Learnability of English diphthongs: One dynamic target vs. two static targets

Anqi Xu[1], *Daniel van Niekerk*[2], *Branislav Gerazov*[3], *Paul Konstantin Krug*[4], *Santitham Prom-on*[5], *Peter Birkholz*[4], *Yi Xu*[2]

[1]School of Humanities and Social Sciences, Harbin Institute of Technology (Shenzhen), China

[2]Department of Speech Hearing and Phonetic Sciences, University College London, UK

[3]Faculty of Electrical Engineering and Information Technologies, Ss. Cyril and Methodius University in Skopje, Skopje, RN Macedonia

[4]Institute of Acoustics and Speech Communication, Technische Universität Dresden, Germany

[5]Computer Engineering Department, King Mongkut's University of Technology Thonburi, Thailand

Abstract

As vowels with intrinsic movements, diphthongs are among the most elusive sounds of speech. Previous research has characterized diphthongs as a combination of two vowels, a vowel followed by a formant transition, or a constant rate of formant change. These accounts are based on acoustic patterns, perceptual cues, and either acoustic or articulatory synthesis, but no consensus has been reached. In this study, we explore the nature of diphthongs by exploring how they can be acquired through vocal learning. The acquisition is simulated by a three-dimensional (3D) vocal tract model with built-in target approximation dynamics, which can learn articulatory targets of phonetic categories under the guidance of a speech recognizer. The simulation attempts to learn to articulate diphthong-embedded monosyllabic English words with either a single dynamic target or two static targets, and the learned synthetic words were presented to native listeners for identification. The results showed that diphthongs learned with dynamic targets were consistently more intelligible across variable durations than those learned with two static targets, with only the exception of /aɪ/. From the perspective of learnability, therefore, English diphthongs are likely unitary vowels with dynamic targets.

Index Terms: diphthongs, computational simulation, 3D vocal tract model, vocal learning, American English

1 Introduction

Diphthongs, a special group of vowels, are featured by having different formant values at their onset and offset, and smooth transitional movements in between (Holbrook & Fairbanks, 1962; Lehiste & Peterson, 1961). Their dynamic quality makes them difficult to characterize, and their nature remains elusive to this day. As complained by Lass (1984:95), "If long vowels produce methodological headaches, diphthongs are a positive migraine." Central to the theoretical uncertainty is whether diphthongs consist of two successive vowels (Lehiste & Peterson, 1961; Trager & Smith, 1951) or a single unitary vowel (Gay, 1968, 1970). Both possibilities, however, have been explored based on evidence from acoustics, articulation and perception studies, as reviewed next.

1.1 Evidence from acoustics and articulation of diphthongs

One of the first observations is that the transcriptions of five English diphthongs (i.e., /aɪ/, /aʊ/, /ɔɪ/, /eɪ/, and /əʊ/) do not correspond well with their actual acoustic properties (Gay, 1968; Holbrook & Fairbanks, 1962; Lehiste & Peterson, 1961; Potter & Peterson, 1948). For instance, although /aɪ/, /eɪ/, /ɔɪ/ are described as having the same ending sound, the final F2 of /eɪ/ is in fact slightly higher than that of /ɔɪ/ and /aɪ/ (Holbrook & Fairbanks, 1962). The initial formants of /aʊ/ and /aɪ/, on the other hand, are reported to be close to several monophthongs such as /ɒ/, /a/ and /æ/ (Holbrook & Fairbanks, 1962; Lehiste & Peterson, 1961). Among the five diphthongs, /eɪ/ and /əʊ/ are sometimes categorized differently because they involve relatively short steady states of formants at their onsets, accompanied by limited formant movements (Lehiste & Peterson, 1961). The durations of /eɪ/ and /əʊ/ are also shorter than those of /aɪ/, /eɪ/, and /ɔɪ/, regardless of speaking rates (Gay, 1968) or stress conditions (Gottfried et al., 1993). The inadequacy of gliding formants and the brief duration of /eɪ/ and /əʊ/ have led to their classification as having a single target, in opposition to /aɪ/, /eɪ/, and /ɔɪ/, which have double targets (Lehiste & Peterson, 1961).

In contrast to Trager & Smith (1951)'s proposal of vowel combinations and Lehiste & Peterson (1961)'s grouping of single and dynamic targets, Gay (1968) investigated the acoustic properties of five American English diphthongs spoken at three speech rates (slow, moderate, and fast). The formant onset of the diphthongs was found to be rather consistent, with the exception of /ɔɪ/, where the F1 and F2 in the slow speech rate had different onset

frequencies compared to the moderate and fast conditions. When sufficient time was available, the F1 and F2 offset values became more extreme, while in fast speech, the final portion of the diphthong could be eliminated. Interestingly, the rate of F2 movement remained consistent across all three speaking rates.

A more recent study by Tasko & Greilick (2010) on careful and conversational speech supports the findings of Gay (1968). Clear speech indeed led to an increase in duration and formant excursion, while F2 slopes were not significantly affected by speaking modes. Furthermore, the loudness of speech was not found to induce changes in the F2 slopes of diphthongs either (Tjaden & Wilding, 2004). The findings in diphthong articulation align well with the acoustics, indicating that the tongue kinematic traces did not show mode-related changes, except for the posterior part of the tongue, which exhibited higher movement speed in clear speech (Tasko & Greilick, 2010). X-ray data has shown that tongue flesh points underwent minimal changes across different speaking rates, with the tongue body, in particular, maintaining invariant velocity (Kent & Moll, 1972). These results largely accord with Thompson and Kim (2019), who investigated the tongue kinematics and acoustic measures of /aɪ/ and /eɪ/ spoken in conversational, clearer, and less clear speaking modes, confirming constant F2 slopes and a strong correlation between acoustics and articulation. This significant correlation was also reported in Dromey et al. (2013) that the tongue movements and formant transitions of diphthongs were highly correlated, despite some exceptions.

The invariant F2 slope of diphthongs in speech production has been nevertheless contested in a number of studies. Weismer (1991) conducted an in-depth investigation into the formant trajectories of diphthongs, in which a native speaker was invited to record /aɪ/ at very fast, conversational, and very slow speech rates. The F1 and F2 transition of /aɪ/ of ‘buy’ within a carrier sentence ‘Buy Bobby a puppy’ were extracted. Contrary to previous findings of unfluctuating F2 movement, F2 slopes appeared to vary with vowel duration. It was reported that the relationship between the duration and the extent of transition was better fitted to a quadratic regression model rather than a linear one. However, the documented F1 and F2 values, in fact, included formant transitions towards the next vowel in the carrying sentence.

Weismer & Berry (2003) also recorded native speakers producing diphthong /ɔɪ/ in a graded speech task ranging from self-determined slowest to fastest speaking rate. Some speakers produced diphthongs as short or long steady-state vowels while maintaining a constant F2 transition at the offset, while others showed no systematic effects on F2 slopes. A possible cause of the inconsistency may be the contextual influence arising from the sounds following the diphthongs, because the monosyllabic target words containing diphthong /ɔɪ/ were embedded in a carrier phrase 'put a [target word] here'. The same experimental paradigm was used again by Tjaden & Weismer (1998) to study the speaking tempo induced F2 changes and the measurements of F2 onset was taken when there was still contextual influence from the preceding vowels. Consequently, previous acoustic measurements of diphthongs may have been compromised due to the carrier sentences used in the recording procedure.

Another piece of evidence challenging the hypothesis that diphthongs are single unitary targets (Gay, 1968, 1970) comes from Dolan & Mimori (1986), who investigated the formant profiles of diphthongs at normal, slow, and fast speech rates. They reported that increased tempo induced fast F2 transition rates. However, the finding is not directly comparable to previous studies as the glide components of diphthongs in this study were defined differently from the conventional approach. Instead of the turning point, the transition onset was selected based on a 15/20-Hz change over 10 ms. In addition, Wouters & Macon (2002) measured spectral transition based on the slopes of the first three formants (F1, F2 and F3) at distinctive speaking rates. Linear regression lines were fitted to the formant slopes and then the spectral changes were measured by the root mean-square errors of the fitted slopes. The spectral changes of diphthongs were found to be reduced in clear speech with prosodic prominence. However, this is likely due to the V-shaped F3 contours of diphthongs (Clermont, 1993). Taken together, the controversy over whether formant slopes remain invariant across speaking rates can be due to the distinctive measurements employed.

Besides speaking rate, the dynamic nature of diphthongs can sometimes be probed in response to linguistic contexts, as the duration of diphthongs can be conditioned by lexical stress, accent, and sentence position (Wouters & Macon, 2002). The spectral rate of change was quantified by the root-mean-square of the slopes for the linear regression lines of F1,

F2 and F3. It was shown that stress, accent, word position and hyperarticulation can induce an increase of the spectral rate of change. What has also been widely studied is diphthongs with different timing before voiced and voiceless consonants. For instance, diphthong /aɪ/ in 'tied' consists of a steady state formant followed by a transitional movement, but the one in 'tight', being shorter in duration, lacks the initial steady state (Moreton, 2004; Thomas, 2000). These context-modulated durational differences triggered similar formant transition patterns as observed in lengthened or shortened utterances with varying speech rates.

1.2 Evidence from perception of diphthongs

The ongoing debate over the relevant acoustic and articulatory features of diphthongs is further complicated by conflicting observations regarding their perception. To investigate what makes diphthongs phonemically distinctive, Gay (1970) created acoustic continua of synthetic diphthongs with variable initial and terminating F2 and F3, along with interpolated formant movements. It was observed that the most prominent perceptual cue for listeners was the F2 movement of the diphthongs rather than the formant onset and offset, which suggests that even for /aɪ/, /eɪ/, and /ɔɪ/, their underlying targets are more likely to constitute the distinct phonetic entities (Gay, 1970).

These results align with more recent studies indicating that the key to the perception of synthetic and natural diphthongs in noise or reverberation is the intensity of F2 transitions (Nábělek et al., 1996). Conversely, some studies suggest that the crucial feature in the identification of manipulated diphthongs is the endpoint rather than the transitional trajectories (Bladon, 1985). Also using synthetic diphthongs, Bond (1978, 1982) approached the question of diphthong identification with an emphasis on transition duration. It was found that long gliding movements ensured a perceptual inclination towards diphthongs, but when the steady-state portion was evident enough, a short formant shift was also adequate. This study similarly underscores the importance of formant transitions in diphthong identification and additionally suggests a potential interaction between the vowel onset and the duration of the glide.

Another line of studies sought to investigate the characteristic acoustic features of diphthongs within a speech corpus. Gottfried et al. (1993) employed a classifier to statistically capture patterns of acoustic changes in diphthongs produced in /bVd/ and /hVd/ contexts, with varying speaking rates and stress locations. It has been found that classification accuracy was comparable whether F1/F2 onsets and slopes, or F1/F2 onsets and offsets were included. Lee et al. (2014) adopted a statistical approach to classify diphthongs produced by speakers of different age and gender. Fisher's discriminant analysis showed that incorporating F1–F3 onset, offset and transition rates yielded the best classification results. Notably, there are methodological differences in how the acoustic landmarks for onsets and offsets were determined. In Gottfried et al. (1993), the landmark was manually determined when there were no significant spectral changes in the first or last 15% of the segment, whereas Lee et al. (2014) used automatic segmentation to determine the onsets and offsets. Different from human perception experiments, more acoustic landmarks are always advantageous than a particular one for machine learning or statistical methods. It could be due to the fact that the large speech datasets used encompass variability in contexts, speakers, speaking rates, and other factors, dissimilar to well-controlled laboratory speech.

1.3 Evidence from modelling studies

Previous simulation studies have sought to model the movements of English diphthongs using a critically damped mass-spring system within the Task Dynamics framework (Browman & Goldstein, 1989, 1986; Saltzman & Munhall, 1989). Hsieh (2017) introduced a gestural coupling model, demonstrating that diphthongs can be represented as two vocalic gestures: ongliding diphthongs involve in-phase coordination of overlapping gestures, whereas offgliding diphthongs require anti-phase coordination of sequential gestures with clear temporal separation. Similarly, Strycharczuk et al. (2024) employed a modified version of Task dynamics proposed by Sorensen & Gafos (2016), to simulate velocity profiles of Tongue Body Constriction Degree (TBCD) for diphthongs. Their model predicts distinct velocity peaks for diphthongs, corresponding to movements toward two articulatory targets, effectively illustrating how diphthongs can be modeled as gesture sequences with two targets. Collectively, these studies highlight the articulatory movements of diphthongs can be effectively captured using a two-gesture framework.

Meanwhile, Stone & Birkholz (2024) extended this research to model not only the articulation of German diphthongs but also their acoustic outcomes. Their simulation demonstrated that German primary diphthongs (/aɪ/, /aʊ/, /ɔʏ/) can be accurately synthesized using static vocal tract shapes derived from monophthongs in a 3D articulatory synthesizer, VocalTractLab. The synthetic diphthongs produced formant transitions that closely matched those of natural diphthongs, particularly for F1 and F2. Crucially, listeners reliably identified these synthesized diphthongs, confirming that their acoustic quality was sufficiently natural for speech perception. This study demonstrates that static targets of monophthongs can generate German diphthongs with natural formant profiles and high perceptual quality. The sufficiency of the two-target approach may be attributed to the more balanced temporal structure of German diphthongs, which emphasizes both the onset and offset steady states. This differs from English diphthongs, which are characterized by a long onset steady state and a short or absent offset steady state (Peeters & Barry, 1989; Peeters, 1996).

Overall, these three simulation studies (Hsieh, 2017; Stone & Birkholz, 2024; Strycharczuk et al., 2024) establish that a two-target approach can effectively model both the articulation and acoustics of diphthongs.

1.4 Missing perspectives

Significant questions remain, however, regarding the nature of the underlying targets of diphthongs. The accounts from previous studies all seem to share one assumption, namely, what is observed from acoustic analysis and perceptual experiments represents the underlying properties of the diphthongs directly. This assumption overlooks two critical aspects that we believe are of importance: (a) articulatory mechanisms, and (b) learnability. Articulatory mechanisms refer to how speech sounds are produced by speakers, which can significantly obscure the mapping between intended and observable speech forms. Learnability refers to whether a proposed/postulated property of a phonetic entity would allow a child or an adult learner to master its articulation, based on the premise that any persistent linguistic feature must be successfully learned by speakers.

