Dynamic acoustic vowel distances within and across dialects

Cynthia G. Clopper[1, a]

[1]*Department of Linguistics, Ohio State University, Columbus, Ohio 43210, USA*

Vowels vary in their acoustic similarity across regional dialects of American English, such that some vowels are more similar to one another in some dialects than others. Acoustic vowel distance measures typically evaluate vowel similarity at a discrete time point, resulting in distance estimates that may not fully capture vowel similarity in formant trajectory dynamics. In the current study, language and accent distance measures, which evaluate acoustic distances between talkers over time, were applied to the evaluation of vowel category similarity within talkers. These vowel category distances were then compared across dialects and their utility in capturing predicted patterns of regional dialect variation in American English was examined. Dynamic time warping of mel-frequency cepstral coefficients was used to assess acoustic distance across the frequency spectrum and captured predicted Southern American English vowel similarity. Root-mean-square distance and generalized additive mixed models were used to assess acoustic distance for selected formant trajectories and captured predicted Southern, New England, and Northern American English vowel similarity. Generalized additive mixed models captured the most predicted variation, but, unlike the other measures, do not return a single acoustic distance value. All three measures are potentially useful for understanding variation in vowel category similarity across dialects.

[a] Email: clopper.1@osu.edu

I. INTRODUCTION

Regional dialects of American English vary in the acoustic-phonetic realization of vowel categories (Labov et al., 2006). This variation leads to vowel categories that are more similar in acoustic-phonetic space in some dialects than others. For example, the Northern Cities vowel shift leads to greater acoustic-phonetic similarity of /ɛ æ/ in the Northern dialect than in the Midland dialect (Clopper and Tamati, 2014), the Southern vowel shift leads to greater acoustic-phonetic similarity of /eɪ ɛ/ in the Southern dialect than in the Western dialect (Farrington et al., 2018), and the low-back merger leads to greater acoustic-phonetic similarity of /ɑ ɔ/ in California English than in New York City English (Nycz and Hall-Lew, 2013).

Acoustic vowel category distance measures[1] are typically defined at a discrete time point within the vowels of interest (e.g., Hay et al., 2006; Kendall and Fridland, 2012; Wassink, 2006). The goal of the current study was to assess dynamic acoustic vowel category distance measures that capture formant trajectories over time within the vowels of interest (e.g., Fox and Jacewicz, 2009; Renwick and Stanley, 2020). These dynamic measures of vowel category distance within regional dialects of American English were evaluated in comparison to qualitative descriptions of regional variation (Labov et al., 2006). The results demonstrated that dynamic vowel category distance measures, including dynamic time warping (DTW) of mel-frequency cepstral coefficients (MFCCs; Bartelds et al., 2020; Lind-Combs et al., 2023; Mielke, 2012), root-mean-square distance (RMSD; Cole et al., 2023; Kaland, 2023), and generalized additive mixed modeling (GAMM; Kirkham et al., 2019; Renwick and Stanley, 2020), can all capture variation in vowel category distances within and across dialects of American English, complementing prior qualitative descriptions.

A. Vowel category distance measures

Regional variation in acoustic-phonetic similarity of vowel categories has been quantified in previous work using Euclidean distances and Pillai scores of formant frequency estimates (see Kelley

and Tucker, 2020; Nycz and Hall-Lew, 2013, for reviews). Euclidean distances between vowel categories are typically defined in the two-dimensional first formant (F1) x second formant (F2) space, using formant estimates from a discrete time point (e.g., vowel midpoint). Euclidean distances between vowels have been used to predict sound change (Wieling et al., 2012), as a proxy for individual participation in ongoing vowel shifts (Farrington et al., 2018; Kendall and Fridland, 2012), to predict accentedness ratings (Gunter et al., 2020), and to assess the effects of lexical competition on dialect variation (Clopper and Tamati, 2014).

Pillai scores are also typically calculated from estimates of F1 and F2 from a discrete time point. Unlike Euclidean distances, Pillai scores quantify the overlap of the vowel category distributions in the F1 x F2 space, instead of simply the distance between the two category means. Pillai scores are used most commonly as a measure of vowel category merger within individuals (Gunter et al., 2020; Hay et al., 2006; Nycz and Hall-Lew, 2013). A number of related measures of multidimensional category overlap have also been proposed. These approaches typically involve estimates of F1 and F2 of the target vowels from a discrete time point, but may also include duration (Wassink, 2006), discrete cosine transformations of the formant frequency trajectories (Elvin et al., 2016), or formant frequencies sampled at multiple discrete time points (Morrison, 2008). These approaches have been used to identify the dimensions of variation that are necessary to distinguish vowel categories within a language or dialect (Elvin et al., 2016; Haynes and Taylor, 2014; Morrison, 2008; Wassink, 2006).

In principle, these vowel category distance measures are appropriate for assessing variation in vowel distances within and across varieties of American English. For example, Euclidean distances between /eɪ ɛ/ have been used to distinguish Southern from Western varieties of American English (Farrington et al., 2018) and Pillai scores for /ɑ ɔ/ have been used to distinguish degrees of the low-back vowel merger in California and New York City varieties (Nycz and Hall-Lew, 2013). However, at least as they are typically implemented, these measures are based on a single sample of each

formant estimated at a discrete time point for each vowel token. This implementation therefore does not capture vowel category distances at other time points in the formant trajectories or over the entire formant trajectory, despite clear evidence that vowels vary within and across dialects of American English in their formant trajectory dynamics (Farrington et al., 2018; Fox and Jacewicz, 2009).

In contrast to this work on vowel category distances, dynamic acoustic distance measures have been used to assess consonant category distances (Mielke, 2012) and lexical distances (Kelley, 2023; Kelley and Tucker, 2022). Mielke (2012) used dynamic time warping of mel-frequency cepstral coefficients, estimated over VCV utterances, to assess the relationship between acoustic and phonological consonant similarity. Similarly, both Kelley (2023) and Kelley and Tucker (2022) used dynamic time warping of mel-frequency cepstral coefficients, estimated over words, to predict lexical competition in speech processing. In the current study, this DTW approach was applied to vowel category distances to assess variation within and across dialects of American English.

Acoustic distances between languages, dialects, and accents have also been quantified in previous work using dynamic acoustic distance measures (e.g., Bartelds et al., 2020; Chernyak et al., 2024; Heeringa et al., 2009). These acoustic language and accent distance measures capture distances between different talkers' productions of the same linguistic content (e.g., words, sentences), whereas the acoustic vowel distance measures, such as Euclidean distances and Pillai scores, capture distances between different vowel categories produced by the same talkers. In the current study, the measures of between-talker language and accent distances were applied to within-talker vowel category distances to assess dynamic vowel category similarity within dialects of American English and to compare those vowel similarities across dialects. These comparisons across dialects were then evaluated in the context of previous qualitative descriptions of regional variation (Labov et al., 2006).

