Published in final edited form as:
J Lang Soc Psychol. 2005 June 1; 24(2): 182–206. doi:10.1177/0261927X05275741.

EFFECTS OF TALKER GENDER ON DIALECT CATEGORIZATION

CYNTHIA G. CLOPPER, **BRIANNA CONREY**, and **DAVID B. PISONI**
Indiana University

Abstract

The identification of the gender of an unfamiliar talker is an easy and automatic process for naïve adult listeners. Sociolinguistic research has consistently revealed gender differences in the production of linguistic variables. Research on the perception of dialect variation, however, has been limited almost exclusively to male talkers. In the present study, naïve participants were asked to categorize unfamiliar talkers by dialect using sentence-length utterances under three presentation conditions: male talkers only, female talkers only, and a mixed gender condition. The results revealed no significant differences in categorization performance across the three presentation conditions. However, a clustering analysis of the listeners' categorization errors revealed significant effects of talker gender on the underlying perceptual similarity spaces. The present findings suggest that naïve listeners are sensitive to gender differences in speech production and are able to use those differences to reliably categorize unfamiliar male and female talkers by dialect.

Keywords

dialect categorization; gender; regional dialect; speech perception; indexical properties

Speech perception involves not only processing of the linguistic message, but also processing and encoding talker-specific properties of the speech signal such as the age, gender, or regional dialect of the talker (Klatt, 1989). The identification of talker gender is a relatively easy task for adult listeners. Lass, Hughes, Bowyer, Waters, and Bourne (1976) reported that naïve adults could identify the gender of unfamiliar talkers based on single vowel utterances with near-ceiling performance. In addition, the listeners were 75% accurate in identifying the gender of the talker using whispered speech, suggesting that gender information is carried in the speech signal by phonological properties other than voicing and fundamental frequency.

More recently, Mullennix and Pisoni (1990) conducted a speeded classification task that provided evidence for the automaticity of talker gender identification. Participants were asked to attend to either the gender of the talker or the linguistic content of the utterance. Mullennix and Pisoni (1990) found that talker gender interfered with performance in the linguistic task more than the linguistic information interfered with the gender identification task, suggesting that gender-specific information is an integral component of the speech perception process.

Gender-specific information in the speech signal is not based exclusively on biological differences between men and women. In her review of the language attitude literature as it pertains to gender differences, Kramarae (1982) noted that although pitch differences between males and females are at least partly the result of biological differences, many other linguistic gender differences appear to be learned. In terms of phonological differences,

women are often thought to produce clearer speech with greater modulation of pitch than men. In one recent study, Namy, Nygaard, and Sauerteig (2002) reported that women were more likely to phonologically accommodate their speech to that of an interlocutor than men. In addition, women accommodated more to male interlocutors than to female interlocutors, but men showed no difference in accommodation based on the gender of the interlocutor. These findings suggest that a talker's gender is a critical component of his or her phonological system.

It is also well known that gender interacts with other social variables that affect phonology, including regional and ethnic dialects. Gender-correlated differences in the production of prestige forms and innovative forms of speech have been reported frequently in the sociolinguistics literature. In an extensive review, Labov (1990, 2001) summarized the observed production differences with the three principles shown below.

1. Women use more prestige forms than men, and, conversely, men use more nonstandard forms than women.
2. Women favor incoming prestige forms in changes from above, which are defined as involving forms associated with a high level of social consciousness.
3. Women also tend to lead in changes from below, which involve variation that has not become stereotyped or associated with particular social groups. However, in a minority of cases such as diphthong centralization on Martha's Vineyard (Labov, 1963), men can lead in changes from below.

Labov (1990, 2001) also discussed the interaction of gender and social class. In general, the second highest status group, which he refers to as the lower middle class (1990) or the upper working class (2001), shows the greatest gender differences in speech production and the least frequent use of stigmatized forms. For instance, the crossover pattern in which the second highest status group uses the standard or conservative form more frequently in careful speech than the highest status group (e.g., Labov, 1966) is more prominent in women than men (Labov, 2001).

Although exceptions exist to these general observations and some of the concepts involved have been called into question (e.g., the definitions of "standard" and "nonstandard" and the methods of assigning social class to women; see Cheshire, 2002), these generalizations have been supported by a number of influential studies in English-speaking Western communities. In an early study, Trudgill (1974) found that women were more conservative than men in the use of nonstandard forms in Norwich, England. For example, men were more likely than women to use the nonstandard form/n/for the standard form/ŋ/. Also, working class speakers were more likely to use the nonstandard/n/than speakers from other social classes. Trudgill hypothesized that women are more status conscious than men and so are more likely to use socially prestigious, linguistically conservative forms and to avoid nonstandard forms like/n/that are typically associated with working class speech.

Following up on this earlier work, Eckert (1989) emphasized the interaction of gender with other socially constructed categories, such as socioeconomic class and social group. In addition, she hypothesized that because women often have less material capital than men, women rely more on phonological variation as "symbolic capital" to signal their group membership. In a study that documented the progress of the Northern Cities Chain Shift among high school students in a Detroit suburb who belonged to one of two social groups, the jocks or the burn-outs, Eckert found that both gender and group membership played a role in speech production. Specifically, the burnouts were further advanced than the jocks in the backing of/ɛ/and/Λ/, the newer changes in the shift that were associated more with Detroit urban speech and potentially with the somewhat subversive, counteradult values

adopted by the burnouts. However, the girls in both groups were more advanced in the fronting of/æ/and/a/and the lowering and fronting of/ɔ/, all of which are older and more established changes in the shift that had already entered the phonological repertoire of most speakers in the community.