1.4.1 Articulatory mechanisms

A number of articulatory mechanisms may significantly limit the production of diphthongs. The first is the well-established fact that any articulatory movement requires a substantial

amount of time. According to Tiffany (1980) and Kent et al. (1987), each segmental movement, on average, needs at least 74 ms. Meanwhile, Nelson et al. (1984) and Y. Xu & Prom-on (2019) report that a unidirectional formant movement would start to asymptote beyond 125 ms. Hence, when a two-vowel sequence lasts longer than 250 ms, it begins to show two distinct movements—one toward each vowel target—as illustrated in Fig. 1A. However, such two-step movements are rarely observed in previous studies.

The general lack of visible two-step movements may suggest an alternative, namely, an underlying articulatory target that is intrinsically dynamic, as illustrated in Fig. 1B. Such dynamic targets are suggested for contour tones like rising and falling tones in Mandarin (Y. Xu, 1997, 1998, 2001), and have been incorporated into the target approximation model for tone and intonation (Prom-on et al., 2009; Y. Xu & Wang, 2001). In this model, both static and dynamic targets can be represented by a simple linear equation, as illustrated in Fig. 1. A static target remains constant over time with a slope of zero (Fig. 1A), whereas a dynamic target has a non-zero slope, i.e., non-zero velocity (Fig. 1B). To articulate such a target, the resulting articulatory and acoustic trajectories would show a relatively constant final velocity that reflects that slope, as depicted in Fig. 1B, unless the articulation is given insufficient time to approach the target, as depicted in Fig. 1C. Cases of constant final velocities have been observed in both diphthongs (Gay, 1968) and contour tones (Y. Xu, 1998, 2001) in formant and f_o trajectories, while the variable final velocities reported in Weismer (1991) and Tjaden & Weismer (1998) likely reflect conditions similar to those illustrated in Fig. 1C. Note also that the dip in the middle of the trajectory in Fig. 1B arises because the approximation of a dynamic target follows a time course of tracking the underlying linear trajectory of the target. This dip occurs as articulation approaches the initial portion of the dynamic target, which is lower than its endpoint.

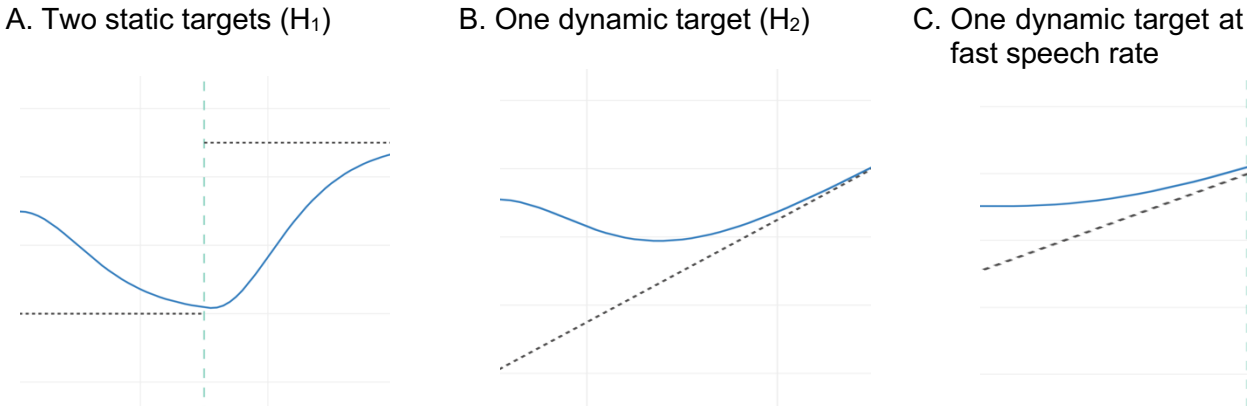


Figure 1: A schematic illustration of the asymptotic approximation of two types of targets resulting in identical surface articulatory trajectories. The solid lines represent surface articulatory contours, while the dotted lines depict the underlying linear targets driving the movement towards the targets. In (A), the vertical line divides the temporal domains of the two static targets. Graphics were generated by quantitative target approximation (qTA) (Prom-on et al., 2009; Y. Xu & Wang, 2001) Demo: https://www.homepages.ucl.ac.uk/~uclyyix/tools.html

Another articulatory mechanism is syllable formation based on the coproduction of consonant and vowel at the syllable onset whereby consonant and vowel cooccur at the onset of the syllable (Bell-Berti & Harris, 1981; Fowler, 1980). It was later proposed that this involves full synchrony of consonant and vowel (Liu et al., 2022; Y. Xu, 2024), as illustrated in Fig. 2, which has now received empirical support (Liu et al., 2022; A. Xu et al., 2019, 2024). This means that the initial opening movement of the vowel or diphthong and the closing movement of the consonant would be fully overlapped with each other. As a result, the initial vowel movements are usually unobserved, because of the interruption of formants induced by the articulatory closure of the consonant. Existing literature tends to focus on the voicing period of diphthongs, while the initial movements have been largely neglected.

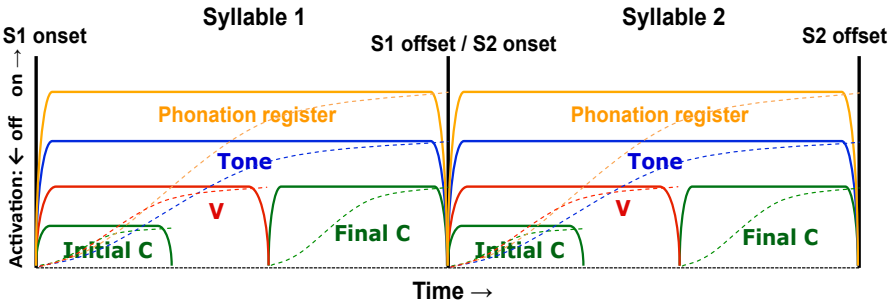


Figure 2: Synchronization model of the syllable. The dashed lines represent target approximation movements toward specific targets (Adapted from Y. Xu & Liu, 2006).

1.4.2 Learnability

Learnability is about whether the proposed properties of a phonetic segment would allow a young child or a second language learner to learn to produce it. This is relevant because if not learnable, the property cannot persist across generations or appear in the language in the first place. Learnability may be closely related to articulatory constraints. For example, a proposed property apparently should not require learners to exceed their maximum speed of articulation, e.g., greater than 13.5 segments/s (Tiffany, 1980). Since 125 ms is needed for a target approximation movement to asymptote (Nelson et al., 1984; Y. Xu & Prom-on, 2019), would it imply that at least 250 ms is needed for a two-vowel-based diphthong? Also, given that the initial portion of vowel target approximation is often obscured by the initial consonant, would the first vowel in a two-vowel-based diphthong be too challenging for language learners to observe?

To address these questions, computational simulations are needed, as behavioral studies alone cannot uncover the underlying learning mechanisms. Furthermore, although previous articulatory modeling of English diphthongs has been effective (Hsieh, 2017; Strycharczuk et al., 2024), it has not tackled the more challenging question of how diphthongs are learned in speech production. In recent research, we have developed a method that can successfully simulate vocal learning of monosyllabic English words by training a 3D articulatory synthesizer with an automatic speech recognizer (van Niekerk et al., 2023; A. Xu et al., 2024). These studies show that learning guided by a speech recognizer is far superior to learning via direct acoustic imitation. This suggests that vocal learning is ultimately about discovering articulatory targets that can generate acoustic patterns that can be perceived

as the intended phonetic categories. Consequently, the learnability of diphthongs would be about whether the postulated properties would allow the learners to discover the articulatory targets that can generate acoustic patterns identifiable as the intended diphthongs by both simulated and real human listeners.

1.5 Current study

In the current study, therefore, we aim to explore the nature of English diphthongs by using computational simulation of vocal learning to examine two hypotheses regarding the underlying articulatory targets for diphthongs: (H1) two consecutive static targets and (H2) a unitary dynamic target, as illustrated in Fig. 1. The plausibility of the two hypotheses will be assessed based on a simulated learning paradigm.

In this paradigm, an articulatory synthesizer will be trained with a 3D vocal tract model to learn American English words containing offglide diphthongs (i.e., /aɪ/, /aʊ/, /ɔɪ/, /eɪ/, and /əʊ/), following the simulation paradigm in Krug et al. (2023), Prom-On et al. (2014), van Niekerk et al. (2023) and A. Xu et al. (2019, 2024). The learning process is guided by a syllable-based phoneme recognizer pre-trained with a deep learning model. At the end of the simulated learning, the words containing the diphthongs will be synthesized using the learned articulatory targets with varying durations to verify their generalizability across different speaking rates. The performance of the two types of articulatory targets will be evaluated based on the following:

1) Intelligibility of the synthesized speech in a listening experiment.
2) Plausibility of the learned articulatory kinematics.
3) Generalizability of the learned articulatory targets at different speech tempos.

2 Method

2.1 Speech materials

Five diphthongs, /aɪ, eɪ, əʊ, aʊ, ɔɪ/, were embedded in real English words with bilabial onset consonants, as listed in Table 1. Using these minimal pairs of real English words ensures that perception experiments can be conducted naturally by native speakers. Since the two target words for "bow" are homographs, hints were included to distinguish them, as indicated in brackets. These same hints were also provided to participants during the listening experiment.

Table 1: Target English words with diphthongs in the simulation.

Diphthongs	/bV/
aɪ	buy
eɪ	bay
əʊ	bow (and arrows)
aʊ	(to) bow
ɔɪ	boy

2.2 Learning framework

We trained a 3D vocal tract model to find optimal articulatory targets for the five English diphthongs using a perception-guided learning paradigm, as shown in Fig. 3. This framework includes both a production and a perception system. Initially, the model explored a set of articulatory targets (Fig. 3A), with kinematic trajectories based on assumptions of either two static targets or one dynamic target (Fig. 3B). These time-varying vocal tract shapes were then converted into cross-sectional area functions to obtain the synthesized speech signals based on acoustic simulation (Fig. 3C). In each learning cycle, the synthetic speech was assessed by the perception system (Fig. 3D) to iteratively search for optimal articulatory targets with minimal perceptual errors. Detailed explanations of each model component will follow in subsequent sections.

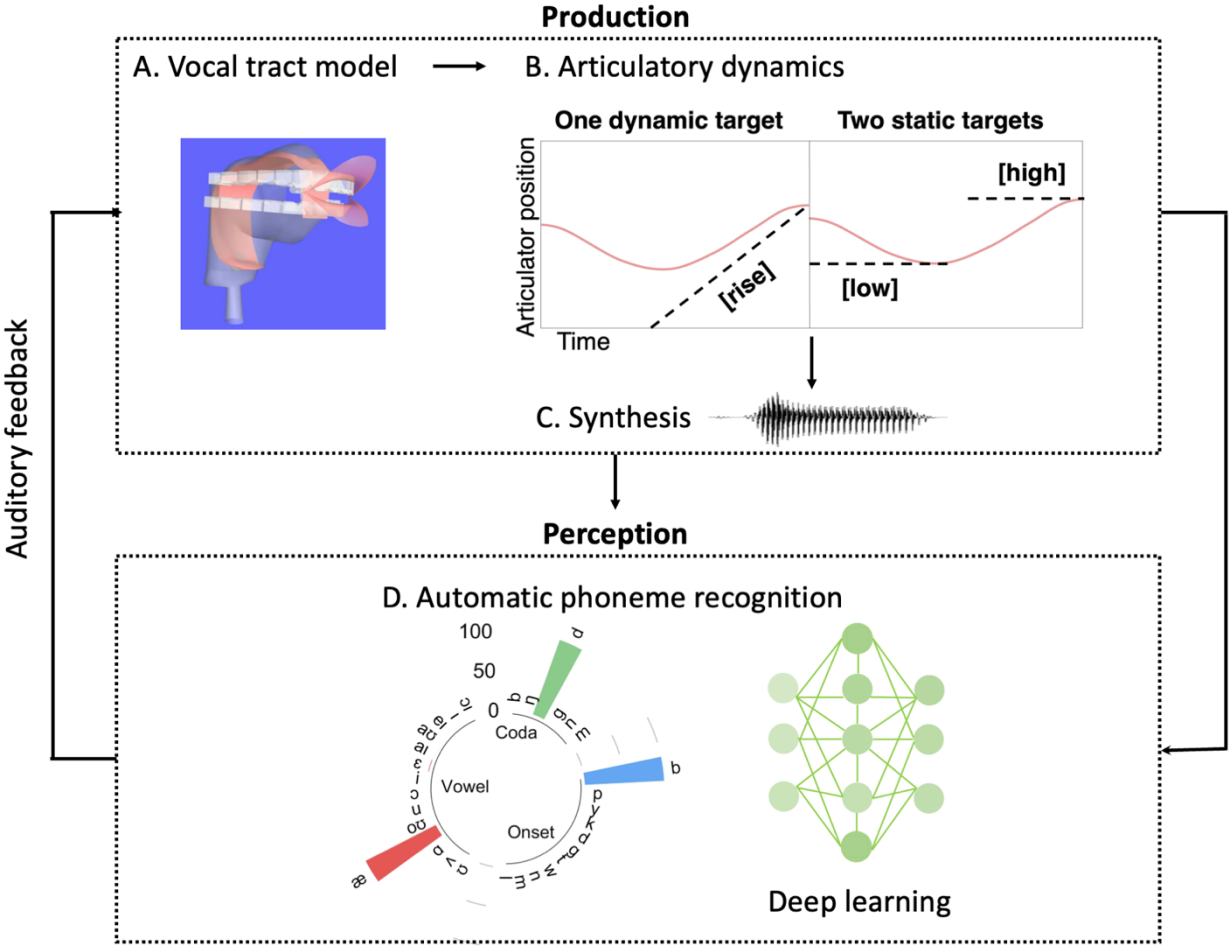


Figure 3: Overview of the learning process.