B. Language and accent distance measures

Three general approaches to quantifying acoustic language and accent distance have been proposed in the literature: dynamic time warping (Bartelds et al., 2020), mean distances over time (Heeringa et al., 2009), and generalized additive mixed modeling (Kirkham et al., 2019). Bartelds et al. (2020) and Lind-Combs et al. (2023) both used dynamic time warping of MFCCs, estimated over words and sentences, respectively, to assess native and non-native accent distances. This approach differs from the Euclidean distance and Pillai score measures of vowel category distance, and from the mean distance and GAMM approaches to language and accent distance, in that it includes a summary representation of the spectrum over the frequency analysis range (i.e., up to the Nyquist frequency), instead of focusing on estimated formant frequencies and their trajectories over time. DTW distances of MFCCs therefore capture dynamic formant trajectory distance, but may also capture distances between consonants, f0 contours, voice quality, and even recording devices and background noise (Bartelds et al., 2020; Lind-Combs et al., 2023). MFCCs have also been used without DTW to assess dialect similarity using multidimensional scaling and clustering techniques (Ferragne and Pellegrino, 2010). DTW of MFCCs was adopted in the current study as a measure of overall vowel distance, including dynamic formant, f0, and voice quality information, within and across dialects. The MFCCs were estimated over the target vowels only, so that consonant effects were minimized, and all distances were calculated within-talker, so that effects of recording device and background noise were also minimized.

Heeringa et al. (2009) used Manhattan (or city-block) distances of formant estimates over time to quantify Norwegian dialect distances. This approach is conceptually similar to Euclidean distance measures of vowel category distance, except that (1) distances were calculated every 10 ms over entire words, including both consonants and vowels, and (2) Manhattan distances were used instead of Euclidean distances. Another conceptually similar measure is root-mean-square distance, in which

distance is defined as the mean Euclidean distance over time. This measure has been used to quantify the similarity of intonation contours (Cole et al., 2023; Kaland, 2023) and was adopted in the current study as a measure of dynamic formant trajectory distance within and across dialects.

Generalized additive mixed models of formant trajectories involve fitting curves to formant estimates over time and then identifying when in time the trajectories overlap and when they diverge. This approach differs from the other distance measures in that a discrete, positive distance measure is not returned. Rather, GAMMs provide an indication of (1) when in time the two formant trajectories differ, and (2) the direction of this difference (i.e., positive or negative). GAMMs have been used to compare vowel-lateral sequences in varieties of UK English (Kirkham et al., 2019), features of Southern American English among white and Black talkers in the American South (Renwick and Stanley, 2020), and Australian vowel change over time (Cox et al., 2024). GAMMs were adopted in the current study to assess the temporal properties of formant trajectory differences within and across regional dialects of American English.

C. The current study

Dynamic time warping of MFCCs, root-mean-square distance, and generalized additive mixed modeling were applied to three small datasets to assess their utility in quantifying dynamic acoustic vowel distances within and across dialects of American English. Each dataset comprised the stimulus materials from a previous perception task designed to examine cross-dialect lexical processing (Clark et al., 2022; Clopper and Walker, 2017; Ross and Clopper, 2023). Each dataset included minimal pair tokens for two vowel contrasts that were expected to differ in their acoustic similarity across two talker dialects, based on previous descriptions of regional variation in American English (Labov et al., 2006). Across datasets, the target vowel contrasts included /æ ɛ/ and /ɪ eɪ/ to capture features of the Northern Cities vowel shift, /ɑɪ ɑ/ and /ɪ ɛ/ to capture features of the Southern vowel shift, and /ɑɹ ɑ/ to capture non-rhoticity in New England. The three dynamic

120 distance measures were evaluated against the predicted patterns of vowel category similarity for
121 these vowel contrasts. The vowel contrasts of interest were not predicted to be merged in any of the
122 target varieties and the central question was therefore how to capture varying non-zero distances
123 between vowel categories across regional dialects.

124 Given that the DTW approach successfully predicts perceptual native and non-native accent
125 distances (Bartelds et al., 2020; Lind-Combs et al., 2023), DTW distances were expected to produce
126 the predicted patterns of differential vowel distance across dialects. However, the cross-dialect vowel
127 distance predictions are based on patterns of overall vowel shifts in one or two dimensions of the F1
128 x F2 vowel space (e.g., raising and fronting of /æ/ in the Northern Cities shift) and/or patterns of
129 vowel formant trajectory dynamics in one or two dimensions of the F1 x F2 x F3 vowel space (e.g.,
130 lowering of F3 in /ɑɹ/ sequences in rhotic varieties). Given that DTW distances were calculated
131 over MFCCs, capturing more spectral information than just one or two formant trajectories, the
132 predicted patterns of differential vowel distance across dialects may be masked by other similarities
133 and differences in the spectra across tokens. For example, predicted differences in F1 might be
134 masked by similarities in F2 and higher formants. The RMSD and GAMM analyses were therefore
135 based on target formant frequencies that were selected to highlight the expected cross-dialect
136 variation. This intentional focus on the target variation was expected to produce strong evidence of
137 the predicted patterns of differential vowel distance across dialects for these measures, as in previous
138 related work (Cox et al., 2024; Kirkham et al., 2019; Renwick and Stanley, 2020). Finally, given that
139 DTW distances and RMSDs were estimated over the entire vowel trajectory, the predicted patterns
140 of differential vowel distance across dialects may not be observed if they are limited to a short
141 temporal span of the trajectory. For example, predicted differences in the offglide of a diphthong
142 might be masked by similarities in the nucleus of the diphthong. The GAMM analysis allowed for a

consideration of this temporal detail and was expected to produce the strongest evidence of the predicted patterns of differential vowel distance across dialects.

II. METHODS

A. Datasets

The first dataset (Midland/Northern) comprises the stimulus materials from Clopper and Walker's (2017) cross-modal lexical decision task. The stimulus talkers were three Midland and three Northern female young adults. The target vowel contrasts were /æ ɛ/ and /ɪ eɪ/. These vowel contrasts were selected to capture variation across the two dialects due to the Northern Cities vowel shift. In particular, in the Northern Cities shift, /æ/ is raised and fronted and /ɛ ɪ/ are lowered and backed (Labov et al., 2006). The raising and fronting of /æ/ and the lowering and backing of /ɛ/ is predicted to lead to a smaller distance between these vowels for Northern talkers than Midland talkers. In both varieties, /ɪ/ is lower and backer than /eɪ/, so the lowering and backing of /ɪ/ in the Northern Cities shift is predicted to lead to a larger distance between these vowels for Northern talkers than Midland talkers. The stimulus materials were 48 minimal pairs (24 per vowel contrast) for each of the six talkers. There were two missing tokens and the minimal pair token for each of these missing tokens was also excluded, leaving 286 minimal pairs (47-48 per talker) for analysis.

The second dataset (Southern/Northern Virginian) comprises the stimulus materials from Clark et al.'s (2022) cross-modal lexical decision task. The stimulus talkers were four Southern and four Northern Virginian female young adults. The target vowel contrasts were /ɑɪ ɑ/ and /ɪ ɛ/. These vowel contrasts were selected to capture variation across the two dialects due to the Southern vowel shift. In particular, in the Southern vowel shift, /ɑɪ/ is monophthongized and /ɪ ɛ/ are raised and fronted, with greater raising and fronting of /ɛ/ than /ɪ/ (Labov et al., 2006). The monophthongization of /ɑɪ/ is predicted to lead to a smaller distance between /ɑɪ ɑ/ for Southern

talkers than Northern Virginian talkers. The more advanced raising and fronting of /ɛ/ than /ɪ/ is likewise predicted to lead to a smaller distance between these vowels for Southern talkers than Northern Virginian talkers. The stimulus materials were 69 word pairs (33 for /ɑɪ ɑ/ and 36 for /ɪ ɛ/) for each of the eight talkers. There were eight missing tokens and the minimal pair token for each of these missing tokens was also excluded, leaving 544 minimal pairs (67-69 per talker) for analysis.

