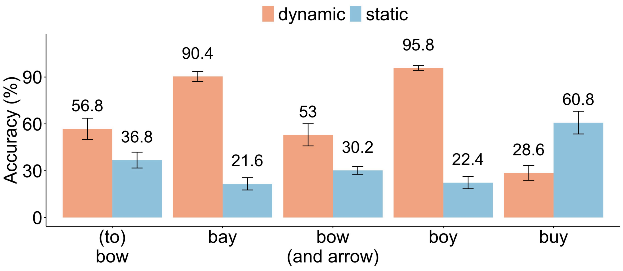

Figure 2: Identification accuracy of words with diphthongs modelled with two static targets or one dynamic target.

The identification accuracy of the diphthongs with different duration is shown in Fig. 3. Again, across all the temporal modulations, not only did the diphthongs with the dynamic target have higher accuracy than those with two static targets for the original duration ($p < .001$), but also for the longer and shorter durations ($p < . 001$).

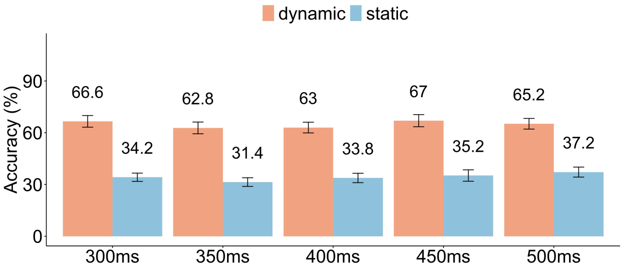

Figure 3: Identification accuracy of words with diphthongs modelled with two static targets or one dynamic target.

Fig. 4 shows the dynamic changes of the learned vocal tract shapes with the lowest recognition error for the target words. The first and second graphs in each row show the starting and ending vocal tract shapes of the CV syllables containing diphthongs. Take /aɪ/ for example, for both the static and dynamic targets, the initial tongue position is relatively low, which later moves towards a higher position. The initial tongue shapes of /eɪ/ seem to be a mid vowel and the terminating tongue position is high for both conditions, whereas the changes in tongue shapes are larger for the dynamic target. /aʊ/ have nearly identical terminating tongue shapes in the two conditions, but the initial tongue position is rather different. The dynamic and static targets both involve little tongue movements for /əʊ/. Finally, in the case of /ɔɪ/, the tongue shapes are retracted in the beginning for both conditions but the dynamic target ends at a higher and more front position. Overall, both the learned articulatory targets based on the static and dynamic targets exhibited beginning and ending vocal tract shapes that resembled two different vowels, whereas the dynamic target showed slightly more plausible vocal tract shapes than the static targets.