In another study, Milroy and Milroy (1993) discussed gender differentiation as it interacts with social class, but they emphasized the role of social network strength in the spread of linguistic change. An individual's social network was defined by how many ties he or she had to others, in how many capacities he or she interacted with them (network multiplexity), and how many of them knew each other (network density). A strong social network is one that is highly dense and multiplex, and an individual's network strength score increased with increases in network density and/or multiplexity. Milroy and Milroy hypothesized that linguistic innovators are individuals who know many people but have weak social networks. They found that women had weak network strength scores and also a low incidence of a particular nonstandard variant, intervocalic fricative deletion; in contrast, men had stronger networks and a higher incidence of the nonstandard variant. Similarly, in his data from Philadelphia, Labov (2001) found that linguistic innovators were most often upwardly mobile and locally well-respected women with many nonlocal ties.

The relationship between gender and linguistic variation in speech requires a great deal more exploration, but gender differentiation within regional dialects has clearly been observed (Eckert, 1989; Labov, 2001). Despite these well-documented differences in production between male and female speech, however, research on the perception of dialect variation has been limited almost exclusively to the study of male talkers. In one of the first studies of dialect perception, Preston (1993) asked naïve participants to listen to samples of speech from nine different talkers and then decide where each talker was from out of a set of nine cities located between Dothan, Alabama, and Saginaw, Michigan. All nine talkers in Preston's study were male.

Purnell, Idsardi, and Baugh (1999) conducted a study of ethnic identification in the San Francisco area using a variant of the matched-guise technique. A single talker left answering machine messages for landlords inquiring about apartments for rent using one of three ethnic guises: Standard American English, African American Vernacular English, and Chicano English. The talker used in the Purnell et al. study was also male.

More recently, researchers in the United Kingdom, the Netherlands, and the United States have conducted forced-choice dialect categorization tasks using multiple talkers from multiple regional varieties of English and Dutch. Williams, Garrett, and Coupland (1999) used male talkers and listeners in their examination of dialect categorization in Wales. Van Bezooijen and Gooskens (1999) also used only male talkers in their studies of dialect categorization in the United Kingdom and the Netherlands. In their dialect categorization research in the United States, Clopper and Pisoni (2004b) used only male talkers as well. In one study that provides an exception to this trend, Van Bezooijen and Ytsma (1999) explored categorization in the Netherlands using a set of 24 female talkers.

Although the precise methodologies and stimulus materials across the dialect categorization tasks have varied considerably, performance in all cases was generally quite poor overall, although above chance. Williams et al. (1999) reported that Welsh adolescents could identify the dialect of other Welsh adolescents in an eight-alternative forced-choice categorization task with approximately 30% accuracy. Adults in the United Kingdom performed somewhat better, accurately categorizing 52% of the talkers by area in a forced-choice categorization task that included identification of the talkers' country (e.g., England or Scotland), region (e.g., North England or Wales), and area (e.g., North Wales or South

Wales; Van Bezooijen & Gooskens, 1999). Meanwhile, adults in the Netherlands accurately identified the province of origin of 40% of the male talkers and 35% of the female talkers (Van Bezooijen & Gooskens, 1999; Van Bezooijen & Ytsma, 1999) in a similar multistage categorization task. Finally, adult listeners in the United States also performed poorly. Clopper and Pisoni (2004b) found that listeners were only 31% correct in a six-alternative forced-choice dialect categorization task.

The similarity between the results reported for male talkers in the Netherlands by Van Bezooijen and Gooskens (1999) and those reported for female talkers by Van Bezooijen and Ytsma (1999) suggests that the gender of the talkers may not have a large effect on perceptual categorization. However, the gender differences in speech production summarized above indicate that an explicit comparison of perceptual performance across talker gender is warranted. The present study was designed to provide such a comparison using three sets of stimulus materials: male talkers only, female talkers only, and a mixed group of male and female talkers. The data from the male talkers have previously been reported by Clopper and Pisoni (2004b) and are summarized briefly here. The data from the female talker and mixed talker groups are reported below in Experiments 1 and 2, respectively.

Data obtained from all three presentation conditions were collected and analyzed using the same experimental methodology. Naïve participants were asked to listen to isolated English sentences and make judgments about where the talkers were from using a six-alternative forced-choice categorization task. The talkers represented six different regional varieties of American English: New England, North, North Midland, South Midland, South, and West. Each listener completed three blocks of trials containing different stimulus materials. In the first block, the listeners heard each talker reading the sentence, “She had your dark suit in greasy wash water all year.” In the second block, they heard each talker reading the sentence, “Don’t ask me to carry an oily rag like that.” Finally, in the third block, each talker read a different, novel sentence.

The data obtained in each experiment were scored for categorization accuracy and submitted to a hierarchical clustering analysis to determine the perceptual similarity spaces of the dialects (Corter, 1982; Nosofsky, 1985). The first column of Table 1 shows the percentage correct accuracy scores for the male talker condition for each of the three experimental blocks (Clopper & Pisoni, 2004b). Chance performance is 17% in a six-alternative forced-choice task. Therefore, although the listeners in this task performed above chance, their overall performance was still quite poor. Performance on Sentence 2 was significantly worse than performance on Sentence 1 and the novel sentences. Categorization performance was the same across the first and novel sentence conditions. Performance was also assessed separately for each dialect. Clopper and Pisoni (2004b) found that listeners were more accurate in categorizing talkers from New England and the South than any of the other four dialect regions. These two regional varieties are perhaps the most marked dialect regions in the United States in terms of both production (Krapp, 1925) and perception (Preston, 1993), so it is not surprising that these two dialects were also the easiest to identify for naïve listeners.

