Journal of Phonetics 107 (2024) 101349

Contents lists available at ScienceDirect

Journal of Phonetics

journal homepage: www.elsevier.com/locate/Phonetics

Towards a dynamical model of English vowels. Evidence from diphthongisation

Patrycja Strycharczuk [a,*], Sam Kirkham [b], Emily Gorman [b], Takayuki Nagamine [b]

[a] Linguistics and English Language, University of Manchester, United Kingdom
[b] Linguistics and English Language, Lancaster University, United Kingdom

ARTICLE INFO

Article history:
Received 7 July 2023
Received in revised form 8 July 2024
Accepted 18 July 2024
Available online xxxx

Keywords:
Vowels
Diphthongs
Representation
Articulatory Phonology
EMA

ABSTRACT

Diphthong vowels exhibit a degree of inherent dynamic change, the extent of which can vary synchronically and diachronically, such that diphthong vowels can become monophthongs and *vice versa*. Modelling this type of change requires defining diphthongs in opposition to monophthongs. However, formulating an explicit definition has proven elusive in acoustics and articulation, as diphthongisation is often gradient in these domains. In this study, we consider whether diphthong vowels form a coherent phonetic category from the articulatory point of view. We present articulometry and acoustic data from six speakers of Northern Anglo-English producing a full set of phonologically long vowels. We analyse several measures of diphthongisation, all of which suggest that diphthongs are not categorically distinct from long monophthongs. We account for this observation with an Articulatory Phonology/Task Dynamic model in which diphthongs and long monophthongs have a common gestural representation, comprising two articulatory targets in each case, but they differ according to gestural constriction and location of the component gestures. We argue that a two-target representation for all long vowels is independently supported by phonological weight, as well as by the nature of historical diphthongisation and present-day dynamic vowel variation in British English.

1. Introduction

Diphthong vowels are characterised by some degree of inherent dynamic change over the course of the vowel, a salient property that needs to be captured by models of vowel production. The traditional theoretical device used to reflect this phonetic property of diphthongs on a more abstract level is compositionality. By 'compositionality' we mean modelling diphthongs as consisting of two component elements, with no further assumptions about the nature of these elements. Note that this is a broader sense than adopted by some other works, such as Hsieh (2017), who defines 'compositionality' as being composed from two elements that function independently within the same system. Compositionality is evoked by the etymology of the word *diphthong* (from Greek *diphthongos*, 'two sounds'), it has been incorporated into phonological models of diphthongs across different theoretical frameworks, it is implicit in IPA transcription, and it is also implicit in the common descriptive practice of using two time points to represent diphthongs. Crucially, the compositional nature of diphthongs is typically conceived of in opposition to monophthongs, which consist of a single component element, represented by a single IPA symbol, and commonly reduced to a single time point in phonetic measurements, normally the acoustic midpoint. While the descriptive tradition may be a matter of convenience, theoretical models that employ a distinction between compositional diphthongs and single component monophthongs predict categorical differences between the two types of sounds.

1.1. Compositionality in models of diphthongs

An example of a phonological model positing a distinct representation for monophthongs and diphthongs is the Autosegmental Phonology model by Goldsmith (1990), who proposes that diphthongs consist of two segments (two root nodes), each linked to a single timing unit, whereas monophthongs consist of a single segment (a single root node), which however, may be linked to two timing units, as is the case with long monophthongs. This constitutes a systematic structural differ-

* Corresponding author at: Department of Linguistics and English Language, University of Manchester, Oxford Rd, Manchester M13 9PL United Kingdom.
E-mail address: patrycja.strycharczuk@manchester.ac.uk (P. Strycharczuk).

https://doi.org/10.1016/j.wocn.2024.101349
0095-4470/

ence between diphthongs and long monopthongs, in that either two or one root nodes are present.

A similar distinction is made by some Articulatory Phonology/ Task Dynamics (AP/TD) models of vowel representation, except the relevant component elements are not segments, but gestures. Within AP/TD, articulatory representations consist of discrete gestures (Browman and Goldstein, 1986; Browman and Goldstein, 1992; Saltzman and Munhall, 1989). Each gesture has a specific spatio-temporal target, whereas transitions arise from movement from one target to another, constrained by a degree of overlap between the gestures. For diphthongs specifically, multiple accounts postulate two component elements. For example, Marin (2007) proposes that complex nuclei effects in diphthongs, such as specific patterns of syllable weight and stress alternations, can be accounted for in AP/TD through a mechanism of gestural coupling originally proposed to explain syllable organisation. She develops a model for Romanian mid diphthongs /ea/ or /oa/, in which the diphthongs are composed of two distinct synchronous (in-phase) vocalic gestures. Variation in the degree of diphthongisation is modelled using blending strength. If the two gestures have an equal blending strength, the output is a monophthong, intermediate between the two gestures. However, increasing the blending strength of one of the two gestures produces a diphthong characterised by observable inherent change. Marin proposes that variation in blending strength arises through additional factors, such as stress or speech rate, and this accounts for the phonetic variation observed in Romanian. This model is not claimed to be a universal representation of all diphthong vowels, but it is suggested that it could be extended to some types of diphthongs in other languages. One of the examples provided is the ongliding sequence /juː/ in American English (as in *beauty* or *few*), which is analysed as a sequence of two vocalic elements that are synchronously coordinated. In contrast, offgliding diphthongs (like PRICE or CHOICE) are said to consist of a nucleus vowel and a coda glide that are coupled sequentially to each other (anti-phase). A key argument for distinguishing between the two types of diphthongs is syllable weight. The onglide /j/ in /juː/ is non-moraic, which supports the in-phase organisation between the two elements. In contrast, offgliding diphthongs are bimoraic, which supports a coda-like sequential organisation.[1] Marin does not specifically discuss the representation of long monophthongs vis-à-vis diphthongs, but she models the offglide /uː/ in *few*, which is long, as corresponding to a single gesture. This suggests that the primary structural difference between diphthongs and long monophthongs is the presence of two vs. one gesture, a gestural equivalent of the autosegmental distinction discussed above.

This type of distinction is also adopted in TADA (Task Dynamics Application), the computational model of AP/TD (Nam et al., 2004). In its current implementation, TADA represents diphthongs, such as PRICE using two targets: a nucleus and an offglide; whereas monophthongs, such as THOUGHT, are represented using a single gestural target. Notably, high long vowels (FLEECE, GOOSE) are represented using two gestures, similarly to canonical diphthongs. We note that this representation may not represent a firm theoretical commitment, and we can only speculate about the rationale behind adopting it, but it is likely that two gestural targets are used for FLEECE and GOOSE to accommodate the small degree of dynamic change that characterises such vowels, i.e. a form of gradient diphthongisation.

1.2. Gradience in diphthongisation

Diphthongisation frequently shows gradient characteristics. Below, we discuss some manifestations of this phenomenon, before considering the relevance of gradient diphthongisation to diphthong compositionality.

Firstly, it is very easy to find examples of vowels that are intermediate between monophthongs and diphthongs, i.e. vowels that show a perceptible degree of dynamic change, which is however small. For example, Sweet (1910, p.11) makes a distinction between 'half-diphthongal' vowel sounds [ei] (FACE) and [ou] GOAT, as opposed to 'fully-diphthongal' [ai] (PRICE) and [oi] (CHOICE). A three-way split according to degree of diphthongisation is also proposed by Lehiste and Peterson (1961) for American English, who use a systematic acoustic approach to quantify degree of diphthongisation. They distinguish between diphthongs, which have two acoustic targets (defined as two acoustic steady states), monophthongs, which have one, and an intermediate category of glides. Glides are characterised by a single target, but also by formant transitions that are inherent to the vowel (as opposed to being transitions from the preceding or into the following consonants). The vowels identified as glides in that way are American English FACE, GOAT and NURSE. In contrast, PRICE, MOUTH and CHOICE are classified by Lehiste and Peterson (1961) as diphthongs. A number of sources adopt a similar three-way distinction that acknowledges a special status of PRICE, MOUTH and CHOICE in English as 'true', 'full' or 'phonemic' diphthongs (Assmann et al., 1982; Jacewicz and Fox, 2012; Morrison and Nearey, 2007). English is not unique in having varying levels of diphthongisation in its vowel inventory. For example, in Dutch, a distinction has been proposed between 'potential diphthongs', /ei/, /øy/ and /ou/, and 'essential' diphthongs, /ɛi/, /œy/ and /ɑu/ (Collins and Mees, 2003).

Since gradient diphthongisation is by definition phonetically small, discrepancies arise between sources, according to which vowels are treated as diphthongs when a binary monophthong-diphthong split is adopted. Consider, for example, the transcription of the FLEECE and FACE vowels in three varieties of English, as captured by the IPA Illustrations of Received Pronunciaction (RP; Roach (2004)), American English (Ladefoged, 1999) and Australian English (Cox and Palethorpe, 2007). The FLEECE vowel is transcribed as monophthongal /i(ː)/ in all three, although Ladefoged (1999) suggests that /iʲ/ is a possible transcription. In comparison, FACE is transcribed as /eɪ/ in Roach (2004), as /e/ in Ladefoged (1999), although /eʲ/ is also mentioned as a possibility, and as /æɪ/ in Cox and Palethorpe (2007). It is somewhat unclear to what extent these discrepancies arise due to differences in transcription conventions, and to what extent they are driven

[1] There is a further complicating factor in that /j/ and /w/ in offgliding diphthongs are considered consonantal in this account. This assumption is somewhat controversial, as Burgdorf and Tilsen (2021) show that difference between vowels and glides can be modelled through their temporal properties rather than an inherent spatial distinction. The issue of whether glides are consonantal or not is not crucial to the idea that ongliding diphthongs differ from offgliding diphthongs in their gestural organisation.

by differences in phonetic quality of these vowels between different varieties. Some diphthongisation of FLEECE is present in all three varieties, and it is quite salient in Australian English (Cox and Docherty, 2023), yet the transcription tends not to reflect it. Thus, transcribing FLEECE as a monophthong seems to be driven by convention, as indeed acknowledged by Ladefoged (1999). In contrast, FACE is transcribed variably with one or two symbols, and here some phonetic differences are likely at play. For example, FACE is clearly more diphthongised in Australian English compared to American English, so this may explain why the vowel would be classified as a diphthong in the former case and as a monophthong in the latter.

Variable diphthongisation is a perennial feature of English, such that the same vowel phoneme may have a monophthongal quality in one variety, but diphthongal in another, and we also find that monophthongs and diphthongs can morph into each other fairly freely. Northern Anglo-English, which we shall focus on in this study, has been noted for a relatively large degree of variation in the degree of diphthongisation. The monophthongal quality of FACE and GOAT is mentioned as a distinguishing accent feature in descriptions of traditional varieties of Lancashire and Yorkshire English (Wells, 1982), and also a possible variant in Newcastle English (Watt, 2002). A monophthongal quality of PRICE can also be a feature in these dialects, as well as in Liverpool English (Watson, 2007). In contrast, FACE, PRICE and GOAT are considered diphthongs in Manchester English (Baranowski and Turton, 2015). Alongside the local varieties, many speakers in the North of England speak a pan-regional variant, General Northern English, in which PRICE, MOUTH, CHOICE, FACE and GOAT are acoustically clearly diphthongal (Strycharczuk et al., 2020). According to Honeybone (2007), SQUARE is generally a monophthong in the North of England, as confirmed by Strycharczuk et al. (2020), who also report a monophthongal quality of NEAR in Manchester English, but not in other urban locations in the North of England. Williams and Escudero (2014) note a diphthongal quality of GOOSE in a sample of speakers from Sheffield. While these reports are based on different criteria and different methodologies, they clearly suggest considerable dynamic variability, as some vowels are likely to cross the divide between monophthongs and diphthongs, depending on the dialect, speaker, and potentially style.

1.3. Gradient diphthongisation as a challenge to diphthong compositionality

So far we have demonstrated that gradient diphthongisation is common in English, which makes it difficult to separate monophthongs from diphthongs. This has practical consequences for choices guiding phonetic analysis, and theoretical consequences for phonological models of vowels.

We have already discussed the issues related to transcription: gradient diphthongisation forces the transcriber to make a somewhat arbitrary choice between using one or two symbols. Phonetic transcription is, of course, to some extent arbitrary, but when it comes to vowel dynamics, it may have analytical consequences. It is common methodological practice to represent vowel quality in diphthongs using two time points, whereas a single time point is used for monophthongs. The IPA illustrations discussed above in Section 1.2 serve as an example. It is also common to follow earlier description and transcription conventions in deciding which vowels are monophthongs and which vowels are diphthongs. In cases where gradiently diphthongised vowels are treated as monophthongs, this results in a loss of important dynamic information. The problem of dynamic reduction has received considerable attention in recent years, with many phonetic studies of vowel reduction moving away from temporal reduction, and not attempting an a priori distinction between monophthongs and diphthongs. For example, many recent studies analyse entire formant trajectories, using techniques such as Generalised Additive Modelling, Discrete Cosine Transformation, or Smoothing Splines ANOVA (see Cox and Docherty (2023) for a recent overview).

While methodological practice can simply avoid dealing with gradient diphthongisation by treating monophthongs and diphthongs alike, representational models of vowels must take a stance on how to distinguish monophthongs from diphthongs, whilst also capturing gradient diphthongisation. If we assume that the structural difference between monophthongs and diphthongs lies in the underlying number of gestures, then diachronic diphthongisation would involve insertion of a gestural target, whereas monophthongisation would involve deletion of a target.[2] Either insertion or deletion would constitute major restructuring, and from that point of view, we might expect such changes to be somewhat constrained. However, as we have already discussed, the opposite is the case: such changes are very common, and it is also common to find stylistic variation within a single speaker between a monophthongal and diphthongal version of the same vowel phoneme. Furthermore, changes in the degree of diphthongisation usually do not entail changes in phonological patterning, such as phonological weight, or phonotactic constraints. In English, long monophthongs and diphthongs function structurally in the same way. Thus, from the point of view of phonology, as well as from the point of view of variation and change, there is no strong argument to distinguish between long monophthongs and diphthongs.

Some existing phonological accounts include gradiently diphthongised vowels in the diphthong category, such that effectively, most long vowels are treated as diphthongised. Popescu and Chitoran (2022) propose that diphthongised long vowels comprise two targets, similarly to diphthongs. They also propose that different degrees of diphthongisation may emerge, depending on how similar the component targets are. When the component targets are different, the vowel is clearly diphthongal. In contrast, when the component targets have a similar constriction location, the result may be a slightly diphthongised vowel. This proposal extends a two-target representation to instances of gradient diphthongisation, similarly to the TADA approach (Section 1.1 above).

Some variationist literature goes one step further. Labov et al. (2006) adopt a two-way split within the North American English vowel inventory: short vs. long vowels. All long vowels are treated as inherently diphthongal, and it is said that a degree of diphthongisation may emerge for all long vowels in

[2] The generalisation and the problem are much the same for an autosegmental representation which uses the number of root nodes to distinguish between monophthongs and diphthongs.

the final position. The diphthongal interpretation of long vowels is also reflected in those vowels being represented by a pair of symbols, whereas a single grapheme is used for short vowels. This system is said to be closely based on Wells (1982), who proposes a split between long and short vowels, with further subdivision of long vowels into upgliding and ingliding ones. While this distinction is not reflected in the Wells's transcription system for American English, this notational decision is attributed by Labov et al. (2006) to convention rather than to mark absence of any diphthongisation. Interestingly, seeds of the idea that diphthongs and long monophthongs are structurally the same may be found even further back. Trubetzkoy (1971, p.173) discusses a case of Slovak vowels, for which he argues that "long nuclei are interpreted as monosyllabic combinations of two like vowels".

An alternative theoretical proposal for dealing with gradient diphthongisation could be enriching the representation with dynamic detail, such that the phonology stores continuous dynamic information rather than abstract targets from which the dynamics emerge in speech. This type of rich representation would not need to make a distinction between diphthongs and monophthongs, instead treating all vowels as dynamic. Xu et al. (2023) present a model along these lines, specifying diphthong representations through a combination of targets and slopes that determine the direction and range of articulatory movement. This type of articulatory modelling represents a radical departure from diphthong compositionality. While the model has not yet been fully developed or validated, it represents a possible conceptual direction for theories of vowel representation that seems compatible with gradient diphthongisation.

The debate concerning diphthong compositionality in articulation closely resembles a similar debate about the status of diphthongs in the perceptual domain. The observations concerning the compositional nature of diphthongs were originally based on auditory properties of vowels, and auditory judgements are still frequently used when classifying vowels as diphthongs or monophthongs. Perceptual research confirms that diphthong vowels can be successfully identified when reduced to two time points, but not one, whereas increasing temporal resolution further does not substantially improve vowel identification (Jibson, 2022). While these findings are apparently in line with the interpretation that diphthongs have two distinct perceptual targets, this advantage of two point models over single point ones may not be not unique to diphthongs: a similar advantage is observed for ostensible monophthongs by Hillenbrand (2013) and Jibson (2022), although Harrington and Cassidy (1994) present divergent findings.

In summary, the idea that diphthongs have two component elements is very well established in phonology and phonetics, as evident from theoretical models and common methodological practice in phonetics. However, any model of vowel representation that incorporates this idea must specify which vowels are diphthongs, and should therefore be modelled as containing two elements. Such a decision may not be straightforward for any given vowel inventory, because diphthongisation appears to be phonetically gradient. Gradient diphthongisation represents a potential area of overlap between monophthongs and diphthongs in which a vowel shows a small degree of inherent change. Based on the proposals in previous literature, there are broadly three ways of capturing gradient diphthongisation in models of vowel representation.

i. Canonical diphthongs are modelled as compositional, as are instances of gradient diphthongisation. In contrast, canonical long monophthongs are modelled as having one component element. This approach incorporates a structural distinction between long monophthongs and diphthongs, and it is consistent with Popescu and Chitoran (2022) and the current TADA model for American English.
ii. All long vowels are modelled as compositional, i.e. all long vowels are inherently diphthongs. This possibility presupposes no categorical distinction between monophthongs and diphthongs, and it is consistent with Labov et al. (2006) for American English.
iii. All vowels are specified for target and trajectory of movement. This approach requires no separation between monophthongs and diphthongs, and it also entails that diphthongs are not inherently compositional.

Our study sets out to inform the discussion about the empirical accuracy of these approaches and their theoretical advantages and disadvantages. The novel perspective we offer comes from a systematic articulatory study of vowel diphthongisation and from articulatory modelling.