The third dataset (New England/Northern) comprises the stimulus materials from Ross and Clopper's (2023) cross-modal lexical decision task. The stimulus talkers were two New England (one female, one male) and two Northern (one female, one non-binary) young adults. The target vowel contrasts were /ɑɪ ɑ/ and /æ ɛ/. These vowel contrasts were selected to capture variation across the two dialects in rhoticity and the Northern Cities shift, respectively. In particular, the New England dialect is non-rhotic (Labov et al., 2006), which is predicted to lead to a smaller distance between /ɑɪ ɑ/ for New England talkers than Northern talkers. As in the Midland/Northern dataset, the raising and fronting of /æ/ and the lowering and backing of /ɛ/ in the Northern Cities shift is predicted to lead to a smaller distance between these vowels for Northern talkers than New England talkers. The stimulus materials were 48 word pairs (24 per vowel contrast) for each of the four talkers. No tokens were missing, so 192 minimal pairs were analyzed.

B. Acoustic distance analysis

For each dataset, the word tokens were stored in separate digital sound files that were segmented to the word onset and offset. These sound files were down-sampled to 16 kHz with 16-bit quantization for analysis. Each token was forced-aligned using the Penn Phonetics Lab Forced Aligner (Yuan and Liberman, 2008) to obtain a preliminary segmentation of the target vowel. These vowel alignments were then hand-corrected, following the segmentation guidance provided by

Peterson and Lehiste (1960). For the New England/Northern dataset, the target vowel /ɑ/ was merged with the following /ɹ/ for all /ɑɹ/ tokens so that /ɑɹ ɑ/ distances could be estimated.

1. Dynamic time warping

The first distance measure was dynamic time warping of mel-frequency cepstral coefficients, following Bartelds et al. (2020) and Lind-Combs et al. (2023). The target vowel was extracted from each token with a Hamming window using the hand-corrected boundaries. MFCCs were then extracted using the *mfcc* function in the *python_speech_features* package in Python (Lyons et al., 2020) from each 25 ms window in each vowel token in 10 ms steps. For each window, 39 coefficients were extracted, including the overall energy and the first 12 cepstral coefficients, along with the first- and second-order derivatives of these measures. The spectrum used to calculate the MFCCs was extracted using a 1024-point FFT with 0.97 pre-emphasis. The MFCCs were z-scored over time separately for each coefficient for each token. Dynamic time warping was then used to estimate the distance between each minimal pair for each talker. The DTW distances were time-normalized to account for differences in overall duration of the paired tokens (Bartelds et al., 2020).

The DTW distances were analyzed using separate linear mixed-effects models for each dataset with talker dialect, vowel contrast, and their interaction as fixed effects. The maximal data-driven by-talker and by-minimal-pair random effects were used for each model. The fixed effects were sum-contrast coded. Statistical significance was assessed using the Satterthwaite approximation of degrees of freedom, as implemented in the *lmerTest* package in R (Kuznetsova et al., 2017). Post-hoc pairwise comparisons of significant effects were conducted using Tukey's HSD tests in the *emmeans* package in R (Lenth, 2024).

2. Root-mean-square distance

The second distance measure was root-mean-square distance of selected formant trajectories over time, following Cole et al.'s (2023) and Kaland's (2023) analyses of intonation contours. For

each token, the first three formant frequencies (F1, F2, and F3) were estimated at 10% intervals over the vowel duration from 0-100% (11 estimates per vowel) using a 12th order LPC analysis (Burg method) in the frequency range of 0-5500 Hz in Praat (Boersma and Weenink, 2023). All formant estimates were converted to Bark to facilitate comparisons across talkers and formants (Traunmüller, 1990). Given a large number of missing formant estimates at vowel onset (0%) and offset (100%), RMSDs were calculated for each minimal pair for each talker over the middle 80% of the vowel duration (10-90%). RMSDs were calculated separately for each formant and one formant was selected for each vowel contrast for analysis. For the Midland/Northern dataset, F1 was selected for both /æ ɛ/ and /ɪ eɪ/ distances to capture effects of raising of /æ/ and lowering of /ɛ ɪ/ in the Northern Cities shift. For the Southern/Northern Virginian dataset, F2 was selected for /ɑɪ ɑ/ distances to capture effects of /ɑɪ/ monophthongization (i.e., reduction of the fronting offglide) in the Southern vowel shift and F1 was selected for /ɪ ɛ/ distances to capture effects of raising of /ɪ ɛ/ in the Southern vowel shift. For the New England/Northern dataset, F3 was selected for /ɑɹ ɑ/ distances to capture effects of non-rhoticity in New England and F1 was selected for /æ ɛ/ distances to capture effects of raising of /æ/ and lowering of /ɛ/ in the Northern Cities shift, as in the Midland/Northern dataset.

RMSDs were analyzed using separate linear mixed-effects models with the same specifications as in the DTW analysis. Due to missing formant estimates, the analysis of the Midland/Northern dataset was based on 283 minimal pairs, the analysis of the Southern/Northern Virginian dataset was based on 543 minimal pairs, and the analysis of the New England/Northern dataset was based on 189 minimal pairs.

3. Generalized additive mixed modeling

The third distance measure involved generalized additive mixed models of selected formant trajectories over time, following Renwick and Stanley (2020). The same selected formant estimates

from the RMSD analysis were used in the GAMMs, except that the entire formant trajectory (0-100%) was used because GAMMs are more robust to missing data than RMSDs. For each minimal pair for each talker, the formant trajectory difference was calculated by subtracting the formant estimate at each time point for one member of the minimal pair from the other. These formant difference trajectories were then analyzed separately for each vowel contrast using GAMMs with a parametric effect of talker dialect and talker dialect smooths over time as fixed effects and with by-talker smooths over time and minimal pair x talker dialect smooths over time as random effects. A correction term for autocorrelation at lag 1 was also included in each model.

III. RESULTS

A. Dynamic time warping

A summary of the mean DTW distances for each talker for each vowel contrast in each of the three datasets is shown in Fig. 1. The by-talker means overlap considerably for both vowel contrasts in the Midland/Northern (Fig. 1a) and New England/Northern (Fig. 1c) datasets, whereas they show clear separation by talker dialect in the Southern/Northern Virginian (Fig. 1b) dataset. As expected, the distances are smaller for both the /ɑɪ ɑ/ and /ɪ ɛ/ contrasts for the Southern talkers than the Northern Virginian talkers.

The linear mixed effects model predicting DTW distances from talker dialect and vowel contrast for the Midland/Northern dataset revealed a marginal main effect of talker dialect (β = .08, F(1, 236.5) = 3.60, p = .059). The Midland talkers had larger DTW distances overall, as expected for the /æ ɛ/ contrast, but contrary to the prediction for the /ɪ eɪ/ contrast. Neither the main effect of vowel contrast nor the interaction were significant. The model for the Southern/Northern Virginian dataset revealed a significant main effect of talker dialect (β = .25, F(1, 7.1) = 29.82, p < .001). The Northern Virginian talkers had larger DTW distances overall, as expected for both vowel contrasts.

Neither the main effect of vowel contrast nor the interaction were significant. The model for the New England/Northern dataset revealed a significant interaction between talker dialect and vowel contrast (β = -0.10, F(1, 140.0) = 4.80, p = .030). The interaction reflects the expected cross-over pattern, in which DTW distances were larger for the Northern talkers than the New England talkers for the /ɑɹ ɑ/ contrast (β = -.29), but larger for the New England talkers than the Northern talkers for the /æ ɛ/ contrast (β = .10). However, post-hoc comparisons of estimated marginal means revealed that the talker dialect effect was not significant for either vowel contrast. The main effects of talker dialect and vowel contrast were also not significant.