Given the high error rate in the categorization task, the stimulus-response confusion matrices were expected to reveal interesting patterns that would reflect the perceptual similarity of the six regional dialects of American English. In particular, the confusion matrices from the male talker condition were submitted to a hierarchical clustering analysis that revealed three major perceptual dialect clusters for the first and novel sentences: New England; South and South Midland; North, North Midland, and West. For the second sentence, a slightly different configuration was found that also consisted of three clusters: New England and

North; South and South Midland; North Midland and West (Clopper & Pisoni, 2004b). The solutions to the hierarchical clustering analysis carried out by Clopper and Pisoni (2004b) are depicted graphically in Figure 1. In these figures, perceptual dissimilarity is proportional to the vertical distance connecting any two nodes. That is, the dissimilarity of any two dialects is the sum of the lengths of the fewest vertical branches connecting them. The three broad dialect clusters obtained in the perceptual similarity analysis were consistent with earlier findings in the sociolinguistics literature on production variation that describe the three major dialects of American English as eastern, southern, and western (Krapp, 1925) or northern, southern, and western (Labov, 1998).

Given the findings reported by Van Bezooijen and her colleagues for Dutch varieties (Van Bezooijen & Gooskens, 1999; Van Bezooijen & Ytsma, 1999), we did not expect to find large differences in performance across the three talker gender conditions in our six-alternative categorization task. However, the differences in production between the male and female talkers that would be predicted based on the variationist literature might lead to significant differences in the underlying perceptual similarity spaces of the dialects for each talker gender condition. For example, we would expect the Northern women to show greater advancement in production of the Northern Cities Chain Shift variables than the northern men. This difference in production should be reflected in the perceptual similarity of the Northern and North Midland dialects: these two dialects should be perceptually more similar for male talkers than for female talkers. Similarly, the Southern dialects (South and South Midland) are generally considered to be less prestigious than the Northern and Western varieties of American English (Preston, 1993), so Southern and South Midland women might exhibit fewer stigmatized features of these dialects in their speech. This difference in production should again be reflected in the perceptual similarity spaces of listeners: Southern and South Midland women should be more similar to women from other regions than Southern and South Midland men are to men from the other dialect regions. These and other differences in production that result from gender differences in linguistic change and prestige form usage should result in different perceptual dialect clusters for male and female talkers.

EXPERIMENT 1

METHOD

Stimulus materials—Stimulus materials consisted of audio recordings of read sentences drawn from the TIMIT Acoustic-Phonetic Continuous Speech Corpus (Fisher, Doddington, & Goudie-Marshall, 1986; Zue, Seneff, & Glass, 1990). The TIMIT corpus includes audio recordings of talkers from eight different dialect regions of the United States. Each talker in the TIMIT corpus was recorded reading 10 sentences. Two of these sentences, the calibration sentences, were the same for each talker and included lexical items and phonetic contexts designed to elicit regional dialect features (Fisher et al., 1986; Zue et al., 1990). These two sentences are shown below. Of the remaining eight sentences for each talker in the TIMIT corpus, five were read by a total of seven talkers in the corpus, and three were read by only a single talker. That is, some of the novel sentences included in the TIMIT corpus were read by multiple talkers and some were not. Although some of the sentence materials were produced by more than one talker on the TIMIT corpus, the novel sentences selected for this experiment were all different for each of the talkers.

1. She had your dark suit in greasy wash water all year (Sentence 1).
2. Don’t ask me to carry an oily rag like that (Sentence 2).

Six of the dialect regions were of interest in this experiment: New England, North, North Midland, South Midland, South, and West. Eight female talkers were selected from each of

these six dialect regions, for a total of 48 different talkers. All of these talkers were White females between the ages of 20 and 29 at the time the recordings were made and were chosen by two phonetically trained listeners (the first and second authors) as the best representatives of their respective regional dialects. Both of the calibration sentences were used from each talker in this experiment. In addition, a third novel sentence was chosen for each talker. These novel sentences were hand-selected to ensure that relevant linguistic features were present in the stimulus materials that the untrained listeners could use to accurately categorize the talkers. Each of the phonetically trained listeners selected one sentence for each talker independently. When the two trained listeners selected different sentences for a given talker, the final sentence selection was made after additional listening to the materials. The novel sentences used in the experiment were, therefore, judged by the phonetically trained listeners to contain lexical items and/or phonetic contexts associated with regional variation documented in the sociolinguistics literature. Above-chance performance by the listeners validates the selections made by the two phonetically trained listeners. None of the novel sentences was ever repeated more than once during the course of the experiment. Each sentence was saved in a separate digital sound file, and all of the sound files were leveled to 55 dB using Level16 (Tice & Carrell, 1998).

Listeners—The listeners in this study were 35 Indiana University undergraduates, all of whom received partial credit in an introductory psychology course for participating. All listeners whose data were used in the final analysis were monolingual native speakers of American English with no history of a hearing or speech disorder. In addition, all listeners performed above chance on at least one of the three blocks of the experiment. Data from two bilingual listeners were discarded as were data from three listeners who performed statistically at chance on all three blocks of the experiment.[1] Thus, the final analysis included data from 30 listeners, 5 males and 25 females. None of these listeners had participated in the male talkers only experiment (Clopper & Pisoni, 2004b). The listeners represented a range of regional dialects, but approximately half (17 of 30) had lived only in Indiana.

Procedure—The experimental procedures used for the dialect categorization task were identical to those used by Clopper and Pisoni (2004b). Listeners were seated at personal computers equipped with headphones and a mouse. Each of the six response alternatives, corresponding to the six dialect regions, was represented by a partial map of the United States, including state boundaries. The maps were arranged on the screen so that they appeared in approximately the correct overall positions but were spatially separated so as to avoid the introduction of response error. The response alternatives are shown in Figure 2. The listeners were given the opportunity to familiarize themselves with the response alternatives and the presentation format prior to beginning the experiment.