1.4. This study

We present a systematic articulatory investigation of diphthongisation across the long vowel subsystem, based on electromagnetic articulography (EMA) data from six speakers of Northern Anglo-English. This variety was chosen because it is rich in potential instances of gradient diphthongisation, as reflected by the considerable dynamic variation. We analyse the dynamic properties of all long vowels in order to quantify their relative degree of diphthongisation, using both articulatory and acoustic metrics. The two main research questions guiding this analysis are as follows.

1. Are articulatory properties of diphthongs consistent with a model in which diphthongs comprise two distinct gestural targets?
2. Are diphthongs categorically distinct from monophthongs, or is diphthongisation gradient in the articulatory domain? We conceive of gradience as phonetic continuity in measures that capture degree of diphthongisation. Thus, we operationalise our question as a classification problem: Can we systematically distinguish monophthongs from diphthongs, using articulatory and acoustic diagnostics?

As far as we know, no study to date has documented articulatory properties of diphthongisation across multiple vowels in the same language/variety. This creates a major gap in our understanding of the nature of diphthongisation, because gradient diphthongisation may in principle arise in acoustics and perception while there is a categorical difference between diphthongs and monophthongs at the articulatory level. A possible conceptualisation of this dichotomy is offered by Strange (1989), who argues that listeners are sensitive to dynamic information that emerges from gestural vowel dynamics, which includes the gestural target, but also the opening and closing phase. In this sense, vowel targets could be seen as dynamic from the perceptual point of view, but categorical at the under-

lying articulatory level, manifested as the presence or absence of component articulatory gestures.

An additional argument for considering articulatory evidence is the fact that the presence or absence of articulatory targets has a stable empirical correlate, under the core assumptions of Articulatory Phonology. In Articulatory Phonology, gestures are abstract units of organisation, which are however, measurable because they have systematic physical correlates. The key correlate of a gesture is movement towards a specific articulatory position (gestural maximum), followed by a change of direction of the articulatory movement (gestural release). This framework also provides a principled way of reconciling categorical and gradient aspects of speech, as it explicitly models continuous speech signals as emerging systematically from a combination of gestural targets, which are categorical units. In this context, we ask whether we can reconstruct the underlying number of gestural targets from the articulatory data, and whether this allows us to classify all vowels as having one or two component targets. This question informs our theoretical perspective on diphthong compositionality, understood as diphthongs being composed of two gestural targets (two independently timed articulatory gestures). We make the distinction between a target and a gesture, because the same sound can be composed of multiple simultaneous gestures, such as the movement of the tongue dorsum and the lips in case of back rounded vowels. While there are two gestures in this case, they overlap closely and can thus be presumed to contribute towards a single target.

Importantly, gestural targets may be present underlyingly but they may not be identifiable in the resulting articulatory signal due to factors such as gestural overlap (one gesture masking another gesture), or gestural undershoot (gesture being reduced). Thus, we supplement our articulatory data with computational modelling manipulating the values of articulatory parameters, and we evaluate various models against the empirical data obtained in the experiment.

2. Materials and method

2.1. Stimuli

The stimuli recorded in the experiment included a full set of English long vowels in an open syllable, preceded by the voiced bilabial stop /b/. The specific words were: *bay, buy, boy, bough, beau, beer, bear, bee, burr, bar, bore, boo*. Note that present-day Northern Anglo-English is generally non-rhotic, with the exception of some areas in Lancashire (Turton and Lennon, 2023). None of our participants pronounced coda rhotics. We used an initial bilabial, because it provides a relatively neutral context for all the vowels, as far as the tongue movement is concerned, and also because all combinations of long vowels with a preceding /b/ correspond to real words in the English lexicon. We used CV words, because in the absence of a following coda, we can be certain that any offglide movement we observe is inherent to the vowel, and not due to coarticulation. Since only phonologically heavy vowels can occur in this context, we did not include short monophthongs. The target words were embedded in the carrier phrase: *She says X*.

2.2. Participants

Six female speakers aged between 19–21 years old ($\overline{x}$ = 19.17, σ = 0.98) took part in the experiment. All participants reported normal speech and hearing, and all were monolingual L1 speakers of English. All speakers were born and grew up in the north of England. Specifically, five speakers lived in the region spanning Lancashire and Greater Manchester from birth until the time of the experiment, while one speaker lived in Sheffield until moving to Lancaster at the age of 18. All speakers had an unambiguously northern English phonological system and all used a notably higher proportion of regionally-marked features than is typical of General Northern English as described in Strycharczuk et al. (2020). Each participant was reimbursed £30 for taking part in the study, which lasted around 1.5 h in total.

2.3. Procedure

Electromagnetic articulography (EMA) data were acquired using a 16-channel Carstens AG501 system, recording at a sampling rate of 1250 Hz. Sensors were attached to the tongue at 1 cm behind the tongue tip (TT), as far back as possible on the tongue dorsum (TD), and an additional sensor located equidistant between the TT and TD sensors. Sensors were also attached to the vermilion border of the upper (UL) and lower (LL) lips, as well as the lower gumline. Reference sensors were attached to the gumline of the upper incisors, bridge of the nose, and on the right and left mastoids behind the ears. All sensors were attached midsagittally, except for the sensors behind the ears. Simultaneous ultrasound tongue imaging data were collected alongside the audio and EMA data, with an additional three sensors attached to the ultrasound probe, but the ultrasound data are not analysed in this study (see Kirkham et al. (2023) for more detail on the co-registration set-up). We recorded the location and orientation of the occlusal plane for each speaker by asking them to bite down onto a bite plate, which also had three EMA sensors attached to it. The audio signal was recorded using a Beyerdynamic Opus 55 microphone attached to a plastic ultrasound probe stabilisation headset, which was being used for simultaneous ultrasound data collection. The microphone signal was pre-amplified using a Grace Design m101 pre-amplifier and digitised at 48 kHz with 16-bit quantisation.

Most speakers produced four repetitions of the stimuli, except f05, who produced five. At the processing stage, it became apparent that the Tongue Dorsum displacement data were considerably out of range for multiple blocks produced by speakers f03 and f06. This suggests that the sensor became detached without the participant or ourselves noticing. We discarded these data from the articulatory analysis, which left a single usable set of repetitions for f03, and two repetitions for f06. The total number of tokens used in the articulatory analysis was 242. We included all the available data (306 tokens) in the acoustic analysis.

2.4. Data processing

EMA recordings were downsampled to 250 Hz and position calculation was carried out using the Carstens default algo-

rithms. Head correction and bite plate rotation were applied to each data sample, with head correction for each speaker optimised based on the best combination of reference sensors that reduced the RMS error across the entire session. Reference sensors were filtered using a 5 Hz low-pass Kaiser-windowed filter, and articulator sensors were filtered using a low-pass Kaiser-windowed filter with 40 Hz pass and 50 Hz stopband edges. We further filtered articulator sensors for the specific analyses reported below in Section 2.5.

2.5. Analysis

The audio data were forced-aligned using the Montreal Forced Aligner (McAuliffe et al., 2017). The boundaries were then manually corrected in Praat (Boersma and Weenink, 2009) by a research assistant, with specific attention paid to the acoustic boundaries of the vowel. The vowel onset was placed at the end of the burst for the preceding plosive, which typically coincided with the onset of a visible formant structure and the offset of voicing. The end of the vowel was marked at the offset of voicing.

The segmentation was used as the basis for articulatory and acoustic analysis. For the articulatory analysis, we focus on the movement of two key sensors: Tongue Dorsum (TD) and Upper Lip (UL). Displacement of the Tongue Dorsum allows for a systematic parsing of vowel gestures, as previously shown by Blackwood Ximenes et al. (2017) and Sotiropoulou et al. (2020). In addition, some vowels are crucially modified by the movement of the lips. In our analysis of the lip movement, we focus on lip protrusion as a correlate of rounding. Lip protrusion can be defined as the horizontal displacement of the lips, although the exact definition varies across the literature (Georgeton and Fougeron, 2014). We focus on the displacement of the upper lip, because we find that the horizontal displacement of the lower lip was systematically affected by the jaw movement, whereas the movement of the upper lip appears more independent. Here, we wanted to use displacement vectors that are as orthogonal as possible to avoid analytical artefacts that could arise from different displacement vectors capturing the same movement, as we combine them into a joint measure. The sensor displacement data were *z*-scored within speaker for normalisation.

The sensor displacement data were extracted for the portion corresponding to the acoustic duration of the vowel, followed by a fixed 75 ms window at the end. We included this window because we find that movement of the tongue and the lips typically continues beyond the offset of voicing at the end of a phrase. Much of our analysis focuses on the first derivative of the sensor displacement data, i.e. velocity. In order to calculate the velocities, we further smoothed the scaled displacement values using a low-pass Butterworth filter (cutoff frequency = 10 Hz). Tangential velocity was then derived, combining the horizontal-vertical Tongue Dorsum sensor displacement, and the horizontal displacement of the Upper Lip.

Acoustic data were analysed in Praat. We extracted formant trajectory values for each vowel, using Fast Track (Barreda, 2021). The settings we used were: lowest analysis frequency = 5 kHz, highest analysis frequency = 7 kHz, Number of steps = 20, Coefficients for formant = 5, Number of formants = 3, Number of bins = 5, Statistic = median. The formant measurements were sampled at every 2 ms (equivalent to 500 Hz). Similar as in the articulatory analysis, we also analysed the first derivative of the formant change. We *z*-scored the formant measurements within speaker, and smoothed the formant trajectories using a low-pass Butterworth filter (cutoff frequency = 10 Hz).

Since our study is concerned with quantifying degree of diphthongisation, it is important to comment on how such a degree can be measured. Several acoustic studies do this by measuring the degree of overall formant displacement. The specific measure that can be used is the Euclidean distance between F1 and F2 values at the onset and offset of the vowel, or an interval equivalent to the 80% portion of the vowel (Fox and Jacewicz, 2009; Haddican et al., 2013; Reed, 2014). This measure can also be refined to account for more complex trajectory shape as a sum of Euclidean distances sampled from a number of windows within the vowel (Fox and Jacewicz, 2009). We use the simpler version, relying on two time points, and we also extend the same approach to articulatory data. We calculate the Euclidean distance between two articulatory positions at pre-defined time points in order to capture the overall degree of articulatory displacement within a vowel. Section 3.1 below provides more detail on how the articulatory Euclidean distance was calculated, and Section 3.3 does the same for the acoustic Euclidean distance.

Euclidean distances can capture the overall degree of acoustic and articulatory change, and they are thus well-suited to capturing gradient aspects of diphthongisation: canonical diphthongs are characterised by more inherent change, compared to canonical monophthongs, with gradient diphthongisation creating in-between patterns. However, since our main research question is whether we can classify all vowels within a system as having one or two component gestural targets, we also need a procedure for identifying the number of gestures and an associated measure.

In Articulatory Phonology, gestural targets are typically identified using the first derivative of displacement data, i.e. velocity. This is based on the observation that the rate of change in movement towards a target is associated with a change of direction in articulatory displacement, which corresponds to a local minimum in the associated tangential velocity profile. The approach underlies gestural parsing, as implemented in MVIEW, widely used software for analysing gestural properties using EMA data (Tiede, 2010). We analyse velocity trajectories to establish whether we find evidence of two component targets for diphthong vowels, and whether we can classify all vowels as having one or two discernible component gestures. Section 3.2 provides a more detailed explanation of the patterns we find.

We also extend the same approach to analysing the rate of change in formant trajectories. Unlike in articulatory studies, the first derivative is not a common measure in acoustics, so we do this largely on an exploratory basis, and to keep our articulatory and acoustic analysis as comparable as possible. However, there are several arguments for analysing formant velocity in addition to formant displacement (which we captured using Euclidean distance). Velocity is inherently a measure of rate of change, and as such, it can straightforwardly reflect the presence of a formant steady state (an interval of

low velocity) vs. change in formant trajectories (a sustained rise in velocity). Furthermore, the measure captures the degree of change irrespective of the direction of movement, and it allows us to combine information from multiple formants into a single value. In order to obtain this measure, we calculated the rate of change per unit time for the smoothed F1 and F2 trajectories. Tangential F1-F2 velocity was calculated based on these measures, defined as the square root of the sum of squared F1 and F2 velocities. Section 3.4 illustrates the resulting patterns, and suggests that tangential velocity is well-suited to quantifying the global rate of change in acoustic dynamics.

2.6. Data availability statement

All the data and code presented in this paper are available as an Open Science Framework repository at https://osf.io/gub32/. The repository also contains the simulations presented in Section 4.

3. Results

3.1. Articulatory displacement

We begin the presentation of results with a visual overview of articulatory displacement. Fig. 1 shows by-speaker mean TD displacement values for each item. The trajectories correspond to the acoustic duration of the vowel. The means were obtained using Generalised Additive Modelling (GAM). As expected, we see the greatest degree of TD displacement for vowels in *boy* and *buy*, followed by *bay*, *beau* and *beer*. Canonical monophthongs like *bar* or *burr* show overall least TD displacement, but there is some TD movement associated with these vowels. There is relatively limited TD displacement in *bough*, despite this vowel's robust acoustic and perceptual diphthongisation. This is likely due to the fact that the perceived change in this vowel is strongly affected by lip rounding, whereas the associated lingual movement is limited.

The role of the lips in diphthongisation is confirmed when we consider the horizontal displacement of the upper lip in normalised time, plotted in Fig. 2. The figure shows by-speaker GAM smoothed mean trajectory of horizontal Upper Lip displacement in normalised time. Overall, the displacement values are higher for rounded vowels like *boo* and *bore*, compared to unrounded vowels like *bee*. We can also see that some vowels are characterised by inherent change in lip displacement. This is especially prominent for *bough*, where we can see a forward movement of the upper lip, consistent with a rounding gesture, and in *boy*, where we find the opposite: the UL sensor moves backwards through the vowel articulation, which can be interpreted as an unrounding gesture.

In order to systematise these observations and to account for the dorsal and labial displacement using a single measure, we calculated the Euclidean distance between 10% and 90% of the vowel in a three-dimensional space defined by horizontal and vertical position of the TD sensor and the horizontal position of the UL sensor. The distributions of the articulatory Euclidean distance values are summarised in Fig. 3, depending on the item. The vowels in *boy* and *buy* have the greatest articulatory Euclidean distance, followed by *bay*, *bough*, *beer*, *beau* and *boo*. In comparison, the articulatory Euclidean distance is low for *bee*, *bore*, *bar*, *bear* and *burr*. These observations are consistent with canonical diphthongs being characterised by greater articulatory displacement, compared to canonical monophthongs. We further note that the distribution of the articulatory Euclidean distance values is fairly continuous, except for a break separating two vowels, *buy* and *boy* from the rest. In addition, some vowels showed considerable variance in the distance values, notably *beer*.

3.2. Articulatory velocity

Let us now examine vowel velocity trajectories from the point of view of diphthongisation. Fig. 4 illustrates the tangential TD-UL velocity profile for two representative tokens, *bar* and *buy*), produced by speaker f04. In *bar*, the velocity trajectory is characterised by an initial velocity peak, followed by a slow decline in velocity, and then another velocity rise. This type of trajectory is consistent with movement towards a single articulatory vowel target, followed by gestural release.

In comparison, *buy*, is markedly different. For this velocity trajectory, we find an initial peak followed by a local minimum, but then velocity rises rapidly in the second half of the vowel, and another local minimum can be seen after acoustic offset of the vowel. This velocity profile is consistent with the vowel having two distinct articulatory targets, and most of the velocity profile is dominated by the movement towards the second target.

These two examples would suggest that at least some vowels have velocity profiles consistent with the hypothesis that monophthongs, like *bar*, have a single gestural target, whereas diphthongs, like *buy*, have two targets. Targets can be clearly discerned, corresponding to local velocity minima. However, the question is whether we can classify the velocity profiles of all vowels as having one or two targets. The velocities for the individual vowels are shown in Fig. 5, plotted in normalised time and overlaid by speaker and by item. As we can see in the figure, many trajectories clearly align with one of the two types; however, the classification is not entirely straightforward in some cases. It is clear that the height of the component peaks can vary considerably, with some peaks being of limited height. However, there is also noise in the data, such that we see a series of small peaks in some cases. If the number of peaks were to be classified manually by a researcher, this would require some potentially arbitrary decisions about what does and what does not count as a peak. In order to avoid this problem, we instead undertook a systematic data-driven approach to parameterising the information about velocity trajectories, using a functional Principal Component Analysis (fPCA; Gubian et al. (2015)).