2.3 Vocal tract model (Fig. 3A)

The articulatory synthesizer, VocalTractLab 2.3 (www.vocaltractlab.de), used in the simulation (Fig. 3A) is based on a geometrical 3D vocal tract model, adapted to MRI data of a German male speaker for the anatomical locations of the articulators. This synthesizer performs one-dimensional aerodynamic-acoustic simulations based on cross-sectional area functions. Table 2 presents sixteen vocal tract parameters used to model the movements of joint muscle forces, all of which were optimized simultaneously during the simulation. Laryngeal articulation control involved setting the vocal folds to be fully adducted with moderate tension for the diphthong targets, while parameters such as the distance between vocal cords, glottis rest area, and relative amplitude for consonant

targets were free parameters. The fundamental frequency (f_0) target of the CV sequence was set to have a falling intonation.

Table 2: Vocal tract parameters involved in the simulation.

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 center positions
VS	Velum shape

2.4 Articulatory dynamics (Fig. 3B)

We used a quantitative target approximation (qTA) model to control the movements of the vocal tract parameters in Table 2 (Prom-on et al., 2009; Y. Xu & Wang, 2001). It provides a mathematical framework for simulating the dynamic process of articulatory movements by describing how articulatory targets are approached during speech production. In this model, each articulatory target is defined by three parameters—position, slope and strength.

- Target position: The desired spatial configuration of the articulators.
- Target slope: The rate of change in target position over time.
    - Static Targets (Fig. 1A): When the slope is zero, the target remains constant over time. The articulators move smoothly toward a fixed position, typical for steady-state sounds.
    - Dynamic Targets (Fig. 1B): When the slope is non-zero, the target shifts linearly over time. This dynamic behavior models changing articulatory states, analogous to rising or falling tonal and intonational contours (see Y. Xu & Wang, 2001 for evidence and justifications).
- Target strength: The rate at which articulatory movements progress toward the target, regardless of whether it is static or dynamic.

As shown in Fig. 3B, similar articulatory curves of the diphthongs can result from either two static targets or one dynamic target. For implementing H1, the two static targets had a slope of zero, which required the optimization of the positions of the sixteen vocal tract parameters, along with the strength (1-dimensional). Additionally, since the duration proportion of the two static targets was underspecified, the duration of each static target was also trained during optimization. For H2, the single dynamic target required the optimization of both the position (16-dimensional), the slope (16-dimensional) and the strength (1-dimensional) of each articulatory target.

Alongside the vowel targets, a consonant target of voiced bilabial stops was optimized concurrently with the diphthong targets. During training, the total duration of the two static targets and the duration of the single dynamic target were set to be identical. Even though there are durational differences between different types of diphthongs (Gay, 1968), we adopted the same duration to ensure that the listeners cannot make use of the temporal cues for identification. The duration of the entire CV syllable is 400 ms, with a voicing duration of approximately 250-300 ms[1]. The actual period of the consonant closure depends on the target position and target strength. As a consequence, the learned utterances may exhibit varying voicing durations after optimization.

In order to generate coarticulated CV sequences, the temporal and spatial movements of the consonant and the diphthong were simulated by synchronized dimension-specific sequential target approximation— a coarticulation model (Liu et al., 2022; A. Xu et al., 2019, 2024; Y. Xu, 2024). In this framework, consonant and diphthong articulations are fully synchronized at syllable onset. Despite the consonant-to-vowel (CV) overlap, for the articulator dimensions that are shared by both the consonant and vowel (Horizontal jaw position[JX], jaw angle [JA] and lip distance [LD] in this study), the execution of the articulatory targets proceeds sequentially. As illustrated in Fig. 4, at the onset of a consonant-vowel (CV) syllable with a bilabial stop, the consonant target (dashed lines)

[1] It was also reported that the duration of /ɔɪ/ was longer than the other four diphthongs (Gay, 1968). Specifically, 'boy' had a mean duration ranging from 274 to 452 ms (Weismer & Berry, 2003), while 'buy' had a mean duration of approximately 250 ms at a conversational speaking rate (Weismer, 1991). For our study, we chose to use a duration of 250-300 ms, which is suitable for all diphthongs.

controls the movement of JA, JX and LD, while the vowel target (solid lines) governs the movement of the rest of the articulatory dimensions, such as the horizontal and vertical tongue tip positions (TTX & TTY). When the interval of the consonant target is over, JX, JA and LD start moving towards the vowel target. We further implemented an oral constriction constraint to make sure that the lips are closed during the consonant target.

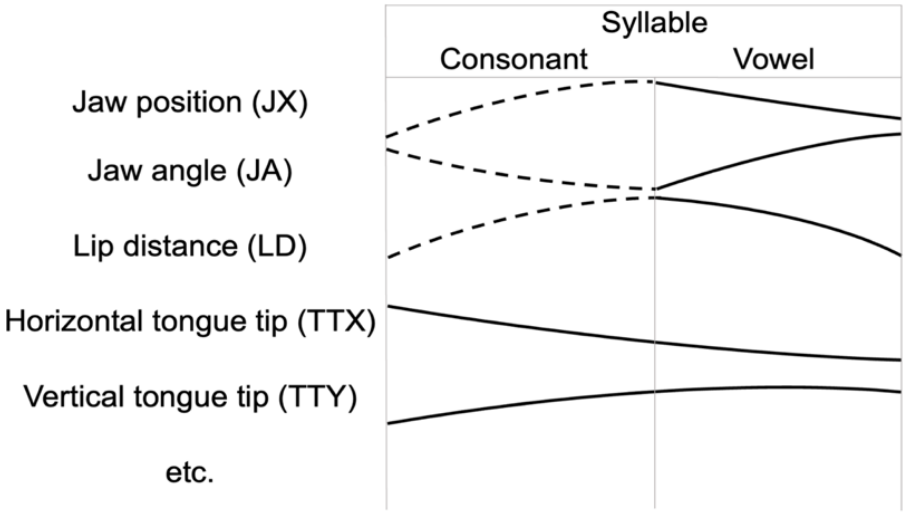


Figure 4: Illustration of the coarticulation model in the case of bilabial stop-vowel sequences. Dashed lines represent the articulatory trajectories of the consonant target and solid lines represent the articulatory trajectories of the vowel target.

2.5 Automatic phoneme recognizer (Fig. 3D)

We employed a deep learning-based speech recognition system (A. Xu et al., 2024) to guide the optimization process, which outputs the recognition rate of each target syllable in terms of an evaluation of the probability of each phoneme in a given speech sequence. The speech data used for training is sourced from the LibriSpeech corpus (Panayotov et al., 2015), comprising recordings of audiobooks by adult male and female speakers of various ages. We extracted 11 onset consonants (/b/, /d/, /g/, /p/, /t/, /k/, /y/, /w/, /n/, /m/, and /l/), 12 vowels, and 5 stressed diphthongs (/aɪ/, /aʊ/, /eɪ/, /oʊ/, /ɔɪ/), along with 6 coda consonants (/b/, /d/, /g/, /n/, /m/, and /ŋ/) from continuous speech in the corpus. The dataset includes speech samples of different syllable types, encompassing 17 vowels, 187 CV syllables, and 1122 CVC words. For training, validation, and testing purposes, the dataset is partitioned into sets containing 116.7, 14.4, and 15 hours of speech, respectively.

During pre-processing, we applied pre-emphasis with a coefficient of 0.97 and computed the log Mel spectrogram using a 25-ms Hamming window with a 5-ms overlap and 26 Mel filters. The input to the deep-learning model consists of log Mel spectrograms with a length of 200 frames (spanning 1 second). The model comprises 8 convolutional layers (Conv) for spectral processing, 6 long short-term memory (LSTM) layers for temporal processing, and 3 dense layers (Dense) for learning the phoneme classification. The model outputs a 34-dimensional vector which represents the probability of each phoneme in the syllable. The vector was then used to estimate the phoneme accuracy of the consonant and the vowel in the CV syllables generated by the vocal tract model.

We initially trained a speech recognition model specifically for diphthongs, using only American English words containing diphthongs. However, this approach proved unsuccessful, as the recognizer struggled to effectively train diphthongs. In both the two static targets and one dynamic target scenarios, the spectrograms of the learned diphthongs showed limited formant movements, resulting in very low intelligibility. Consequently, we opted to train the speech recognizer on *all* onset consonants and vowels in English. This broader approach enabled processing of the contrasting phonological differences in complex contexts.

2.6 Optimization algorithm

To simulate the learning of the articulatory parameters, we employed simulated annealing (Kirkpatrick et al., 1983) to optimize both vocal tract and glottis parameters through trial and error. This stochastic algorithm finds optimal solutions by gradually reducing the temperature which controls the acceptance rate for candidate targets, and refining the target search criteria from coarse to fine. Simulated annealing is well-suited for optimizing models with numerous degrees of freedom, such as speech production. To stabilize the learning outcomes, we implemented simulated annealing in two stages, illustrated in Fig. 5. Initially, the process began with a neutral position (schwa), followed by random adjustments of vocal tract parameters. We ran 10 processes in parallel for each target word, each comprising 2000 iterations. Subsequently, the articulatory target with the lowest recognition error from each of these 10 processes was selected for further, more localized optimization. In the second stage, these selected sets of articulatory targets were explored by the 10 processes,

each undergoing 1000 iterations of random adjustments. We then refined the top 10 articulatory targets through an additional 1000 iterations of fine-tuning.

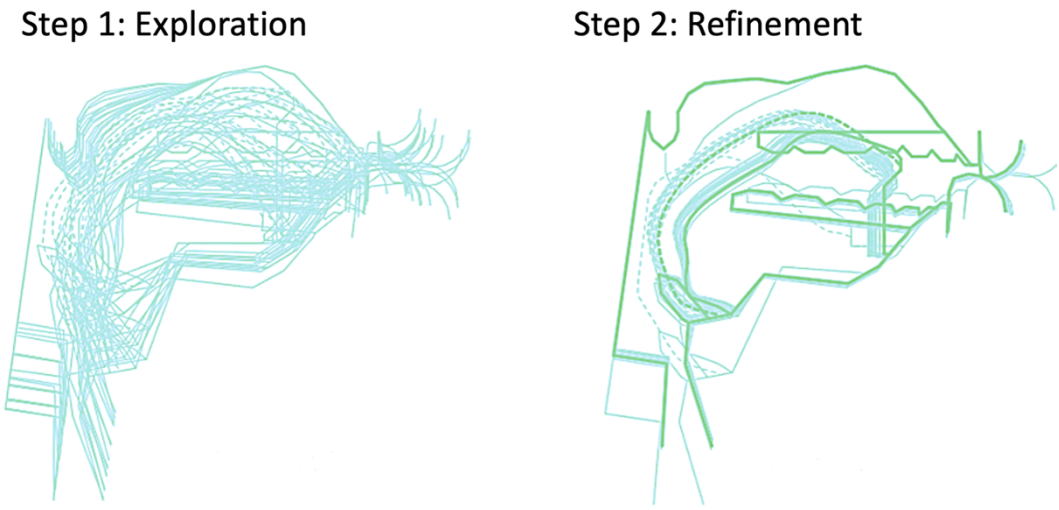


Figure 5: Optimization processes in two steps

2.7 Listening experiment

The purpose of the listening experiment is to evaluate the learnability of underlying articulatory targets. Successful acquisition is demonstrated when listeners can accurately identify the learned synthetic words containing the intended diphthongs. The speech materials used in the listening experiments included the English words learned by the vocal tract model, as well as regenerated words with shorter or longer durations. After optimization, 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 synthesized the target words with longer durations (450ms and 500ms) and shorter durations (350ms and 300ms) to examine generalizability across speaking rates. For the static targets, we proportionally increased or decreased the learned duration of the two static targets while maintaining the duration ratio and articulatory parameters. For the dynamic targets, we only adjusted the duration of the syllable to match the new duration. In total, 250 stimuli were evaluated in the listening experiment.

The listeners were 20 native American English speakers (12 male; mean age: 36) recruited and screened via Prolific[1]. The stimuli were randomized and presented to the participants using Gorilla[2]. Before the experiment, participants completed a brief questionnaire on demographic and language background. Listeners were instructed to conduct the experiment on a computer in a quiet environment wearing headphones. A headphone screening (Woods et al., 2017) was administered, followed by five practice trials. During the experiment, participants were asked to listen to each audio clip carefully, up to five times, and select the word from the five options. The experiment lasted approximately 20 minutes.

2.8 Statistical analysis

In order to compare the modeling performance of the two types of articulatory targets, we analyzed the perceptual accuracy and reaction time of the synthetic diphthongs in the listening experiment. We used generalized linear mixed models (GLMMs) to analyze whether the listeners correctly identified the target diphthongs, treated as a binary variable (TRUE or FALSE). The target type (dynamic and static), diphthong type (/aʊ/, /eɪ/, /əʊ/, /ɔɪ/, and /aɪ/), and duration (300ms, 350ms, 400ms, 450ms, and 500ms) were treated as categorical predictors. Starting with a simple model with the participant as a random intercept, we iteratively added all main effects and interactions of the fixed effects if they significantly improved the model fit, as judged by likelihood ratio tests. We used the same principle to construct a model for reaction time, which was included as a continuous variable. A series of post-hoc comparisons were conducted to examine if different levels within the significant fixed effects and interaction effects differed from each other. Tukey corrections were applied when comparing multiple estimates within a factor. The analysis was performed in R (R Core Team, 2024) using package lme4' for GLMMs (Bates et al., 2015) and emmeans (Searle et al., 1980) for post-hoc comparisons. A demonstration video, stimuli used in the perception experiment and the codes used for computational modeling and statistical analyzes can be found in https://gitlab.com/Anqi_Xu/dynamic_diphthongs.

