Acoustic cues and linguistic experience as factors in regional dialect classification

Steven Alcorn,1,2,a) Kirsten Meemann,2 Cynthia G. Clopper,3 and Rajka Smiljanic2
1Department of Spanish and Portuguese, Indiana University, Bloomington, Indiana 47405, USA
2Department of Linguistics, The University of Texas at Austin, Austin, Texas 78712, USA
3Department of Linguistics, The Ohio State University, Columbus, Ohio 43210, USA

ABSTRACT:
Listeners are able to classify talkers by regional dialect of their native language when provided with even short speech samples. However, the way in which American English listeners use segmental and prosodic information to make such decisions is largely unknown. This study used a free classification task to assess native American English listeners’ ability to group together talkers from six major dialect regions of American English. Listeners residing in Ohio and Texas were provided with a sentence-long (experiment 1) or paragraph-long (experiment 2) speech sample produced by talkers from each of the six regions presented in one of three conditions: unmodified, monotonized (i.e., flattened F0), and low-pass filtered (i.e., spectral information above 400 Hz removed). In both experiments, listeners in the unmodified and monotonized conditions made more accurate groupings, reflecting their reliance on segmental properties for classifying regional variation. Accuracy was highest for Northern and Western talkers (experiment 1) and Mid-Atlantic talkers (experiment 2). Listeners with experience with multiple dialects as a result of geographic mobility did not show increased accuracy, suggesting a complex relationship between linguistic experience and the perception of available acoustic cues to socioindexical variation. © 2020 Acoustical Society of America.

https://doi.org/10.1121/10.0000551
(Received 31 March 2019; revised 9 October 2019; accepted 31 October 2019; published online 31 January 2020)

[Editor: Irina A. Shport]

I. INTRODUCTION

While segmental features that differentiate the major dialects of American English have traditionally been the focus of work in the area of sociophonetic variation (e.g., Labov et al., 2006), both segmental and prosodic characteristics have been shown to vary across dialects (e.g., Byrd, 1994; Clopper and Smiljanic, 2011, 2015; Jacewicz et al., 2010), and work from a variety of languages has shown that both types of information influence listeners’ perception and classification of dialects (e.g., Fuchs, 2016; Leemann et al., 2018; Peters et al., 2002; Purnell et al., 1999; van Bezooijen and Gooskens, 1999). The primary goal of the current study was to explore the nature of American English listeners’ reliance on segmental and prosodic cues in a dialect classification task by manipulating which cues were available in the acoustic signal. A secondary objective was to empirically test the extent to which listeners’ experience with their own regional dialect facilitates classification of that dialect relative to other dialects of American English.

A. Segmental and prosodic cues in dialect classification

The effect of different types of phonetic cues on listeners’ perception of dialects and languages has traditionally been investigated using identification tasks. van Bezooijen and Gooskens (1999) found that for dialects of Dutch and UK English, listeners’ identification of regional accents was most accurate when both segmental and prosodic cues were available but was still good when segmental information with flattened intonation was provided, and was considerably less accurate when speech was low-pass filtered at 350 Hz with only prosodic information (i.e., intonation and rhythm) available. Moreover, for at least one of the regional dialects in each language, listeners were more accurate when segmental information with flattened intonation was provided than when both segmental and prosodic cues were available, pointing to the critical importance of segmental information and the relatively minor role of prosodic information for dialect identification in these languages. However, Ikeno and Hansen (2006) showed that American English dialect identification was diminished when listeners were presented with flattened F0 compared to the original F0 contour, suggesting a significant role for prosody in dialect identification.

Subsequent studies have provided additional evidence for asymmetric attention to segmental cues over prosodic cues by listeners from languages including English (Baker et al., 2009; Clopper and Bradlow, 2009), German (Leemann et al., 2018; Ruch, 2018), and Norwegian (Gooskens, 2005). For example, Leemann et al. (2018) tested the ability of Swiss German listeners from the Zurich dialect to identify two dialects of German (Valais and Bern Swiss German), which are known to vary in terms of both segmental and prosodic properties. They created three conditions: unmodified, swapped speech rhythm (in which syllable durations of speakers from one dialect were replaced with those of speakers from the

—

a)Electronic mail: stalcorn@indiana.edu
other dialect), and swapped speech rhythm and intonation (in which syllable durations were replaced as in the swapped rhythm condition, and the intonation contours of speakers from one dialect were mapped onto the speech of speakers from the other dialect). Identification accuracy was higher in the unmodified condition than in the swapped conditions, and performance was well above chance for all three conditions, highlighting listeners’ reliance on primarily segmental cues to identify the two dialects, even in the face of potentially misleading prosodic information.

This preference for segmental cues extends to the identification of other talker characteristics, such as ethnicity. Thomas and Reaser (2004) found that listeners from Hyde County (in eastern North Carolina) were significantly more accurate at deciding whether a talker was European-American or African-American when given unmodified or monotonized speech samples than when they were given low-pass filtered samples, which removed all spectral information above 330 Hz, preserving intonational and rhythmic cues but eliminating segmental cues such as vowel quality. While the authors argued that prosodic information was not irrelevant for identifying a talker’s ethnic background, it was not a sufficiently reliable cue by itself.

Notwithstanding the crucial role of segmental information, previous studies have shown that when listeners are given auditory stimuli in which only prosodic cues are available, they are able to distinguish between dialects of the same language at an accuracy level above chance (e.g., Barkat et al., 1999; Fuchs, 2016), and can also accurately classify speakers of different ethnicities based solely on prosodic information (Szakay, 2008a). However, evidence from discrimination and identification studies on listeners’ ability to distinguish between accents of a language, as well as between two different languages, suggests a complex interplay of different prosodic cues that likely varies from one specific combination of languages/accents to the next. Frota et al. (2002) discovered that for European and Brazilian Portuguese, two varieties of the same language that differ substantially in terms of their prosodic structure (including both intonation and rhythm), native European Portuguese listeners could discriminate between the two when F0 was present in the stimuli (i.e., low-pass filtered), but not when listening to monotonized stimuli, suggesting that intonation is a useful cue for discrimination, at least for this pair of language varieties. A subsequent experiment revealed that native European Portuguese listeners were able to tell European Portuguese and Dutch apart, even when F0 was flattened to remove intonational cues. Although European Portuguese and Dutch are both considered to be stress-timed languages, the listeners attended to a rhythmic property that served to distinguish the two languages, namely the proportion of vocalic intervals in the speech sample (%V), and ignored the variability in the duration of the consonantal intervals (ΔC), which would not have been a reliable cue for discrimination.