FPCA is a statistical method for reducing variance in time-varying measurements to orthogonal principal components. The tangential velocity profiles were input to the analysis for a time window corresponding to the acoustic duration of the vowel followed by a fixed 75 ms window. Based on this analysis, 98% of variance related to the spatio-temporal information in the velocity data can be reduced to four Principal Components. These components are illustrated in Appendix A. For the purpose of our analysis, we focus on the first Principal Component, PC1, which captured 61% of the variance. The

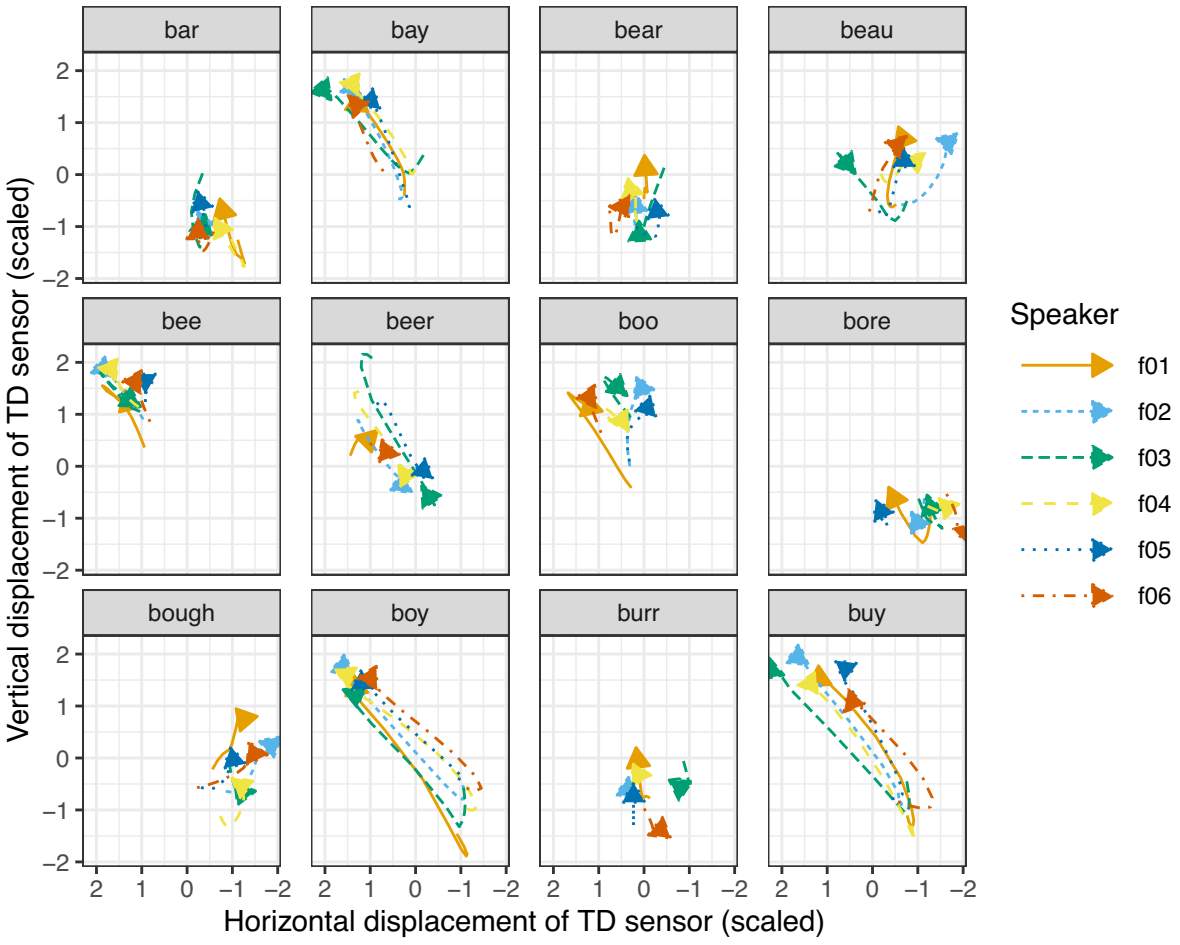

Fig. 1. Mean by-speaker trajectory of TD sensor displacement for each vowel. The beginning and the end of each arrow correspond to the acoustic onset and offset of the vowel respectively.

left panel of Fig. 6 shows a perturbation plot, which illustrates how variation in the PC1 score affects the shape of the velocity trajectory. The right panel of this figure shows the effect of item on PC1 score.

As we can see from the perturbation plot in Fig. 6, most of the dynamic variance associated with variation in PC1 occurs around 0.75 of normalised vowel duration. An increase in PC1 corresponds to an increase of the velocity value at this time point, which is also the local velocity peak. Conversely, a decrease in PC1 score is associated with a lower velocity values at the same time point, creating a local velocity trough as PC1 falls further below 0. As a result of this perturbation, two local velocity minima emerge in the corresponding velocity trajectory when PC1 is positive. From the right panel of Fig. 6, we can see that such positive scores are typically found for *bay*, *bough*, *buy* and *boy*. In contrast, low PC1 scores correspond to velocity trajectories in which a single local minimum can be discerned. Cross-referencing with the right panel of Fig. 6, such low scores are found for canonical long monophthongs like *burr*, *bore* and *bar*. Based on this interpretation, we can take PC1 as a proxy for diphthongisation. As we can see from the right panel of Fig. 6, the distribution of the PC1 scores largely corresponds to an expectation that we might have about monophthongs and diphthongs. We find the highest PC1 scores for *boy, buy, bough* and *bay*, followed by *beau, boo, beer* and *bee*, and then the median PC1 scores tail off, and are low for *bar, bear, bore* and *burr*. Crucially, the distribution of PC1 is entirely continuous – while it may not be unimodal, PC1 scores in all ranges are attested, suggesting an intermediate degree of diphthongisation for the vowels in the middle of the distribution, such as vowels in *beau*, *boo* and *beer*.

Let us consider what the intermediate PC scores mean in terms of velocity trajectories. In principle, intermediate PC1 scores could arise from some vowels varying categorically between a one-minimum vs two-minima type of trajectory (categorical variation between a monophthong and a diphthong), or they could genuinely represent a trajectory that is intermediate between the two prototypical categories. In order to explore this question further, we reconstructed the velocity trajectories based on mean by-item and by-speaker PC1 scores, following the procedure in Cronenberg et al. (2020). The reconstructed trajectories, shown in Fig. 7, suggest a mixture of categorical and gradient variation in the trajectory shape, as captured by PC1. The vowels with intermediate PC1 scores, *beau*, *boo* and *beer*, alternate between a monophthongal one-minimum trajectory with a trough in the second half and a diphthongal trajectory with two minima and a peak in the second half. For example, *bee* is monophthongal for f04 and f05, but diphthongal for the remaining speakers. However, we also observe more gradient variation in the height of the second peak. This

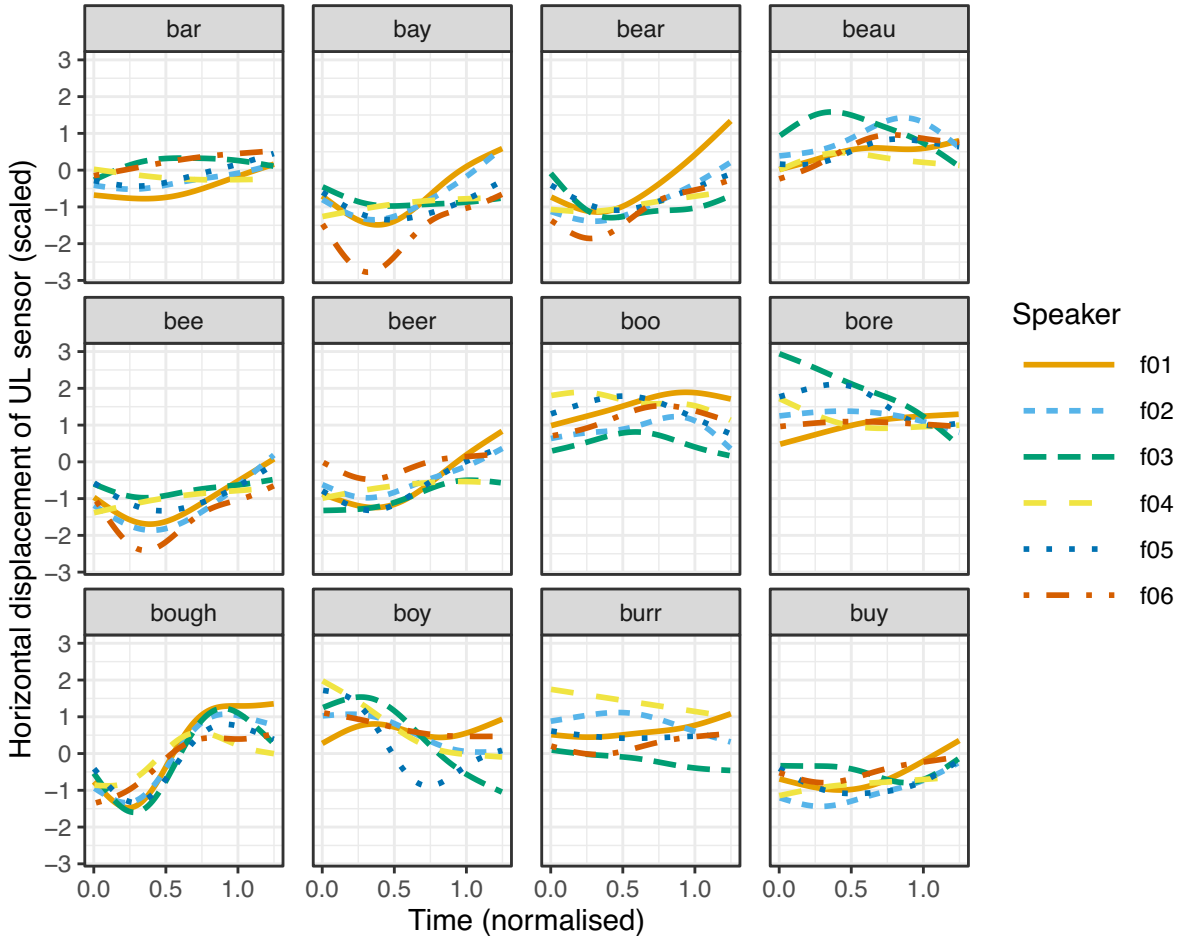

Fig. 2. Mean by-speaker trajectory of UL sensor displacement for each vowel, relative to normalised time.

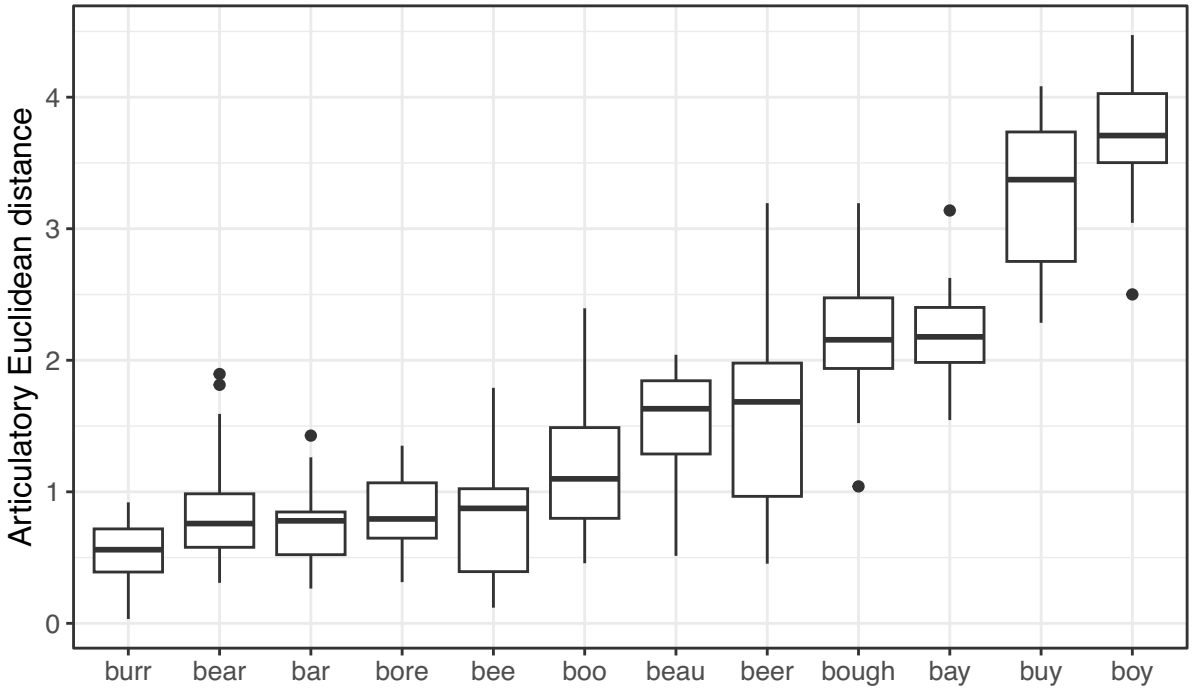

Fig. 3. The articulatory Euclidean distance depending on the vowel.

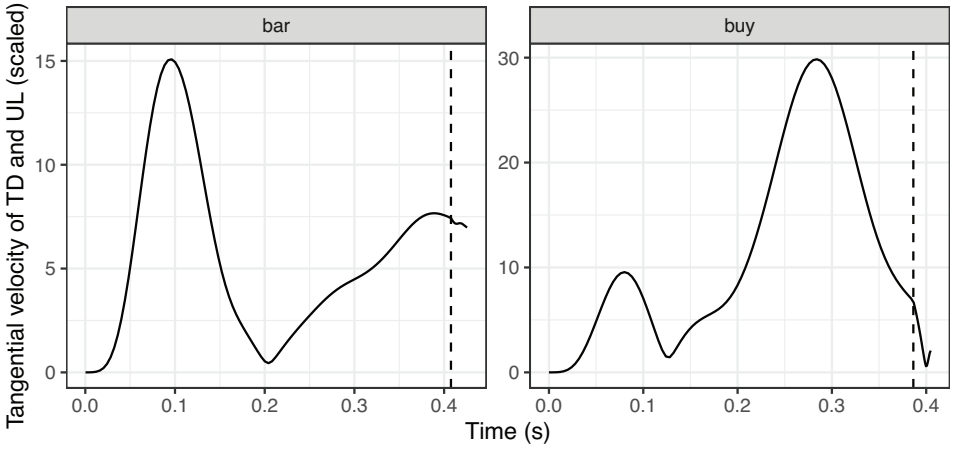

Fig. 4. TD-UL velocity for two example tokens pronounced by speaker f04. The dashed line represents the acoustic offset of the vowel.

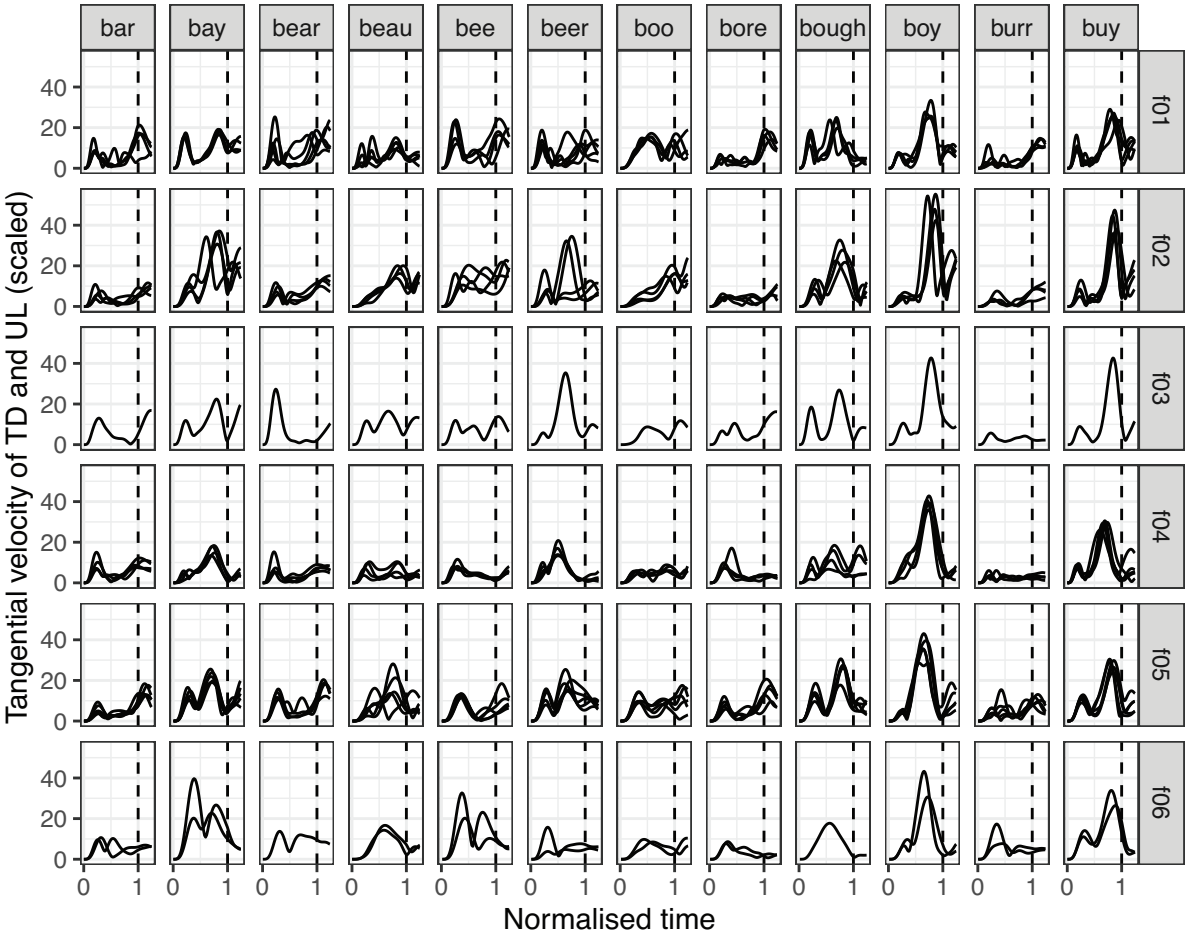

Fig. 5. TD-UL velocity profiles for all the individual vowel tokens.

is evident in vowels with intermediate values of PC1, such as *beer*, but also in canonical diphthongs like *boy* or *buy*. In this case, greater height of the second peak can be interpreted as a correlated of increased diphthongisation, related to greater distance between the targets within a diphthong.

3.3. Formant displacement

In order to compare our articulatory findings to a more familiar acoustic measure, we analysed the vowel formant trajectories. Mean F1 and F2 trajectories for each speaker (GAM-smoothed) are plotted in Fig. 8, depending on the item. The trajectories represent the mid 90% of the vowel rather than the

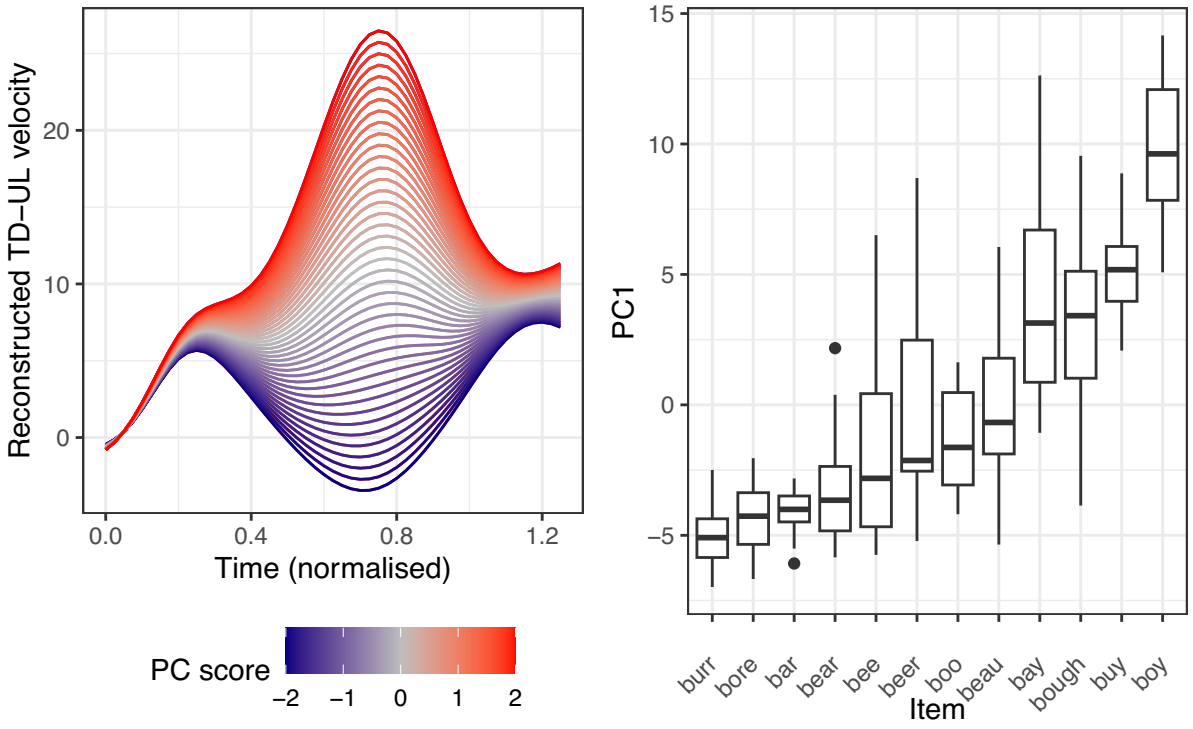

Fig. 6. Left: Perturbation of the articulatory velocity trajectory depending on PC1 score. Right: PC1 score depending on the item.

entire duration, due to difficulty of obtaining reliable formant measurements at the edges of the vowel.