[1] www.prolific.com

[2] gorilla.sc

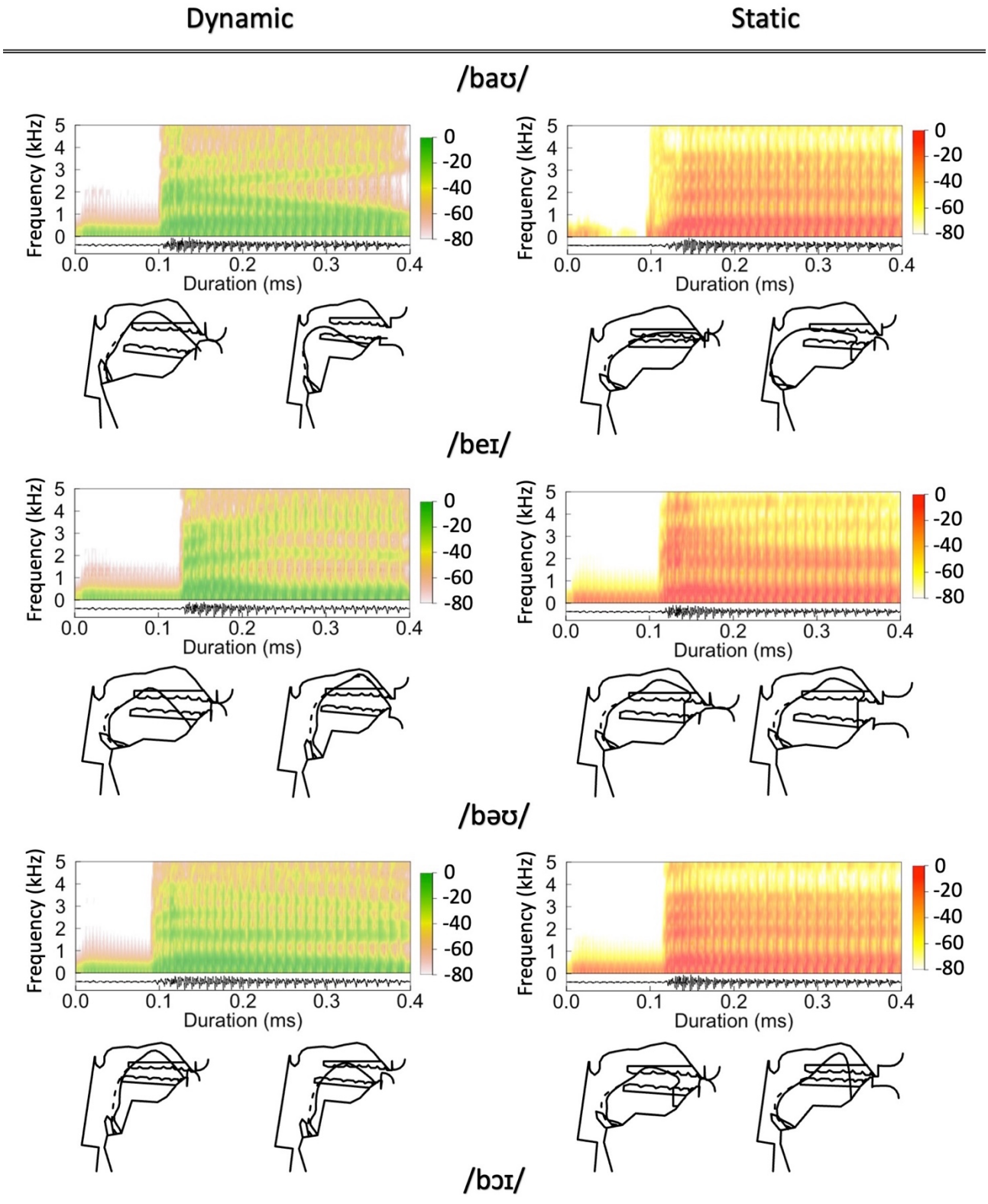
3 Results

3.1 Acoustic and articulatory analysis

We will first report the acoustic characteristics and the articulatory dynamics of the learned diphthongs synthesized by a single dynamic target and two static targets. We used the diphthongs with the lowest recognition error for each target word as examples, as illustrated in Fig. 6-8. In the spectrograms, it can be observed that the formants of /baʊ/, /beɪ/, and /bɔɪ/ based on a single dynamic target exhibit more transitional changes compared to those based on two static targets. Both /bəʊ/ and /baɪ/, regardless of the underlying target type, show deficiencies in formant movements. Nevertheless, articulations synthesized using a single dynamic target exhibited greater variation in the shape of active articulators compared to those synthesized with two static targets.

Fig. 6 also illustrates the articulatory dynamics of the learned vocal tract shapes. The first and second graphs in each row show the starting and ending vocal tract shapes of the CV syllables containing diphthongs. For example, in the case of /aʊ/, the terminating tongue shapes are alike in both conditions, but the initial tongue positions differ remarkably, with the dynamic target showing more backward movement. For the diphthong /eɪ/, the initial tongue configuration resembles that of a mid vowel, while the terminal positions are elevated in both conditions. However, the magnitude of tongue body height change is greater for the dynamic target. Both dynamic and static targets involve minimal tongue movement for /əʊ/. For /ɔɪ/, the tongue shapes are retracted at the beginning in both conditions, but the dynamic target ends at a higher and more forward position. Finally, for /aɪ/, in both static and dynamic targets, the tongue rises to the roof of the mouth or the alveolar ridge. However, the initial tongue position for /aɪ/ synthesized by the dynamic target is not as low as the one based on the two static targets.

Overall, the learned articulatory targets, both static and dynamic, exhibited starting and ending vocal tract shapes that resembled two different vowels. For the diphthongs /eɪ/ and /ɔɪ/, dynamic targets resulted in slightly greater changes in vocal tract shape compared to static targets. However, for /aɪ/, static targets led to greater articulatory movement. In contrast, the learned articulatory targets for /aʊ/ and /əʊ/ exhibited minimal movement in both conditions.


Dynamic
Static
/baʊ/
/beɪ/
/bəʊ/
/bɔɪ/
Frequency (kHz)
Duration (ms)
0
-20
-40
-60
-80

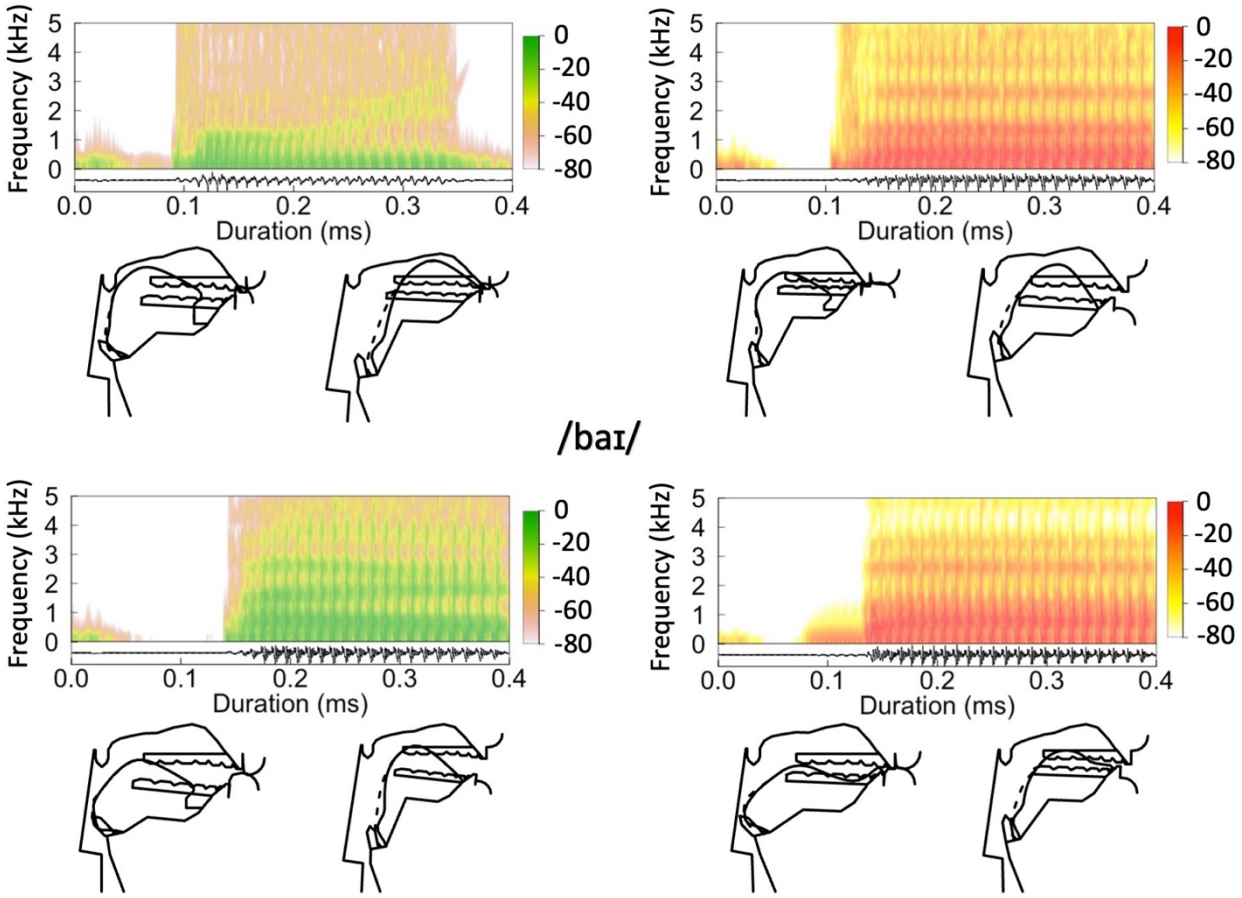


Figure 6: Learned diphthongs with the lowest recognition error by one dynamic target (left) or two static targets (right). For each diphthong, the upper panels show spectrograms and waveforms and the lower panels show vocal tract shapes at the beginning and the end of the speech utterances. The dotted line shows the lateral tongue positions.

We further analyzed simulated articulatory trajectories for diphthongs synthesized using either a single dynamic target or two static targets. The articulatory movements of sixteen vocal tract parameters are detailed in Appendix Figure A. The trajectories of /eɪ/ and /ɔɪ/ synthesized with a dynamic target exhibited substantial changes across all dimensions, whereas those of /eɪ/ synthesized with static targets showed considerably less variation, consistent with the vocal tract shapes illustrated in Fig. 6.

We also compared the horizontal and vertical tongue body positions, which are crucial for determining vowel qualities (Blackwood Ximenes et al., 2017). Fig. 7 presents the simulated articulatory trajectories of five diphthongs synthesized with either a single dynamic target or two static targets. The articulatory trajectories show that dynamic targets generally produce more continuous and fluid articulatory movements, whereas static targets result in flatter

trajectories, indicating less movement. Notably, for /eɪ/, the dynamic trajectories exhibit a larger shift in both horizontal and vertical dimensions. In contrast, the static targets tend to maintain a relatively stable tongue position, particularly evident in /eɪ/ and /əʊ/, where minimal movement is observed. These findings suggest that dynamic targets better capture the natural kinematics of diphthong production compared to static targets.

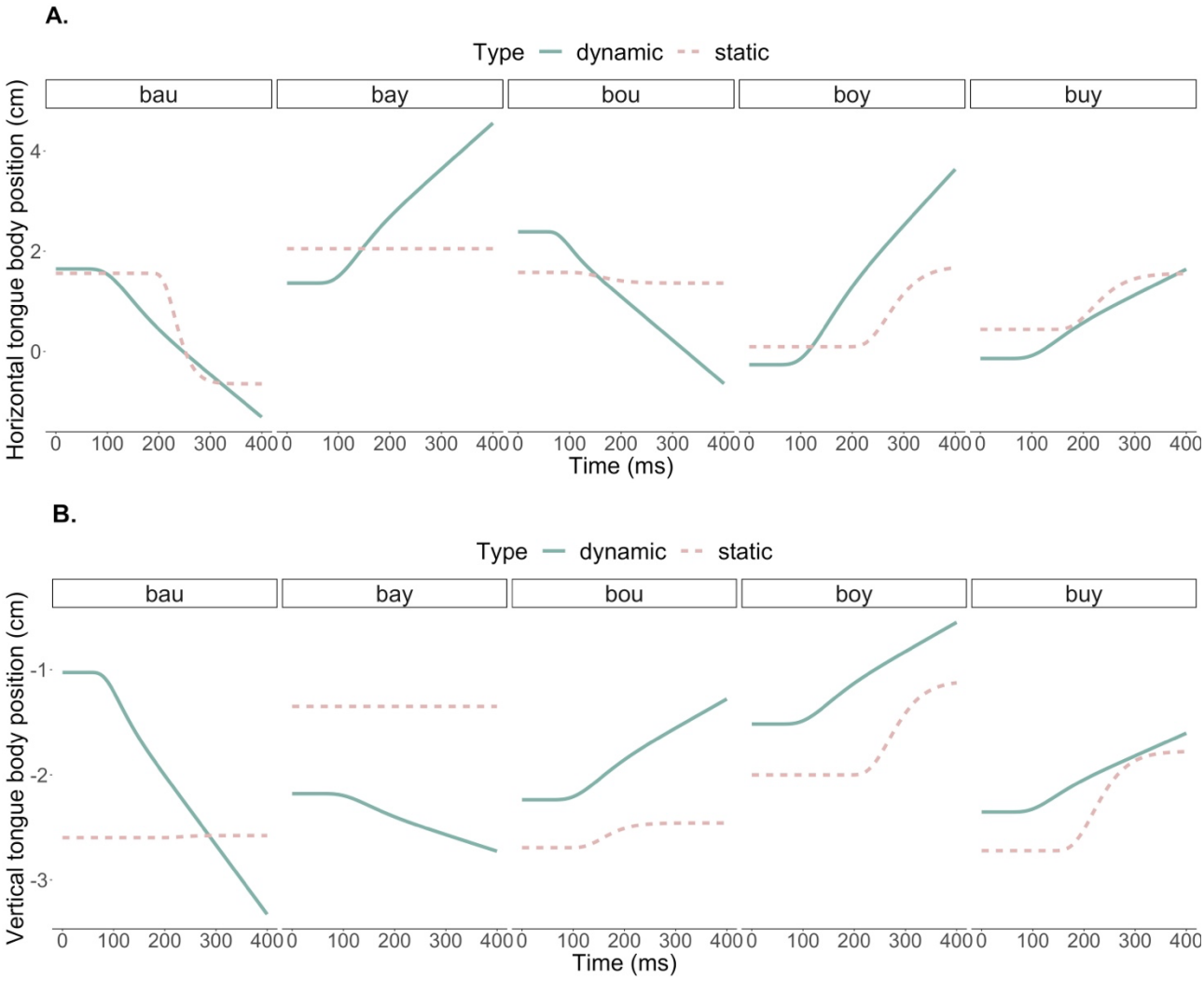


Figure 7: Simulated articulatory trajectories of five diphthongs synthesized using either a single dynamic target or two static targets. (A) shows the trajectories of the horizontal tongue body position, while (B) presents the vertical tongue body position.