Building on this work, Vicenik and Sundara (2013) tested American English listeners on their ability to discriminate between two prosodically similar languages (English vs German) or two English accents (American vs Australian) solely on the basis of prosodic information. Three conditions were created to tease apart the influence of rhythm and intonation. In the first condition, speech samples were low-pass filtered at 400 Hz to remove segmental information but preserve both rhythmic and intonational information; in the second condition, the samples were synthesized to remove segmental and intonational information, leaving only rhythm; and in the third condition, the segmental and rhythmic properties were removed and only the intonational quality was preserved. The listeners were asked to identify only whether each sample was spoken in American English or something else. In the English vs German discrimination task, listeners were able to use intonation and rhythm separately to discriminate the two varieties with above chance accuracy, and there was no additional benefit when intonation and rhythm were available simultaneously. In the American vs Australian English discrimination task, listeners were able to distinguish between the two varieties with above chance accuracy when presented with rhythm and intonation together, as well as with only rhythm, but not with only intonation. It should be noted that both discrimination tasks were difficult, with an average accuracy score just above chance (53%) across conditions, suggesting that accent and language discrimination relies heavily on segmental cues in addition to prosodic information.

Taken together, the results of these studies suggest that listeners rely more on segmental cues than prosodic cues when identifying or classifying regional dialects, and although prosodic cues can be sufficient for performing these tasks, accuracy tends to be lower when segmental cues are not available. In addition to this asymmetry in the use of prosodic and segmental cues, however, listener-specific experience with accent and dialect variation acquired throughout the lifespan plays a role in the perception of regional variation.

B. Listener experience and dialect classification

Regardless of how listeners use available acoustic cues to recognize, identify, discriminate between, and group talkers based on regional dialect, it is clear that exposure to regional variation is crucial for developing these abilities. Experience acquired as a result of living in a region where a specific dialect is spoken confers an advantage in identifying that dialect; by the same token, a lack of experience with a dialect makes it more difficult to identify, even for listeners of the same native language (e.g., American English listeners identifying dialects of British English; Ikeno and Hansen, 2007). This experience begins to accumulate early in life: while previous studies have demonstrated that very young children lack representations of regional dialects (Floccia et al., 2009; Girard et al., 2008), even listeners as young as 4–5 years old can distinguish between New England talkers and those from other regions of the United States (Jones et al., 2017). Dialect classification accuracy improves steadily throughout childhood and adolescence as children are exposed to variation and
become aware of the indexical value of dialect variation (Jones et al., 2017).

Notions of cultural boundaries and language ideologies, often based on highly salient features associated with the speech of a particular region, are also formed and serve as the basis for judgments about regional dialect variation, especially for listeners who lack experience with particular dialects (Bucholtz et al., 2007; Evans, 2011; Preston, 1993). In a study of local and non-local listeners’ identification of Utah speech, Baker et al. (2009) found that listeners from areas closer to Utah were more accurate at identifying Utah speakers than listeners from more distant regions, and they also attended to less stereotypical phonetic characteristics of Utahns. Similarly, Yan (2015) elicited perceptual judgments from Chinese listeners about regional differences within one prefecture in central China using tasks designed to tap into language ideologies (i.e., a hand-drawn map task and a dialect difference rating task), as well as an identification task using speech samples from 12 talkers. She found that the long-term representations of dialect differences largely aligned with performance on the identification task, which suggests a linguistic basis for listeners’ beliefs about dialect variation.

At the same time, representations of regional dialects are not static but can be influenced by a listener’s geographic mobility, which leads to exposure to and experience with regional varieties other than one’s own throughout the lifespan. Clopper and Pisoni (2004) tested dialect identification accuracy among geographically mobile and non-mobile listeners and found that experience with multiple regional dialects during childhood and adolescence led to enhanced dialect identification abilities when compared to listeners who had lived in only one dialect region. Listeners who had resided in a specific region were also better at identifying talkers from that region. Additionally, although both listener groups revealed similar overall patterns of dialect identification accuracy, listeners with more varied linguistic experience demonstrated increased perceptual distance between talkers from different regions. In the study on the perception of Utah English by Baker et al. (2009), listeners from other regions of the United States with more than two years of experience in Utah were better able to correctly identify Utah talkers as being from Utah than their counterparts with less experience in the region. Williams et al. (1999) likewise demonstrated the beneficial effect of geographic mobility. They asked Welsh adolescents and adults to categorize samples of spontaneous speech representing six different regional varieties of Welsh English. The adults outperformed the adolescents, and the authors attributed this difference to the adults’ greater experience with other varieties as a result of their travels. Exposure to other regional dialects through one’s social networks, without physical geographical mobility, may also lead to more robust representations of regional phonetic variation (Diaz-Campos and Navarro-Galisteo, 2009).

C. The current study

In this study, we aimed to shed light on listeners’ use of segmental and prosodic cues in regional dialect perception, as well as the effect of linguistic experience on knowledge of regional variation. We conducted two free classification experiments that systematically manipulated the acoustic cues available to the listeners. Our primary goal was to examine listeners’ reliance on segmental, intonational, and rhythmic information in dialect classification. In experiment 1, listeners were asked to sort 60 talkers, comprising 10 talkers from each of the 6 main dialect regions of American English (Mid-Atlantic, Midland, New England, Northern, Southern, and Western; Labov et al., 2006), based on sentence-long read speech samples, whereas in experiment 2 participants were asked to sort 18 talkers (3 talkers from each of the 6 regions) based on paragraph-long read speech samples. Sentence-length speech samples were used in experiment 1 as in previous free classification tasks (Clopper and Bradlow, 2008; Clopper and Pisoni, 2007; Clopper et al., 2013; Jones et al., 2017). Longer speech samples were used in experiment 2 to provide participants with additional segmental and prosodic cues (including pauses, which were not present in the sentence-long stimuli) on which to base their classifications. Due to the longer stimuli, we included fewer talkers in experiment 2.

A second goal of the study was to examine the potential benefit of dialect experience on regional dialect classification when the available segmental and prosodic cues were systematically manipulated. To that end, listeners from two different locations, Ohio and Texas, were recruited. Most of Ohio (including where the experiments were conducted) is located in the Midland dialect region, whereas Texas is part of the Southern dialect region (Labov et al., 2006).