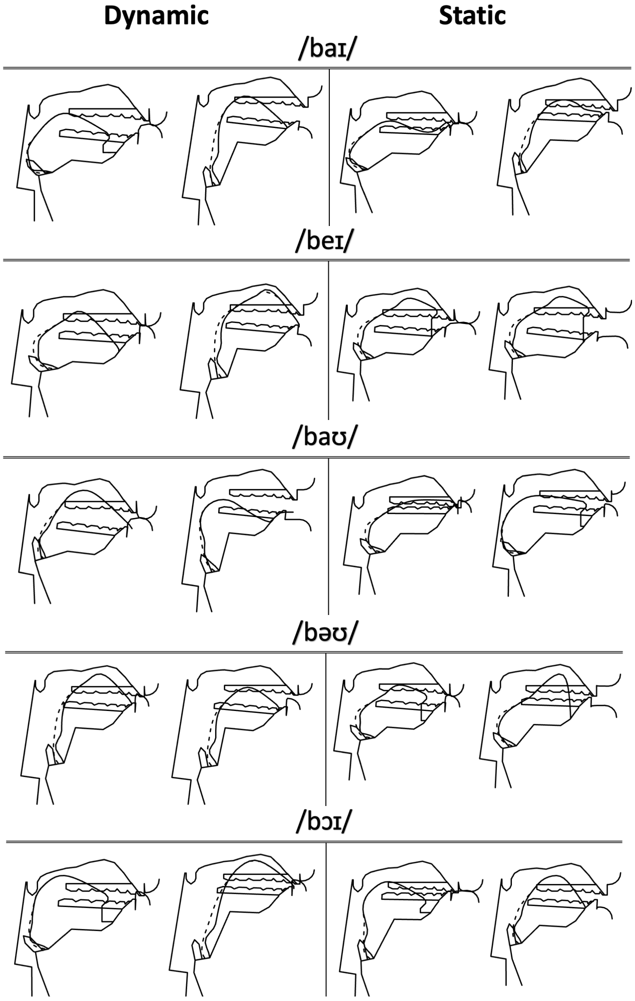

Figure 4: Midsagittal sections of the learned vocal tract shapes with the lowest recognition error.

4. DISCUSSION

We have adopted a new approach to probe the nature of diphthongs via computational simulation of vocal learning. We tested the hypothesis that diphthongs are either a single dynamic articulatory target or two static targets by training a vocal tract model to learn English diphthongs embedded in real words under the guidance of a phoneme recogniser. The results show that unitary dynamic targets can generate more intelligible speech than consecutive static targets. When synthesising words with longer and shorter durations, the dynamic targets also showed more advantage. The learning simulation thus offers new evidence that diphthongs are likely to be independent phonetic entities with unitary dynamic targets.

The theoretical account of diphthongs as unit phonemes was originally proposed on the basis that formant transition stayed constant in varying speech rates [9], but counterevidence emerged subsequently [14]. We have used a new methodology to address the controversy by emulating the dynamic movement of diphthongs under different temporal modulations. The dynamic targets generated more intelligible speech than the static targets, except for /aɪ/. The benefit of having a dynamic target held true even when the speaking rate was modified, which is consistent with the unfluctuating formant slopes observed in [4], [9], [10], [12]. These findings show that diphthongs may have underlying dynamic targets, supporting the proposal of Gay [9].

A main novelty of this study is to investigate diphthongs by simulating its articulatory learning. Previous research has identified various auditory signatures of diphthongs, such as F2 transition rates [4], [5], diphthong endpoints [6] , F1-F2 onset and offset [7] and all of the above [8]. But the lack of consensus suggests that different studies may have observed different manifestations of the diphthongs due to not only differences in methodology, but also cross-speaker variations. During the acquisition of diphthongs, learners necessarily have to deal with cross-speaker variations. A successful learning strategy is likely one that has focused on the most critical property of diphthongs. Thus, learning simulation may be an effective way of discovering the core property of diphthongs. The power of learning simulation as a means of discovering the core mechanisms of speech has been previously demonstrated [15]–[17]. The present study is a further confirmation of its effectiveness.

One source of weakness of this study is that the speech data for training the phoneme recogniser is not balanced across all the speech sequences. This imbalance may have led to varied identification accuracy of the recogniser which was a possible source of the uneven learning performance of the diphthongs. Also, the scope of this study is limited to English. Given that there are noticeable cross-linguistic differences in both the perception and production of diphthongs [18], [19], further research can explore how diphthongs in other languages should be modelled. These limitations notwithstanding, the present study contributes insights into the dynamic nature of English diphthongs. The computational approach opens a new path towards examining theoretical constructs in speech production and perception.

5. ACKNOWLEDGEMENTS

This work has been funded by the Leverhulme Trust Research Project Grant RPG-2019-241: "High quality simulation of early vocal learning".

6. REFERENCES

[1] I. Lehiste and G. E. Peterson, “Transitions, glides, and diphthongs,” *J Acoust Soc Am*, vol. 33, no. 3, pp. 268–277, Mar. 1961, doi: 10.1121/1.1908638.

[2] A. Holbrook and G. Fairbanks, “Diphthong formants and their movements,” *J Speech Hear Res*, vol. 5, no. 1, pp. 38–58, Mar. 1962, doi: 10.1044/jshr.0501.38.

[3] G. L. Trager and H. L. Smith, *An outline of English Structure*. Norman, Oklahoma: Battenburg Press, 1951.