Fig. 8 presents the velocity profiles of the articulatory movements shown in Fig. 7. In both the horizontal (A) and vertical (B) tongue body velocity trajectories, the dynamic targets exhibit smoother and more continuous velocity changes, whereas the static targets produce abrupt shifts characterized by discrete peaks and plateaus. Especially /aʊ/ and /ɔɪ/, the

dynamic targets result in a more fluid and sustained velocity pattern, whereas the static targets generate sharp velocity peaks followed by sudden deceleration. These suggest that the dynamic targets facilitate more natural and coordinated tongue movements, while the static targets impose more abrupt transitions between articulatory states.

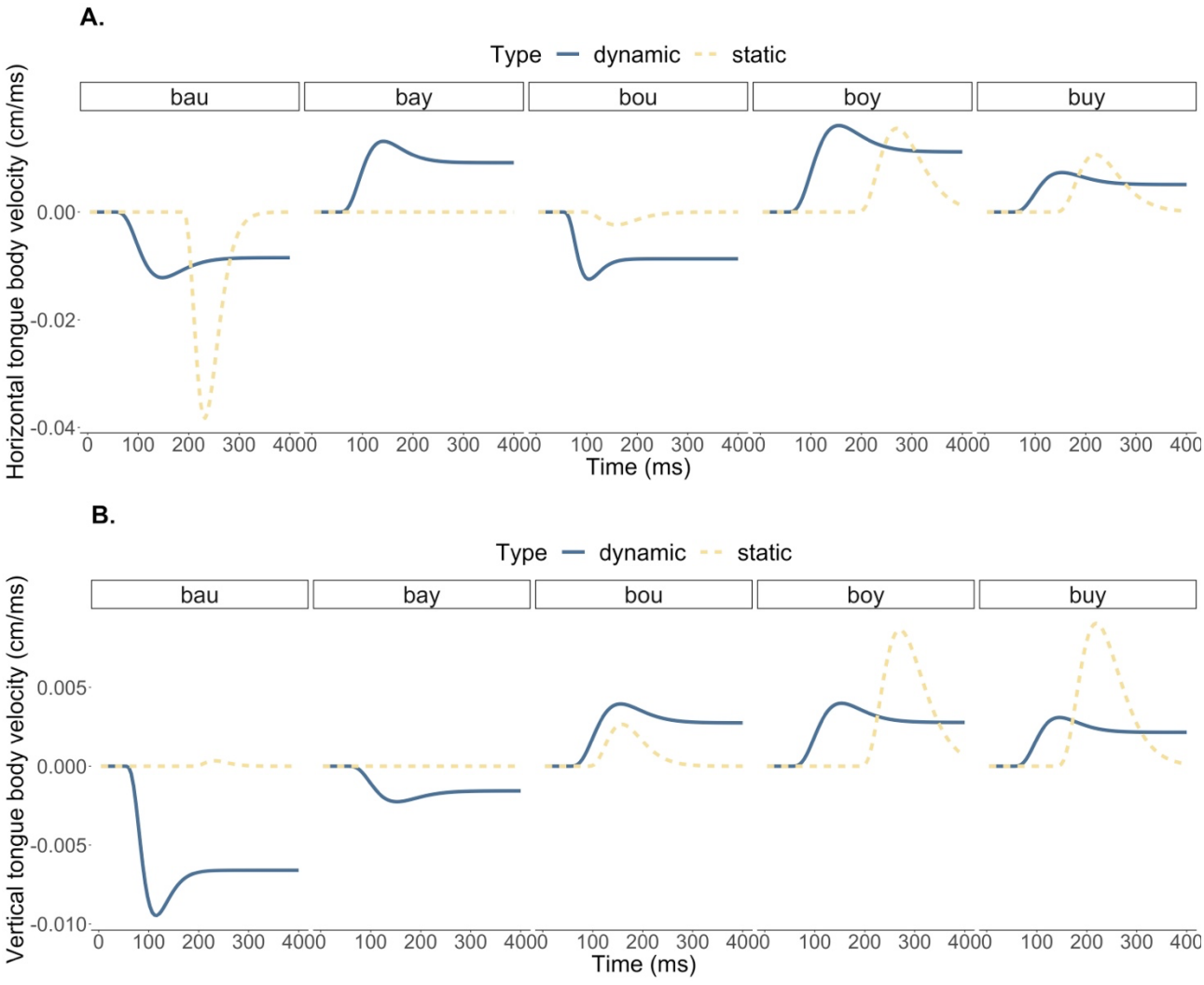


Figure 8: Simulated velocity trajectories of five diphthongs synthesized using either a single dynamic target or two static targets. (A) shows the velocity of the horizontal tongue body position, while (B) presents the velocity of the vertical tongue body position.

3.2 Intelligibility analysis

The identification rates of the learned diphthongs across target words are shown in Fig. 9. The average accuracy was 64.92% for diphthongs synthesized with one dynamic target and 34.36% for two static targets, respectively. The single dynamic target yielded diphthongs that were significantly more intelligible than those synthesized with two static targets except

for /aɪ/. GLMM showed that the main effect of target type was significant (X^2 = 493.37, *df* = 1, *p* < .001). So, the dynamic target was more advantageous than the static targets during the modeling of diphthongs. We also found that the diphthong type had a significant effect on perceptual accuracy (X^2 = 104.93, *df* = 4, *p* < .001). The accuracy was highest for /eɪ/ and /ɔɪ/, and the difference between the two was not significant (*p* = .021). Besides, /əʊ/ and /aɪ/ did not differ significantly in terms of the accuracy (*p* = .670). /aʊ/ had similar perceptual accuracy to /əʊ/ (*p* = .088) and /aɪ/ (*p* = .785). The difference between the rest of the diphthong pairs was all significant (*p* < .001). The interaction between target type and diphthong type was significant as well (X^2 = 941.22, *df* = 4, *p* < .001). /eɪ/, /əʊ/, /aʊ/ and /ɔɪ/ with a dynamic target had fairly high accuracy than the ones with two static targets (*p* < .001). In contract, the two static targets had the higher accuracy than the single dynamic target for /aɪ/ (*p* < .001).

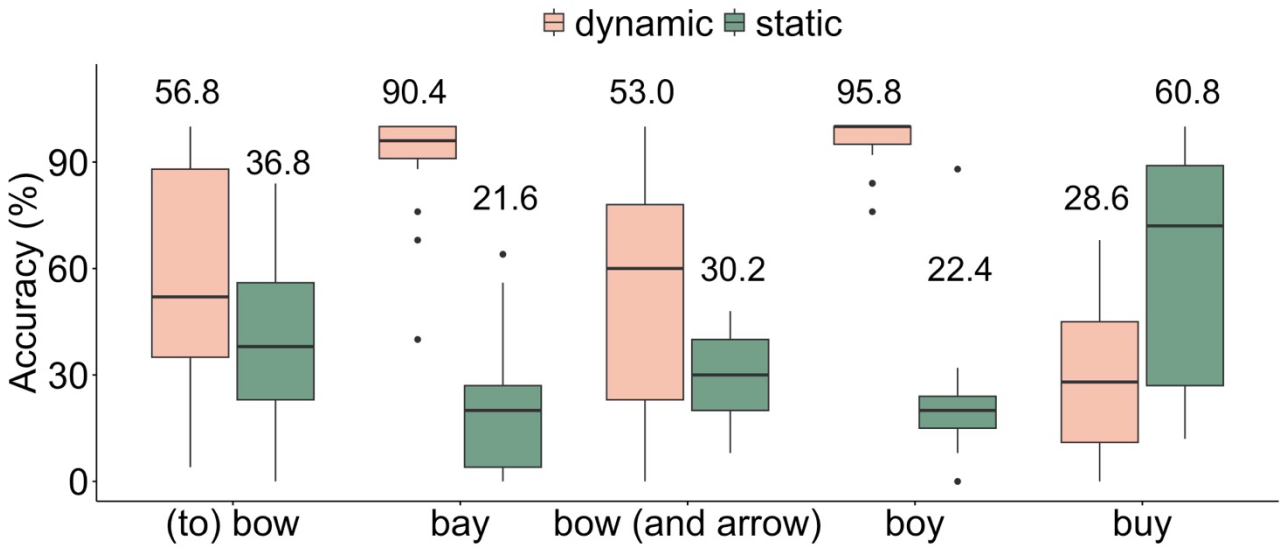


Figure 9: By-subject identification accuracy of words with different diphthongs modeled with two static targets or one dynamic target. The numbers show the mean perceptual accuracy under the two conditions.

To examine whether the learned articulatory targets can be generalized to varying speaking rates, we reused the learned static or dynamic targets to synthesize new speech utterances with different durations. The identification accuracy of the diphthongs with different duration is shown in Fig. 10. Regardless of syllable duration, the synthetic diphthongs based on a

single dynamic target performed better than the ones with two static targets. The statistical analysis also confirmed that the main effect of duration was not significant (X^2 = 1.803, *df* = 4, *p* = .772). Furthermore, both the interaction between duration and target type (X^2 = 2.914, *df* = 8, *p* = 0.940) and the interaction between duration and diphthong type (X^2 = 13.923, *df* = 20, *p* = 0.834) were non-significant. Likewise, the three-way interaction between duration, target type and diphthong type was non-significant (X^2 = 35.509, *df* = 40, *p* = 0.673).

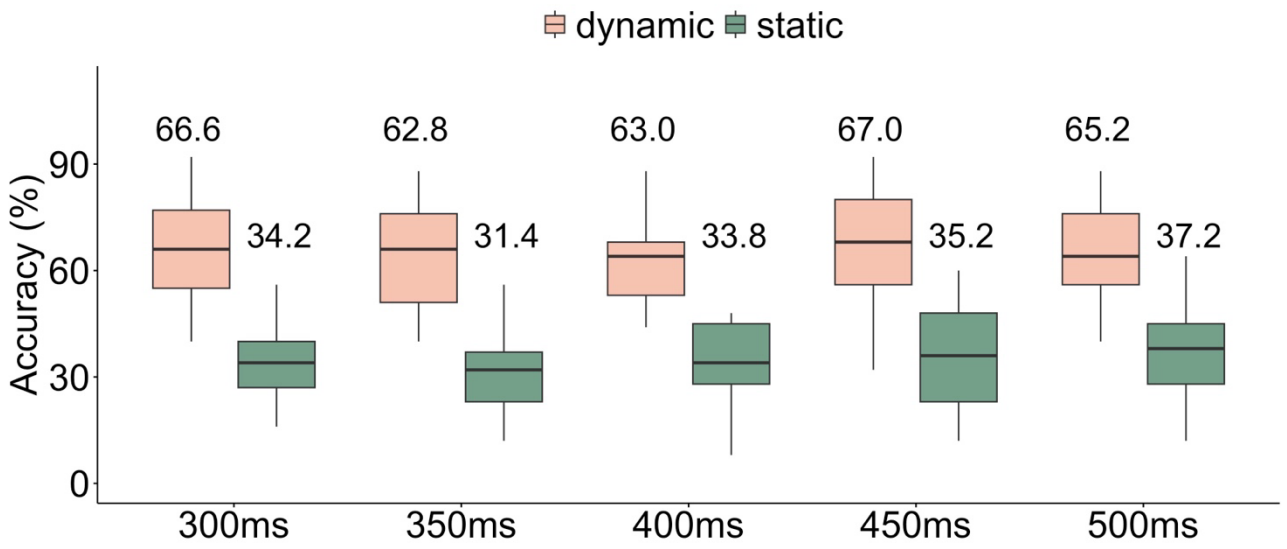


Figure 10: By-subject identification accuracy of words with diphthongs modeled with one dynamic target and two static targets across different syllable durations. 400ms was the original syllable duration and the rest of the speech utterances were synthesized using the learned articulatory targets. The numbers show the mean perceptual accuracy under the two learning conditions.

A confusion matrix of the listening experiment is shown in Fig. 11. With dynamic targets, /eɪ/ and /ɔɪ/ were nearly always correctly identified. /aʊ/ was sometimes mistaken as /əʊ/; and /əʊ/ was heard as /eɪ/ or /aʊ/. Nearly half of /aɪ/ were judged as /eɪ/ by the native listeners, while only 29% of /aɪ/ was correctly identified. In contract, more than half of /aɪ/ synthesized with two static targets was regarded as the correct diphthong. /aʊ/ was often mistaken as all the rest of the diphthongs. Participants tended to judge /eɪ/ as /əʊ/ and /aʊ/, while /əʊ/ was sometimes heard as /eɪ/. Most of /ɔɪ/ was identified as /aʊ/ and few of them was regarded as /aɪ/ or /aʊ/.

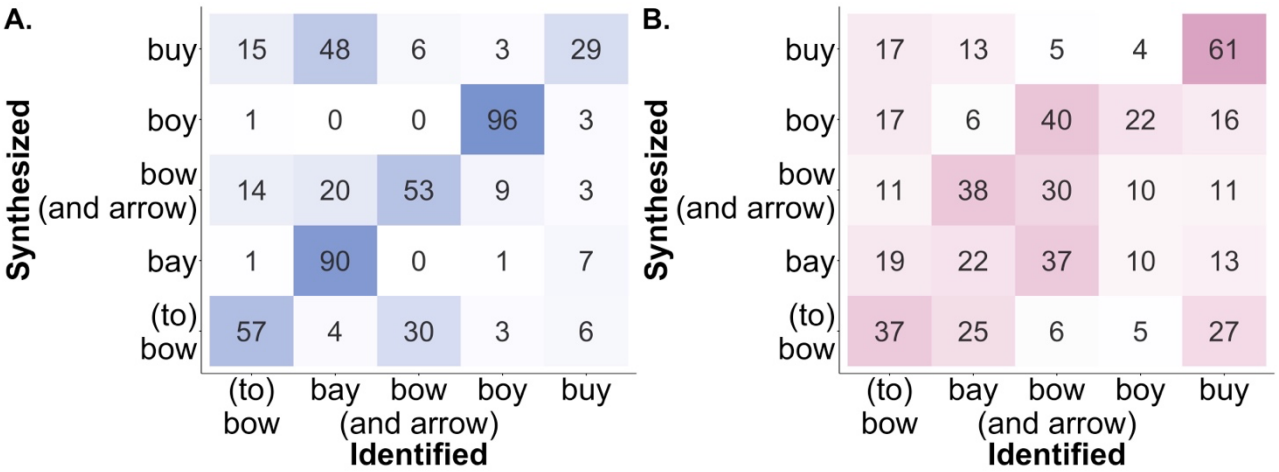


Figure 11: Confusion matrix of synthetic words with diphthongs distinguished by native listeners (A: one dynamic target; B: two static targets). The numbers indicate the percentage of correctly identified diphthongs. Darker colors indicate higher identification accuracy.