In general, we find a clear and pronounced formant excursion for *buy, boy* and *bough*, and slightly lesser but still clear change for *beau* and *bay*. For *beer*, *boo* and *bee*, there is variable intermediate degree of formant change, whereas *bar, burr, bore* and *bear* show very little change. These generalisations are broadly consistent with the Euclidean distance between formant values at 10 and 90% of the vowel, as shown in Fig. 9.

3.4. Formant velocity

In this part of the analysis, we focus on the rate of change in the formant trajectories, operationalised as tangential velocity of F1 and F2 displacement. The formant velocities for an example token of *bar* and *buy* are plotted in Fig. 10. All the individual velocities are shown in Appendix B. In this case, we visualise the mid 90% of the vowel trajectory in normalised time.

The velocity profiles bear some similarities to the articulatory velocities presented in Section 3.2. Notably, we can see a rapid rise in velocity in the second half of *buy*. In contrast to articulatory data, the first half of the trajectory in this vowel does not show a clear initial peak. For *bar*, the entire trajectory is relatively flat, with no discernible initial peak. The absence of such a peak is consistent with there being a steady state in the formants.

We conducted an fPCA on the formant velocity data, in order to derive a numerical measure of formant change. The first Principal Component, illustrated in Fig. 11, captured 63% of variance. As we can see in the perturbation plot in Fig. 11, PC1 is correlated with the presence of a peak in the second half of the vowel formant velocity trajectory. The higher the PC1 scores, the steeper the rise. Negative PC1 scores correspond to a trajectory that drops off slowly following an initial peak. We can generalise that PC1 is a measure of acoustic diphthongisation (presence vs. absence of a second velocity peak). Appendix C presents a perturbation plot for the first four PCs.

The right panel of Fig. 11 shows the distribution of PC1 scores, depending on item. The relative values of acoustic PC1 scores closely resemble the results for articulatory PC1 (compare to Fig. 6), whereas some small differences emerge between this measure and the formant Euclidean distance, shown in Fig. 9. Specifically, *beau* and *bough* show more intermediate Euclidean distance values, whereas they are more diphthongal, according to the acoustic PC1. The overall distribution of the acoustic PC1 is continuous, and intermediate values are represented.

3.5. Hierarchical clustering

So far, it would seem that different measures of diphthongisation generally converge in distinguishing between vowels like *boy* or *buy* on the one hand, and vowels like *bar* or *burr* on the other. However, all the measures we have considered also yield intermediate values, with vowels in *bee, boo, beau* and *beer* tending towards the middle.

We performed a clustering analysis in order to explore whether any clusters of vowel emerge from the measures of diphthongisation that we have considered, and whether these clusters differ between articulatory and acoustic measures. We used hierarchical cluster analysis with Ward's clustering

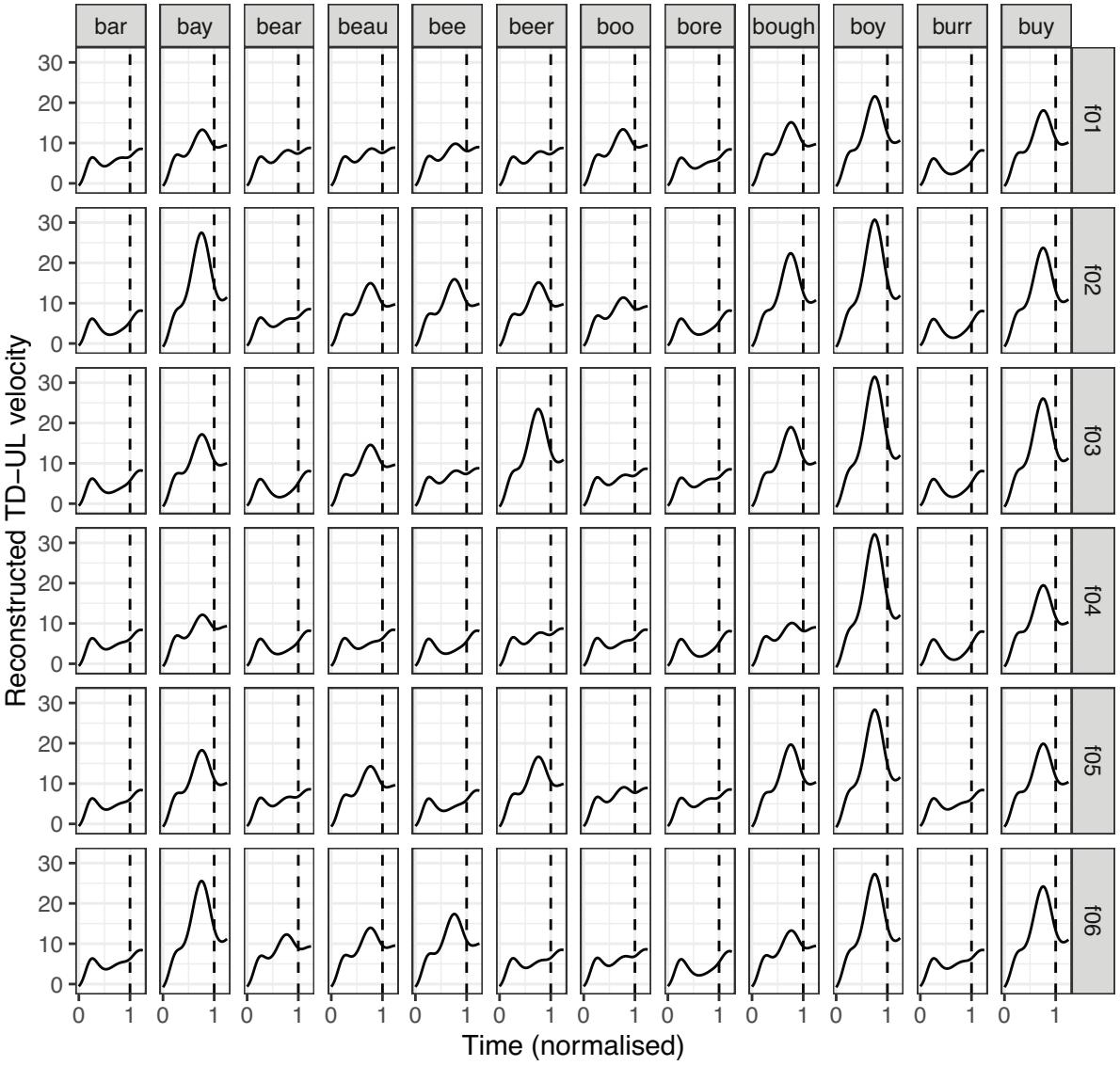

Fig. 7. Reconstructed TD-UL velocity trajectories based on by-vowel, by-speaker mean value of PC1.

criterion (Murtagh and Legendre, 2014; Ward, 1963). The input to clustering was a distance matrix based on by-vowel means for the four diphthongisation measures we have described: articulatory Euclidean distance, articulatory PC1, formant Euclidean distance and acoustic PC1. The correlations between these measures are included in Appendix D.

Fig. 12 summarises the results of clustering, depending on different combinations of diphthongisation measures. Articulatory clustering was based on articulatory Euclidean distance and articulatory PC1. Acoustic clustering was based on formant Euclidean distance and acoustic PC1. Finally, a clustering was also performed on all the measures combined.

In all cases, two to three clusters emerge from the data. The vowels in *buy, boy, bough* and *bay* always form a cluster. Additionally, the vowel in *beau* is grouped with these vowels in the acoustics, but not in articulation, where it patterns with *bee*, *boo* and *beer*. It is difficult to determine whether this vowel presents a case where articulatory displacement and acoustic change diverge systematically, or whether the discrepancy is an artefact of the measures we used. Another cluster is formed consistently by vowels *bar*, *burr*, *bore* and *bear*. A third cluster is formed by *bee*, *beer*, *boo* and (only in articulation) *beau*.

Based on previous analysis, we may have an expectation of the dynamic properties of the three clusters, as showing different degrees of diphthongisation. In order to verify them, we plotted the four measures of diphthongisation as a function of cluster. The relevant plots are in Fig. 13. Combined measures clustering was taken as the basis for these plots. For the articulatory and acoustic Euclidean distances, we plotted their values by cluster. For the articulatory velocity and the formant velocity, we calculated the by-cluster mean values of the articulatory PC1 and acoustic PC1, and we reconstructed the relevant velocity trajectories for each cluster, based on these means.

Cluster 1 comprises the vowels *bar, burr, bore, bear*. These are canonical monophthongs for which we can see a single articulatory target, a formant velocity consistent with a steady state (little formant change in the second part of the vowel)

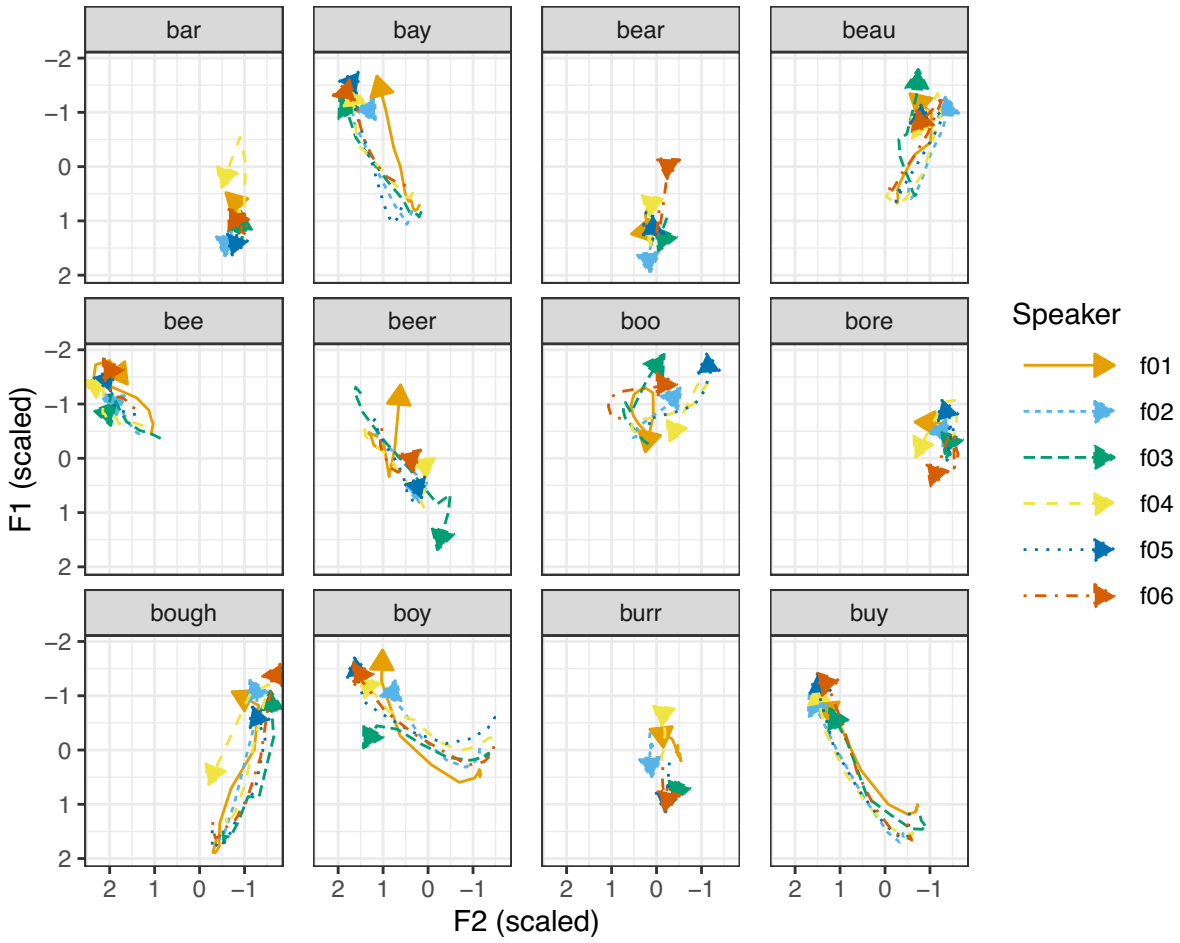

Fig. 8. Mean F1 and F2 displacement, depending on speaker and item. The trajectories represent the mid 90% of the vowel.

and a small degree of articulatory and acoustic displacement. Cluster 2 is formed by *buy, boy, bough* and *bay*. We can interpret this group as canonical diphthongs that show a two-target articulatory velocity, a clear late peak in formant velocity and a large degree of articulatory and acoustic displacement. Cluster 3 includes vowels *bee, boo, beau* and *beer*. They form an in-between category characterised by some articulatory and acoustic displacement, as reflected by the values of the Euclidean distances. The average TD-UL velocity trajectory for this cluster shows the presence of a second velocity peak, which is however limited in height. However, to some extent, the intermediate nature of this category is due to inter-speaker variation. Comparing Fig. 7, vowels like *bee* and *beer* may show a single peak for some speakers, but two peaks for other speakers. Acoustically, this produces an intermediate mean degree of change: greater than in canonical monophthongs, but less than in canonical diphthongs.

3.6. Summary of the results

In response to our research questions, two key observations emerge from our combined articulatory and acoustic analysis of diphthongisation in vowels. Firstly, for some vowels, (e.g. *buy*, *boy*), we can very clearly discern two distinct articulatory targets, consistent with the predictions of a compositional model of diphthongs. At the other end of the scale, some vowels (e.g. *burr*, *bar*) only show one discernible articulatory target. Secondly however, we observe that it is not possible to draw a boundary between two-target and one-target vowels. All the potential measures of diphthongisation that we have considered show gradience, and they commonly include intermediate values that fall between canonical diphthongs and canonical monophthongs. Such intermediate values are typical of vowels in *bee*, *boo*, *beer* and *beau*. These vowels can vary between a one-target and two-target trajectory type, depending on the speaker, but they also show more gradient variation, with varying height of the second velocity peak. This type of trajectory corresponds to intermediate degrees of articulatory displacement and acoustic formant displacement.

To account for these facts, a representational model must reconcile some aspects of categoricity and gradience. The key question for articulatory modelling is how to capture gradient articulatory variation between one-target and two-target vowels. While the phonetic manifestation of the variation shows gradience, the underlying number of targets is not continuous: an articulatory target is either present or absent. We must therefore consider which articulatory parameters can give rise to the kind of variation we find. We address this question through simulation, using a task dynamic model of gestural coordination (Saltzman and Munhall, 1989; Browman and Goldstein, 1992; Sorensen and Gafos, 2016).

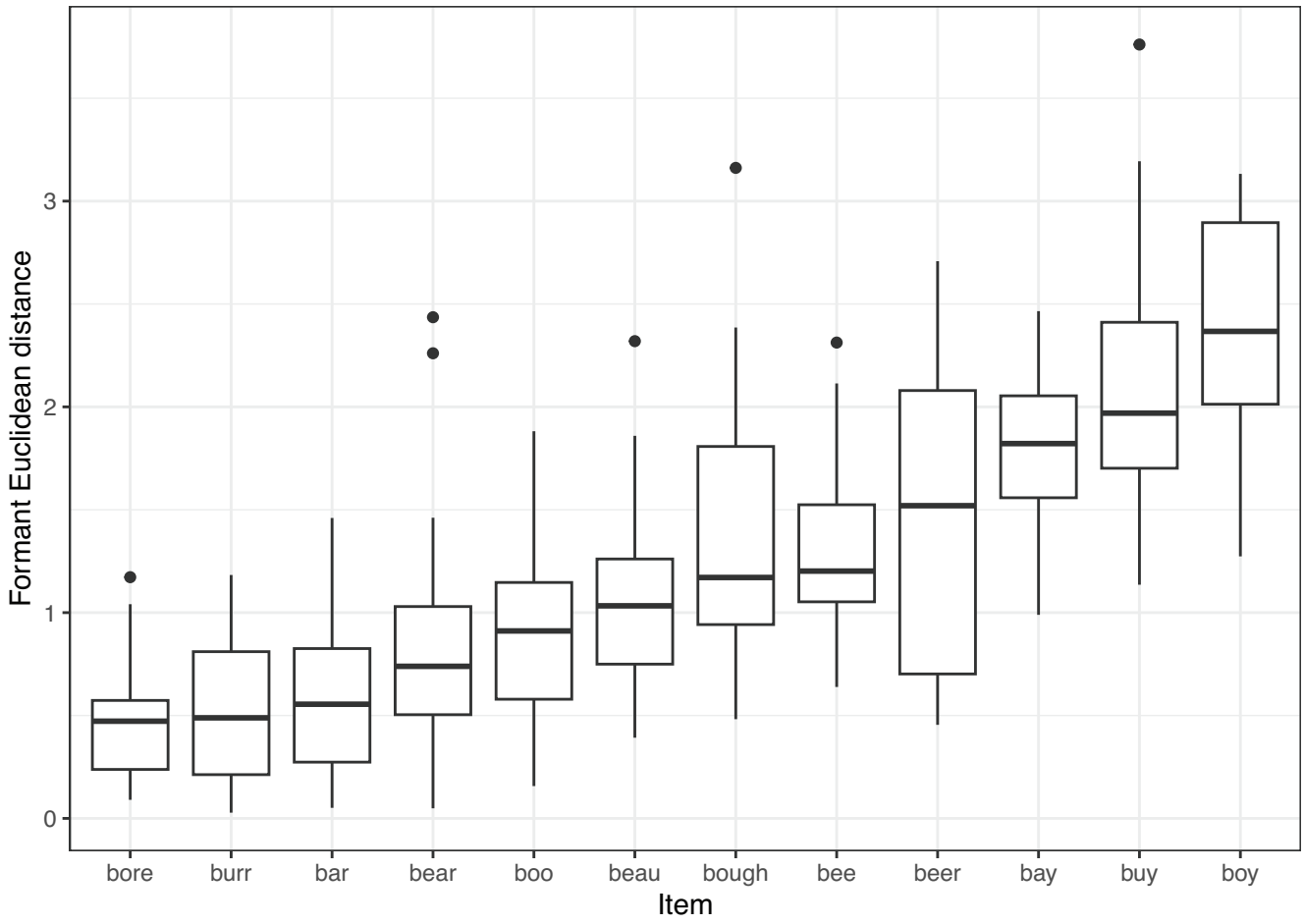

Fig. 9. Acoustic Euclidean distance (F1 and F2 displacement between 10 and 90% of the vowel) by item.