Overall, the DTW results are suggestive of the predicted patterns, but are only statistically robust with these small samples in the Southern/Northern Virginian dataset. Moreover, the marginal result in the Midland/Northern dataset is contrary to predictions.

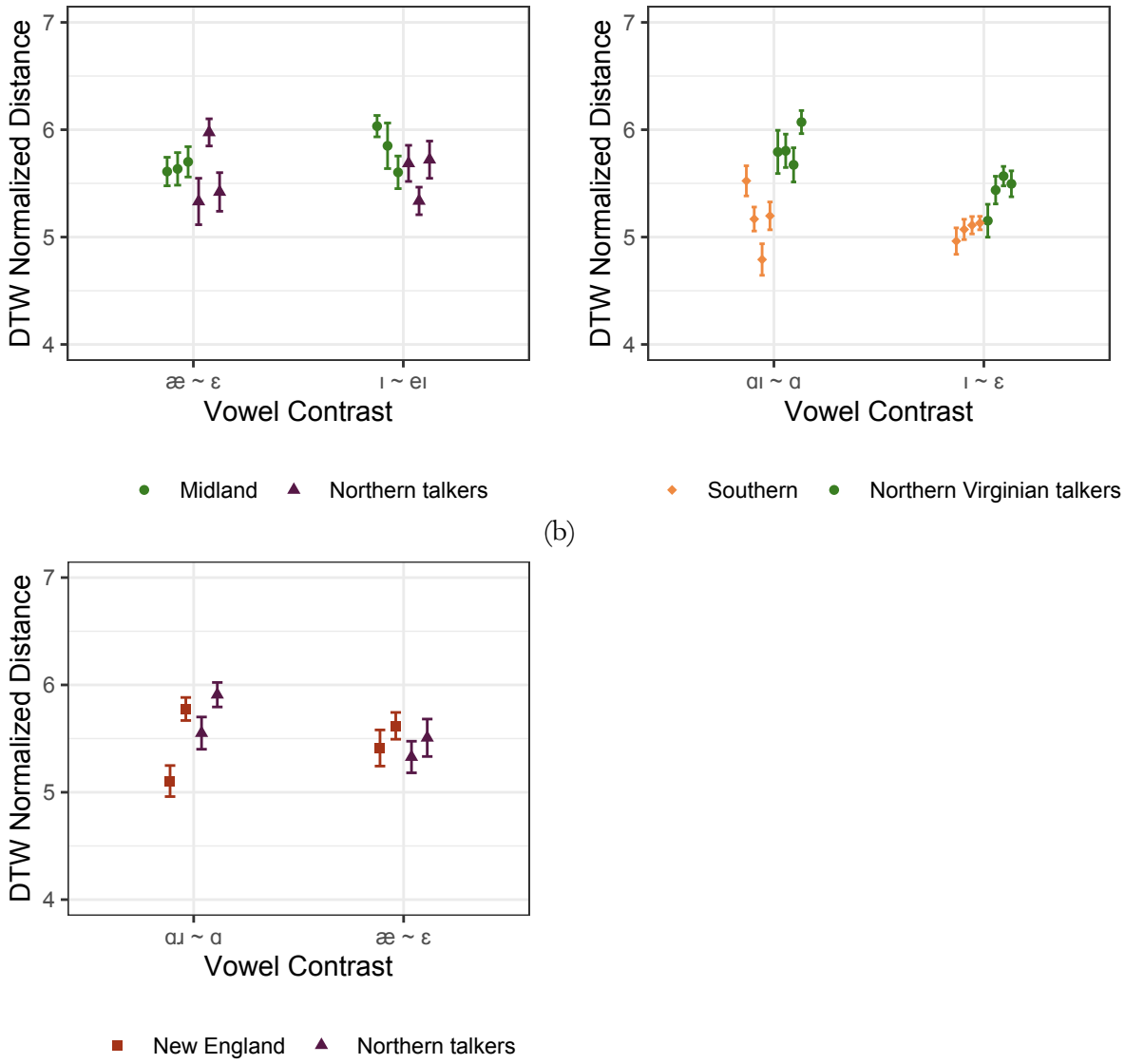


FIG 1. (Color online). Mean DTW distances for each talker for each vowel contrast for the Midland/Northern (a), Southern/Northern Virginian (b), and New England/Northern (c) datasets. Error bars are standard errors.

B. Root-mean-square distance

A summary of the mean RMSDs for each talker for each vowel contrast in each of the three datasets is shown in Fig. 2. The by-talker means show the expected talker dialect differences for the /ɪ eɪ/ contrast in the Midland/Northern (Fig. 2a) dataset, for both vowel contrasts in the

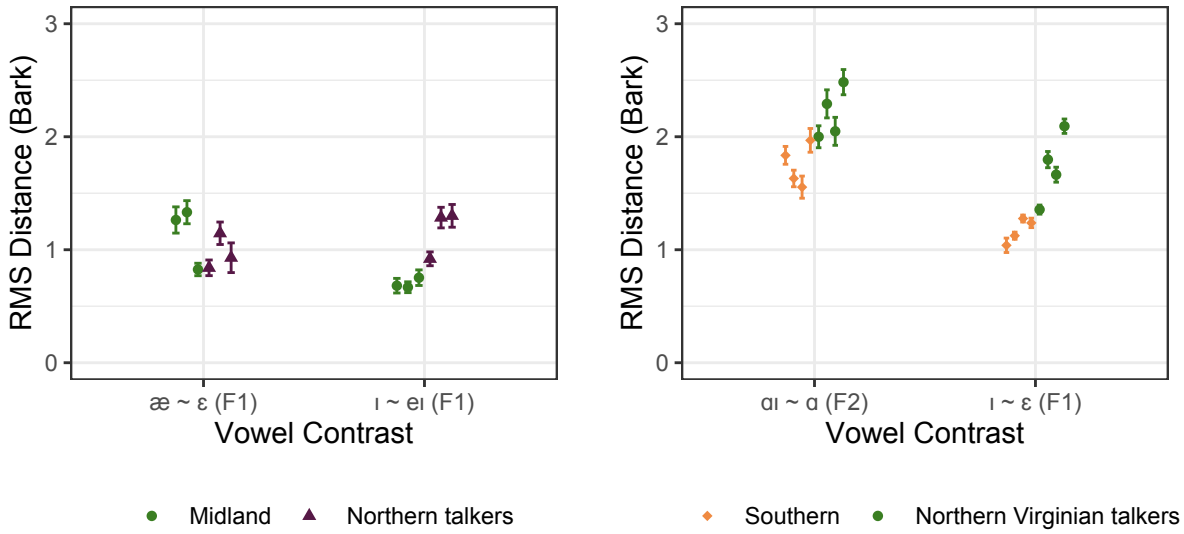
Southern/Northern Virginian (Fig. 2b) dataset, and for the /ɑɪ ɑ/ contrast in the New England/Northern (Fig. 2c) dataset. The by-talker means overlap considerably for the /æ ɛ/ contrast in both the Midland/Northern and New England/Northern datasets.

The linear mixed effects model predicting RMSDs from talker dialect and vowel contrast for the Midland/Northern dataset revealed a significant interaction between talker dialect and vowel contrast (β = -.16, F(1, 5.2) = 8.54, p = .032). Post-hoc comparisons of estimated marginal means for each vowel contrast revealed a significant effect of talker dialect for the /ɪ eɪ/ contrast (β = -.46, t(5.5) = -3.34, p = .018). As expected, the Northern talkers had larger RMSDs than the Midland talkers. The talker dialect effect was not significant for the /æ ɛ/ contrast. The main effects of talker dialect and vowel contrast were also not significant.