The experiment itself consisted of three blocks of test trials that were completed by all listeners in the same order. In the first block, the listeners heard all 48 female talkers read the first calibration sentence once. The presentation order of the talkers was random and was unrelated to dialect region. After the presentation of each sentence, the listeners were asked to indicate which region they thought the talker was from by clicking on that region with the mouse. No feedback was given as to the accuracy of the listener responses. The second

[1]A strict exclusion criterion was used to ensure that all participants included in the final data analysis were trying to perform the task accurately. The source of at-chance performance on a categorization task like the one used here cannot be determined: Participants who failed to perform above chance on any of the three blocks of trials may not have been attending to the task or may have been unable to perform the task despite their best efforts. The data from these participants were, therefore, excluded. Using this strict criterion, fewer than 5% of the total participants were excluded. The question of individual differences in performance is beyond the scope of this study, but remains an interesting and important question for future research on dialect categorization.

block was similar to the first, except that the listeners heard the second calibration sentence. The third block was the same as the first two, but the novel sentences were presented. Throughout the experiment, the sentences were presented at an average signal level of 70 dB SPL over Beyerdynamic DT100 headphones.

RESULTS AND DISCUSSION

Perceptual categorization

Overall categorization performance on the female talkers was consistent with the previous research on male talkers reported by Clopper and Pisoni (2004b). As shown in the second column of Table 1, the listeners were able to correctly categorize 31%, 28%, and 31% of the female talkers in the first, second, and third experimental blocks, respectively.

A repeated-measures ANOVA on talker dialect (New England, North, North Midland, South Midland, South, or West) and experimental block (first, second, or novel sentences) for the female talker condition revealed a significant effect of talker dialect, $F(5, 174) = 17.35$, $p < .001$, and a significant Dialect × Block interaction, $F(10, 174) = 6.33$, $p < .001$. The main effect of experimental block was not significant. Posthoc Tukey tests on talker dialect revealed better overall performance on New England and Southern talkers than talkers from the other four regions (all ps < .001). Performance on North Midland talkers was also significantly better than performance on the Western talkers ($p < .001$). None of the other pairwise comparisons were significant.

As shown in Figure 3, the Dialect × Block interaction is the result of increased performance on Southern talkers in the novel sentence block and decreased performance on the New England talkers in the second block of trials. These results are consistent with the earlier findings reported by Clopper and Pisoni (2004b) for male talkers, which also revealed better categorization performance for the New England and Southern talkers and an effect of experimental block for those same groups of talkers.

A repeated-measures ANOVA on listener gender (male or female) and experimental block (first, second, or novel sentences) revealed no significant main effects or interactions. Categorization performance was not significantly affected by the gender of the listeners.

Perceptual similarity

Although overall categorization performance was quite poor, a closer examination of the listeners' responses in the female talker condition revealed that listeners were not responding randomly. Their errors were structured based on the perceptual similarity of the different dialects. To assess the listeners' perceptual con-fusions, one stimulus-response matrix was constructed for each experimental block, collapsed across all of the listeners. These matrices were then submitted to the Similarity Choice Model (SCM; Nosofsky, 1985) to extract similarity matrices for use in a hierarchical clustering analysis of perceptual similarity. Two SCM analyses were computed across the three error matrices for the female talker condition. In the first analysis, the similarity parameters were allowed to vary freely across each of the three matrices. In the second, restricted analysis, the similarity parameters were held constant across all three matrices. A comparison of the model fit for the two solutions (unrestricted and restricted by similarity) revealed a significantly better fit for the unrestricted model than the restricted model. This difference in model fit indicates that the different perceptual similarity models in the unrestricted analysis are not due to chance and that different models are necessary to account for the confusion data for each of the three sentence conditions (Sentence 1, Sentence 2, and novel sentences). Clopper and Pisoni (2004b) also found that different perceptual similarity models were necessary to account for the data from the three different experimental blocks in the male talker condition.

The similarity matrices resulting from the three individual SCM analyses were then submitted to a hierarchical clustering scheme, ADDTREE (Corter, 1982). The ADDTREE analysis computed a hierarchical structure based on the similarity input that is represented graphically in Figure 4 for the female talkers for each of the three sentence blocks. As in Figure 1, perceptual dissimilarity is indicated by the vertical lengths of the branches connecting any two dialects.

The similarity structure of the dialects for the female talkers is similar overall to the results found for the male talkers (Clopper & Pisoni, 2004b). For the first sentence, the listeners perceived three major dialect clusters: New England; South and South Midland; North, North Midland, and West. For the second and novel sentences, the listeners also perceived three major clusters with a slightly different composition: New England and North; South and South Midland; North Midland and West.

Comparison to the male talker condition

As discussed above, performance in the female talker condition was qualitatively similar to performance in the male talker condition. To quantify the cross-condition comparisons, we conducted a repeated measures ANOVA on talker dialect, experimental block, and talker gender (male or female). The results of the ANOVA revealed a significant main effect of talker dialect, $F(5, 276) = 29.13$, $p < .001$, a significant main effect of experimental block, $F(2, 552) = 5.40$, $p < .01$, and a significant Dialect × Block interaction, $F(10, 276) = 9.39$, $p < .001$. In confirmation of our qualitative assessment of the consistency of the categorization results across gender conditions, neither the main effect of gender nor any of the interactions involving talker gender were significant.