To examine these questions, a free classification paradigm was employed to allow listeners to form their own groups of talkers based on their evaluation of the speech produced by those talkers. Free classification tasks have been adopted in previous work examining perception of dialect variation in American English (e.g., Clopper and Bradlow, 2008; Clopper and Pisoni, 2007; Jones et al., 2017), cross-linguistic similarity (e.g., Bradlow et al., 2010), and non-native varieties of English (e.g., Atagi and Bent, 2013; Bent et al., 2016; McCullough and Clopper, 2016). The advantage of free classification over other task modalities, like forced-choice identification, is that in free classification, no labels are provided to listeners so they must work inductively (i.e., starting with individual talkers and from there creating groups based on perceived similarities) rather than deductively (i.e., starting with a given category label and then considering the features of a given talker’s speech). Thus, the groups that listeners produce reflect their own perceptual category structure rather than a category structure imposed by the experimenter. By giving listeners minimal restrictions, any patterns or biases that emerge can be assumed to reflect listeners’ perceptual dialect categories, rather than an artifact of the experimental design or linguistic stereotypes activated by the presence of specific regional labels. Previous work has demonstrated that free classification tasks can reduce or even eliminate response biases that are present in forced-choice identification tasks (Clopper and Pisoni, 2007).
By testing listeners on six different dialects of American English, we sought to extend previous findings regarding the role of segmental and prosodic cues in the perceptual identification of international varieties of English (e.g., Ikeno and Hansen, 2007; Vicenik and Sundara, 2013), as well as regional varieties of languages other than English (e.g., Leemann et al., 2018; van Bezooijen and Gooskens, 1999). Previous work has examined dialect identification and classification for unmodified stimuli in American English (e.g., Clopper and Pisoni, 2004, 2007), but listeners’ performance for modified stimuli that isolate different acoustic cues remains understudied. Ikeno and Hansen (2006) examined listeners’ identification of three highly stereotyped dialects of American English (i.e., New York, Mississippi, and California) for stimuli with various combinations of acoustic cue manipulations. The current study extends this work by including regional dialects with less stereotypical representations as well.

II. EXPERIMENT 1

A. Methods

1. Participants

One hundred twenty monolingual native speakers of American English were recruited from undergraduate student subject pools at The Ohio State University (n = 60) and The University of Texas at Austin (n = 60). The mean age of the Ohio listeners was 20.2 years old (range 18–50 years old), and the mean age of the Texas listeners was 19.3 years old (range 18–26 years old). Data from 44 additional listeners were discarded due to early knowledge of a foreign language (n = 33), reported history of a speech, language, or hearing impairment (n = 5), or experimenter error (n = 6). None of the listeners included in the analysis reported exposure to a language other than English before age 6 years old, and none had advanced proficiency in a foreign language. Listeners received partial course credit for their participation in the study.

Each listener provided a list of all of the places where they had resided. This information was used to classify listeners into three groups: those from Ohio and Texas who had only resided in the Midland or in the South, respectively, and for both locations, those who had resided in more than one region, including the region where they were currently attending university. Thus, each listener’s residential history included all of the places they had lived for any amount of time prior to their participation in the experiment. These classifications were based on the main regional divisions of American English proposed by Labov et al. (2006). In the Ohio group, 22 listeners had resided in only the Midland region, and 38 listeners had lived in multiple regions. In the Texas group, 43 listeners had lived only in the South, and 17 listeners had lived in multiple regions.

2. Stimuli

One sentence-long utterance from each of 60 different talkers was used as the auditory stimulus material in experiment 1. Each talker produced the sentence “They lived in a cottage deep in the woods,” while reading the Goldilocks passage (Stockwell, 2002), which was recorded as part of the Nationwide Speech Project (NSP) corpus (Clopper and Pisoni, 2006). This sentence was chosen because it contains both segmental and prosodic features that can distinguish regional dialects of American English, and was produced without disfluencies by all 60 talkers in the NSP corpus. Ten talkers (five male, five female) were recorded in the NSP corpus from each of six major dialect regions of the United States, as defined by Labov et al. (2006): Mid-Atlantic, Midland, New England, Northern, Southern, and Western. All talkers were native monolingual speakers of American English who had resided exclusively in their home dialect region until at least age 18 years old. The talkers ranged in age from 18 to 25 years old. The mean duration of the sentence stimuli was 1.77 s [standard deviation (SD) = 0.16 s, range 1.37–2.25 s].

To investigate which cues listeners use when classifying talkers by regional dialect, three test conditions were created. All sentences were first equalized for root mean square (RMS) amplitude in Praat (Boersma and Weenink, 2019). In the unmodified condition, listeners were presented with these amplitude-normalized, but otherwise natural, sentences. This condition served as a control condition. In the monotonized condition, the F0 contour of each sentence was flattened to the mean F0 of the sentence using Time-Domain Pitch Synchronous Overlap and Add (TD-PSOLA) in Praat. In this condition, segmental and rhythmic information (which includes speech rate, frequency and duration of pauses, and the relative duration of consonant and vowel intervals) was preserved, but intonational cues were made uninformative. No unexpected processing artifacts were noted in the monotonized stimuli. In the low-pass filtered condition, stimuli were created using a low-pass filter with a 400 Hz cutoff with 50 Hz smoothing in Praat. Several previous studies have used low-pass filtered speech for dialect identification tasks (e.g., Frota et al., 2002; Ikeno, 2005; Jacewicz et al., 2015; van Leyden and van Hueven, 2006; Szakay, 2008b; Thomas and Reaser, 2002; Vicenik and Sundara, 2013). Most of these studies used a cutoff frequency between 300 and 400 Hz, although Ikeno (2005) used a lower cutoff of 225 Hz. The low-pass filtered stimuli were rendered unintelligible because the segmental information concentrated in the higher frequencies was removed. At the same time, syllabic rhythm and intonation were preserved, including for the female talkers whose F0 may rise above the lower cutoff frequencies used in previous studies. As the filtering resulted in a muffled sound, the loudness was adjusted for this condition to a comfortable listening level.

In the free classification task, participants were asked to group talkers together on a grid displayed on a computer screen. The auditory stimuli were therefore converted to digital movies with a static visual display of the talker’s initials. In cases where two talkers shared the same initials, one of the talkers was assigned random initials within the constraints of typical American names (e.g., none of the
initials were “X”). This free classification design follows previous studies that have used a similar overall setup (e.g., Clopper and Bradlow, 2009; Clopper and Pisoni, 2007).

3. Procedure

Listeners were randomly assigned to one of three listening conditions (unmodified, monotonized, or low-pass filtered) in a between-subject design. The same number of listeners ($n = 20$) was assigned to each listening condition in each testing location. Because listeners’ demographic information was not assessed until after they had been assigned to a condition, the conditions were not balanced for residential history. Of the 38 Ohio listeners who had resided in multiple regions, 13 were assigned to the unmodified condition, 13 to the monotonized condition, and 12 to the low-pass filtered condition. Of the 17 Texas listeners who had lived in multiple regions, 2 participated in the unmodified condition, 7 in the monotonized condition, and 8 in the low-pass filtered condition.

Listeners were presented with a PowerPoint file (Microsoft, Redmond, WA) that contained the talker icons arranged in alphabetical order on the left-hand side and a blank $22 \times 22$ grid on the right of the slide, as shown in Fig. 1. Listeners could click on an icon to play the sentence read by the talker whose initials appeared on the icon or click and hold the icon to drag it onto the grid. Listeners were instructed to arrange the icons on the grid based on where they thought the talkers were from. They were told that there were no restrictions on the size or number of groups they could make, but to put all talkers from the same place in a group together. Importantly, specific regional labels were not given to the listeners.