The model for the Southern/Northern Virginian dataset revealed significant main effects of talker dialect (β = .25, F(1, 6.4) = 12.75, p = .011) and vowel contrast (β = .26, F(1, 25.1) = 28.73, p < .001). As expected, the Northern Virginian talkers had larger RMSDs than the Southern talkers overall. The overall RMSDs were also larger for the /ɑɪ ɑ/ contrast than the /ɪ ɛ/ contrast. This vowel contrast effect likely reflects inherent differences between the two contrasts: whereas the /ɑɪ ɑ/ contrast involves a difference between a diphthong and a monophthong, where larger distances might be expected, the /ɪ ɛ/ contrast involves a difference between two lax vowels, where shorter distances might be expected. The interaction between talker dialect and vowel contrast was not significant.

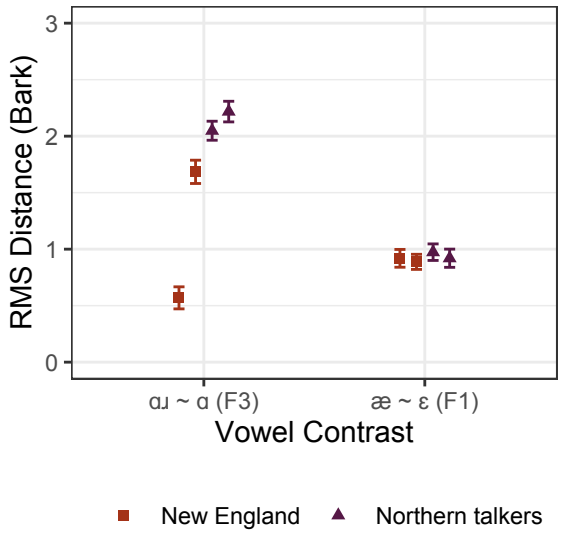
The model for the New England/Northern dataset revealed a significant main effect of vowel contrast (β = .35, F(1, 183.0) = 107.27, p < .001) and a significant interaction between talker dialect and vowel contrast (β = -.24, F(1, 183.0) = 52.23, p < .001). The overall RMSDs were larger for the /ɑɪ ɑ/ contrast than the /æ ɛ/ contrast. As in the Southern/Northern Virginian dataset, this effect likely reflects inherent properties of the two contrasts, including a vowel-rhotic sequence in the /ɑɪ

ɑ/ contrast versus two lax vowels in the /æ ɛ/ contrast. Post-hoc comparisons of estimated marginal means for each vowel contrast revealed a marginal effect of talker dialect for the /ɑɹ ɑ/ contrast (β = -1.01, t(2.2) = -3.51, p = .062). As expected, the Northern talkers had larger RMSDs than the New England talkers. The talker dialect effect was not significant for the /æ ɛ/ contrast. The main effect of talker dialect was also not significant.



(a)

(b)



(c)

311 FIG 2. (Color online). Mean RMSDs for each talker for each vowel contrast for the
312 Midland/Northern (a), Southern/Northern Virginian (b), and New England/Northern (c) datasets.
313 Error bars are standard errors.

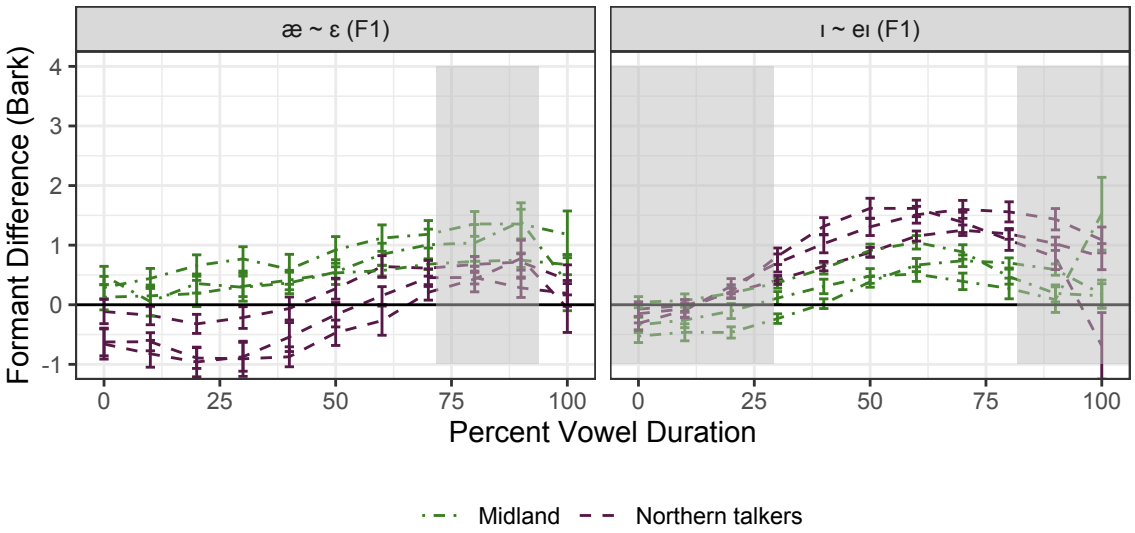
314 Like the DTW results, the RMSD results are suggestive of the predicted patterns, although more
315 of the predicted patterns are statistically robust and there is no evidence of unexpected patterns. In
316 particular, in the RMSD analysis, significant talker dialect effects in the predicted direction are
317 observed for the /ɪ eɪ/ contrast in the Midland/Northern dataset and for both vowel contrasts in
318 the Southern/Northern Virginian dataset. The predicted effect is marginal for the /ɑɪ ɑ/ contrast in
319 the New England/Northern dataset.

320 **C. Generalized additive mixed modeling**

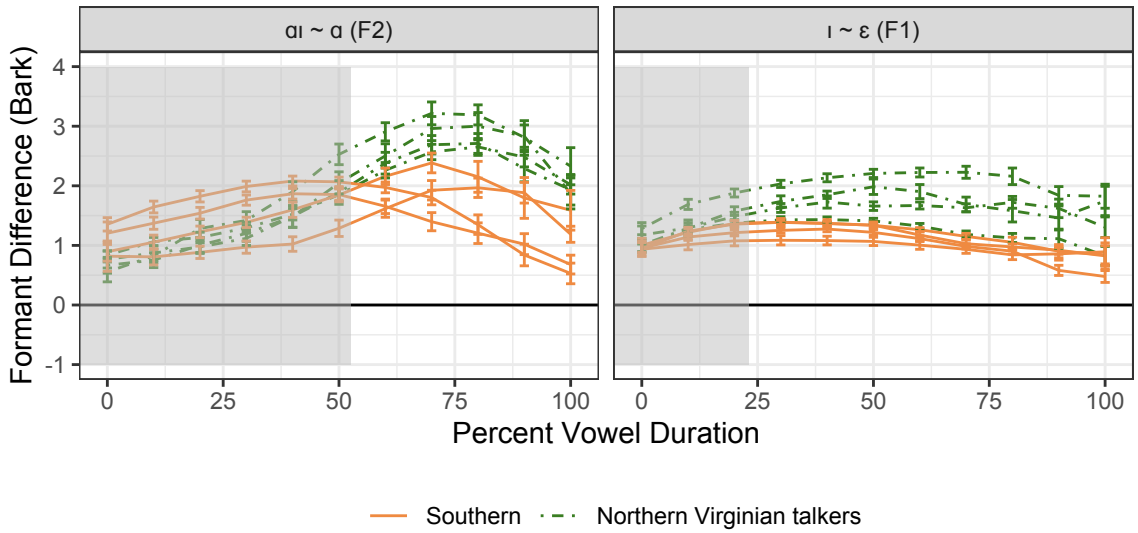
321 A summary of the mean formant trajectory differences for each talker for each vowel contrast in
322 each of the three datasets is shown in Fig. 3. The by-talker means show the expected talker dialect
323 differences for both contrasts in the Midland/Northern (Fig. 3a) dataset, for both vowel contrasts in
324 the Southern/Northern Virginian (Fig. 3b) dataset, and for the /ɑɪ ɑ/ contrast in the New
325 England/Northern (Fig. 3c) dataset. The by-talker means overlap considerably for the /æ ɛ/
326 contrast in the New England/Northern dataset.