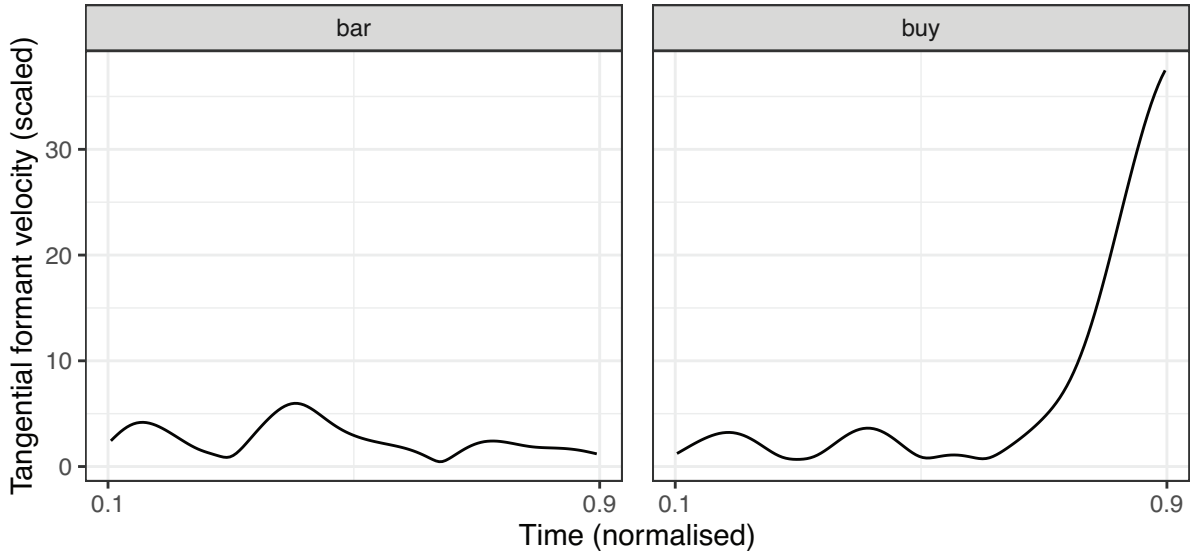

Fig. 10. Formant velocity profiles fortwo example tokens.

4. Computational modelling

4.1. Aims

This section presents simulations of articulatory dynamics in vowels using a task dynamic model. This allows us to generate quantitative predictions based on an explicit set of theoretical assumptions. In doing so, we can test the predictions of the AP/TD model by comparing the output of the simulation to the empirical data (see Burgdorf and Tilsen (2021); Hsieh (2017); Marin (2007) for a similar approach for developing AP models of vowels). Our specific aim is to evaluate the fol-

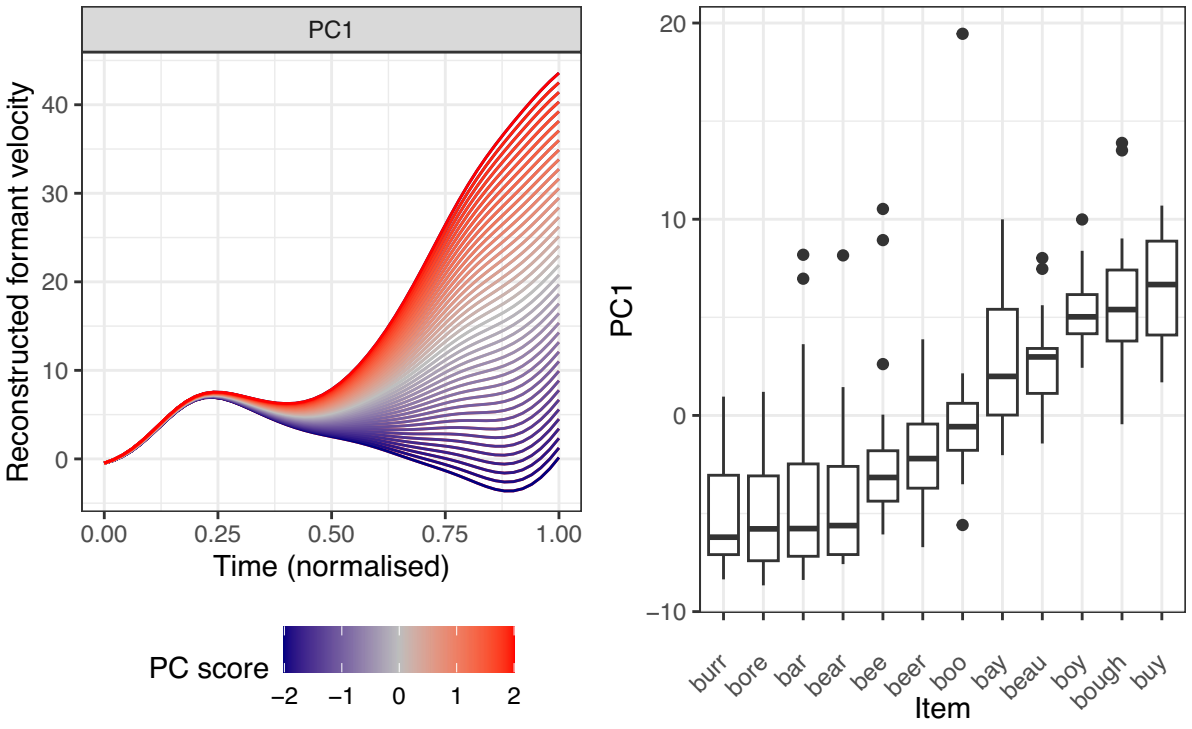

Fig. 11. Left: Perturbation of the formant velocity trajectory depending on PC1. Right: PC1 score depending on the item.

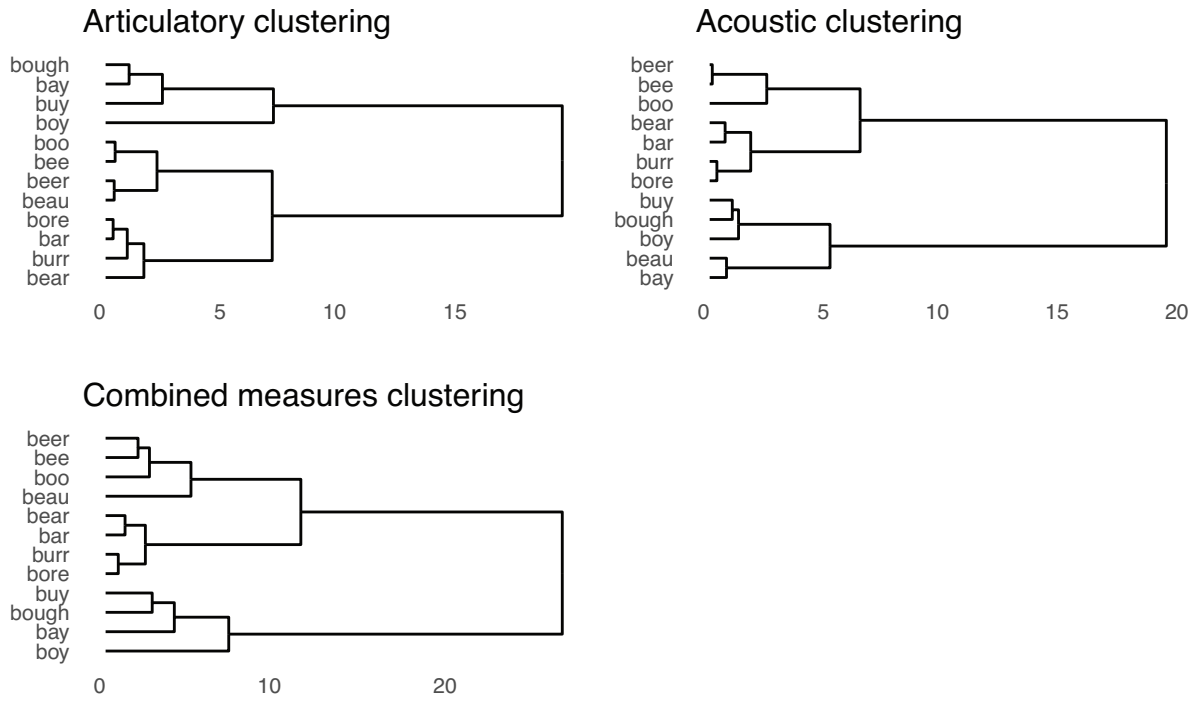

Fig. 12. Results of hierarchical clustering based on different measures of diphthongisation.

lowing two proposals that arise from earlier theoretical proposals discussed in Sections 1.1–1.3.

i. Diphthongs have two targets, and long monophthongs have a single target (i.e. diphthongs are compositional, and monophthongs are not). In evaluating this proposal we focus especially on the issue of vowel duration, and discuss several possible mechanisms to capture the relevant dynamic and duration facts.

ii. A two-target model of all long vowels (i.e. both diphthongs and monophthongs are compositional). In this view, a long monophthong comprises two (near-) identical targets and a diphthong comprises two distinct targets. The difference between a long monophthong and a diphthong can, therefore, be modelled as a gradient change in a target's gestural parameters. This type of model has been sketched out in previous literature, but it

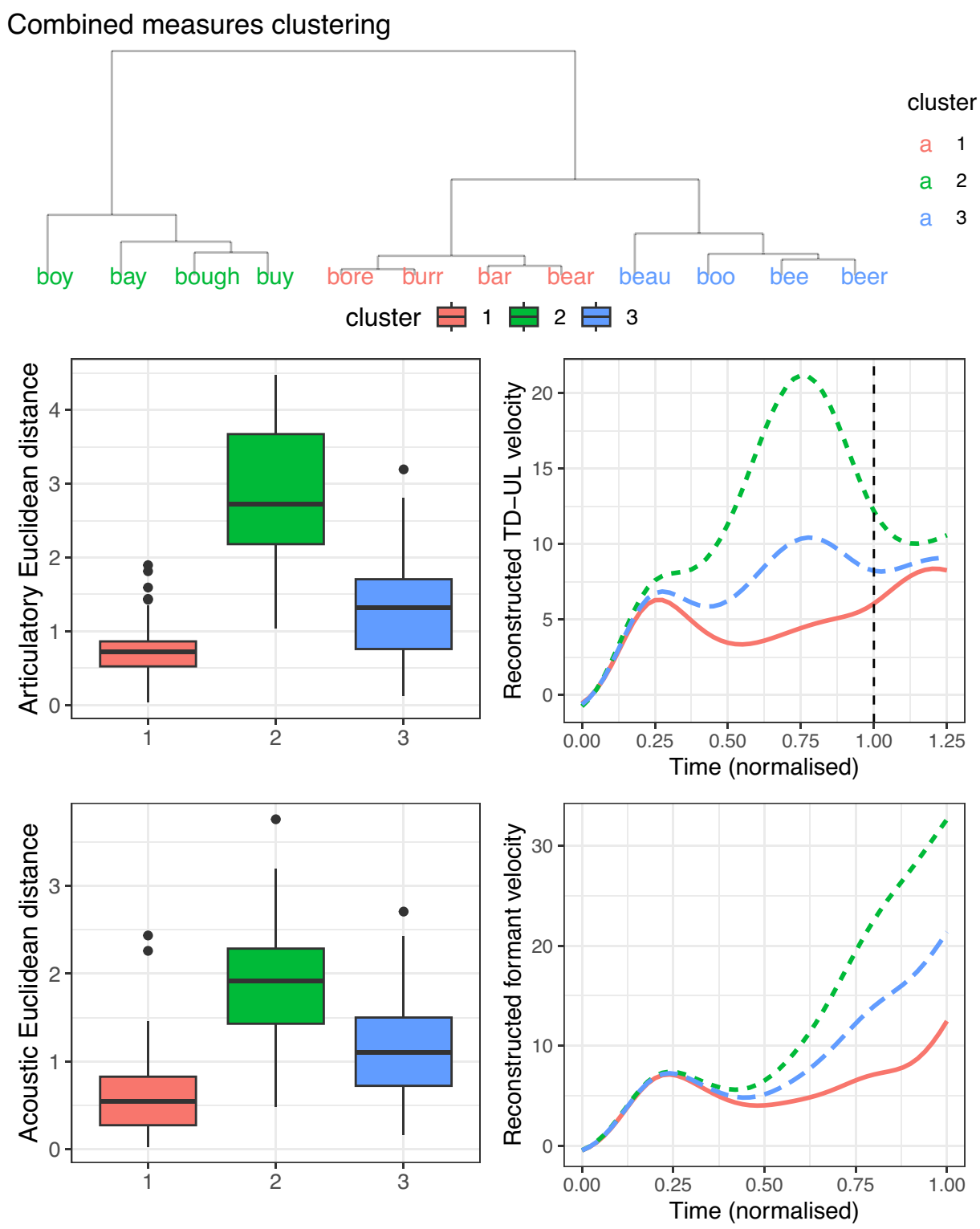

Fig. 13. The effect of cluster on the four measures of diphthongisation.

has not been shown that a two-target model can generate realistic articulatory data, especially for monophthongs and for gradient diphthongisation.

4.2. Method

We simulated gestural dynamics using Eq. 1, which is a modified task dynamic model proposed by Sorensen and Gafos (2016). This differs from the classic Saltzman and Munhall (1989) model in the addition of a cubic term dx^3, which acts as a non-linear restoring force on the spring in the damped mass-spring model. This corrects for the overly short time-to-peak velocity and asymmetric velocity profiles in previous models. Values of $d > 0$ increase the strength of the non-linear restoring force. In our simulations, a uniform value of d adequately reproduces the qualitative velocity distinctions that are central to our predictions, but we here optimise d sepa-

rately for each simulation in order to show that the model is also capable of generating empirically-realistic velocity profiles. Additionally, b is a damping parameter defined as $b = 2\sqrt{mk}$, where k is stiffness and m is a mass parameter that is always equal to 1.

$$\ddot{x} + b\dot{x} + kx - dx^3 = 0 \quad (1)$$

In all models, we simulate the tract variable Tongue Body Constriction Degree (TBCD) as a proxy for the tongue dorsum constriction, which we define in the normalised coordinate space [0,1], where x = 0 is minimally constricted (representing a maximally open constriction) and x = 1 is maximally constricted (representing complete closure) (Burgdorf and Tilsen, 2021). In all cases, we assume all vowel gestures to have uniform stiffness (k). Note that this is not an essential feature of our model, as variation in stiffness can form part of gestural representation, but we do this to show that our predictions do not intrinsically rely on stiffness variation.

The timing between two vowel gestures was defined using a coupled oscillator model of gestural coordination in Eq. 2 (Tilsen, 2018). Φ_{ij} is the relative phase between oscillators i, j, such that $\Phi_{ij} = \theta_i - \theta_j$. C_{ij} is a matrix of coupling strengths between oscillators i, j, where $C_{ij} > 0$ is in-phase and $C_{ij} < 0$ is anti-phase.

$$\dot{\theta}_i = 2\pi f_i + \Sigma_j C_{ij} \sin(\Phi_{ij}) \quad (2)$$

We make the simplifying assumption that a vowel gesture is defined as a 250 ms period of gestural activation and all gestures have the same oscillator frequency f = 4 Hz. Anti-phase coupling between two vowel gestures results in a 125 ms lag when oscillator frequencies are 4 Hz, so a vowel with two 250 ms gestures coupled anti-phase is 375 ms in duration. The 250 ms duration of gestural activation intervals were set by hand based on the average pattern in the empirical data. This is a simplifying heuristic for the purposes of illustration; a feedback-based suppression mechanism is instead a more likely approximation of how speakers control gestural deactivation and, therefore, the timing of gestural offsets (Tilsen, 2022).

All simulations were conducted using the Python programming language. We solved for the velocity of TBCD using the Explicit Runge–Kutta method with order 5(4) via SciPy's `integrate.solve_ivp` function (Virtanen et al., 2020). In all cases, we solve using a time step of Δt = 0.001 s and all simulations have uniform stiffness (k = 2000) and damping ($b = 2\sqrt{km}$, where m = 1). The value of d is manually specified for each gesture to produce velocity trajectories similar to those observed in empirical data. Further details can be found in the online documentation at https://osf.io/gub32/.

4.3. Modelling one-target monophthongs vs. two-target diphthongs

We begin by simulating canonical examples of a monophthong and a diphthong, represented by *bar* and *buy* respectively. We selected these examples, because *bar* consistently shows a single velocity minimum for all speakers in our data, whereas *buy* consistently shows two minima. In addition, the vowels in *bar* and *buy* are unrounded, which allows us to focus on TBCD as a primary correlate of articulatory movement inherent to the vowel. We simulated the TBCD velocity trajectories for *bar* and *buy* assuming the former consists of a single dorsal gesture (TBCD = 0.3), and the latter consists of two distinct dorsal gestures coupled anti-phase (TBCD = {0.3, 0.9}), as hypothesised in Section 1.

The top panel of Fig. 14 shows the results of the simulation. The TBCD velocity trajectories pattern as expected in terms of shape: the one-gesture trajectory shows a single peak followed by a local minimum, whereas the two-target trajectory consists of two velocity peaks. Note that there is no final gestural release in the simulated data. Notably, the model predicts a durational difference between the two vowels: the one-target monophthong is shorter than the two-target diphthong (250 ms vs. 375 ms respectively).