3.3 Reaction time analysis

In addition, we analyzed the reaction time of the listeners while judging the synthetic speech. The reaction time of each target diphthong synthesized either by a single dynamic target or two static targets is shown in Fig. 12. The participants spent less time judging the synthetic diphthongs based on a single dynamic target than the ones with two static targets. The statistical analysis confirmed that the main effect of target type was significant ($X^2 = 86.384$, $df = 1$, $p < .001$). We also found that the listeners spent different time identifying different types of diphthongs ($X^2 = 14.045$, $df = 4$, $p = .007$). The diphthong pairs having significant differences were the same as the ones that were statistically different in terms of perceptual accuracy. The participants needed more time to identify /aʊ/ than /eɪ/ ($p = .023$). The reaction time of /eɪ/ was also significantly shorter than /ɔɪ/ ($p = .041$). The difference between the rest of the diphthong pairs was all non-significant ($p > .050$). As shown in Fig. 12, the reaction time of static or dynamic targets was variable across the five diphthongs. The statistical analysis suggested that interaction between target type and diphthong type was significant ($X^2 = 40.628$, $df = 4$, $p < .001$). The listeners spent around the same time on identifying /aʊ/ ($p = .147$) and /əʊ/ ($p = .065$) synthesized by two static targets or one dynamic target. In contrast, for words containing /ɔɪ/ ($p < .001$), /eɪ/ ($p < .001$), and /aɪ/ ($p = .010$), the participants responded faster when judging the diphthongs synthesized by dynamic target.

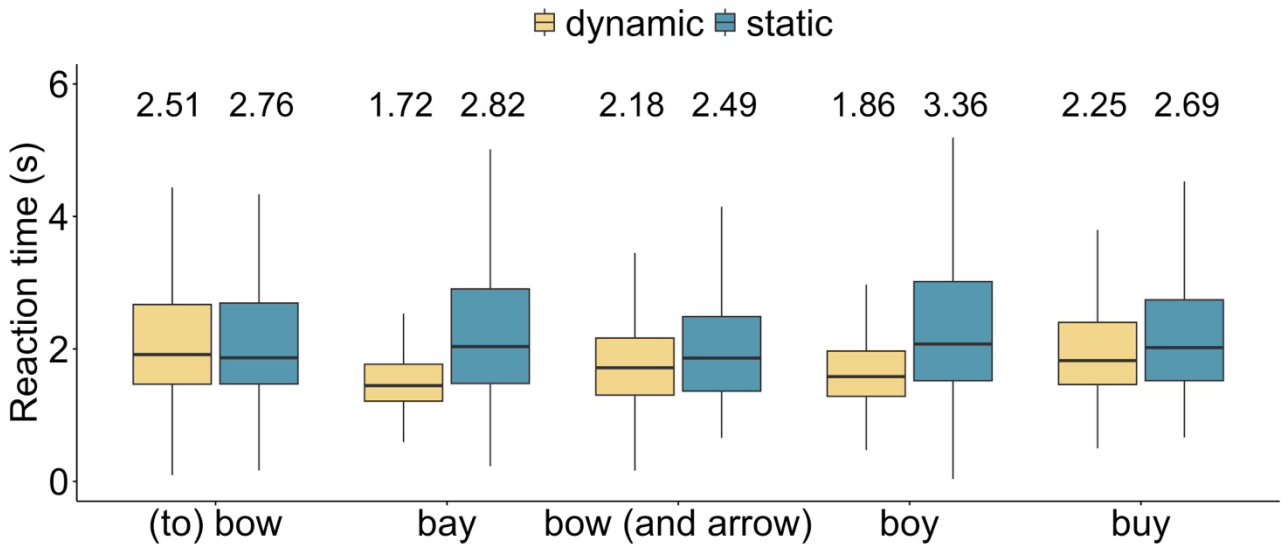


Figure 12: Reaction time of words with different diphthongs modeled with two static targets or one dynamic target. The numbers show the mean perceptual reaction time of each listening trial under the two conditions.

Again, across all the duration modulations, not only did the diphthongs with the dynamic target had shorter reaction time than those with two static targets for the original duration, but also for the longer and shorter durations. Fig. 13 shows the distribution of reaction time of participants while distinguishing synthetic diphthongs with and without durational changes. The statistical analysis showed that duration of the diphthong did not seem to affect the reaction time (X^2 = 1.334, df = 4, p = .856). Neither the interaction between duration and target type (X^2 = 3.808, df = 8, p = .874), nor the interaction between duration and diphthong type (X^2 = 12.235, df = 20, p = .908) was significant. Likewise, the three-way interaction between duration, diphthong type and target type was non-significant (X^2 = 33.112, df = 40, p = .772). To mitigate potential variability introduced by the equipment participants used, we applied a z-score transformation to each participant's data. However, this transformation did not alter the overall pattern. Additional analyses are provided in the Appendix.

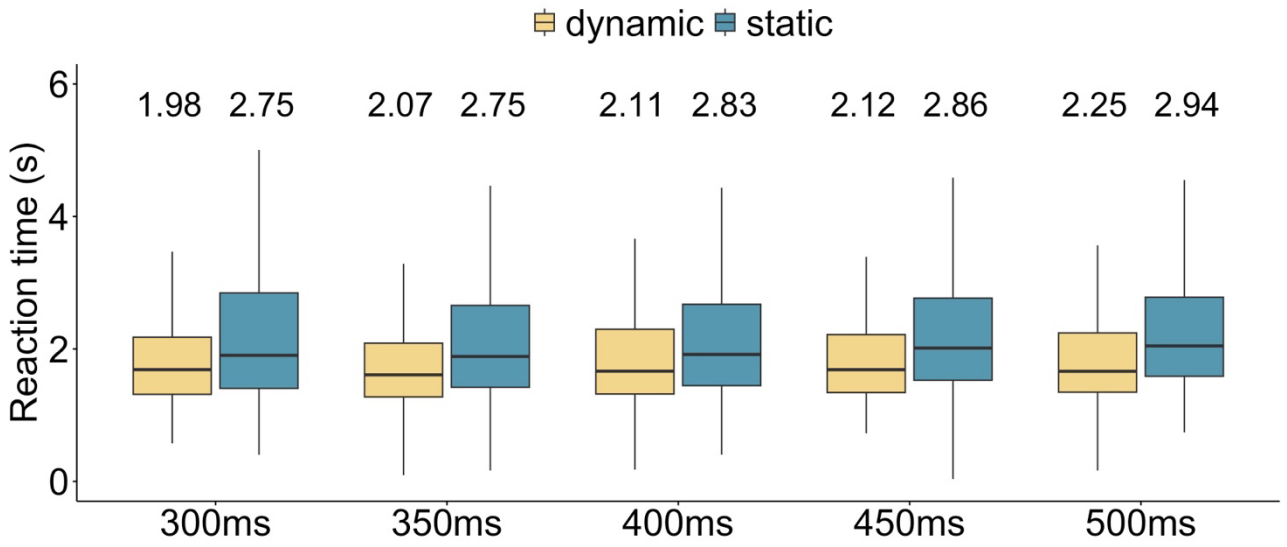


Figure 13: Reaction time of words with diphthongs modeled with two static targets or one dynamic target across different syllable duration. 400ms was the original syllable duration and the rest of the speech utterances were synthesized using the learned articulatory targets (shortened durations: 300ms and 350ms; lengthened durations: 450ms and 500ms). The numbers show the mean perceptual reaction time of each listening trial under the two conditions.

4 Discussion

We adopted a novel approach to investigate the nature of diphthongs by evaluating their learnability through computational simulations of vocal learning. With this method, we tested two hypotheses: diphthongs are articulated either with a single dynamic target or with two static targets. A vocal tract model was trained to learn English diphthongs embedded in real words, guided by a speech recognizer. The results show that unitary dynamic targets produced on average more intelligible speech with more plausible articulatory and acoustic characteristics compared to consecutive static targets, except for /aɪ/. Furthermore, when durations were used to synthesize the words with the learned articulatory parameters, the dynamic targets demonstrated consistent superiority and quicker reaction times. The simulation results suggest that dynamic targets are more easily acquired by learners, thereby providing tentative support for the hypothesis that English diphthongs are produced with unitary dynamic articulatory targets.

When analyzing the samples of the learned speech, we observed that diphthongs synthesized by dynamic targets exhibited greater modulation of formants in the spectrograms, with the exception of /aɪ/ (Fig. 6). The acoustic patterns largely correspond to the marked articulatory dynamics associated with dynamic targets. Clear gliding formants and articulatory movements were evident in /eɪ/ and /ɔɪ/ for the dynamic-target versions of the learned syllables, but not for the static-target versions. Additionally, for /əʊ/ synthesized under both conditions, we noted marginal formant movements and minimal changes in the shape of the vocal tracts, supporting previous observations by Gay (1968) and Lehiste & Peterson (1961). These results align with previous findings that acoustics and articulation are highly correlated in the production of diphthongs (Dromey et al., 2013). Furthermore, the perceptual accuracy and shorter reaction time in the listening experiment confirm that a single dynamic target is more plausible than two static targets, with only the exception of /aɪ/. Furthermore, the results also show that the diphthongs learned with single dynamic targets had better generalizability than those learned with two static targets. Under five durational conditions (300ms, 350ms, 400ms, 450ms, and 500ms), the single dynamic target exhibited higher overall intelligibility and shorter reaction times compared to the two static targets. This is consistent with the unvarying formant slopes observed by Gay (1968, 1970); Kent & Moll (1972) and Tasko & Greilick (2010).

The earliest theoretical account of diphthongs as single-unit phonemes was based on the observation that formant transitions remain relatively stable across varying speech rates (Gay, 1968). However, subsequent research revealed that the spectral rate of change can vary with linguistic prominence (Wouters & Macon, 2002), indicating that two successive vowel targets can also produce transitions that appear relatively consistent. In response to these mixed findings, several studies have employed computational simulations to model diphthong production, either through articulatory approaches (Hsieh, 2017; Strycharczuk et al., 2024) or acoustic approaches (Stone & Birkholz, 2024). Yet it remains unclear whether the specific underlying targets proposed by these models are successfully acquired by language learners. The present study addresses this issue from a learnability perspective—namely, how vocal learners develop the skill to produce intelligible diphthongs. This approach rests on the assumption that only phonetic properties which are learnable can be

maintained in a language, since unlearnable properties cannot be transmitted across generations.

We tested this learnability hypothesis using a recently developed vocal learning modeling paradigm (Krug et al., 2023; van Niekerk et al., 2023; A. Xu et al., 2024). This paradigm integrates a state-of-the-art articulatory synthesizer, which incorporates a target approximation model and consonant–vowel (CV) co-production dynamics to simulate production, with a deep-learning-based speech recognizer to provide perceptual training guidance. Through this integrated approach, we can systematically examine hypotheses with realistic speech input and output, and investigate complex interactions between production and perception—factors that are difficult to isolate or observe in behavioral studies. Our findings demonstrate that unitary dynamic targets enable the simulated learning of English diphthongs, thereby supporting the single-phoneme hypothesis by illustrating both the efficiency and feasibility of adopting a single dynamic target in speech acquisition.

There are a number of other reasons for the difficulty of simulating the learning of English diphthongs with two static vowels. First, given the 400 ms syllable duration used in the simulation, there is plenty of time for the simulated speaker to approach two successive vowel targets, as each would need a minimal time of only 125 ms (Nelson et al., 1984; Y. Xu & Prom-on, 2019). However, due to C-V coproduction (Bell-Berti & Harris, 1981; Fowler, 1980; Liu et al., 2022; Y. Xu, 2024), cf. Fig. 2 implemented in our model, the formant movements toward the first vowel are largely masked by the voiceless consonant with its long closure duration (≈100–130 ms), cf. Fig. 6. This may render the diphthong identification rate by our speech recognizer less informative for the optimization of the articulatory target parameters. Another possibility is that a single dynamic target simplifies the control process by reducing degrees of freedom—especially regarding timing in articulatory gestures. Rather than coordinating two discrete targets and managing the timing of transitions between them, learners only need maintain one overarching control scheme, thereby decreasing complexity. Regardless of the precise reason, nevertheless, the ease of finding an optimal single dynamic vowel target for diphthongs suggests that learners may not have to deal with those difficulties in the first place.

The difficulty of learning a single dynamic target for /aɪ/ is intriguing. Upon closer examination of its acoustics and articulation, we observed that the diphthongal transitions were subtle under both conditions (Fig. 6). The lack of transitional movements is surprising, as /aɪ/ typically involves more dynamic changes (Gay, 1968; Lehiste & Peterson, 1961). This anomaly could be due to the speech recognizer's high tolerance to synthetic tokens of /aɪ/ with little diphthongal formant transitions. This is likely due to the fact that the input to the recognizer was not well-controlled for accent variations (Panayotov et al., 2015) which would have allowed speakers from the southern areas of the United States to be included in the Librispeech corpus. Southern accented /aɪ/ is often spoken with /a/, resulting in shorter duration and restricted diphthongal formant movements (Weil et al., 2000; Wise et al., 1954). Additionally, /aɪ/ is sometimes realized as /aɛ/ or /a:/ in some other regional dialects (Fox & Jacewicz, 2009; Moreton, 2021), and as /ɛɪ/ by speakers from certain social groups (Crane, 1977). This variability in the speech corpus may have biased the performance of the recognizer, leading to the unexpected guidance. This may explain why /aɪ/ synthesized by the dynamic target was frequently mistaken for /eɪ/. The static targets were less negatively impacted due to their lack of formant shifts in the synthetic utterances, which resemble the static version of /aɪ/ (Fig. 6) that are acceptable to the recognizer.