327 The GAMMs predicting formant trajectory differences over time for the Midland/Northern
328 dataset revealed significant main effects of talker dialect for both the /æ ɛ/ contrast (β = -.77, t = -
329 3.00, p = .003) and the /ɪ eɪ/ contrast (β = .48, t = 3.21, p = .001). Significant talker dialect
330 differences in the trajectories were observed for the /æ ɛ/ contrast between 0% and 72% and
331 between 94% and 100% of the vowel duration and for the /ɪ eɪ/ contrast between 29% and 82% of
332 the vowel duration. Formant differences were larger for the Midland talkers than the Northern
333 talkers for the /æ ɛ/ contrast and for the Northern talkers than the Midland talkers for the /ɪ eɪ/

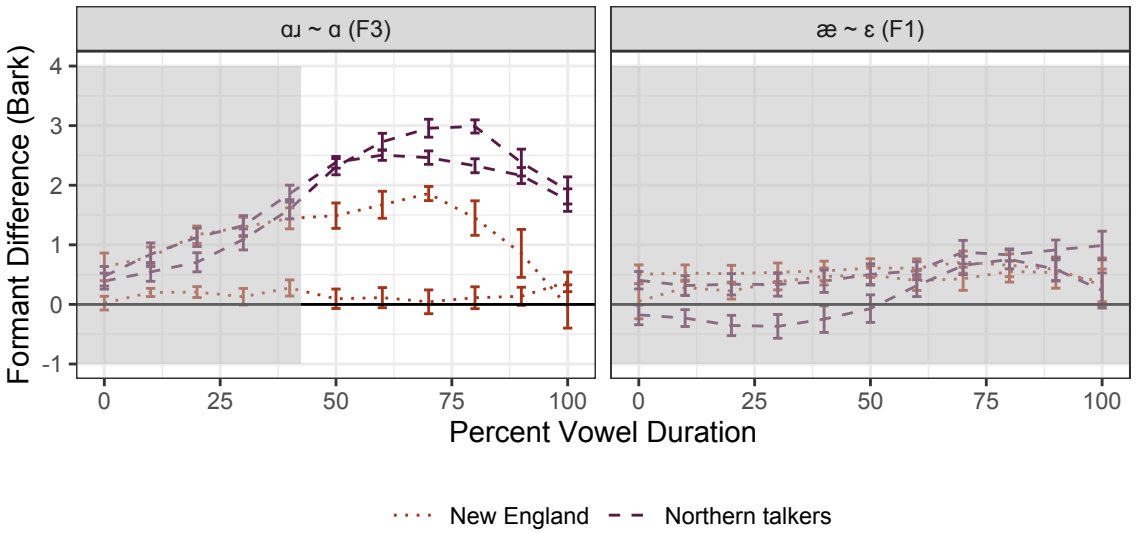
contrast, as expected. The GAMMs for the Southern/Northern Virginian dataset revealed significant main effects of talker dialect for both the /ɑɪ ɑ/ contrast (β = -.39, t = -2.39, p = .017) and the /ɪ ɛ/ contrast (β = -.51, t = -3.17, p = .002). Significant talker dialect differences in the trajectories were observed for the /ɑɪ ɑ/ contrast between 53% and 100% of the vowel duration and for the /ɪ ɛ/ contrast between 23% and 100% of the vowel duration. Formant trajectory differences were larger for the Northern Virginian talkers than the Southern talkers for both vowel contrasts, as expected. The GAMMs for the New England/Northern dataset revealed a significant main effect of talker dialect for the /ɑɪ ɑ/ contrast (β = 1.11, t = 2.13, p = .034). Significant talker dialect differences in the trajectories were observed for the /ɑɪ ɑ/ contrast between 42% and 100% of the vowel duration. The talker dialect effect was not significant for the /æ ɛ/ contrast. Formant trajectory differences were larger for the Northern talkers than the New England talkers for the /ɑɪ ɑ/ contrast, as expected.



(a)



(b)



(c)

FIG 3. (Color online). Mean formant trajectory differences for each talker for each vowel contrast for the Midland/Northern (a), Southern/Northern Virginian (b), and New England/Northern (c) datasets. Error bars are standard errors. Temporal regions with non-significant talker dialect differences in the GAMMs are shaded in gray.

Like the DTW and RMSD results, the GAMMs are suggestive of the predicted patterns, although more of the predicted patterns are statistically robust and there is no evidence of unexpected patterns. In particular, in the GAMM analysis, significant talker dialect effects in the

predicted direction are observed for both vowel contrasts in the Midland/Northern dataset, for both vowel contrasts in the Southern/Northern Virginian dataset, and for the /ɑɪ ɑ/ contrast in the New England/Northern dataset.

IV. DISCUSSION

All three of the acoustic distance measures considered in this study—DTW, RMSD, and GAMMs—led to predicted patterns of vowel category distance within and across dialects. DTWs revealed significant effects in the expected direction for both the /ɑɪ ɑ/ and /ɪ ɛ/ contrasts in the Southern/Northern Virginian dataset. RMSDs revealed significant effects in the expected direction for these two contrasts, as well as for the /ɪ eɪ/ contrast in the Midland/Northern dataset. GAMMs revealed significant effects in the expected direction for these three contrasts, as well as for the /æ ɛ/ contrast in the Midland/Northern dataset and the /ɑɪ ɑ/ contrast in the New England/Northern dataset. The only predicted effect that was not observed across any of the distance measures was for the /æ ɛ/ contrast in the New England/Northern dataset. This null result likely reflects the variability between the two Northern talkers in this dataset. As shown in Fig. 3c, one of the Northern talkers produced formant trajectory differences for the /æ ɛ/ contrast that are similar to the Northern talkers in the Midland/Northern dataset (see Fig. 3a), whereas the other Northern talker's mean formant trajectory difference for the /æ ɛ/ contrast was similar to the two New England talkers in the New England/Northern dataset.

Although the three datasets that were analyzed in the current study are all rather small, the results suggest that dynamic acoustic distance measures that have been used to assess language and accent distances in previous work (e.g., Bartelds et al., 2020; Heeringa et al., 2009; Renwick and Stanley, 2020) can usefully be applied to assessments of vowel category distance within and across dialects. This quantification of dynamic vowel category distance in the current study complements

previous qualitative descriptions of regional vowel variation in American English (e.g., Labov et al., 2006), as well as previous measures of vowel category distance at discrete time points (e.g., Hay et al., 2006; Kendall and Fridland, 2012; Wassink, 2006).