A comparison of the male and female talker conditions for the perceptual similarity analysis did, however, reveal significant differences between the two conditions for each of the three experimental blocks. As in the analysis above for the sentences within the female talker condition, two SCM analyses were conducted using data from the male talker condition and the female talker condition. The model fit of the restricted model, in which similarity parameters were held constant across the two talker conditions, was significantly worse than the model fit for the unrestricted model, in which the similarity parameters varied freely across the two conditions. This difference in model fit suggests that the differences between talker conditions are not due to chance, but reflect an underlying difference in perceptual spaces for male and female talkers.

A visual inspection of Figures 1 and 4 provides some insight into these differences. For Sentence 1 and Sentence 2, the perceptual dialect clustering solutions are virtually identical in structure across the two gender conditions and the difference between the male and female talkers can be attributed to differences in the lengths of the individual branches. In perceptual terms, these differences in branch lengths reflect differences in dialect discriminability. For the novel sentences, the differences between the two gender conditions are more apparent. The Northern male talkers cluster with the North Midland and Western talkers, whereas the Northern female talkers cluster with the New England talkers. Despite the overall similarity in categorization performance as measured in terms of accuracy, significant differences in the perception of male and female talkers were revealed in the similarity space analysis. Experiment 2 examined the perception of dialect variation of male and female talkers more directly using a single set of listeners and a mixed group of talkers.

EXPERIMENT 2

METHOD

Stimulus materials—The stimulus materials were again drawn from the TIMIT Acoustic-Phonetic Continuous Speech Corpus (Fisher et al., 1986; Zue et al., 1990). For each dialect region, the six female talkers for whom listener categorization performance was best in Experiment 1 and the six male talkers for whom categorization performance was best in the previous study (Clopper & Pisoni, 2004b) were chosen for use in the mixed talkers experiment. The stimuli in the mixed talkers condition consisted of the two calibration sentences and one novel sentence from each of 72 talkers, 36 of whom were female and 36 male. For most talkers, the novel sentence was the same one used in the previous categorization experiments. In the few cases in which the previously chosen novel sentence was the same for a male and a female talker, two phonetically trained listeners (the first and second authors) replaced the novel sentence for one of the talkers with a different novel sentence judged to be similar in its potential to elicit regional variation in pronunciation.

Listeners—The listeners were 32 Indiana University undergraduates, all of whom received partial credit in an introductory psychology course for participating. There was a data recording error for one listener and another was bilingual, so the final analysis was performed on the data from 30 listeners, 10 males and 20 females. All of these listeners were monolingual native speakers of American English with no history of a hearing or speech disorder. The listeners represented a range of regional varieties, but Indiana was the most commonly represented state with 13 lifetime residents out of 30 participants. None of these listeners had participated in the earlier experiments.

Procedure—The procedure was identical to that used in Experiment 1, except that the talkers consisted of the mixed group of 36 males and 36 females. It should be noted that in all three blocks of the experiment, the gender of the talker on a particular trial was selected randomly and was not predictable from previous trials.

RESULTS AND DISCUSSION

Perceptual categorization

Categorization performance by naïve listeners on a mixed set of male and female talkers was consistent with previous research on male talkers only (Clopper & Pisoni, 2004b). As shown in the last column of Table 1, the listeners correctly categorized 33%, 28%, and 34% of the mixed talkers in the first, second, and third experimental blocks, respectively. Performance was slightly higher overall in this experiment than in the single-gender experiments, but this difference was not significant, as discussed below.

A repeated measures ANOVA on talker dialect and experimental block for the mixed talker condition revealed a significant main effect of talker dialect, $F(5, 174) = 20.19$, $p < .001$, a significant main effect of experimental block, $F(2, 348) = 10.66$, $p < .001$, and a significant Dialect × Block interaction, $F(10, 174) = 6.21$, $p < .001$. As in Experiment 1, posthoc Tukey tests revealed better overall categorization performance on New England and Southern talkers than the other four talker groups (all ps < .001). In addition, performance on South Midland talkers was significantly better than Western talkers ($p < .001$). None of the other pairwise comparisons were significant. Posthoc Bonferroni analyses on experimental block confirmed that categorization performance on the first sentence and the novel sentences was significantly better than performance on the second sentence (both ps < .001). Performance on the first sentence and novel sentences was not significantly different.

As shown in Figure 5, the Dialect × Block interaction can be explained by worse performance on New England talkers in the second block and better performance on the Southern talkers in the final block. These results are consistent with the findings reported by Clopper and Pisoni (2004b) for male talkers and the data reported in Experiment 1 for female talkers. Once again, performance was better for the New England and Southern talkers, and the effects of experimental block were most pronounced for those two talker groups. Unlike the results for the female talkers, however, we also uncovered an overall effect of experimental block in the mixed talker condition that matches the pattern reported for male talkers only by Clopper and Pisoni (2004b).

A repeated-measures ANOVA on listener gender (male or female) and experimental block (first, second, or novel sentences) also revealed a significant main effect of experimental block, $F(2, 27) = 10.52$, $p < .001$. However, neither the main effect of listener gender nor the Gender × Block interaction was significant. Again, listener gender did not affect performance on the categorization task.

Perceptual similarity—The stimulus-response confusion matrices from each of the three experimental blocks were submitted to the SCM analysis (Nosofsky, 1985) described above. The analysis again revealed significant differences in similarity for each of the three sentence conditions (first, second, or novel), so the three resulting similarity matrices were submitted to the hierarchical clustering scheme, ADDTREE (Corter, 1982). The results of the ADDTREE analysis for the mixed talker condition are shown graphically in Figure 6. As in Experiment 1, the vertical lengths of the branches connecting any two dialect nodes reflect the perceptual dissimilarity of those two dialects.