The current study represents only preliminary work in using learnability to explore the nature of diphthongs, and several limitations remain. One of the limitations is that the speech data used for training the phoneme recognizer is not balanced across all speech sequences. This imbalance may have resulted in varied identification accuracy of the recognizer, potentially contributing to the uneven learning performance of the diphthongs. Another limitation is the lack of control over the duration of diphthong samples in the corpus used to train the recognizer. Many of the samples may be too short to allow the targets, whether static or dynamic, to approach their asymptotes. The exact effect of the resulting undershoot is therefore unknown. English diphthongs exhibit substantial dialectal variation, and some lack dynamic formant movements altogether (Haddican et al., 2013), suggesting that such diphthongs might be more effectively modeled with static targets. Fourth, diphthongs in certain languages function as vowel–vowel sequences (Trager & Smith, 1951), as in German, where they can be synthesized by combining monophthongal vowels (Stone & Birkholz, 2024). Applying the present method to German would be valuable in determining

whether diphthongs in that language are learnable with successive vowel targets; similar cross-linguistic extensions could also be explored in future studies. Finally, further research is necessary to clarify how different articulatory targets are encoded and stored in the brain.

5 Conclusion

We investigated whether English diphthongs have a single dynamic target or two static targets by testing their learnability in a simulated vocal learning paradigm. We used VocalTractLab, a 3D vocal tract model with built-in target approximation dynamics, and a C-V coproduction model to simulate the articulation system, and a deep-learning-based speech recognizer to simulate perceptual guidance. We simulated the learning process as optimization of articulatory parameters guided by perceptual recognition. The results of the simulations showed that diphthongs learned with dynamic targets were consistently more intelligible across variable durations than those learned with two static targets. From the perspective of learnability, therefore, we may conclude that English diphthongs are likely unitary vowels with dynamic targets rather than combinations of monophthongal vowels.

6 Acknowledgements

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

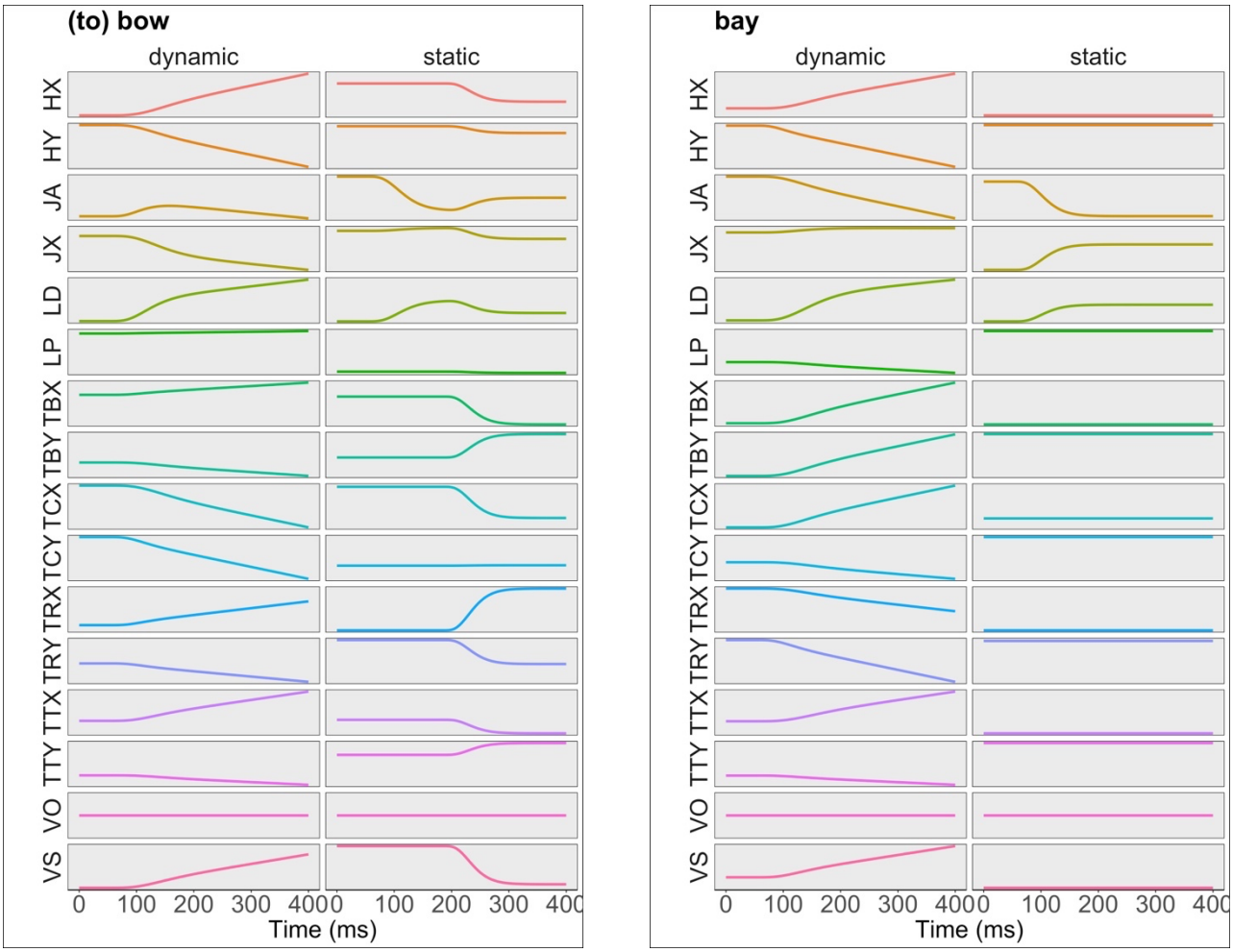
7 Appendix

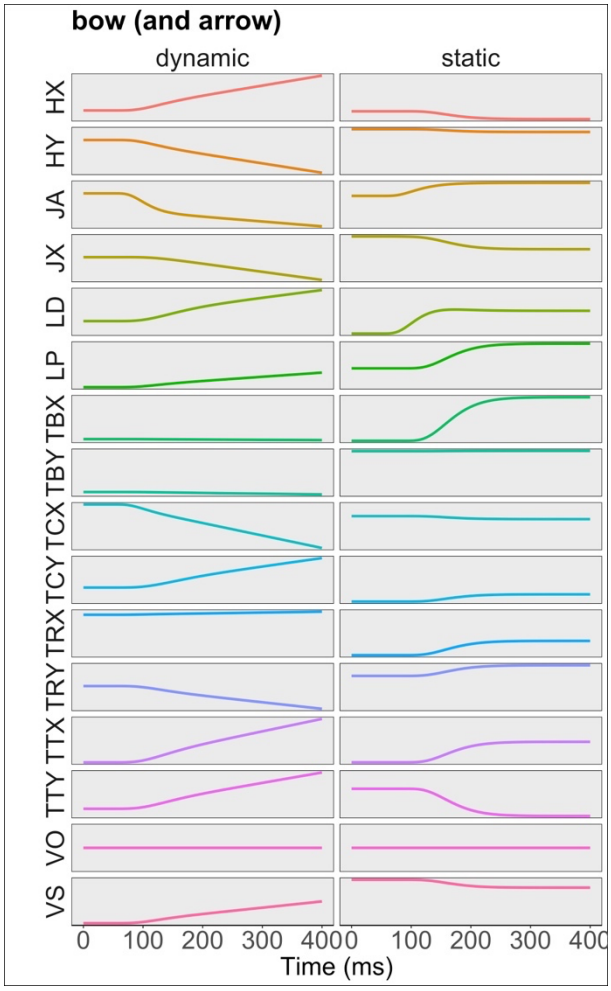
Additional analysis of reaction time

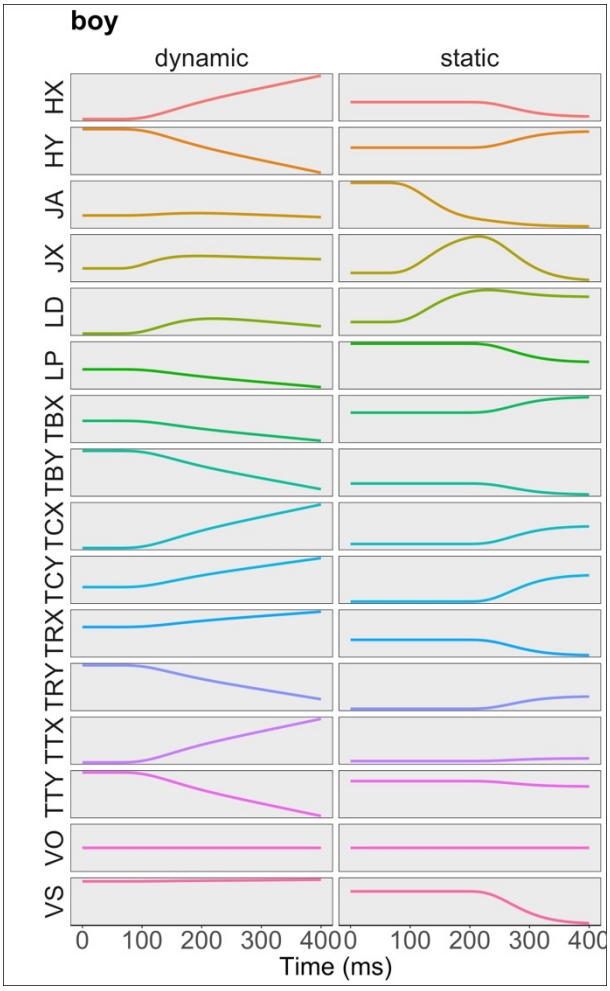
In order to counteract the possible effects of each participant's operating system and browser, we converted each individual's reaction time into z-scores and then re-examined the data. Analyses affirmed that target type remained significant ($X^2 = 160.06$, $df = 1$, $p < .001$). Furthermore, the results revealed that listeners needed different intervals to recognize distinct diphthong categories ($X^2 = 51.103$, $df = 4$, $p < .001$). Participants took more time to identify /aʊ/ than /eɪ/ and /əʊ/ ($p < .001$). Likewise, they recognized /eɪ/ faster than /ɔɪ/ ($p = .001$) and /aɪ/ ($p = .044$). Additionally, listeners spent less time identifying /əʊ/ compared to /ɔɪ/ ($p = .001$) and /aɪ/ ($p = .040$). No substantial differences emerged between other diphthong combinations ($p > .050$). The re-analysis also found a noteworthy interaction of target type with diphthong type ($X^2 = 94.41$, $df = 4$, $p < .001$). Specifically, participants

752 took comparable amount of time to identify /aʊ/ ($p = .219$) whether it was synthesized using
753 two static targets or one dynamic target. Conversely, for words containing /əʊ/ ($p = .012$),
754 /ɔɪ/ ($p < .001$), /eɪ/ ($p < .001$), and /aɪ/ ($p < .001$), responses were quicker when the
755 diphthongs were synthesized with a dynamic target.
756 The statistical analysis suggested that the duration of diphthongs did not significantly affect
757 reaction time ($X^2 = 3.193$, $df = 4$, $p = .526$). Likewise, neither the relationship between
758 duration and target type ($X^2 = 4.388$, $df = 8$, $p = .821$), nor that between duration and
759 diphthong type ($X^2 = 23.786$, $df = 20$, $p = .252$), reached significance. In addition, the three-
760 way combination of duration, diphthong type, and target type also proved non-significant (X^2
761 $= 39.075$, $df = 40$, $p = .512$).
762

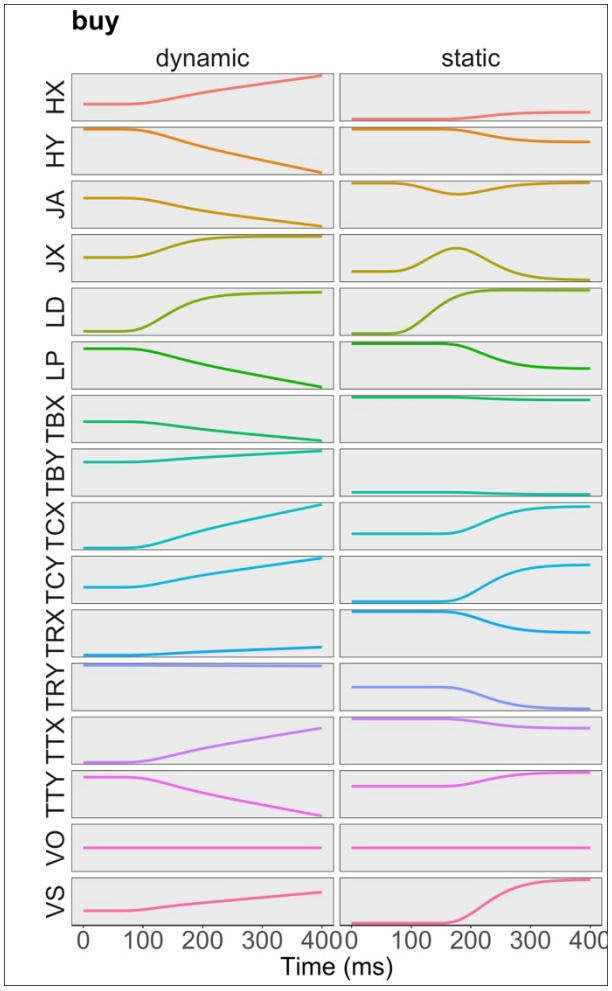
Figure A. Articulatory movements of 400-ms diphthongs synthesized using dynamic and static targets. Abbreviations for the sixteen vocal tract parameters are listed in Table 2. The y-axis represents scaled distances to normalize differences in the ranges of vocal tract parameters.




bow (and arrow)
dynamic
static
HX
HY
JA
JX
LD
LP
TBX
TBY
TCX
TCY
TRX
TRY
TTX
TTY
VO
VS
0
100
200
300
400
Time (ms)


boy
dynamic
static
HX
HY
JA
JX
LD
LP
TBX
TBY
TCX
TCY
TRX
TRY
TTX
TTY
VO
VS
0
100
200
300
400
Time (ms)

768



8 References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Bell-Berti, F., & Harris, K. S. (1981). A Temporal Model of Speech Production. *Phonetica*, *38*(1–3), 9–20. https://doi.org/10.1159/000260011

Blackwood Ximenes, A., Shaw, J. A., & Carignan, C. (2017). A comparison of acoustic and articulatory methods for analyzing vowel differences across dialects: Data from American and Australian English. *The Journal of the Acoustical Society of America*, *142*(1), 363–377. https://doi.org/10.1121/1.4991346

Bladon, A. (1985). Diphthongs: A case study of dynamic auditory processing. *Speech Communication*, *4*(1–3), 145–154. https://doi.org/10.1016/0167-6393(85)90042-1

Bond, Z. S. (1978). The Effects of Varying Glide Durations On Diphthong Identification. *Language and Speech*, *21*(3), 253–263. https://doi.org/10.1177/002383097802100304

Bond, Z. S. (1982). Experiments with synthetic diphthongs. *Journal of Phonetics*, *10*(3), 259–264. https://doi.org/10.1016/S0095-4470(19)30987-8

Browman, C. P., & Goldstein, L. (1989). Articulatory gestures as phonological units. *Phonology*, *6*(2), 201–251. https://doi.org/10.1017/S0952675700001019

Browman, C. P., & Goldstein, L. M. (1986). Towards an articulatory phonology. *Phonology Yearbook*, *3*, 219–252. https://doi.org/10.1017/s0952675700000658

Clermont, F. (1993). Spectro-temporal description of dipthongs in F1−F2−F3 space. *Speech Communication*, *13*(3–4), 377–390. https://doi.org/10.1016/0167-6393(93)90036-K

Crane, L. B. (1977). The social stratification of /ai/ in Tuscaloosa, Alabama. In D. L. Shores & C. P. Hines (Eds.), *Papers in language variation* (pp. 180–200). University of Alabama Press.