B. Results

1. Number of groups

Given the minimal instructions that listeners received about how to create their groupings, the number of talker groups that the listeners made in each condition and for each location was first analyzed. Ohio listeners made an average of 9.1 (SD = 5.3) groups in the unmodified condition, 7.9 (SD = 3.8) groups in the monotonized condition, and 9.4 (SD = 5.1) groups in the low-pass filtered condition. Texas listeners made an average of 8.4 (SD = 5.1) groups in the unmodified condition, 8.6 (SD = 5.2) groups in the monotonized condition, and 8.6 (SD = 5.2) groups in the low-pass filtered condition. A two-way analysis of variance (ANOVA) with condition (unmodified, monotonized, low-pass filtered), and listener location (Ohio, Texas) as independent variables and the number of groups produced by each listener as the dependent variable revealed no significant main effects. The interaction between condition and listener location was also not significant.

2. Accuracy analysis

Accuracy results for all three conditions and all six dialects are shown in Fig. 2. The pairwise accuracy rate was determined by calculating the number of times each listener placed talkers from the same dialect region in the same

![FIG. 1. Screenshot of the free classification task presentation.](image-url)
group divided by the total possible number of same-region pairings. There were a total of 270 possible same-region pairings (45 within-region pairings × 6 dialects = 270 same-region pairings). The pairwise talker error rate was calculated for each listener as the number of times the listener placed talkers from different dialect regions in the same group divided by the total possible number of different-region pairings. There were a total of 1500 different-region pairings (50 different-region talkers × 60 talkers, divided by 2 because talker order is irrelevant for identifying pairings). The same-region and different-region pairings mirror the concepts of hits and false alarms, respectively, in signal detection theory (SDT; Macmillan and Creelman, 2005). The talker error rate was subtracted from the pairwise accuracy rate to yield a difference score, similar to a d’ score in SDT. This difference score was used as the primary measure of accuracy because it normalizes somewhat for the varied numbers of groups that listeners made. Listeners who made many small groups had both relatively few same-region pairings and relatively few different-region pairings, whereas listeners who made few large groups had both relatively many same-region pairings and relatively many different-region pairings.

A repeated-measures ANOVA was conducted with the accuracy difference score as the dependent variable, talker dialect as a within-subject variable, and condition and listener location as between-subject variables. The analysis revealed significant main effects of condition \( F(2,114) = 14.155, \ p < 0.0001 \) and talker dialect \( F(5,570) = 10.638, \ p < 0.0001 \), as well as a significant condition \( \times \) talker dialect interaction \( F(10,570) = 3.289, \ p < 0.001 \). There was no significant main effect of listener location, and none of the other interactions were significant.

Pairwise comparisons were conducted using t-tests with Bonferroni correction to further explore the significant main effects and interaction. For condition, independent- sample t-tests showed that compared to performance in the low-pass filtered condition, listeners were more accurate in the unmodified \( t(78) = 5.104, \ p < 0.0001, \ M_{\text{diff}} = 0.0359, \ 95\% \text{ confidence interval (CI)} (0.030,0.067) \) and monotonized \( t(78) = 5.304, \ p < 0.0001, \ M_{\text{diff}} = 0.0392, \ 95\% \text{ CI} (0.032,0.071) \) conditions. However, there was no difference between the unmodified and monotonized conditions. The pairwise comparisons for talker dialect, conducted as paired-sample t-tests, revealed seven significant differences. Listeners grouped Northern talkers together more accurately than Mid-Atlantic \( t(119) = 3.433, \ p < 0.001, \ M_{\text{diff}} = 0.0370, \ 95\% \text{ CI} (0.016,0.058) \), New England \( t(119) = 3.823, \ p < 0.001, \ M_{\text{diff}} = 0.0470, \ 95\% \text{ CI} (0.023,0.071) \), and Southern \( t(119) = 4.884, \ p < 0.0001, \ M_{\text{diff}} = 0.069, \ 95\% \text{ CI} (0.041,0.097) \) talkers. They also grouped Western talkers together more accurately than Mid-Atlantic \( t(119) = 3.559, \ p < 0.001, \ M_{\text{diff}} = 0.047, \ 95\% \text{ CI} (0.021,0.073) \), New England \( t(119) = 4.10, \ p < 0.0001, \ M_{\text{diff}} = 0.057, \ 95\% \text{ CI} (0.029,0.085) \), and Southern \( t(119) = 5.051, \ p < 0.0001, \ M_{\text{diff}} = 0.079, \ 95\% \text{ CI} (0.048,0.110) \) talkers. Accuracy was also higher for Midland talkers than Southern talkers \( t(119) = 4.10, \ p < 0.0001, \ M_{\text{diff}} = 0.039, \ 95\% \text{ CI} (0.020,0.058) \).

Finally, a series of independent-sample t-tests with Bonferroni correction was conducted to examine the significant condition \( \times \) talker dialect interaction. Since the primary goal of this study was to explore how listeners use segmental and prosodic information for dialect classification, the t-tests compared accuracy in the three conditions within each talker dialect. Four of these comparisons were significant. Listeners were more accurate for the Northern talkers in the unmodified condition compared to the low-pass filtered condition \( t(78) = 4.70, \ p < 0.0001, \ M_{\text{diff}} = 0.121, \ 95\% \text{ CI} (0.070,0.173) \) and in the monotonized condition compared to the low-pass filtered condition \( t(78) = 4.374, \ p < 0.0001, \ M_{\text{diff}} = 0.098, \ 95\% \text{ CI} (0.053,0.143) \). Listeners were also more accurate for the Western talkers in the unmodified condition compared to the low-pass filtered condition \( t(78) = 3.323, \ p < 0.01, \ M_{\text{diff}} = 0.107, \ 95\% \text{ CI} (0.043,0.171) \) and in the monotonized condition compared to the low-pass filtered condition \( t(78) = 3.925, \ p < 0.001, \ M_{\text{diff}} = 0.093, \ 95\% \text{ CI} (0.046,0.139) \). Thus, the main effect of condition was driven primarily by the Northern and Western talkers for whom accuracy was highest overall.