[4] T. Gay, “A perceptual study of American English diphthongs,” *Lang Speech*, vol. 13, no. 2, pp. 65–88, Apr. 1970, doi: 10.1177/002383097001300201.

[5] A. K. Nábělek, A. Ovchinnikov, Z. Czyzewski, and H. J. Crowley, “Cues for perception of synthetic and natural diphthongs in either noise or reverberation,” *J Acoust Soc Am*, vol. 99, no. 3, pp. 1742–1753, Mar. 1996, doi: 10.1121/1.415238.

[6] A. Bladon, “Diphthongs: A case study of dynamic auditory processing,” *Speech Commun*, vol. 4, pp. 145–154, 1985.

[7] M. Gottfried, J. D. Miller, and D. J. Meyer, “Three approaches to the classification of American English diphthongs,” *J Phon*, vol. 21, no. 3, pp. 205–229, Jul. 1993, doi: 10.1016/S0095-4470(19)31337-3.

[8] S. Lee, A. Potamianos, and S. Narayanan, “Developmental acoustic study of American English diphthongs,” *J Acoust Soc Am*, vol. 136, no. 4, pp. 1880–1894, Oct. 2014, doi: 10.1121/1.4894799.

[9] T. Gay, “Effect of speaking rate on diphthong formant movements,” *J Acoust Soc Am*, vol. 44, no. 6, pp. 1570–1573, Dec. 1968, doi: 10.1121/1.1911298.

[10] S. M. Tasko and K. Greilick, “Acoustic and articulatory features of diphthong production: A speech clarity study,” *Journal of Speech, Language, and Hearing Research*, vol. 53, no. 1, pp. 84–99, Feb. 2010, doi: 10.1044/1092-4388(2009/08-0124).

[11] K. Tjaden and G. E. Wilding, “Rate and loudness manipulations in dysarthria: acoustic and perceptual findings,” *Journal of Speech, Language and Hearing Research*, vol. 47, pp. 766–783, 2004.

[12] R. D. Kent and K. L. Moll, “Tongue body articulation during vowel and diphthong gestures,” *Folia phoniat*, vol. 24, pp. 278–300, 1972.

[13] C. Dromey, G. O. Jang, and K. Hollis, “Assessing correlations between lingual movements and formants,” *Speech Commun*, vol. 55, no. 2, pp. 315–328, Feb. 2013, doi: 10.1016/j.specom.2012.09.001.

[14] J. Wouters and M. W. Macon, “Effects of prosodic factors on spectral dynamics. I. Analysis,” *J Acoust Soc Am*, vol. 111, no. 1, pp. 417–427, Jan. 2002, doi: 10.1121/1.1428262.

[15] A. Xu, P. Birkholz, and Y. Xu, “Coarticulation as synchronized dimension-specific sequential target approximation: An articulatory synthesis simulation,” in *Proceedings of the international congress of phonetic sciences (ICPhS)*, 2019. [Online]. Available: www.vocaltractlab.de

[16] D. R. van Niekerk, A. Xu, B. Gerazov, P. K. Krug, P. Birkholz, and Y. Xu, “Finding Intelligible Consonant-Vowel Sounds Using High-Quality Articulatory Synthesis,” in *Interspeech 2020*, Oct. 2020, pp. 4457–4461. doi: 10.21437/Interspeech.2020-2545.

[17] S. Prom-On, P. Birkholz, and Y. Xu, “Identifying underlying articulatory targets of Thai vowels from acoustic data based on an analysis-by-synthesis approach,” *EURASIP J Audio Speech Music Process*, vol. 2014, no. 1, p. 23, 2014, doi: 10.1186/1687-4722-2014-23.

[18] W. J. M. Peeters, “Dipthong dynamics: A cross-linguistic perceptual analysis of temporal patterns in Dutch, English, and German,” 1996.

[19] W. J. M. Peeters and W. J. Barry, “Diphthong dynamics: production and perception in Southern British Englsih,” in *EUROSPEECH*, 1989, pp. 1055–1058.