As in the male-only talker condition, the perceptual similarity analysis of Sentence 1 and the novel sentences revealed three dialect clusters: New England; South and South Midland; North, North Midland, and West. The clustering results obtained for Sentence 2 in the mixed talker condition were somewhat anomalous, with New England, North, and West forming one cluster, South and South Midland a second cluster, and North Midland connecting late in the tree to the southern cluster. Given the consistently poorer performance on the second sentence than the other two sentence blocks across all three gender conditions, this somewhat anomalous similarity space may simply be due to the large number of errors made in this part of the experiment.

Comparison to the single-gender conditions—A repeated measures ANOVA on talker dialect, experimental block, and talker gender (male, female, or mixed) was again used to quantitatively assess the effects of talker gender on categorization accuracy across Experiment 1, Experiment 2, and Clopper and Pisoni (2004b). The results again revealed a significant main effect of dialect, $F(5, 450) = 47.16$, $p < .001$, a significant main effect of block, $F(2, 900) = 13.19$, $p < .001$, and a significant Dialect × Block interaction, $F(10, 450) = 15.18$, $p < .001$. In support of our qualitative assessment of the lack of effect of talker gender on categorization performance, neither the main effect of gender nor any of the interactions involving gender were significant (see Table 1).

A comparison of the similarity spaces across the three gender conditions confirmed significant differences for each of the three sentence blocks as a result of talker gender. Similarity Choice Model analyses confirmed significant differences between the similarity parameters for the male talker, female talker, and mixed talker conditions that are not due to chance. For Sentence 1, all three gender conditions resulted in a similarity space that included the three major clusters: New England; South and South Midland; North, North Midland, and West. For Sentence 2, the single-gender conditions revealed a slightly different configuration of dialects with the three clusters: New England and North; South

and South Midland; North Midland and West; whereas the mixed gender condition revealed an anomalous two-cluster solution with New England, North, and West; South, South Midland, and North Midland. Finally, for the novel sentences the male and mixed gender talker conditions again provided a solution with the three major clusters: New England; South and South Midland; North, North Midland, and West; whereas the female talker condition provided a solution with New England and North; South and South Midland; and North Midland and West.

Taken together, these results suggest that the perceptual similarity of Northern talkers is fluid across gender and sentence conditions with respect to the New England dialect and the North Midland and Western dialects. In general, however, the earlier findings reported by Clopper and Pisoni (2004b) for three major perceptual dialects of American English were replicated in the current experiments using female only and mixed gender talker groups.

GENERAL DISCUSSION

In terms of categorization performance, we observed no effect of talker gender, and all three listener groups performed at roughly 31% correct on the six-alternative forced-choice task. These categorization results are similar to those reported by Van Bezooijen and her colleagues (Van Bezooijen & Gooskens, 1999; Van Bezooijen & Ytsma, 1999) who found roughly equivalent performance on male and female Dutch talkers in forced-choice categorization tasks. Thus, although the research on perceptual categorization of dialects has in the past relied heavily on male talkers, the data presented here suggest that replication of the studies with female talkers would provide comparable data sets. Despite the well-documented differences in speech production between males and females (e.g., Labov, 1990, 2001; Trudgill, 1974), naïve listeners are nevertheless able to categorize both males and females by dialect with similar levels of performance in a six-alternative forced-choice categorization task. That is, gender differences do not interfere with the accurate categorization of regional dialect.

Overall categorization performance in our task was significantly above chance but was somewhat lower than performance by the Dutch listeners (Van Bezooijen & Gooskens, 1999; Van Bezooijen & Ytsma, 1999) or the Welsh listeners (Williams et al., 1999). The use of read stimulus materials for our study may explain this difference in performance for our participants (who were approximately 31% correct overall) and the higher performance reported by Williams et al. (1999) and Van Bezooijen and Gooskens (1999) for their participants who based their responses on longer segments of conversational speech. As is well known in both the speech perception and sociolinguistic fields, the setting in which spoken language is recorded impacts production. Read speech materials are less likely to exhibit marked dialect forms than are conversational speech materials (Labov, 1994). However, the use of read materials permitted us to control for prosodic, syntactic, and lexical variation and to focus mainly on phonological variation. Thus, the perceptual similarity spaces obtained from the listeners reflect the phonological similarity of the dialects included in this study. Despite the controlled nature of the stimulus materials, the experiments still produced interesting and interpretable results about the role of gender in the categorization of talkers by regional dialect.

In examining the underlying perceptual similarity spaces of the dialects, we found significant differences based on talker gender. In particular, although the similarity spaces for all three listener groups were highly similar using Sentence 1, there were significant differences in the clusters for Sentence 2 and the novel sentences. As mentioned above, performance on Sentence 2 seemed to be somewhat anomalous, given that it was typically worse in terms of overall categorization accuracy than performance on the other sentences

and in terms of the unexpected similarity space that resulted for this sentence in the mixed gender condition of Experiment 2. Interpretation of the results of Sentence 2 is therefore difficult.

The gender differences found in the similarity spaces for the novel sentences are more interesting. In particular, for the female talker condition, we found three major clusters: New England and North; South and South Midland; and North Midland and West. In contrast, in the male and mixed gender talker conditions, the Northern talkers clustered with the North Midland and Western talkers to produce the three major clusters: New England; South and South Midland; and North, North Midland, and West.

Although the precise cause of this difference in similarity structure is unknown at this time, one possible explanation is that the Northern women were more advanced in the Northern Cities Chain Shift than the Northern men. Recall that the TIMIT recordings were made in the late 1980s and that Eckert (1989) reported gender differences in the advancement of the Northern Cities Chain Shift for high school–aged males and females. We might expect similar gender differences to be present between males and females in their 20s during the same time period. In particular, like Eckert's (1989) talkers, the female talkers used in the current study may be more advanced in the fronting of/æ/and/a/than the male talkers. These two variables are present in many of the novel sentences for the Northern talkers, so listeners could have used the pronunciation of these vowels as reliable perceptual cues to dialect affiliation.