To further explore the effect of listener experience on dialect classification, the accuracy of listeners with residency experience in a single dialect region (either the Midland or the South) was compared with the accuracy of listeners who had resided in multiple regions using a repeated-measures ANOVA in which accuracy was the dependent variable, talker dialect was a within-subject variable, and condition and residence (Midland, South, multiple) were between-subject variables. As in the ANOVA reported above, condition and talker dialect yielded significant main effects, and there was a significant condition \( \times \) talker dialect interaction. However, there was no significant difference in accuracy between the three residency groups, nor did residence interact with any of the other variables.
3. Multidimensional scaling analysis

While the accuracy results are informative regarding the listeners’ ability to group talkers from the same macro-regions, they do not capture the perceived similarity between talkers of the same region, nor how the degree of perceived similarity may fluctuate when listeners do not have access to certain cues in the speech signal. To further examine how the listeners grouped talkers and what these groupings indicate about listeners’ perceptual dialect categories, a 60 \times 60 talker similarity matrix was generated for all listeners in each of the three listening conditions. The listeners from Ohio and Texas were combined for these analyses because they did not show a statistically significant difference in the accuracy analysis. Each matrix contained the number of times that each talker was paired with every other talker by all of the listeners in each condition. Thus, the potential values for each talker pairing ranged from 0 (the 2 talkers were not placed in the same group by any listener in that condition) to 40 (the 2 talkers were placed in the same group by all of the listeners in that condition). These matrices were then submitted to multidimensional scaling (MDS) analyses to visualize the perceptual talker similarity spaces. Analyses were conducted for each condition in 1, 2, and 3 dimensions. Inspection of scree plots revealed that the two-dimensional solutions provided the best fit in each of the analyses.

The two-dimensional solution for the unmodified condition is shown in Fig. 3. Each point in the space represents one talker, and the distance between any two points in the space is proportional to the perceptual distance between those two talkers. Dimension 1 in this solution can be interpreted as corresponding to talker dialect with New England (triangles) and Western (diamonds) talkers aligned on the left and Northern (plus signs) and Mid-Atlantic (squares) talkers aligned on the right. The Midland (circles) and Southern (\( \times \) symbols) talkers are spread throughout the center of the space. Dimension 2 represents talker gender: females (grey) are concentrated in the top half of the plot and males (black) in the bottom half with very little overlap between the two. Thus, although participants were asked to group the talkers by regional dialect, they attended to both dialect and gender in making their classifications.

The two-dimensional solution for the monotonized condition is shown in Fig. 4 and is similar to the solution for the unmodified condition. Talker dialect is again represented by dimension 1 with New England and Western talkers grouped together on the left and Northern and Mid-Atlantic talkers on the right. Dimension 2, again, represents talker gender with only a small degree of overlap between the females in the top half and the males in the bottom half of the space.

The two-dimensional solution for the low-pass filtered condition is shown in Fig. 5. This plot was rotated 60 deg counterclockwise so that the orientation of the dimension corresponding to gender would be the same as in Figs. 3 and 4. Talker gender is again one of the relevant dimensions, and the division between males and females is clear with virtually no overlap. However, the orthogonal dimension cannot be interpreted as talker dialect, since the clusters of same-dialect talkers observed in Figs. 3 and 4 are not present in Fig. 5. This result is not unexpected given that the low-pass filtered condition yielded the lowest overall accuracy rate of all three conditions. The interpretation of
the second dimension of Fig. 5 is not immediately obvious. Thus, in the low-pass filtered condition, the listeners were attending to talker gender but were not able to use the available cues to successfully classify the talkers by regional dialect.

C. Discussion

The results of experiment 1 showed that listeners were most accurate in their groupings when segmental information was available (unmodified and monotonized conditions) compared to when only prosodic cues were available (low-pass filtered), consistent with previous work (e.g., Leemann et al., 2018; van Bezooijen and Gooskens, 1999). Listeners performed equally well when given access to all acoustic cues present in the original stimuli in the unmodified condition or when given access only to segmental and rhythmic information in the monotonized condition. In other words, removing the intonational properties of the speech sample did not significantly impact the listeners’ ability to group the talkers by dialect. However, the removal of segmental features in the low-pass filtered condition led to significantly lower accuracy.

The MDS solutions for the three listening conditions provide further evidence for these results. First, when segmental information was available in the unmodified and monotonized conditions, the dialects for which listeners were most accurate had their talkers grouped in closer proximity to one another (i.e., the Northern and Western talkers) than the dialects with lower accuracy. Figures 3 and 4 show a distinct split between the Northern and Western talkers along dimension 1 with the Northern talkers concentrated in the positive values and the Western talkers concentrated in the negative values. Such a clear separation is not observed for any of the other four dialects in the unmodified and monotonized conditions. Second, the separation in dimension 1 of New England and Western talkers from Northern and Mid-Atlantic talkers likely reflects segmental differences between the dialects. Clopper et al. (2005) found that both Northern and Mid-Atlantic talkers produced relatively fronted /a/, as in “cottage” in the stimulus sentence in this experiment, whereas New England and Western talkers produced very few features that distinguished them from the other dialects. Fronting of /a/ may therefore have been a critical segmental cue for the listeners in the unmodified and monotonized conditions in experiment 1.

The MDS solutions also revealed that talker gender was relevant for groupings in all three conditions since clear separation with little overlap between male and female talkers was observed. This result is not surprising since listeners can reliably detect a talker’s gender from both the average F0 of the talker’s voice and the production of single segments (Honorof and Whalen, 2010), and previous studies using free classification tasks have also noted that listeners rely on gender for forming groups even when instructed to use other criteria (e.g., regional dialect) for their decisions (Clopper and Pisoni, 2007).

Unexpectedly, listener location was not a significant predictor of listeners’ ability to classify the talkers by dialect, and no listener location × talker dialect interaction was observed. Thus, neither Ohio nor Texas listeners benefited when classifying talkers from their own region (Midland and Southern, respectively) or any other specific dialect. A second analysis examining listeners with single-region vs multiple-region backgrounds likewise revealed no effect of residence and no residence × talker dialect interaction. This similar performance of listeners from the two locations (Ohio and Texas) and single-region vs multiple-region residential backgrounds differs from previous findings, demonstrating benefits of listener experience with their own dialect and geographic mobility and experience with other dialects (Baker et al., 2009; Clopper and Pisoni, 2004, 2007; Williams et al., 1999) for correct dialect identification and classification. The lack of observed effects of listener experience in the current study may reflect sufficient exposure to variation for all of the participants as a result of their attending large universities with diverse populations of students, given that The Ohio State University and The University of Texas at Austin are two of the largest universities in the United States.

In experiment 1, listeners heard only a short speech sample from each of the 60 talkers ($M_{\text{duration}} = 1.77$ s). The results showed that listeners were at least partially successful in identifying and classifying dialects based on a single sentence, and accuracy varied depending on the modifications made to the signal as in previous work (Clopper and Pisoni, 2004; Ikeno and Hansen, 2006, 2007; Leemann et al., 2018). Nevertheless, the use of such a short stimulus sentence may have provided limited dialect-specific segmental, rhythmic, and intonational information from each

FIG. 5. MDS solution for the low-pass filtered condition in experiment 1. AT, Mid-Atlantic; MI, Midland; NE, New England; NO, North; SO, South; WE, West.
speaker. Indeed, the paucity of dialect-specific information may have led listeners to rely more heavily on other socio-indexical cues, such as those to talker gender, to make their classifications. Longer stimuli containing more dialect-specific segmental and prosodic cues may allow listeners to refine their classification decisions, leading to higher accuracy. To explore this possibility, a second free classification experiment was conducted using a paragraph-long reading sample produced by a smaller number of talkers.