In general, diphthongs have not been reported to be longer than monophthongs in English (Peterson and Lehiste, 1960; Lehiste and Peterson, 1961). Nevertheless, duration ought to be investigated more systematically in light of the modelling outcome. Fig. 15 shows the distributions of vowel duration, depending on the item. As we can see, the duration of the vowel in *buy* is not systematically longer than that in *bar*. More generally, it is also not the case the canonical diphthongs are longer than canonical monophthongs. While there is some variation in vowel duration, it mainly seems correlated with vowel height, in line with higher vowels being inherently shorter than low vowels (Peterson and Lehiste, 1960).

Let us consider some alternative models that do not predict a duration difference between long monophthongs and diphthongs. One possibility is to model the two component gestures in diphthongs as being coupled in-phase. As previously discussed in Section 1.1, in-phase coupling has previously been proposed for ongliding diphthongs, such as /juː/, in contrast to offgliding diphthongs such as PRICE (Marin, 2007; Hsieh, 2017). A key difference between these two sets of vowels is syllable weight: the glide does not contribute to syllable weight in ongliding diphthongs, but it does so in the case of offgliding diphthongs. This can be captured through a coupling asymmetry between the two sets of vowels that mirrors the weight asymmetry between onsets and codas: onset consonants, coupled in-phase to the vowel, are transparent to syllable weight, whereas coda consonants, coupled anti-phase to the vowel, carry syllable weight. Thus, modelling diphthong gestures as coupled in-phase predicts that diphthongs are phonologically light, which is incorrect for offgliding diphthongs such as *buy*.

An alternative proposal is that long monophthongs have a single long gesture, whereas diphthongs are a composition of two short gestures. In this view, both classes can have similar duration, but it arises from different sources. To illustrate this, the middle panel of Fig. 14 shows *bar* modelled as a single long target at TBCD = 0.3, with 375 ms duration. This model seems to provide a good empirical fit, generating a velocity trajectory that has the desirable dynamic properties (presence of a single velocity peak) and the expected duration equivalent to diphthong vowels.

4.4. Modelling long monophthongs as two-target vowels

Let us now consider whether realistic articulatory vowel dynamics can be generated by a model in which diphthongs are modelled as comprising two targets coupled anti-phase, and the same is true for long monophthongs. In case of long

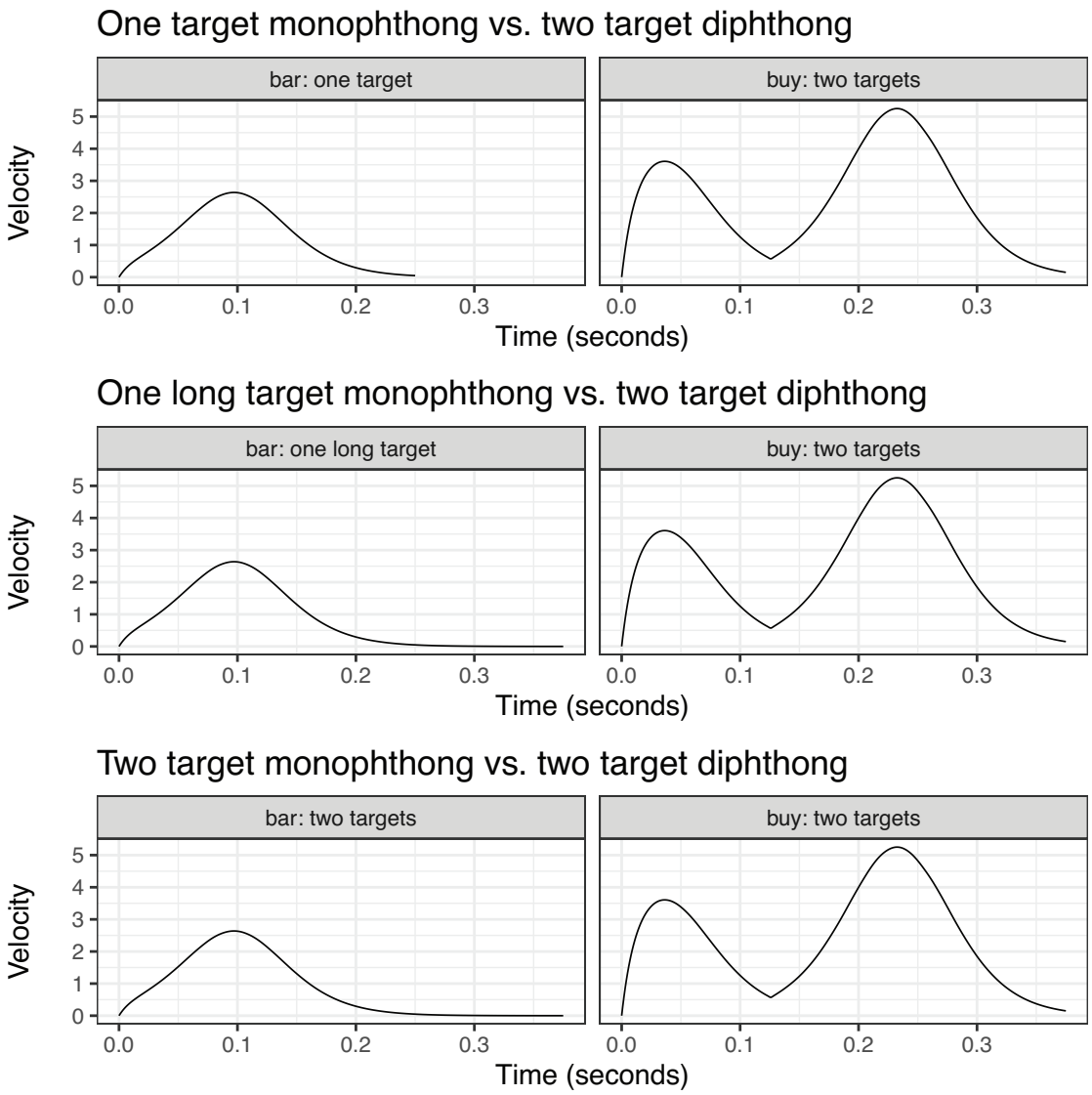

Fig. 14. Simulated TBCD velocities for *bar* and *buy* under three sets of assumptions.

monophthongs, the component targets are identical in terms of gestural constriction and location, whereas for diphthongs, the constriction and location of the component targets differ, resulting in inherent vowel change. This is in line with a previous proposal by Popescu and Chitoran (2022). The main modification we propose is that all phonologically long vowels in English have two articulatory targets, and not just the vowels characterised by some degree of audible or measurable diphthongisation. Note that in this model we do not consider glides to have a separate status and instead model diphthongs as the composition of two short monophthong targets.

Example simulations of *bar* and *buy* are in the bottom panel of Fig. 14. Specifically, *bar* was modelled as having two vocalic gestures, with the same TBCD targets for each gesture (TBCD = {0.3, 0.3}), which were coupled anti-phase to one another (we refer to the first as the nucleus and the second as the offglide). Note that the gestural parameters of *bar* are identical to one-target *bar* in the middle panel of Fig. 14, except here we have two monophthongal gestures coupled anti-phase, rather than a single long gesture. The vowel in *buy* has two vocalic gestures with different targets (TBCD = {0.3, 0.9}). The blending ratio was set at 1:100 in favour of the offglide. As we can see, a monophthongal trajectory with a single local minimum (a single target) emerges from two underlying identical targets, but the duration of the monophthong increases, matching the predicted duration of a diphthong.

So far, we have seen that the two-target model produces a plausible velocity trajectory for canonical monophthongs, and it correctly predicts that diphthongs and long monophthongs are phonologically heavy and have a similar phonetic duration. We now further consider whether such a model can also capture intermediate degrees of diphthongisation and variable diphthongisation, characteristic of vowels such as in *bee*.

We modelled the long monophthong /i/ comprising two targets, which we refer to as the nucleus [i] and glide [j]. Burgdorf and Tilsen (2021) show that the differences between high vowels and glides can result from syllable organisation (vowels occur as nuclei; glides occur as onsets/codas) rather than different gestural specifications, but we also allow for the possibility of variability between them. Specifically, we

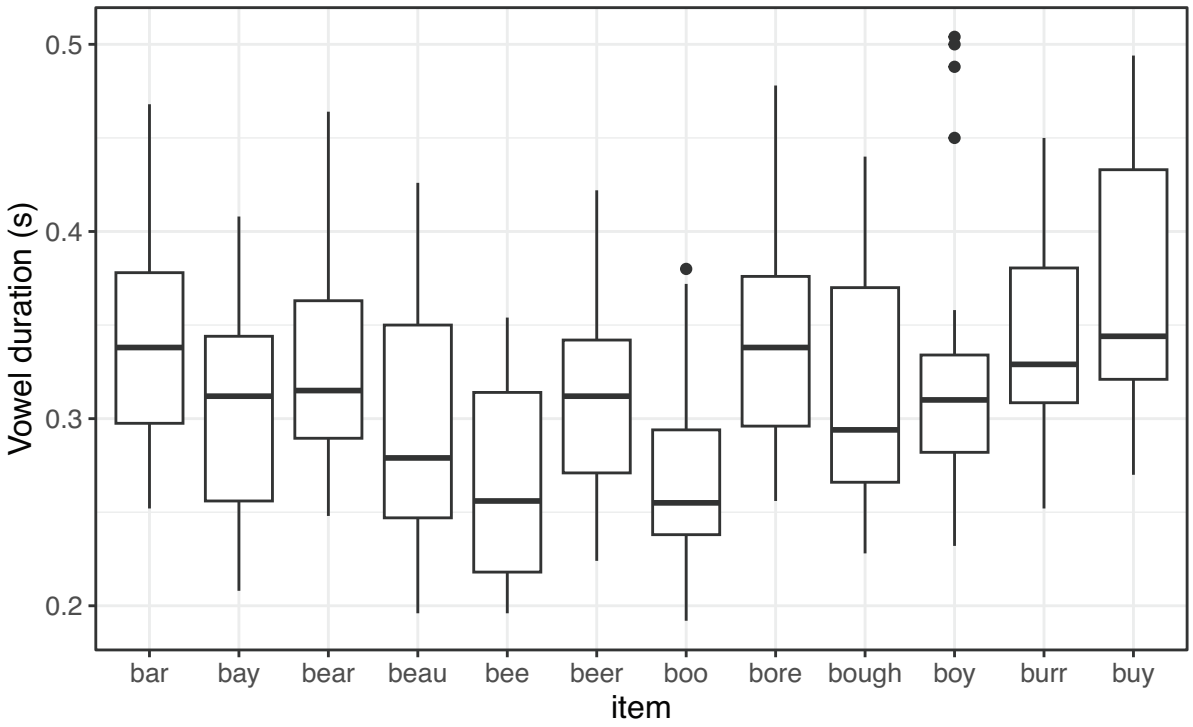

Fig. 15. Vowel duration by item.

explore the possibility that variation in diphthongisation of /i/ arises through variation in the nucleus TBCD target, with no variation in other articulatory parameters.

First, we fix the TBCD target for [j] at 0.9, representing a palatal constriction, but allow the TBCD target for [i] to vary across the range {0.9, 0.8, 0.7, 0.6}. For example, a TBCD target of [i] = 0.9 and [j] = 0.9 gives identical targets for nucleus and offglide, whereas a value of [i] = 0.6 and [j] = 0.9 represents a more open constriction for the nucleus than the glide. As in our previous simulations of two-target vowels, we use a blending ratio of 1:100 in favour of the offglide and uniform stiffness for both gestures. We note that the blending ratio and stiffness values are not a requirement of our model and qualitatively similar results can be obtained using lower blending ratios or different stiffness values for nucleus and coda.

Fig. 16 shows that when TBCD targets are identical (TBCD = 0.9) the model produces a one-target velocity profile for /i/. Changing the nucleus TBCD target to a lower (more open) value produces a successively larger second velocity peak, demonstrating that /i/ becomes more diphthongal as the nucleus target diverges further from the offglide target. The variation in the height of the second velocity peak is qualitatively similar to the variation observed in the empirical data. As shown in Fig. 5, individuals vary in their production of *bee* with respect to how high the second peak is. For some of them, there is no second peak, whereas for others, the second peak is present, but it is not as high as in prototypical diphthongs. The model confirms that this variation can arise from varying the gestural constriction of the nucleus, while the remaining parameters remain unchanged. This would be impossible with the 'one long target' model, without recourse to an additional mechanism for adding targets or splitting an existing target.

4.5. Summary of modelling

In sum, there are different ways of modelling the distinction between canonical diphthongs and canonical monophthongs. For diphthongs, a compositional model with two component targets generates two velocity peaks, same as we find in the empirical data. For monophthongs, a single velocity peak can be modelled either using a single long vowel target, or as a combination of two identical component targets. These two representations yield the same empirical predictions for canonical cases, and so modelling long monophthongs as one long target versus two coupled targets cannot be distinguished from the data alone. Nevertheless, we propose that the two-target model is preferred on theoretical grounds. We discuss this in Section 5 below, with special attention paid to the case of gradient diphthongisation, which can be captured using two similar targets, as shown by our modelling.

5. Discussion

In this study, we analysed the articulatory nature of diphthong production, in comparison to long monophthongs. In doing so, we have focused on two questions. The first question was whether diphthongs can be consistently analysed as comprising two articulatory targets, in contrast to monophthongs that could be characterised as comprising one target. Diphthong compositionality, the idea that diphthongs have two articulatory targets, finds support in the articulatory data. The movement of the articulators, as we have observed it, can be characterised as movement from one target to another. The key empirical reflection of the two targets in the abstract representation is the presence of two articulatory maxima (local velocity minima) in the dynamic trajectories of tongue dorsum

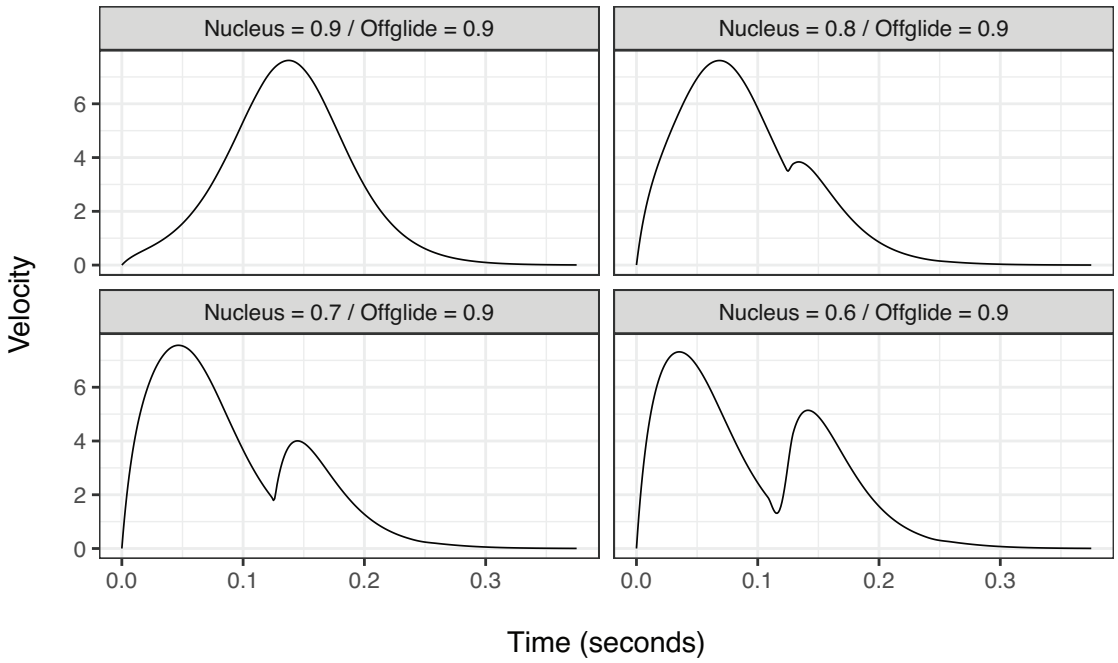

Fig. 16. Simulated TBCD velocity for *bee* across different [i] TBCD target values. Nucleus = 0.9 corresponds with a classic one-target long monophthong. In all cases, the offglide target is TBCD = 0.9.

movement and lip protrusion. Note that the second target may not always be reached, as proposed by Lindblom and Studdert-Kennedy (1967), and as confirmed here by tokens in which the second target is delayed beyond the acoustic offset of the vowel.

The second question concerns the nature of the distinction between diphthongs and monophthongs. Traditionally, compositional diphthongs are represented in opposition to non-compositional monophthongs, which entails that the two are distinct phonological categories, as discussed in Section 1.1. Contrary to this prediction, we find that diphthongisation is gradient in the articulatory and acoustic domain. The clustering analysis in Section 3.5 shows that regardless of whether we consider articulatory or acoustic measures of diphthongisation, an intermediate category arises between canonical monophthongs and canonical diphthongs, and this category includes the vowels with variable or limited degree of diphthongisation: *bear, bee* and *boo*. The observation is not trivial considering non-linearities between articulation and acoustics (Stevens and Keyser, 1989), which could produce gradient acoustic change from categorical articulatory shifts. While this is of course still possible in some cases, not all instances of intermediate diphthongisation can be explained in this way.

Modelling presented in Section 4 shows that the articulatory properties of canonical diphthongs are well captured in a compositional model in which diphthongs consist of two articulatory targets coupled anti-phase. A two-target model is also well-suited to capturing the phenomenon of gradient diphthongisation. Our simulation shows that the difference between a long monophthong with no discernible diphthongisation and one with variable diphthongisation can be modelled as a change in the nucleus TBCD target (Fig. 16). As a consequence, small changes in the degree of diphthongisation emerge in the system from variation in simple gestural parameters, with no change in the underlying structural organisation of gestures. When it comes to modelling long monophthongs, we seem to have two empirically equivalent alternatives. The dynamic and durational properties of long monophthongs can be accurately predicted by a model in which long monophthongs consist of two identical articulatory targets. However, they are equally well predicted by a single-target model assuming long target duration for monophthongs. Even though empirically, these models are equivalent, the two-target model is arguably preferred from the point of view of phonological theory, as well as from the patterns of variation and change.

A compositional two-target model of long monophthongs is theoretically appealing in establishing a systematic correspondence between phonological structure and syllable weight. Like diphthongs, but in contrast to short monophthongs, long monophthongs are phonologically heavy. The weight distinction between long and short monophthongs can be captured in our model using the tools independently proposed for coda consonants. In case of coda consonants, syllable weight follows from the presence of a consonant gesture coupled anti-phase to the vowel gesture. In a two-target model of long monophthongs, the structure is much the same, except the anti-phase coordination holds between two component vowel gestures, rather than a vocalic and a consonantal one. In contrast to long monophthongs, we can model short monophthongs as having only a single underlying target. An added benefit is that the difference in phonological structure and phonotactic behaviour of long and short monophthongs is in this case also systematically correlated with phonetic duration: one-target monophthongs are phonetically shorter than two-target monophthongs.