In addition, Clopper and Pisoni (2003) have shown that perceptual similarity is related to the dimension of markedness in linguistic variables. In particular, the more linguistically unmarked regions such as the North Midland and the West tend to be found at one end of a perceptual continuum, whereas the more marked regions such as the South and New England tend to be found at the other end. The difference in the advancement of the Northern Cities Chain Shift between men and women might have caused the clustering of the Northern women with the more linguistically marked New England women in the novel sentences instead of with the less linguistically marked North Midland and Western women.

We also predicted that the Southern and South Midland women might be more similar to the women from the unmarked dialects than their male counterparts. Two variables that are related to Southern speech are/aɪ/monophthongization and/u/fronting in the words *like* and *suit*, respectively. Because women tend to avoid socially marked forms, we might expect them to produce less monophthongization of/aɪ/and a more backed/u/, making them more similar to Northern women. The clustering analyses did not reveal any evidence to support this hypothesis, however. In all of the solutions, the South/South Midland node was connected to the other major clusters relatively late, meaning that the Southern and South Midland talkers were perceptually distinct from the other talkers in all three talker gender conditions. These results may reflect a "covert prestige" associated with Southern American English varieties for both male and female talkers. Labov (1994) has described covert prestige as reflecting loyalty to local norms. In addition, Preston (1993) has repeatedly found that participants identify local varieties as being the most "pleasant" form of English spoken in the United States. Thus, the male and female Southern talkers may have exhibited the same degree of the Southern variants as a show of pride or loyalty to the South.

Previous analyses of the TIMIT corpus provided evidence for differences in production due to gender and dialect differences. Byrd (1994) explored the role of gender and regional dialect for all 630 talkers included in the TIMIT corpus and found significant effects of both variables on speaking rate and the production of a number of segmental properties such as flaps, glottal stops, reduced vowels, and palatalization. Clopper and Pisoni (2004b) reported

significant differences due to regional dialect for a number of segmental properties, including r-lessness and vowel diphthongization for the set of male talkers used in their categorization study.

These acoustic studies confirm that both gender and dialect differences in phonological production exist for the talkers included in the corpus. The results of the clustering analysis are, therefore, due in part to the accurate perception of gender differences in production within a regional variety. That is, naïve listeners know that male and female talkers from the same region might produce the same word or phoneme in a reliably different way and that a difference in production does not necessarily indicate a difference in regional variety. The "same" difference in production may, however, indicate a difference in regional variety when two talkers of the same gender are compared.

The gender and regional dialect of the listeners are also important aspects of the perceptual categorization task. The analyses summarized above revealed no significant gender differences in categorization accuracy in either of the two experiments. Due to the small number of male listeners, we were not able to compute separate perceptual similarity spaces for male and female listeners in this study. However, given the interesting perceptual similarity differences due to talker gender, we might expect to find similar differences due to listener gender. A more balanced group of listeners would allow for such an investigation that could explore both listener gender issues as well as cross-gender versus same-gender perception (e.g., male listeners and female talkers vs. male listeners and male talkers).

Listener differences due to residential history were also not explored in this experiment. Previous research has shown that listeners who have lived in a given region can more accurately identify talkers from that region than listeners who have not lived there (Clopper & Pisoni, 2004a). Perceptual similarity is also affected by residential history. Future research should examine the interaction of the listeners' gender and residential history on the perception of dialect variation. In addition, perception research using discrimination paradigms such as same/different tasks or similarity ratings tasks will be useful in teasing apart the roles of the talkers and the listeners in the perceptual similarity of regional dialects.

CONCLUSION

Although the variationist sociolinguistics literature describing gender differences in speech production is fairly extensive (e.g., Eckert, 1989; Labov, 1990, 2001; Trudgill, 1974), much less research has been carried out on the perception of gender differences. In the field of perceptual dialectology, Preston's (1993) map-drawing and correctness ratings tasks were not designed to explicitly differentiate male and female speech. Research on attitude judgments about language, on the other hand, has investigated the role of talker gender in the assignment of attitudes (Ryan & Giles, 1982). However, these studies typically focus on subjective judgments of femininity, power, and prestige and not on the explicit identification of the dialect of the speaker. The results of the two perceptual categorization experiments reported here provide new insights into our understanding of the role of gender in variation. In particular, we were able to explore the effects of talker gender on the accuracy of naïve listeners' categorization performance as well as the effects of gender on the underlying perceptual similarity of regional varieties of American English. Our findings demonstrate that although overall dialect categorization of male and female talkers was comparable, and unaffected by manipulations of the experimental context, the underlying similarity spaces for male and female talkers differed selectively in subtle ways based on dialect region. Naïve listeners are able to detect these dialect differences as revealed by the clustering analyses of the perceptual confusions.

Acknowledgments

This work was supported by NIH NIDCD R01 research grant DC00111 and NIH NIDCD T32 training grant DC00012 to Indiana University. The second author was also supported by an Indiana University Chancellor's Fellowship and an NSF Graduate Research Fellowship. The authors would like to thank Luis Hernandez for his technical advice and support and the audience at the 48th annual meeting of the International Linguistics Association for their comments.

References

Byrd D. Relations of sex and dialect to reduction. Speech Communication. 1994; 15:39–54.