### III. EXPERIMENT 2

#### A. Methods

1. **Participants**

   One hundred and twenty native American English listeners residing in Ohio \((n = 60)\) and Texas \((n = 60)\) participated in experiment 2. These were different individuals from the ones in experiment 1. The mean age of the Ohio listeners was 21.0 years old (range 18–31 years old), and the mean age of the Texas listeners was 19.5 years old (range 18–27 years old). Data from 41 additional listeners were excluded from the analysis due to early knowledge of a foreign language \((n = 32)\); reported history of a speech, language, or hearing impairment \((n = 1)\); failure to provide their residential history \((n = 6)\); or experimenter error \((n = 2)\). None of the remaining listeners had exposure to a language other than English before age 6 years old, and none had advanced proficiency in a foreign language. Participants received partial course credit for their participation.

   The same coding scheme as in experiment 1 was used to place listeners into two groups within each testing location based on their residential history: only Midland or only South, for Ohio and Texas listeners, respectively, or multiple regions. In the Ohio group, 21 listeners had resided only in the Midland region, and 39 listeners had lived in more than one region. In the Texas group, 42 listeners had lived only in the South and 18 listeners had lived in multiple regions.

2. **Stimuli**

   A paragraph-long speech sample was taken from 18 female talkers who read the Goldilocks passage (Stockwell, 2002) as part of the NSP corpus (Clopper and Pisoni, 2006). The extracted portion for all talkers was “One morning, Mommy Bear had made some porridge for breakfast, but it was too hot to eat at once. ‘Let’s go for a walk while it cools down,’ said Daddy Bear. ‘What a good idea,’ exclaimed Mommy Bear. And, with their bear coats and bear shoes on, they all set off for a short walk in the woods.” The talkers selected for experiment 2 were a subset of the 60 talkers used in experiment 1 and included 3 talkers from each of the 6 regions. The number of talkers was reduced so that the task could be completed in a reasonable amount of time, despite the longer speech samples. All female talkers were selected to eliminate the observed effect of talker gender on performance in experiment 1. The three talkers within each region were selected for their reading fluency and dialect-appropriate segmental and prosodic features (Clopper and Smiljanic, 2015; Labov et al., 2006). The mean duration of the paragraph, including pauses, was 14.72 s \((SD = 1.22\) s, range 12.39–16.99 s). All stimuli were equated for RMS amplitude using Praat (Boersma and Weenink, 2019). Three listening conditions (unmodified, monotonized, and low-pass filtered) were created for the paragraph-long stimuli following the same procedure for the sentence-long stimuli in experiment 1.

#### 3. Procedure

Except for the longer stimuli and reduced number of talkers, all procedures were the same as in experiment 1. Because listeners’ demographic information was not available when assigning them to a condition, the three listening conditions were not balanced for residential history. Of the 39 Ohio listeners who had resided in multiple regions, 11 were assigned to the unmodified condition, 15 to the monotonized condition, and 13 to the low-pass filtered condition. Of the 18 Texas listeners who were coded as having lived in multiple regions, 8 were assigned to the unmodified condition, 6 to the monotonized condition, and 4 to the low-pass filtered condition.

#### B. Results

1. **Number of groups**

   Ohio listeners made an average of 5.9 \((SD = 1.6)\) groups in the unmodified condition, 5.5 \((SD = 2.2)\) groups in the monotonized condition, and 4.9 \((SD = 1.4)\) groups in the low-pass filtered condition. Texas listeners made an average of 5.9 \((SD = 1.6)\) groups in the unmodified condition, 6.5 \((SD = 2.0)\) groups in the monotonized condition, and 5.7 \((SD = 2.4)\) groups in the low-pass filtered condition. A two-way ANOVA with number of groups as the dependent variable and listener location and condition as independent variables revealed no significant main effects. The listener location \(\times\) condition interaction was also not significant.

2. **Accuracy analysis**

   Figure 6 shows the mean accuracy for each of the six dialects in each of the three listening conditions. Accuracy was defined as in experiment 1, except that there were only 3 possible correct pairings per dialect and 45 possible incorrect pairings per dialect due to the reduced number of talkers (3 per dialect). A repeated-measures ANOVA was conducted with the accuracy difference score (i.e., pairwise accuracy rate–pairwise error rate) as the dependent variable, talker dialect as a within-subject variable, and condition and listener location as between-subject variables. The ANOVA revealed significant main effects of condition \([F(2,114) = 11.939, p < 0.0001]\) and talker dialect \([F(5,570) = 16.566, p < 0.0001]\), as well as a significant condition \(\times\) talker dialect interaction \([F(10,570) = 4.244, p < 0.0001]\). The main effects of listener location and all other interactions were not significant.
To compare accuracy across the three listening conditions, three independent-sample t-tests with Bonferroni correction were performed. Accuracy was lower in the low-pass filtered condition when compared to both the unmodified condition \( t(78) = 5.101, p < 0.0001, \ M_{diff} = 0.124, 95\% \ CI (0.075,0.172) \) and the monotonized condition \( t(78) = 3.016, p < 0.01, \ M_{diff} = 0.073, 95\% \ CI (0.025,0.122) \). There was no significant difference between the unmodified and monotonized conditions. As in experiment 1, these results suggest that the segmental properties available in the unmodified and monotonized conditions favor accurate classification, while prosodic and rhythmic cues alone lead to less accurate groupings.

To explore differences in performance across the six talker dialects, a series of pairwise comparisons was performed using paired-samples t-tests. The results revealed the following significant differences in accuracy: listeners were more accurate at grouping Mid-Atlantic talkers than Midland \( t(119) = 7.581, p < 0.0001, \ M_{diff} = 0.296, 95\% \ CI (0.219,0.373) \), New England \( t(119) = 6.03, p < 0.0001, \ M_{diff} = 0.240, 95\% \ CI (0.161, 0.312) \), Northern \( t(119) = 4.624, p < 0.0001, \ M_{diff} = 0.191, 95\% \ CI (0.109,0.273) \), Southern \( t(119) = 5.521, p < 0.0001, \ M_{diff} = 0.235, 95\% \ CI (0.151,0.320) \), and Western \( t(119) = 4.261, p < 0.0001, \ M_{diff} = 0.181, 95\% \ CI (0.097,0.265) \) talkers. Additionally, accuracy was higher for Western talkers than Midland talkers \( t(119) = 3.778, p < 0.001, \ M_{diff} = 0.115, 95\% \ CI (0.055,0.176) \).