Furthermore, while it is possible to model long monophthongs as a single long target, this requires additional assumptions, such as long monophthongs having an inherently

different phonological specification from all other vowels. For example, one conceptualisation of activation interval duration is by mapping target achievement to a particular phase of a virtual oscillator cycle, such as 270° (Browman and Goldstein, 1995), with differences in duration arising as a consequence of oscillator frequencies (see Tilsen (2018) for further illustration). A related conceptualisation of duration is time-to-target achievement, which is modulated by stiffness (Ratko et al., 2023). In both cases, long monophthongs must have a different gestural specification from short vowels and diphthongs. While not inherently problematic, this requires extra assumptions that must be justified over a simpler model.

Timing mechanisms in the standard model of AP/TD have come under intense critique (Turk and Shattuck-Hufnagel, 2020), so we also comment on the issue of long monophthongs versus diphthongs in light of newer developments in AP/TD. Selection-coordination theory poses a feedback-induced suppression mechanism for the timing of gestural offsets, whereby speakers learn to use feedback to control gestural duration (Tilsen, 2016). This could facilitate different gestural durations very easily, as speakers can use feedback in different ways to suppress a gesture depending on the communicative demands. However, this framework still supports a two-target model for long vowels: Tilsen (2016) explicitly states that long vowels are the result of 'intentional reselection' of the same gesture and provides a developmental explanation for the emergence of such patterns. By extension, the gradient transition from a long monophthong to a diphthong can be viewed as gradual dissociation of the two component gestures, such that one gesture gradually takes on a different target value from the other gesture.

Having considered the phonological issues, let us now turn to the predictions made by the one-target model in terms of variation and change. There is no way for the one long target to become a two target vowel (i.e. undergo diphthongisation), without the categorical addition of another target, or some sort of 'splitting' of the long target. If this is the case, then diphthongisation of long vowels should be somewhat constrained, because it requires a structural reorganisation, but the evidence presented in Section 1.3 suggests that gradient diphthongisation is pervasive in English, both synchronically and diachronically. In comparison, the two-target model can produce gradient change from a long monophthong to a diphthong via variation in only the nucleus target value. This affords a historically accurate model of the gestural representation of English vowels. As shown for the *bee* example in Fig. 16, variation in the degree of diphthongisation can be captured through variation in the spatial properties of one of the component vowel targets – it represents variation in vowel quality, rather than variation in underlying vowel structure. Variation in vowel quality is ubiquitous, and from that point of view, we might expect that variation in the degree of diphthongisation is also common. This expectation is consistent with observed variation in English. In the Northern English data presented here, there are several vowels that vary between more monophthongal and more diphthongal realisations. This is the case for *bee, boo* and *beer*, and the same mechanism would account for variation between monophthongal *bay* and *beau* for some Northern English speakers in contrast to a more diphthongal pronunciation in the pan-regional standard. Similar variation also exists in other varieties of English (e.g. Standard Southern British English; Lindsey (2019)).

Following the same reasoning, historical changes such as diphthongisation and monophthongisation fall under a wider type of changes in vowel quality. In contrast, changes affecting vowel length are distinct, as they require major phonological reorganisation in terms of inserting or deleting a vowel target. The history of English illustrates a potential diachronic connection between phonological length and the presence of an underlying articulatory gesture. In non-rhotic varieties of English, many of the long monophthongs emerged historically from sequences of a short vowel followed by /r/. A plausible diachronic account of the changes is as follows. The coda /r/ initially involves a combination of a vocalic and a consonantal gesture, as is a common characteristic of liquids (Proctor, 2011; Proctor et al., 2019). The final liquid then undergoes a gestural reorganisation in which the consonantal gesture is gradually reduced. The vocalic gesture of the liquid, however, remains in place, becoming a diphthong offglide. This type of representation can capture some of the modern-day remnants of historical /r/ codas, as in the diphthongal variants of the NEAR vowel. The offglide may then gradually assimilate to the preceding nucleus, as we see for instance with monophthongal variants of NEAR, such that the only remnants of the offglide gesture are phonological length and phonetic duration. While not all long monophthongs descend from coda liquids, historical coda /r/ vowels, like START, NURSE and NEAR exemplify cases in which vowel length descends from a distinct gesture, potentially reinforcing analysis of vowel length in terms of articulatory sequences of vowels.

Some of the theoretical advantages of the two-target model of long monophthongs also set it apart from a third alternative sketched out in Section 1.3, namely a model that treats both diphthongs and long monophthongs as inherently dynamic but not compositional, along the lines of Xu et al. (2023). We have not modelled this scenario explicitly, as doing so would represent a major computational challenge and a considerable leap relative to extant models. However, we wish to note some arguments why this added computational complexity might not provide an improvement over a simpler two-target model. A key difference between a two-target model and a target + slope model is that the vowel offglide is a phonological object in the former case, but not in the latter. In a target model, offglide is a target, i.e. it corresponds to a phonological object. In a target + slope model, an offglide arises as a phonetic realisation of a gradient slope parameter setting, but it is not a category in itself. Diphthong offglides, however, frequently show some category-like behaviour. As shown by Hsieh (2017), many (but not all) diphthongs can be modelled as a composition of targets that function independently within the same vowel system. Indeed, in our own model, the offglide /i/ target has the same basic parameters as the onglide target of FLEECE with further gradient variation in the constriction degree, which mirrors such variation arising due to independent phonetic factors. Furthermore, diphthong offglides have a tendency for systemic behaviour. A striking example for this comes from the pattern of changes affecting the GOOSE and GOAT vowels in English. As noted by Labov (1994, p. 208), fronting of the GOOSE vowel frequently triggers subsequent fronting of GOAT. In a two-target model, this can be straightforwardly captured as a generalisa-

tion of fronting: fronting of the GOOSE offglide is generalised to the GOAT offglide. In contrast, it is not so clear how this type of systemic pressure could be captured in a model where vowel dynamics are derived from adjustment to the slope.

We do acknowledge that our articulatory model we have presented also has some limitations. First, we do not claim to propose optimal parameters for capturing the dynamics of long monophthongs, but instead show that we can reproduce qualitative distinctions that are evidenced in empirical data. Note also that our model assumes a single idealised speaker, but gestural parameters such as TBCD are not necessarily speaker-invariant or language-invariant. The model we propose here allows variation in gestural targets as a mechanism for gradient diphthongisation. We note that such a phenomenon is well captured by dynamical phonological models in which gestural parameters are assigned for an utterance via the evolution of dynamic fields, which could correspond to lexical representations (Kirov and Gafos, 2007). A proof-of-concept demonstrating a dynamic neural field model for gradient diphthongisation is presented in Kirkham and Strycharczuk (2024). Such a perspective facilitates gradient word-specific phonetic realisations, as well as the potential for dynamic change in phonological representations within and across speakers. We also acknowledge that the key argument for our proposal rests on the assumption that articulatory gestures have intrinsic timing. While a foundational assumption in AP/TD, this is an assumption nonetheless, and it has been called into question (Turk and Shattuck-Hufnagel, 2020).

We would argue that a two-target model of long vowels is promising in that it provides a way of reconciling categorical and gradient aspects of diphthongisation, and it also captures a range of phonetic, phonological and variationist facts. A reanalysis of variation and change affecting degree of diphthongisation in terms of changes to the quality of vowel target is yet to be verified through more systematic modelling. It remains to be seen whether manipulating gestural location and constriction can yield a plausible model of synchronic and diachronic differences in diphthongisation. Another aspect which requires refinement is that of exact timing of the component gestures. A reviewer points out that some varieties of English are distinguished by the timing of the diphthongs, e.g. for some speakers of Australian English *beet* and *boot* are produced with a long onglide, such that the target is reached late. Therefore, a complete model of vowel representation would need to capture the variation not just in the quality of the component gestures, but also in their relative timing. Another compelling question concerns distinguishing between diphthongs and vowel hiatus sequences in articulatory terms. Sequences of adjacent vowels may be heterosyllabic, showing distinct acoustic and phonological properties from diphthongs (Chitoran and Hualde, 2007). Diphthongs pattern distinctly from segment sequences, acting as unit; for example, they are not reversed in backward talking (Cowan et al., 1985). Shaw et al. (2021) propose an articulatory framework for distinguishing between complex segments and segment sequences, but their proposal involves in-phase coupling for complex segments, and as such, it is not readily extendable to the diphthong case, where the relevant component gestures are coordinated anti-phase.

Furthermore, we may ask to what extent the current model extends to models of diphthongs in other languages. The model we have proposed is intended for English, as it is motivated by language-specific phonetic and phonological behaviour. A similar model could be appropriate for long monophthongs in other languages that share phonological and phonetic similarities with diphthongs in the same language. However, there are languages where long monophthongs and diphthongs differ in crucial ways. For instance, Dutch /i/ and /u/ are phonologically long, as shown by their phonotactic behaviour, but they are phonetically short (Adank et al., 2004). This differentiates them from /i/ and /u/ in English, as well as from canonical diphthongs in Dutch, for which increased phonetic duration goes hand-in–hand with phonological length. As such, Dutch present a compelling future case study.

A consequence of our model is that diphthongs are not a distinct phonological category from monophthongs, because diphthongs and long monophthongs share the same phonological structure. In this view, diphthongs can be defined as a sequence of vowels coupled anti-phase that differ in their gestural specifications, whereas the component vowel targets for long monophthongs share the same gestural properties. This is well-supported on the representational level by our phonological modelling. However, identifying individual vowels as diphthongs and monophthongs may prove difficult in practice, because the spatial difference between two component targets can be phonetically very small, corresponding to a small degree of inherent change. There is a question of how much change is needed to be taken as meaningful, and it is not clear that a principled approach is possible, or indeed required in the absence of distinct phonological behaviour that would separate diphthongs from long monophthongs. Accordingly, the terms ‘monophthong’ and ‘diphthong’ become relevant only in the context of phonetic description - they can be used gradiently (‘more or less monophthongal / diphthongal’), or to denote typical examples.

A practical consequence is that it is not advisable to assume a distinction between long monophthongs and diphthongs, as is common methodological practice, for instance in acoustic studies of vowel variation. Performing acoustic measurements of vowels frequently involves a degree of dynamic reduction, and it might seem sensible to approach such reduction differently, depending on how much inherent change there is in a vowel. A common approach is taking measurements at a single timepoint for monophthongs and at two timepoints for diphthongs. However, the in-between cases as documented here demonstrate that dynamic differences are likely to emerge within a vowel category. A more principled approach is to use a minimum of two time point measurements for all long vowels, or using a data-driven approach to dynamic reduction (such as fPCA used here or Discrete Cosine Transformation). This may be superfluous in some cases, since a categorical distinction between monophthongs and diphthongs may emerge in some systems, but such a distinction should not be taken for granted.

6. Conclusion

This study set out to establish whether diphthong vowels can be distinguished from monophthongs by the number of underlying articulatory gestures: two in the case of diphthongs,

one in the case of monophthongs. The two-target representation for diphthongs is consistent with the kinematic properties of vowels as documented in this paper, and it is also supported by the results of computational simulations we presented. However, we have argued that the presence of two targets does not in itself entail diphthongisation, but rather, its main consequences are phonological weight and phonetic duration. Therefore, all phonologically long vowels can be modelled as compositional, i.e. having two targets. Within long vowels, degrees of diphthongisation can then emerge from the level of dissimilarity between the two component articulatory targets. The implication of this proposal is that long monophthongs and diphthongs are structurally identical, and therefore they are also predicted to share the same phonological behaviour. Another prediction is that any long vowel could in principle diphthongise, and that diphthongs and long monophthongs can morph into each other fairly freely. These predictions are largely consistent with our current knowledge of diphthongs in British English, but a systematic validation of the proposal would have to take the form of a detailed simulation of variation within the bounds of a particular vowel system showing that varying gestural specific parameters without manipulating the number of gestures can produce realistic patterns of variation. Another compelling avenue for follow-up research is a typological survey of diphthongisation, informed by articulation as well as phonological patterning and historical change, since the interplay between gestural specification, phonological weight and phonetic duration are likely to be language-specific.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank the participants for contributing to our study. We also thank the *Journal of Phonetics* editor, Taehong Cho, and reviewers for insightful comments on an earlier version of this manuscript. We are grateful to Ricardo Bermúdez-Otero, who provided helpful ideas and references, and to Darin Flynn for bringing the case of backward talking to our attention. We are also grateful to the audiences at the Third Angus McIntosh Centre Symposium, the 30th Manchester Phonology Meeting, and the 14th Language Variation and Change Conference for their questions and suggestions. Any remaining errors are our own. The research was funded by the Arts and Humanities Research Council Grant No. AH/S011900/1.

Appendix A

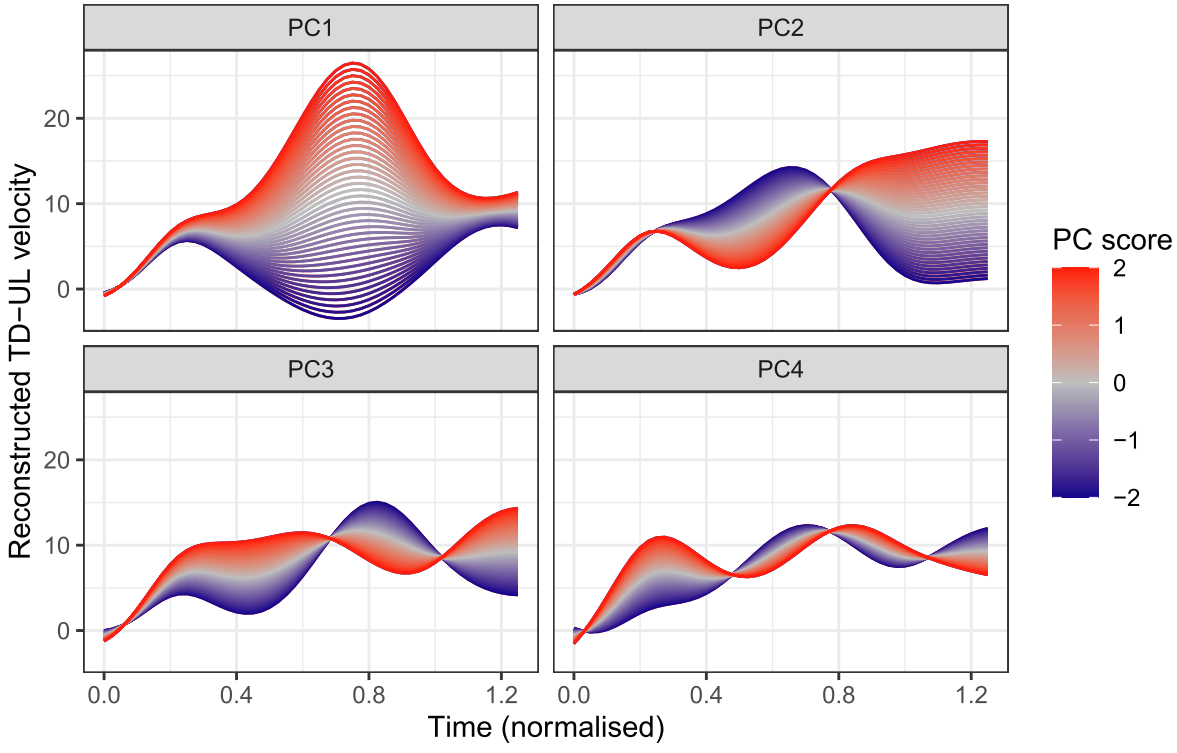

Fig. 17: Perturbation arising in the TD-UL velocity trajectory as a result of variation in the first four PC scores. The cumulative variance explained with the inclusion of subsequent PCs was: 0.63 (PC1), 0.85 (PC2), 0.97 (PC3), 0.98 (PC4).

Appendix B

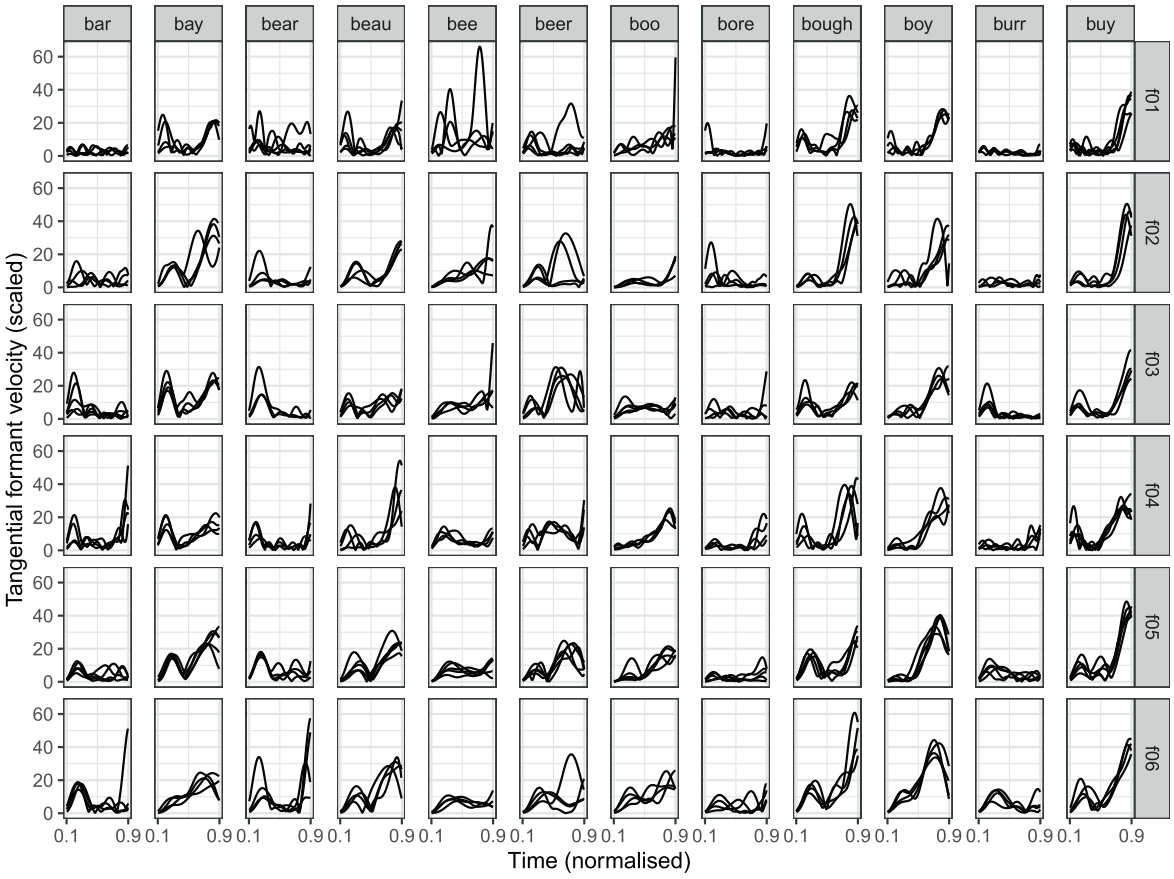

Fig. 18: Formant velocity profiles for all the individual vowel tokens.