Moreover, although the magnitude of the DTW distances are not readily interpretable, the RMSD and GAMM analyses were conducted in Bark and the magnitudes of the effects are therefore interpretable from the model coefficients. The significant talker dialect effects ranged in magnitude from .25 Bark in the RMSD analysis of the Southern/Northern Virginian dataset to 1.11 Bark in the GAMM analysis of the New England/Northern dataset. These differences are likely perceptible, given vowel discrimination thresholds of approximately .28 Bark for American English listeners (Kewley-Port and Zheng, 1999). Thus, just as these kinds of distance measures predict perception of language and accent distances (e.g., Heeringa et al., 2009; Gunter et al., 2020; Porretta et al., 2015) and performance in cross-accent speech processing tasks (e.g., Chernyak et al., 2024; Hay et al., 2006), the dynamic measures examined in this study are also likely to predict judgments about regional accents and cross-dialect speech processing performance. Exploring the relationship between these dynamic vowel category distance measures and the perception of regional dialects is a critical next step for linking acoustic distances between vowel categories to speech processing performance (see also Bent et al., 2021; Kelley, 2023; Kelley and Tucker, 2022).

The increase in the number of predicted patterns that were observed from the DTW analysis to the RMSD analysis to the GAMM analysis is consistent with the information that is used to estimate distance in each approach. DTW of MFCCs captures more information in the spectrum than the formant-based RMSD and GAMM analyses, including voice quality, f0, and higher formant structure. The MFCCs also include first- and second-order derivatives of the cepstral coefficients, which explicitly capture the slope and rate of change of these coefficients over time, in addition to the time-varying coefficients themselves. These sources of additional information may mask the

403 predicted differences, which are characterized by one or two time-varying formants, leading to fewer
404 predicted results. However, listeners also have access to this broader array of information in the
405 spectrum and so DTW distances may better predict perception than RMSD or GAMMs. Future
406 research exploring the relationship between these measures and the perception of regional dialects is
407 critical for understanding the roles of general acoustic distance and specific formant trajectory
408 distance in cross-dialect speech processing.

409 The RMSD and GAMM analyses in the current study captured formant trajectory information
410 from a single formant that was selected for each vowel contrast to highlight the predicted variation
411 in vowel category distance within and across dialects. Despite this similarity, these two analyses
412 differ in two primary ways. First, RMSDs return a discrete measure of distance, whereas GAMMs
413 assess difference throughout the formant trajectory. Second, RMSDs return a positive distance
414 value, which can mask variation in the direction of the formant trajectory difference, whereas
415 GAMMs assess signed (positive and negative) differences. The former difference in the two
416 approaches likely explains the significant difference that was observed for the /ɑɹ ɑ/ contrast in the
417 New England/Northern dataset in the GAMM analysis, but not in the RMSD analysis. The formant
418 trajectory differences were only significantly different for the two talker dialects in the second half of
419 the contrast, consistent with the contrast between the presence vs. absence of the following rhotic
420 for Northern vs. New England talkers, respectively (see Fig. 3c). This difference is captured by the
421 GAMM analysis, but lost in the RMSD analysis, which returns just a single distance value, collapsed
422 over time. However, this single distance value makes RMSDs more suitable than GAMMs for
423 predicting perception because RMSDs can be straightforwardly entered as predictors in statistical
424 models.

425 The latter difference between the two approaches likely explains the significant difference that
426 was observed for the /æ ɛ/ contrast in the Midland/Northern dataset in the GAMM analysis, but

not in the RMSD analysis. As shown in Fig. 3a, whereas the Midland talkers had a positive formant trajectory difference for this contrast throughout the entire trajectory, the Northern talkers had a negative formant trajectory difference in the first half of the contrast and a positive formant trajectory difference in the second half of the contrast. The overall magnitudes of the formant trajectory differences were comparable across dialects, so that the absolute, positive distance measure in the RMSD analysis could not distinguish the two trajectory differences, whereas the signed differences emerge as significant in the GAMM analysis. Future research exploring the relationship between these measures and the perception of regional dialects is critical for understanding the roles of overall formant trajectory distance and the direction and magnitude of formant trajectory difference in cross-dialect speech processing.

The three analyses that were conducted in this study are a subset of the possible applications of dynamic distance measures to vowel category distances within and across dialects. First, one formant was selected for analysis in the RMSD and GAMM analyses, whereas other formants may also exhibit interesting differences in distance for these vowel contrasts. For example, the analysis of the /ɑɪ ɑ/ contrast in the Southern/Northern Virginian analysis focused on F2 to capture fronting of the offglide, but it would also be reasonable to analyze F1 to capture raising of the offglide. Second, the DTW analysis was based on MFCCs, whereas the RMSD and GAMM analyses were based on formant estimates. A DTW analysis could also be conducted on the formant trajectories to disentangle DTW as a distance measure from the information about the spectrum included in the analysis.[2] Finally, several recent studies have used discrete cosine transformations (DCTs) to capture formant trajectories (e.g., Cox et al., 2024; Elvin et al., 2016). Euclidean distance could be applied to DCTs to further disentangle the distance measures from the characterization of the spectrum.

In summary, vowel categories are acoustically more similar to one another in some regional dialects of American English than in others, even in the absence of vowel category mergers (Labov

et al., 2006). The application of dynamic acoustic distance measures, including DTW, RMSD, and GAMMs, can capture this variation in vowel category acoustic distance within and across dialects. The analyses that focused on specific formant trajectory distance (i.e., RMSD, GAMMs) captured more of these predicted patterns than the analysis that focused on overall acoustic distance (i.e., DTW), although overall acoustic distance may be a better predictor of perception. Within the analyses that focused on specific formant trajectory distance, the analysis that focused on formant trajectory difference (i.e., GAMMs) captured more of these predicted patterns than the analysis that focused on formant trajectory distance (i.e., RMSD), although RMSD may be more useful for predicting perception because it returns a discrete distance value for each minimal pair. These approaches therefore address varying aspects of the nature of dynamic vowel category distance and may be most useful for distinct kinds of questions. The relationship between these acoustic distance measures and human perception remains to be explored.

ACKNOWLEDGMENTS

This work was partially supported by the National Science Foundation (BCS-1843454).

AUTHOR DECLARATIONS

Conflict of Interest

The author has no conflicts to disclose.

Ethics Approval

The analysis involved existing datasets. The Midland/Northern and Southern/Northern Virginian datasets are de-identified. The analysis of the New England/Northern dataset was approved by the Institutional Review Board at Ohio State University (2017B0576) and informed consent was obtained from all participants.

DATA AVAILABILITY

The data that support the findings of this study are available on the OSF repository for this project: https://osf.io/hp6xc/

ENDNOTES

[1] Although many of the approaches discussed here are not distance metrics in the mathematical sense (including dynamic time warping, Pillai scores, and generalized additive mixed models), they can usefully be applied to questions related to vowel category acoustic distance and so are referred to here as "distance measures," as in previous work (e.g., Bartelds et al., 2020; Mielke, 2012).

[2] I would like to thank Matthew Kelley for this suggestion.

REFERENCES

Bartelds, M., Richter, C., Liberman, M., and Wieling, M. (**2020**). "A new acoustic-based pronunciation distance measure," *Front. Artif. Intell.* **3**(39), 1–10.

Bent, T., Holt, R. F., Van Engen, K. J., Jamsek, I. A., Arzbecker, L. J., Liang, L., and Brown, E. (**2021**). "How pronunciation distance impacts word recognition in children and adults," *J. Acoust. Soc. Am.* **150**, 4103–4117.

Boersma, P., and Weenink, D. (**2023**). "Praat: Doing phonetics by computer," Computer program, Version 6.4.01.