The significant condition \( \times \) talker dialect interaction was explored with a series of independent-sample t-tests with Bonferroni correction comparing performance across conditions for each talker dialect. Listeners were more accurate at grouping Western talkers together in the unmodified condition compared to the low-pass filtered condition \( t(78) = 4.654, p < 0.0001, \ M_{diff} = 0.269, 95\% \ CI (0.154,0.383) \) and in the monotonized condition compared to the low-pass filtered condition \( t(78) = 3.122, p < 0.01, \ M_{diff} = 0.164, 95\% \ CI (0.059,0.268) \). Additionally, accuracy was higher for Northern talkers in the unmodified condition than the low-pass filtered condition \( t(78) = 4.589, p < 0.0001, \ M_{diff} = 0.310, 95\% \ CI (0.176,0.444) \). Although visual inspection of Fig. 6 shows a large difference between listening conditions for the Mid-Atlantic talkers, the post hoc t-test comparing the unmodified and low-pass filtered condition fell short of the Bonferroni-corrected significance level \( t(78) = 3.06, p = 0.003, \ M_{diff} = 0.257, 95\% \ CI (0.090,0.425) \). Thus, as in experiment 1, the main effect of condition was driven primarily by the Northern and Western talkers for whom performance was relatively accurate overall.

Last, the accuracy of listeners with residential experience in a single region (either the Midland or the South) was compared with that of listeners who had resided in multiple regions. A repeated-measures ANOVA was run with accuracy as the dependent variable, talker dialect as a within-subject variable, and residence (Midland, South, multiple) as between-subject variables. The main effects of condition and talker dialect reported above, as well as the condition \( \times \) talker dialect interaction, were significant. As in experiment 1, there was no significant main effect of residence, nor did residence interact with any other variables.

3. MDS analysis

As in experiment 1, MDS analyses were conducted for each listening condition. Given the smaller number of talkers, MDS analyses were only conducted in one and two dimensions to avoid model over-parameterization. Scree plots were visually inspected, and it was determined that the two-dimensional solution was most appropriate for each condition. The three MDS plots in Figs. 7–9 offer a visual representation of the perceptual distances between talkers as classified by listeners in each listening condition. Recall that unlike experiment 1, where the MDS plots revealed a clear separation of male and female talkers, all talkers in experiment 2 were female and therefore gender was not a possible interpretation for either dimension. In the solution for the unmodified condition, shown in Fig. 7, dimension 1 corresponds to talker dialect, with Mid-Atlantic talkers (squares) on the left and Western talkers (diamonds) on the right. Dimension 2 may also correspond to a different aspect of talker dialect, with Northern talkers (plus signs) at the top and two of the three Southern talkers (\( \times \) symbols) at the bottom. The solution for the monotonized condition, shown in Fig. 8, can be interpreted similarly. Dimension 1 shows the Mid-Atlantic and Western talkers to the left and the right of the space, respectively, as well as the Northern and Southern talkers to the top and the bottom of the space, respectively, although talkers are more spread out in dimension 2 in Fig. 8 than in Fig. 7. In Fig. 9, which corresponds to the groupings in the low-pass filtered condition, neither dimension corresponds clearly to talker dialect. Moreover, no dialect has all three of its talkers grouped closely together, although for some dialects (e.g., Mid-Atlantic,
New England, and Southern) two of the three talkers are located in close proximity to each other.

C. Discussion

Similar to experiment 1, listeners classifying paragraph-long stimuli were more accurate when they had access to all speech characteristics in the unmodified condition or to segmental and rhythmic properties but not intonational information in the monotonized condition, relative to when the only cues available for classification were intonation and rhythm in the low-pass filtered condition. This higher accuracy in the unmodified and monotonized conditions was observed even though the talker dialect with the highest accuracy in experiment 2 (Mid-Atlantic) was not the same as the most accurately classified dialects in experiment 1 (Northern and Western). This difference across experiments in which dialects were most accurately classified likely reflects differences in the segmental features available in the stimuli. For example, in the experiment 2 stimuli, the three talkers from the Mid-Atlantic dialect produced a lower vowel in the first syllable of “porridge” than the talkers from the other regions. This salient segmental cue could have led to the high accuracy for the Mid-Atlantic dialect in experiment 2. This cue was not available in the stimulus sentence in experiment 1 for the listeners to rely on, leading to lower overall accuracy for the Mid-Atlantic talkers. Note that in the low-pass filtered condition, where this vowel would not have been perceptible to listeners, accuracy for the Mid-Atlantic dialect is more in line with the other dialects. This interpretation is supported by participant responses during an informal debriefing exercise after completing the classification task in experiment 2, in which some listeners reported attending to specific words when making their groupings, including the word “porridge.” It would be useful for free classification tasks in the future to include a more structured debriefing to probe which features listeners report using when making their classifications. This information would be useful not only for the purpose of understanding performance on the task itself but would also shed light on language ideologies and linguistic stereotypes, at least for the young college students in these studies.
Turning to the question of the effect of listener experience on dialect classification accuracy, experiment 2 did not reveal a significant benefit of residential history or geographic mobility since Ohio and Texas listeners did not exhibit differential performance on Midland and Southern talkers, and listeners who had lived in multiple regions did not create more accurate groupings than listeners who had lived in a single region (either the Midland for Ohio listeners or the South for Texas listeners). We provide some possible reasons for these findings in the general discussion in Sec. IV.

IV. GENERAL DISCUSSION

Taken together, the findings from experiments 1 and 2 provide evidence that American English listeners rely primarily on segmental information when classifying talkers by regional dialect. The information available in the unmodified and monotonized conditions in both experiments allowed listeners to make more accurate groupings than in the low-pass filtered condition, which removed the segmental information that makes speech intelligible while leaving prosodic cues intact. This asymmetry in favor of segmental properties aligns well with the findings from other studies conducted on sociophonetic variation and identification for English and other languages (e.g., Leemann et al., 2018; Thomas and Reaser, 2004; van Bezooijen and Gooskens, 1999).

While we cannot directly compare the accuracy scores from experiments 1 and 2 because of the different number of talkers included in each experiment (and thus the different number of possible correct and incorrect talker pairings), it is noteworthy that even in experiment 2, which contained a paragraph-long speech sample from each talker, accuracy in the unmodified condition was near or below zero for three dialects (Midland, New England, and Southern). That is, the paragraph-length stimuli did not provide sufficient information, even when unmodified, to allow successful dialect classification performance. Thus, providing longer stretches of speech from each talker does not necessarily lead to more accurate performance in a free classification task. One explanation for this result is that the longer stimuli impose additional processing costs that offset the benefits of additional dialect-specific information in the speech signal. In particular, to perform this task, listeners must identify features in the speech sample that are regionally distinctive and ignore sources of talker-specific or gender-specific variation. Listeners must also store this information in working memory while they listen for other talkers with similar features. The considerable cognitive load imposed by this task might favor the use of simpler strategies, such as identifying a single word or phrase that exhibits variation among the talkers and then using that word or phrase as the basis for classification, rather than constructing a more holistic regional dialect representation for each talker based on the full speech sample. As noted in the discussion of the experiment 2 results, some participants reported attending to particular words when they were asked about their classification strategies, consistent with this proposal.