Appendix C

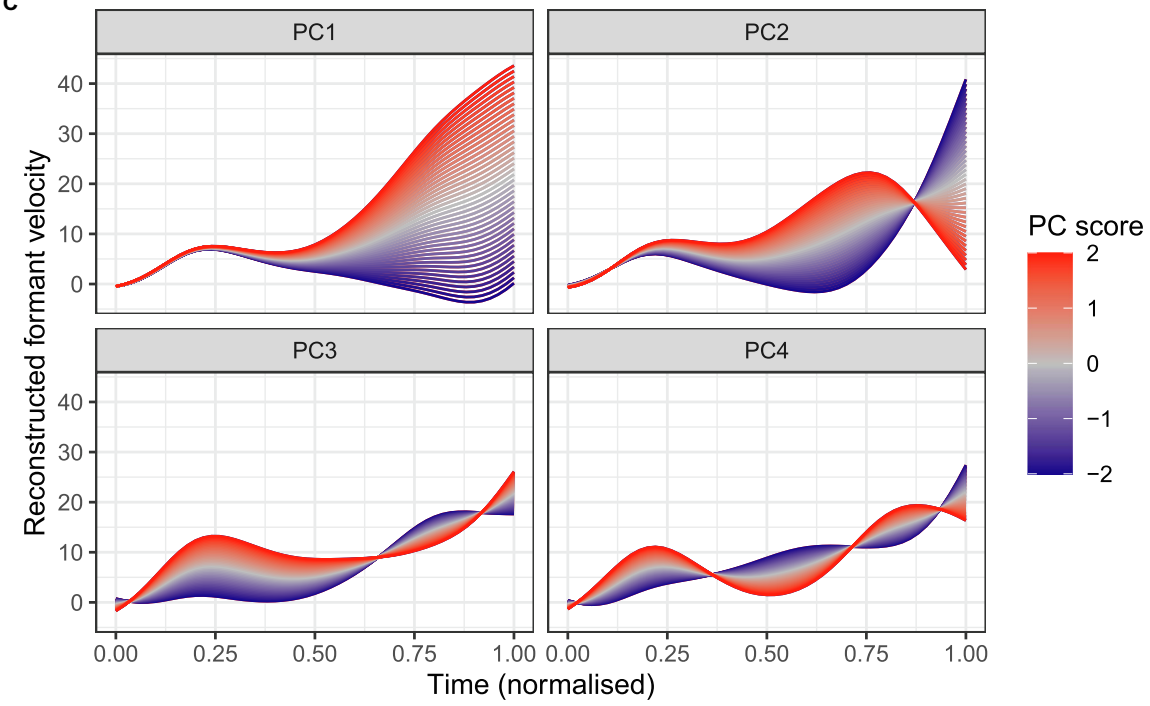

Fig. 19: Perturbation arising in the formant velocity trajectory as a result of variation in the first four PCs. The cumulative variance explained with the inclusion of subsequent PCs was: 0.63 (PC1), 0.87 (PC2), 0.95 (PC3), 0.99 (PC4).

Appendix D

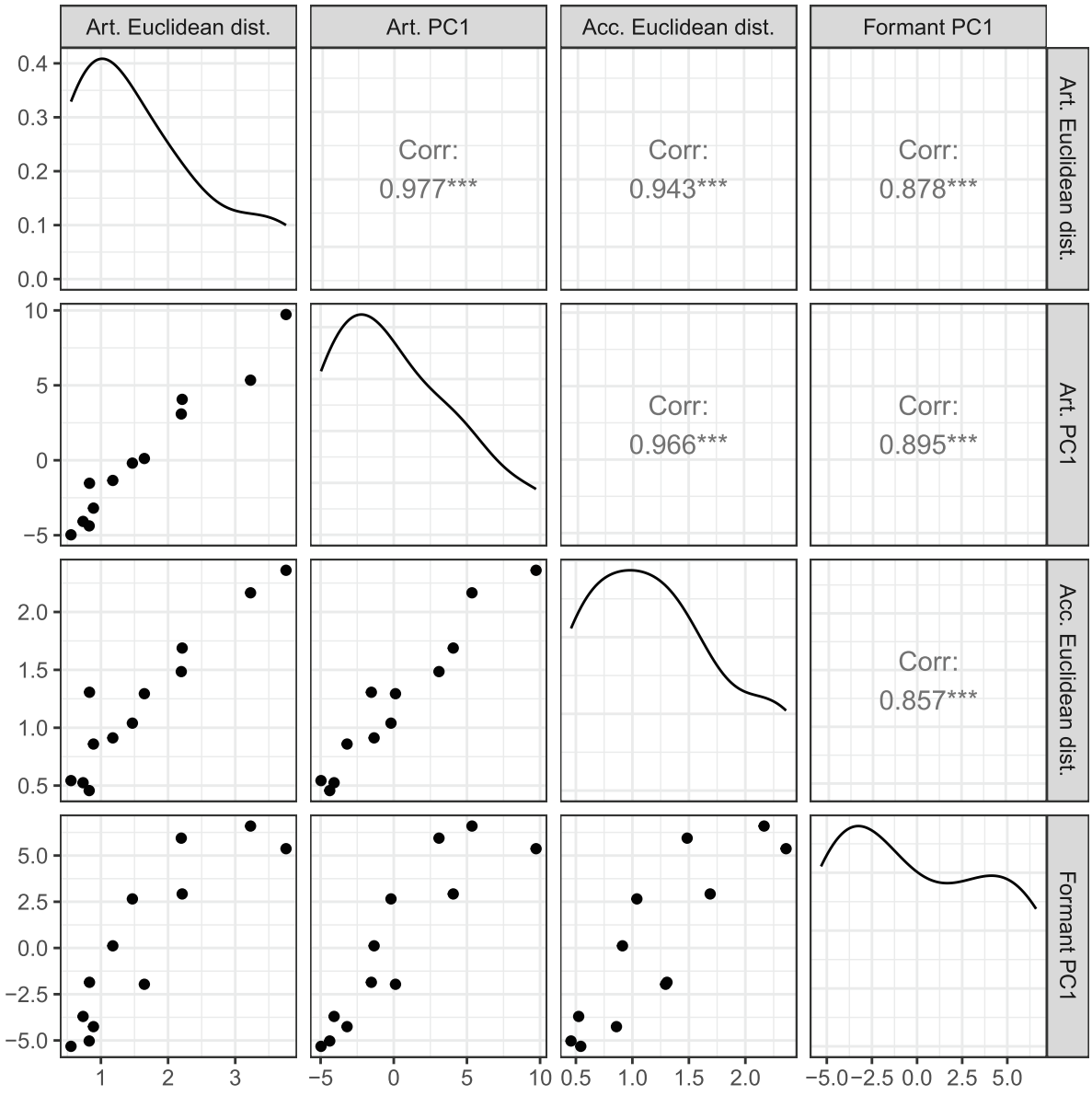

Fig. 20: Correlations between the by-item means of the four diphthongisation measures.

References

Adank, P., Van Hout, R., & Smits, R. (2004). An acoustic description of the vowels of Northern and Southern Standard Dutch. *The Journal of the Acoustical Society of America, 116*, 1729–1738.

Assmann, P. F., Nearey, T. M., & Hogan, J. T. (1982). Vowel identification: Orthographic, perceptual, and acoustic aspects. *The Journal of the Acoustical Society of America, 71*, 975–989.

Baranowski, M., & Turton, D. (2015). Manchester English. In R. Hickey (Ed.), *Researching Northern Englishes* (pp. 293–316). Amsterdam and Philadelphia: John Benjamins.

Barreda, S. (2021). Fast track: fast (nearly) automatic formant-tracking using Praat. *Linguistics Vanguard, 7*.

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*, 363–377.

Boersma, P., & Weenink, D. (2009). *Praat: Doing phonetics by computer [Computer programme]*. Version 6.2.14 http://www.praat.org/.

Browman, C. P., & Goldstein, L. (1986). Towards an articulatory phonology. *Phonology yearbook, 3*, 219–252.

Browman, C. P., & Goldstein, L. (1992). Articulatory phonology: An overview. *Phonetica, 49*, 155–180.

Browman, C. P., & Goldstein, L. (1995). Gestural syllable position effects in American English. In F. Bell-Berti & L. J. Raphael (Eds.), *Producing speech: Contemporary issues* (pp. 19–33). Springer.

Burgdorf, D. C., & Tilsen, S. (2021). Temporal differences between high vowels and glides are more robust than spatial differences. *Journal of Phonetics, 88* 101073.

Chitoran, I., & Hualde, J. I. (2007). From hiatus to diphthong: The evolution of vowel sequences in Romance. *Phonology, 24*, 37–75.

Collins, B., & Mees, I. M. (2003). *Practical phonetics and phonology: A resource book for students*. London: Routledge.

Cowan, N., Braine, M. D., & Leavitt, L. A. (1985). The phonological and metaphonological representation of speech: Evidence from fluent backward talkers. *Journal of Memory and Language, 24*, 679–698.

Cox, F., & Docherty, G. (2023). Sociophonetics and vowels. In C. Strelluf (Ed.), *The Routledge Handbook of Sociophonetics* (pp. 114–142). Routledge.

Cox, F., & Palethorpe, S. (2007). Australian English. *Journal of the International Phonetic Association, 37*, 341–350.

Cronenberg, J., Gubian, M., Harrington, J., & Ruch, H. (2020). A dynamic model of the change from pre-to post-aspiration in Andalusian Spanish. *Journal of Phonetics, 83* 101016.

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*, 2603–2618.

Georgeton, L., & Fougeron, C. (2014). Domain-initial strengthening on French vowels and phonological contrasts: Evidence from lip articulation and spectral variation. *Journal of Phonetics, 44*, 83–95.

Goldsmith, J. A. (1990). *Autosegmental and Metrical Phonology.* Basil Blackwell.

Gubian, M., Torreira, F., & Boves, L. (2015). Using Functional Data Analysis for investigating multidimensional dynamic phonetic contrasts. *Journal of Phonetics, 49*, 16–40.

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*, 371–403.

Harrington, J., & Cassidy, S. (1994). Dynamic and target theories of vowel classification: Evidence from monophthongs and diphthongs in Australian English. *Language and Speech, 37*, 357–373.

Hillenbrand, J. M., & Docherty, G. (2013). Static and dynamic approaches to vowel perception. In G. S. Morrison & P. F. Assman (Eds.), *Vowel Inherent Spectral Change* (pp. 9–30). Springer.

Honeybone, P. (2007). New dialect formation in nineteenth century Liverpool: a brief history of Scouse. In A. Grant & C. Grey (Eds). *The Mersey sound: Liverpool's language people and places* (pp. 106-140). Liverpool: Open House Press.

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

Jacewicz, E., & Fox, R. A. (2012). The effects of cross-generational and cross-dialectal variation on vowel identification and classification. *The Journal of the Acoustical Society of America, 131*, 1413–1433.

Jibson, J. (2022). *Temporal Formant Detail as a Function of Language Task* Ph.D. thesis. The University of Wisconsin-Madison.

Kirkham, S., & Strycharczuk, P. (2024). A dynamic neural field model of vowel diphthongisation. In *Proc. ISSP 2024–13th International Seminar on Speech Production* (pp. 205–208).

Kirkham, S., Strycharczuk, P., Gorman, E., Nagamine, T., & Wrench, A. (2023). Co-registration of simultaneous high speed ultrasound and electromagnetic articulography for speech production research. In R. Skarnitzl & J. Volín (Eds.), *Proceedings of the 20th International Congress of Phonetic Sciences* (pp. 918–922). Guarant International..

Kirov, C., & Gafos, A. I. (2007). Dynamic phonetic detail in lexical representations. In *Proceedings of the 16th International Congress of Phonetic Sciences* (pp. 637–640).

Labov, W. (1994). *Principles of Linguistic Change: Internal Factors* (vol. 1) Oxford: Blackwell.

Labov, W., Ash, S., & Boberg, C. (2006). *The atlas of North American English: Phonetics, Phonology and Sound Change*. Mouton de Gruyter.

Ladefoged, P. American English. (1999). *Handbook of the International Phonetic Association. A guide to the use of the International Phonetic Alphabet*.

Lehiste, I., & Peterson, G. E. (1961). Transitions, glides, and diphthongs. *The Journal of the Acoustical Society of America, 33*, 268–277.

Lindblom, B. E., & Studdert-Kennedy, M. (1967). On the role of formant transitions in vowel recognition. *The Journal of the Acoustical Society of America, 42*, 830–843.

Lindsey, G. (2019). *English after RP: Standard British Pronunciation Today.* Springer.

Marin, S. N. (2007). *Vowel to vowel coordination, diphthongs and Articulatory Phonology. Ph.D. thesis*. Yale University.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. *Proc. Interspeech, 2017*, 498–502.

Morrison, G. S., & Nearey, T. M. (2007). Testing theories of vowel inherent spectral change. *The Journal of the Acoustical Society of America, 122*, EL15–EL22.

Murtagh, F., & Legendre, P. (2014). Ward's hierarchical agglomerative clustering method: Which algorithms implement Ward's criterion? *Journal of Classification, 31*, 274–295.

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

Peterson, G. E., & Lehiste, I. (1960). Duration of syllable nuclei in English. *Journal of the Acoustical Society of America, 32*, 693–703.

Popescu, A., & Chitoran, I. (2022). Linking gestural representations to syllable count judgments: A cross-language test. *Laboratory Phonology, 13*.

Proctor, M. (2011). Towards a gestural characterization of liquids: Evidence from Spanish and Russian. *Laboratory Phonology, 2*, 451–485.

Proctor, M., Walker, R., Smith, C., Szalay, T., Goldstein, L., & Narayanan, S. (2019). Articulatory characterization of English liquid-final rimes. *Journal of Phonetics, 77* 100921.

Ratko, L., Proctor, M., & Cox, F. (2023). Gestural characterisation of vowel length contrasts in Australian English. *Journal of Phonetics, 98* 101237.

Reed, P. E. (2014). Inter-and intra-generational /æɪ/ monophthongization, indexicality, and Southern Appalachian identity. *Southern Journal of Linguistics, 38*, 159–194.

Roach, P. (2004). British English: Received Pronunciation. *Journal of the International Phonetic Association, 34*, 239–245.

Saltzman, E. L., & Munhall, K. G. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology, 1*, 333–382.

Shaw, J. A., Oh, S., Durvasula, K., & Kochetov, A. (2021). Articulatory coordination distinguishes complex segments from segment sequences. *Phonology, 38*, 437–477.

Sorensen, T., & Gafos, A. I. (2016). The gesture as an autonomous nonlinear dynamical system. *Ecological Psychology, 28*(4), 188–215.

Sotiropoulou, S., Gibson, M., & Gafos, A. (2020). Global organization in Spanish onsets. *Journal of Phonetics, 82*, 100995.

Stevens, K. N., & Keyser, S. J. (1989). Primary features and their enhancement in consonants. *Language, 65*, 81–106.

Strange, W. (1989). Evolving theories of vowel perception. *The Journal of the Acoustical Society of America, 85*, 2081–2087.

Strycharczuk, P., López-Ibáñez, M., Brown, G., & Leemann, A. (Eds.). (2020). *General Northern English. Exploring regional variation in the North of England with machine learning. Frontiers in Artificial Intelligence* (pp. 48).

Sweet, H. (1910). *The sounds of English: An introduction to phonetics*. Clarendon Press.

Tiede, M. (2010). MVIEW: Multi-channel visualization application for displaying dynamic sensor movements. *Computer programme*.

Tilsen, S. (2016). Selection and coordination: The articulatory basis for the emergence of phonological structure. *Journal of Phonetics, 55*, 53–77.

Tilsen, S. (2018). Three mechanisms for modeling articulation: Selection, coordination, and intention. *Cornell Working Papers in Phonetics and Phonology*, 1–49.

Tilsen, S. (2022). An informal logic of feedback-based temporal control. *Frontiers in Human Neuroscience, 16*(851991), 1–29.

Trubetzkoy, N.S. (1971). *Principles of phonology.* Translation of *Grundzüge der Phonologie* by Christiane A.M. Baltaxe, Second edition.

Turk, A., & Shattuck-Hufnagel, S. (2020). *Speech timing: implications for theories of phonology, speech production, and speech motor control*. Oxford: Oxford University Press.

Turton, D., & Lennon, R. (2023). An acoustic analysis of rhoticity in Lancashire, England. *Journal of Phonetics, 101* 101280.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., & van Mulbregt, P. SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods, 17*, 261–272.

Ward, J. H. Jr., (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association, 58*, 236–244.

Watson, K. (2007). Liverpool English. *Journal of the International Phonetic Association, 37*, 351–360.

Watt, D. (2002). 'I don't speak with a Geordie accent, I speak, like, the Northern accent': contact-induced levelling in the Tyneside vowel system. *Journal of Sociolinguistics, 6*, 44–63.

Wells, J. (1982). *Accents of English 2: The British Isles* (vol. 2) Cambridge: Cambridge University Press.

Wells, J. (1982). *Accents of English 3: Beyond the British Isles* (vol. 3) Cambridge: Cambridge University Press.

Williams, D., & Escudero, P. (2014). A cross-dialectal acoustic comparison of vowels in Northern and Southern British English. *The Journal of the Acoustical Society of America, 136*, 2751–2761.

Xu, A., Gerazov, B., van Niekerk, D., Krug, P. K., Prom-on, S., Birkholz, P., & Xu, Y. (2023). Computational models for articulatory learning of English diphthongs: One dynamic target vs. two static targets. In R. Skarnitzl & J. Volín (Eds.), *Proceedings of the 20th International Congress of Phonetic Sciences* (pp. 4140–4144). Prague: Guarant International.