Dolan, W., & Mimori, Y. (1986). Rate-independent variability in English and Japanese complex F2 transitions. *UCLA Working Papers in Phonetics*, *63*, 125–153.

Dromey, C., Jang, G. O., & Hollis, K. (2013). Assessing correlations between lingual movements and formants. *Speech Communication*, *55*(2), 315–328. https://doi.org/10.1016/j.specom.2012.09.001

Fowler, Carol. A. (1980). Coarticulation and theories of extrinsic timing control. *Journal of Phonetic*, *8*, 113–133. https://doi.org/doi.org/10.1016/S0095-4470(19)31446-9

Fox, R. A., & Jacewicz, E. (2009). Cross-dialectal variation in formant dynamics of American English vowels. *The Journal of the Acoustical Society of America*, *126*(5), 2603–2618. https://doi.org/10.1121/1.3212921

Gay, T. (1968). Effect of speaking rate on diphthong formant movements. *The Journal of the Acoustical Society of America*, *44*(6), 1570–1573. https://doi.org/10.1121/1.1911298

Gay, T. (1970). A perceptual study of American English diphthongs. *Language and Speech*, *13*(2), 65–88. https://doi.org/10.1177/002383097001300201

Gottfried, M., Miller, J. D., & Meyer, D. J. (1993). Three approaches to the classification of American English diphthongs. *Journal of Phonetics*, *21*(3), 205–229. https://doi.org/10.1016/S0095-4470(19)31337-3

Haddican, B., Foulkes, P., Hughes, V., & Richards, H. (2013). Interaction of social and linguistic constraints on two vowel changes in northern England. *Language Variation and Change*, *25*(3), 371–403. https://doi.org/10.1017/S0954394513000197

Holbrook, A., & Fairbanks, G. (1962). Diphthong formants and their movements. *Journal of Speech and Hearing Research*, *5*(1), 38–58. https://doi.org/10.1044/jshr.0501.38

Hsieh, F.-Y. (2017). *A gestural approach to the phonological representation of English diphthongs* [Ph.D. thesis]. University of Southern California.

Kent, R. D., Kent, J. F., & Rosenbek, J. C. (1987). Maximum Performance Tests of Speech Production. *Journal of Speech and Hearing Disorders*, *52*(4), 367–387. https://doi.org/10.1044/jshd.5204.367

Kent, R. D., & Moll, K. L. (1972). Tongue Body Articulation during Vowel and Diphthong Gestures. *Folia Phoniatrica et Logopaedica*, *24*(4), 278–300. https://doi.org/10.1159/000263574

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, *220*(4598), 671–680. https://doi.org/10.1126/science.220.4598.671

Krug, P. K., Birkholz, P., Gerazov, B., van Niekerk, D. R., Xu, A., & Xu, Y. (2023). Artificial Vocal Learning Guided by Phoneme Recognition and Visual Information. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *31*, 1734–1744. https://doi.org/10.1109/TASLP.2023.3264454

Lass, R. (1984). Vowel system universals and typology: prologue to theory. *Phonology Yearbook*, *1*, 75–111. https://doi.org/10.1017/S0952675700000300

Lee, S., Potamianos, A., & Narayanan, S. (2014). Developmental acoustic study of American English diphthongs. *The Journal of the Acoustical Society of America*, *136*(4), 1880–1894. https://doi.org/10.1121/1.4894799

Lehiste, I., & Peterson, G. E. (1961). Transitions, glides, and diphthongs. *The Journal of the Acoustical Society of America*, *33*(3), 268–277. https://doi.org/10.1121/1.1908638

Liu, Zirui., Xu, Yi., & Hsieh, F. (2022). Coarticulation as synchronised CV co-onset – Parallel evidence from articulation and acoustics. *Journal of Phonetics*, *90*, 101116. doi.org/10.1016/j.wocn.2021.101116

Moreton, E. (2004). Realization of the English postvocalic [voice] contrast in F1 and F2. *Journal of Phonetics*, *32*(1), 1–33. https://doi.org/10.1016/S0095-4470(03)00004-4

Moreton, E. (2021). 2. Phonological Abstractness In English Diphthong Raising. *The Publication of the American Dialect Society*, *106*(1), 13–44. https://doi.org/10.1215/00031283-9551267

Nábělek, A. K., Ovchinnikov, A., Czyzewski, Z., & Crowley, H. J. (1996). Cues for perception of synthetic and natural diphthongs in either noise or reverberation. *The Journal of the Acoustical Society of America*, *99*(3), 1742–1753. https://doi.org/10.1121/1.415238

Nelson, W. L., Perkell, J. S., & Westbury, J. R. (1984). Mandible movements during increasingly rapid articulations of single syllables: Preliminary observations. *The Journal of the Acoustical Society of America*, *75*(3), 945–951. https://doi.org/10.1121/1.390559

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5206–5210.

Peeters, W. J. M., & Barry, W. J. (1989). Diphthong dynamics: production and perception in Southern British Englsih. *EUROSPEECH*, 1055–1058.

Peeters, W. J. W. (1996). *Dipthong dynamics: A cross-linguistic perceptual analysis of temporal patterns in Dutch, English, and German* [Ph.D. thesis]. Universiteit Utrecht.

Potter, R. K., & Peterson, G. E. (1948). The Representation of Vowels and Their Movements. *The Journal of the Acoustical Society of America*, *20*(4), 528–535. https://doi.org/10.1121/1.1906406

Prom-On, S., Birkholz, P., & Xu, Y. (2014). Identifying underlying articulatory targets of Thai vowels from acoustic data based on an analysis-by-synthesis approach. *Eurasip Journal on Audio, Speech, and Music Processing*, *2014*(1), 23. https://doi.org/10.1186/1687-4722-2014-23

Prom-on, S., Xu, Y., & Thipakorn, B. (2009). Modeling tone and intonation in Mandarin and English as a process of target approximation. *The Journal of the Acoustical Society of America*, *125*(1), 405–424. https://doi.org/10.1121/1.3037222

R Core Team. (2024). *R: A language and environment for statistical computing*. Http://Www.R-Project.Org/.

876 Saltzman, E. L., & Munhall, K. G. (1989). A Dynamical Approach to Gestural Patterning in
877 Speech Production. *Ecological Psychology*, *1*(4), 333–382.
878 https://doi.org/10.1207/s15326969eco0104_2
879 Searle, S. R., Speed, F. M., & Milliken, G. A. (1980). Population Marginal Means in the
880 Linear Model: An Alternative to Least Squares Means. *The American Statistician*,
881 *34*(4), 216–221. https://doi.org/10.1080/00031305.1980.10483031
882 Sorensen, T., & Gafos, A. (2016). The Gesture as an Autonomous Nonlinear Dynamical
883 System. *Ecological Psychology*, *28*(4), 188–215.
884 https://doi.org/10.1080/10407413.2016.1230368
885 Stone, S., & Birkholz, P. (2024). Monophthong vocal tract shapes are sufficient for
886 articulatory synthesis of German primary diphthongs. *Speech Communication*, *157*,
887 103041. https://doi.org/10.1016/j.specom.2024.103041
888 Strycharczuk, P., Kirkham, S., Gorman, E., & Nagamine, T. (2024). Towards a dynamical
889 model of English vowels. Evidence from diphthongisation. *Journal of Phonetics*, *107*.
890 https://doi.org/10.1016/j.wocn.2024.101349
891 Tasko, S. M., & Greilick, K. (2010). Acoustic and articulatory features of diphthong
892 production: A speech clarity study. *Journal of Speech, Language, and Hearing*
893 *Research*, *53*(1), 84–99. https://doi.org/10.1044/1092-4388(2009/08-0124)
894 Thomas, E. R. (2000). Spectral differences in /ai/ offsets conditioned by voicing of the
895 following consonant. *Journal of Phonetics*, *28*, 1–25. https://doi.org/10.006/jpho
896 Thompson, A., & Kim, Y. (2019). Relation of second formant trajectories to tongue
897 kinematics. *The Journal of the Acoustical Society of America*, *145*(4), EL323–EL328.
898 https://doi.org/10.1121/1.5099163
899 Tiffany, W. R. (1980). The Effects of Syllable Structure on Diadochokinetic and Reading
900 Rates. *Journal of Speech, Language, and Hearing Research*, *23*(4), 894–908.
901 https://doi.org/10.1044/jshr.2304.894
902 Tjaden, K., & Weismer, G. (1998). Speaking-rate-induced variability in F2 trajectories.
903 *Journal of Speech, Language, and Hearing Research*, *41*(5), 976–989.
904 https://doi.org/10.1044/jslhr.4105.976
905 Tjaden, K., & Wilding, G. E. (2004). Rate and Loudness Manipulations in Dysarthria.
906 *Journal of Speech, Language, and Hearing Research*, *47*(4), 766–783.
907 https://doi.org/10.1044/1092-4388(2004/058)

Trager, G. L., & Smith, H. L. (1951). *An outline of English Structure*. Battenburg Press.

van Niekerk, D. R., Xu, A., Gerazov, B., Krug, P. K., Birkholz, P., Halliday, L., Prom-on, S., & Xu, Y. (2023). Simulating vocal learning of spoken language: Beyond imitation. *Speech Communication*, *147*, 51–62. https://doi.org/10.1016/j.specom.2023.01.003

Weil, K. S., Fitch, J. L., & Wolfe, V. I. (2000). Diphthong changes in style shifting from southern english to standard american english. *Journal of Communication Disorders*, *33*(2), 151–163. https://doi.org/10.1016/S0021-9924(99)00029-5

Weismer, G. (1991). Assessment of articulatory timing. In *Assessment of speech and voice production: Research and clinical applications* (NIDCD Monograph, Vol. 1, pp. 84–95). National Institute on Deafness and Other Communication Disorders.

Weismer, G., & Berry, J. (2003). Effects of speaking rate on second formant trajectories of selected vocalic nuclei. *The Journal of the Acoustical Society of America*, *113*(6), 3362. https://doi.org/10.1121/1.1572142

Wise, C. M., Nobles, W. S., & Metz, H. (1954). The southern American diphthong /aI/. *Southern Speech Journal*, *19*, 304–312.

Woods, K. J. P., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, and Psychophysics*, *79*(7), 2064–2072. https://doi.org/10.3758/s13414-017-1361-2

Wouters, J., & Macon, M. W. (2002). Effects of prosodic factors on spectral dynamics. I. Analysis. *The Journal of the Acoustical Society of America*, *111*(1), 417–427. https://doi.org/10.1121/1.1428262

Xu, A., Birkholz, P., & Xu, Y. (2019). Coarticulation as synchronized dimension-specific sequential target approximation: An articulatory synthesis simulation. *Proceedings of the International Congress of Phonetic Sciences (ICPhS)*. https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2019/papers/ICPhS_254.pdf

Xu, A., van Niekerk, D. R., Gerazov, B., Krug, P. K., Birkholz, P., Prom-on, S., Halliday, L. F., & Xu, Y. (2024). Artificial vocal learning guided by speech recognition: What it may tell us about how children learn to speak. *Journal of Phonetics*, *105*, 101338. https://doi.org/10.1016/j.wocn.2024.101338

Xu, Y. (1997). Contextual tonal variations in Mandarin. *Journal of Phonetics*, *25*(1), 61–83. https://doi.org/10.1006/jpho.1996.0034

Xu, Y. (1998). Consistency of Tone-Syllable Alignment across Different Syllable Structures and Speaking Rates. *Phonetica*, *55*(4), 179–203. https://doi.org/10.1159/000028432

Xu, Y. (2001). Fundamental Frequency Peak Delay in Mandarin. *Phonetica*, *58*(1–2), 26–52. https://doi.org/10.1159/000028487

Xu, Y. (2024). Syllable as a Synchronization Mechanism That Makes Human Speech Possible. *Brain Sciences*, *15*(1), 33. https://doi.org/10.3390/brainsci15010033

Xu, Y., & Liu, F. (2006). Tonal alignment, syllable structure and coarticulation: Toward an integrated model. *Italian Journal of Linguistics*, *18*, 125–159.

Xu, Y., & Prom-on, S. (2019). Economy of Effort or Maximum Rate of Information? Exploring Basic Principles of Articulatory Dynamics. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.02469

Xu, Y., & Wang, E. Q. (2001). Pitch targets and their realization: Evidence from Mandarin Chinese. *Speech Communication*, *33*(4), 319–337. https://doi.org/10.1016/S0167-6393(00)00063-7