In addition, in the current study both the number of talkers and the length of stimuli varied between the two experiments because we anticipated that presenting listeners with a paragraph-long stimulus read by 60 different talkers would result in an excessively demanding task. Thus, the listeners were exposed to greater between-talker variability (i.e., shorter speech samples produced by a larger number of talkers identifying with two genders) in experiment 1 and to greater within-talker variability (i.e., longer speech samples produced by fewer talkers identifying with one gender) in experiment 2. This difference in the nature of the variability in the stimulus materials may have affected classification performance. Future studies should examine more closely how task demands, including cognitive effort, talker-related variability, and the availability of salient acoustic cues, impact classification accuracy.

A second goal of this study was to examine the effect of residential history on dialect classification accuracy in light of previous work that has found benefits associated with experience with particular dialects and, more generally, with living in multiple dialect regions (Baker et al., 2009; Clopper and Pisoni, 2004; Williams et al., 1999). In both experiments in the current study, listeners did not classify talkers from their local dialect region (Ohio or Texas) more accurately than talkers from other regions, and listeners who had lived in multiple regions were not more accurate than those who had lived in a single region. As noted above, this lack of an effect may reflect the participants’ shared exposure to dialect variation on their large and diverse university campuses. A second potential explanation relates to the degree and timing of the listeners’ exposure to multiple dialect regions. Clopper and Pisoni (2004) found that residential history predicts performance on a dialect categorization task, but they compared a geographically mobile group of so-called “army brats” who had lived in at least three different states before age 18 years old to a group of non-mobile participants. By contrast, our multiple-residence listeners had lived in a minimum of just two regions, and there was no age requirement placed on their exposure; in fact, some listeners’ first residence outside of their home region was around age 18 years old when they began attending university. Just as early exposure plays a crucial role in phonological learning during first language acquisition (e.g., Kuhl, 2000; Werker and Teras, 1984), representations of regional dialects may be less malleable and more affected by linguistic exposure earlier in life. When Jones et al. (2017) studied perception of American English regional dialects (with New England, Northern, Midland, and Southern talkers) in a cross-sectional sample of children aged 8–17 years old using a free classification task, they found that accuracy improved as age increased. However, the 16- and 17-year-old listeners’ classifications were as accurate as those made by adult listeners, even though the adults were more likely to have lived in multiple regions.
than the adolescents. The findings by Jones et al. (2017) provide support for the proposal that the listeners in our study (who failed to show effects of residence or mobility on performance) may not have had sufficient experience with other dialects early enough in their lives for effects of residential history to emerge.

The low accuracy in all listening conditions and both experiments for the Southern talkers, whose dialect is perhaps the most stereotyped of all American English dialects (Niedzielski and Preston, 2000), was also unexpected. Previous work has examined the phonetic properties of Southern American English in both the segmental and prosodic domains, demonstrating that Southern speakers engage in monophthongization of diphthongs (Thomas, 2003), speak more slowly (Clopper and Smiljanic, 2015), pause more frequently and for longer durations (Clopper and Smiljanic, 2011; Jacewicz et al., 2009), and exhibit more variable vowel durations (Clopper and Smiljanic, 2015) compared to the speakers of other American English regional dialects. Nevertheless, it is possible that representations of the Southern dialect for American English listeners do not include all of these characteristics. In fact, listeners from outside of a given dialect region tend to rely on a limited set of stereotypical features to identify the dialect of that region (Baker et al., 2009). The Southern talkers who produced the stimuli for both experiments 1 and 2 may not have exhibited the most strongly stereotyped features, leading to lower classification accuracy. The Southern talkers also may not have exhibited as many of the phonetic characteristics of Southern American English as they would have in spontaneous speech, especially with interlocutors with whom they would feel more at ease in using their native regional variety. Read speech produced in the laboratory may have favored a more formal speech style in which stereotyped regional features were reduced (Labov, 1972). Further research is needed to understand which specific segmental and prosodic cues listeners attend to when classifying talkers by regional dialect.

In summary, the present findings provide important insights into the perceptual consequences of the availability of segmental and prosodic information in the speech of talkers from different regions, namely, that American English listeners rely primarily on segmental information when grouping talkers from the same regional dialect together in a free classification task. Segmental information appears to be necessary, although not always sufficient, for classifying regional dialects, given that the conditions containing segmental information led to more accurate groupings than the condition without segmental information. Listeners were able to use this segmental information to accurately classify talkers from some regions (i.e., the Northern and Western talkers in both experiments and the Mid-Atlantic talkers in experiment 2). However, even with the segmental information available, classification of some dialects was poor (i.e., the Midland, New England, and Southern talkers in both experiments).

The results of this study cannot tell us which salient segmental features led to the observed patterns of accuracy and perceptual similarity across dialects. A detailed study teasing apart the relative contribution of specific segmental and prosodic features to classification accuracy remains a pressing goal. Similarly, the difference in the length of the speech samples and the number of talkers included in the two experiments may be partially responsible for the sharp increase in accuracy for Mid-Atlantic talkers and the small decline in accuracy for Midland talkers from experiment 1 to experiment 2. Future studies should manipulate the variables of speech sample length and number of talkers independently to pinpoint the source of these differences.

Another contribution of this work is that the results call into question the nature of the advantage of dialect experience or exposure on dialect classification performance. Listeners from Ohio and Texas were not better at grouping talkers from their own dialect, and listeners who had resided in multiple dialect regions were not more accurate than those who had only resided in a single dialect region. These results suggest that greater experience and geographic mobility do not guarantee increased accuracy in a dialect classification task. Future work should further examine the nature of listener experience, especially with respect to the amount and timing of exposure, and how it relates to the classification accuracy of regional dialects of English and other languages.

ACKNOWLEDGMENTS

We are grateful to Frida Ballard, Sarah Blackshear, Luis De La Cruz, Maria Gavino, Sandie Keerstock, Kevin Lilley, Olivia Naworol, Gabriela Oppenheimer, Jai-Lyn Peña, Hannah Ray, Gabrielle Shaddock, Irene Smith, Liliana Terry, and Erin Walpole for assistance with data collection. This manuscript also benefited from feedback received at the 173rd and 174th Meetings of the Acoustical Society of America in Boston and New Orleans, respectively, and the 4th Experimental and Theoretical Advances in Prosody conference at the University of Massachusetts at Amherst.

1More detailed information about the frequency or intensity of exposure to each dialect for the multiple-region listeners is not available from the data we collected.