Chernyak, B. R., Bradlow, A. R., Keshet, J., and Goldrick, M. (**2024**). "A perceptual similarity space for speech based on self-supervised speech representations," *J. Acoust. Soc. Am.* **155**, 3915–3929.

Clark, H., Bissell, M., Clopper, C. G., and Walker, A. (**2022**). "Effects of lexical competition, dialect familiarity, and dialect exposure on lexical processing," Poster presented at Laboratory Phonology 18, virtual, June 23–25.

496 Clopper, C. G., and Tamati, T. N. (**2014**). "Effects of local lexical competition and regional dialect
497 on vowel production," *J. Acoust. Soc. Am.* **136**, 1–4.
498 Clopper, C. G., and Walker, A. (**2017**). "Effects of lexical competition and dialect exposure on
499 phonological priming," *Lang. Speech* **60**, 85–109.
500 Cole, J., Steffman, J., Shattuck-Hufnagel, S., and Tilsen, S. (**2023**). "Hierarchical distinctions in the
501 production and perception of nuclear tunes in American English," *Lab. Phonol.* **14**(1), 1–51.
502 Cox, F., Penney, J., and Palethorpe, S. (**2024**). "Australian English monophthong change across 50
503 years: Static vs. dynamic measures," *Languages* **9**(99), 1–35.
504 Elvin, J., Williams, D., and Escudero, P. (**2016**). "Dynamic acoustic properties of monophthongs
505 and diphthongs in Western Sydney Australian English," *J. Acoust. Soc. Am.* **140**, 576–581.
506 Farrington, C., Kendall, T., and Fridland, V. (**2018**). "Vowel dynamics in the Southern Vowel Shift,"
507 *Am. Speech* **93**, 186–222.
508 Ferragne, E., and Pellegrino, F. (**2010**). "Vowel systems and accent similarity in the British Isles:
509 Exploiting multidimensional acoustic distances in phonetics," *J. Phonetics* **38**, 526–539.
510 Fox, R. A., and Jacewicz, E. (**2009**). "Cross-dialectal variation in formant dynamics of American
511 English vowels," *J. Acoust. Soc. Am.* **126**, 2603–2618.
512 Gunter, K. M., Vaughn, C. R., and Kendall, T. S. (**2020**). "Perceiving Southernness: Vowel
513 categories and acoustic cues in Southernness ratings," *J. Acoust. Soc. Am.* **147**, 643–656.
514 Hay, J., Warren, P., and Drager, K. (**2006**). "Factors influencing speech perception in the context of
515 a merger-in-progress," *J. Phonetics* **34**, 458–484.
516 Haynes, E. F., and Taylor, M. (**2014**). "An assessment of acoustic contrast between long and short
517 vowels using convex hulls," *J. Acoust. Soc. Am.* **136**, 883–891.
518 Heeringa, W., Johnson, K., and Gooskens, C. (**2009**). "Measuring Norwegian dialect distances using
519 acoustic features," *Speech Comm.* **51**, 167–183.

Kaland, C. (**2023**). "Intonation contour similarity: f0 representations and distance measures compared to human perception in two languages," *J. Acoust. Soc. Am.* **154**, 95–107.

Kelley, M. C. (**2023**). "Acoustic absement in detail: Quantifying acoustic differences across time-series representations of speech data," *Proc. 20th Int. Congress Phonetic Sci.* 679–683.

Kelley, M. C., and Tucker, B. V. (**2020**). "A comparison of four vowel overlap measures," *J. Acoust. Soc. Am.* **147**, 137–145.

Kelley, M. C., and Tucker, B. V. (**2022**). "Using acoustic distance and acoustic absement to quantify lexical competition," *J. Acoust. Soc. Am.* **151**, 1367–1379.

Kendall, T., and Fridland, V. (**2012**). "Variation in perception and production of mid front vowels in the U.S. Southern Vowel Shift," *J. Phonetics* **40**, 289–306.

Kewley-Port, D. K., and Zheng, Y. (**1999**). "Vowel formant discrimination: Towards more ordinary listening conditions," *J. Acoust. Soc. Am.* **106**, 2945–2958.

Kirkham, S., Nance, C., Littlewood, B., Lightfoot, K., and Groarke. E. (**2019**). "Dialect variation in formant dynamics: The acoustics of lateral and vowel sequences in Manchester and Liverpool English," *J. Acoust. Soc. Am.* **145**, 784–794.

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (**2017**). "lmerTest package: Tests in linear mixed effects models," *J. Stat. Software* **82**(13), 1–26.

Labov, W., Ash, S., and Boberg, C. (**2006**). *Atlas of North American English* (Mouton de Gruyter, New York).

Lenth, R. (**2024**). "emmeans: Estimated marginal means, aka least-squares means," R package, Version 1.10.0.

Lind-Combs, H. C., Bent, T., Holt, R. F., Clopper, C. G., and Brown, E. (**2023**). "Comparing Levenshtein distance and dynamic time warping in predicting listeners' judgments of accent distance," *Speech Comm.* **155**(102987), 1–14.

Lyons, J., Wang, D. Y.-B., Gianluca, Shteingart, H., Mavrinac, E., Gaurkar, Y., Watcharawisetkul, W., Birch, S., Lu. Z., Hölzl, J., Lesinskis, J., Almér, H., Lord, C., and Stark, A. (**2020**). "jameslyons/python_speech_features: release v0.6.1," *Zenodo*. https://doi.org/10.5281/zenodo.3607820.

Mielke, J. (**2012**). "A phonetically based metric of sound similarity," *Lingua* **122**, 145–163.

Morrison, G. S. (**2008**). "Comment on 'A geometric representation of spectral and temporal vowel features: Quantification of vowel overlap in three linguistic varieties' [J. Acoust. Soc. Am. 119, 2334–2350 (2006)] (L)," *J. Acoust. Soc. Am.* **123**, 37–40.

Nycz, J., and Hall-Lew, L. (**2013**). "Best practices in measuring vowel merger," *Proc. Mtgs. Acoust.* **20**(060008), 1–19.

Peterson, G. E., and Lehiste, I. (**1960**). "Duration of syllable nuclei in English," *J. Acoust. Soc. Am.* **32**, 693–703.

Porretta, V., Kyröläinen, A.-J., and Tucker, B. V. (**2015**). "Perceived foreign accentedness: Acoustic distances and lexical properties," *Atten. Percept. Psychophys.* **77**, 2438–2451.

Renwick, M. E. L., and Stanley, J. A. (**2020**). "Modeling dynamic trajectories of front vowels in the American South," *J. Acoust. Soc. Am.* **147**, 579–595.

Ross, J., and Clopper, C. G. (**2023**). "Talker variability in cross-dialect lexical processing," *Proc. 20th Int. Congress Phonetic Sci.* 152–156.

Traunmüller, H. (**1990**). "Analytical expressions for the tonotopic sensory scale," *J. Acoust. Soc. Am.* **88**, 97–100.

Wassink, A. B. (**2006**). "A geometric representation of spectral and temporal vowel features: Quantification of vowel overlap in three linguistic varieties," *J. Acoust. Soc. Am.* **119**, 2334–2350.

567 Wieling, M., Margaretha, E., and Nerbonne, J. (**2012**). "Inducing a measure of phonetic similarity
568 from pronunciation variation," *J. Phonetics* **40**, 307–314.
569 Yuan, J. and M. Liberman. (**2008**). "Speaker identification on the SCOTUS corpus," *Proc. Mtgs.*
570 *Acoust. 2008,* 5687–5690.