Cheshire, J. Sex and gender in variationist research. In: Chambers, JK.; Trudgill, P.; Schilling-Estes, N., editors. Handbook of language variation and change. Malden, MA: Blackwell; 2002. p. 423-443.

Clopper, CG.; Pisoni, DB. Free classification of regional varieties of American English. Poster presented at the annual meeting of New Ways of Analyzing Variation in English (NWAVE); Philadelphia, PA. 2003 October.

Clopper CG, Pisoni DB. Homebodies and army brats: Some effects of early linguistic experience and residential history on dialect categorization. Language Variation and Change. 2004a; 16:31–48.

Clopper CG, Pisoni DB. Some acoustic cues for the perceptual categorization of American English regional dialects. Journal of Phonetics. 2004b; 32:111–140.

Corter JE. ADDTREE/P: A PASCAL program for fitting additive trees based on Sattath and Tversky's ADDTREE algorithm. Behavior Research Methods and Instrumentation. 1982; 14:353–354.

Eckert P. The whole woman: Sex and gender differences in variation. Language Variation and Change. 1989; 1:245–267.

Fisher, WM.; Doddington, GR.; Goudie-Marshall, KM. The DARPA speech recognition research database: Specifications and status. Proceedings of the DARPA Speech Recognition Workshop; 1986. p. 93-99.

Klatt, DH. Review of selected models of speech perception. In: Marslen-Wilson, W., editor. Lexical representation and process. Cambridge, MA: MIT Press; 1989. p. 169-226.

Kramarae, C. Gender: How she speaks. In: Ryan, EB.; Giles, H., editors. Attitudes towards language variation. London: Edward Arnold; 1982. p. 84-98.

Krapp, GP. The English language in America. New York: Frederick Ungar; 1925.

Labov W. The social motivation of a sound change. Word. 1963; 19:273–309.

Labov, W. The social stratification of English in New York City. Washington, DC: Center for Applied Linguistics; 1966.

Labov W. The intersection of sex and social class in the course of linguistic change. Language Variation and Change. 1990; 2:205–254.

Labov, W. Principles of linguistic change: Internal factors. Cambridge, MA: Blackwell; 1994.

Labov, W. The three dialects of English. In: Linn, MD., editor. Handbook of dialects and language variation. San Diego, CA: Academic Press; 1998. p. 39-81.

Labov, W. Principles of linguistic change: Social factors. Malden, MA: Blackwell; 2001.

Lass NJ, Hughes KR, Bowyer MD, Waters LT, Bourne VT. Speaker sex identification from voiced, whispered, and filtered isolated vowels. Journal of the Acoustical Society of America. 1976; 59:675–678. [PubMed: 1254794]

Milroy J, Milroy L. Mechanisms of change in urban dialects: The role of class, social network and gender. International Journal of Applied Linguistics. 1993; 3:57–77.

Mullennix JW, Pisoni DB. Stimulus variability and processing dependencies in speech perception. Perception and Psychophysics. 1990; 47:379–390. [PubMed: 2345691]

Namy LL, Nygaard LC, Sauerteig D. Gender differences in vocal accommodation: The role of perception. Journal of Language and Social Psychology. 2002; 21:422–432.

Nosofsky R. Overall similarity and the identification of separable-dimension stimuli: A choice-model analysis. Perception and Psychophysics. 1985; 38:415–432. [PubMed: 3831920]

Preston, D. Folk dialectology. In: Preston, D., editor. American dialect research. Philadelphia: John Benjamins; 1993. p. 333-378.

Purnell T, Idsardi W, Baugh J. Perceptual and phonetic experiments on American English dialect identification. Journal of Language and Social Psychology. 1999; 18:10–30.

Ryan, EB.; Giles, H., editors. Attitudes towards language variation. London: Edward Arnold; 1982.

Tice, R.; Carrell, T. Level16 (Version 2.0.3) [Computer software]. Lincoln: University of Nebraska; 1998.

Trudgill, P. The social differentiation of English in Norwich. Cambridge, UK: Cambridge University Press; 1974.

Van Bezooijen R, Gooskens C. Identification of language varieties: The contribution of different linguistic levels. Journal of Language and Social Psychology. 1999; 18:31–48.

Van Bezooijen R, Ytsma J. Accents of Dutch: Personality impression, divergence, and identifiability. Belgian Journal of Linguistics. 1999; 13:105–129.

Williams, A.; Garrett, P.; Coupland, N. Dialect recognition. In: Preston, DR., editor. Handbook of perceptual dialectology. Philadelphia: John Benjamins; 1999. p. 345-358.

Zue V, Seneff S, Glass J. Speech database development at MIT: TIMIT and beyond. Speech Communication. 1990; 9:351–356.

Biographies

Cynthia G. Clopper recevied a Ph.D. in linguistics and cognitive science from Indiana University in 2004. She received a B.A. in linguistics and Russian from Duke University in 1999 and an M.A. in linguistics from Indiana University in 2001. She is currently an NIH postdoctoral fellow, working under the direction of David Pisoni.

Brianna Conrey is a graduate student in psychology at Indiana University. She received a B.A. in linguistics from Rice University in 2002. She is currently an NSF graduate fellow.

David B. Pisoni is one of the leading figures in the field of speech perception and spoken language processing. He received a Ph.D. from the University of Michigan in 1971 and did postdoctoral work at MIT under the direction of Kenneth Stevens. He joined the faculty at Indiana University in 1971 and was promoted to professor of psychology in 1977. He is currently chancellors' professor of psychology and cognitive science. He also holds adjunct appointments in linguistics and otolaryngology, head and neck surgery at the Indiana University School of Medicine in Indianapolis. He has been program director of the NIH-sponsored training program in speech, hearing, and sensory communication at Indiana University since 1978.

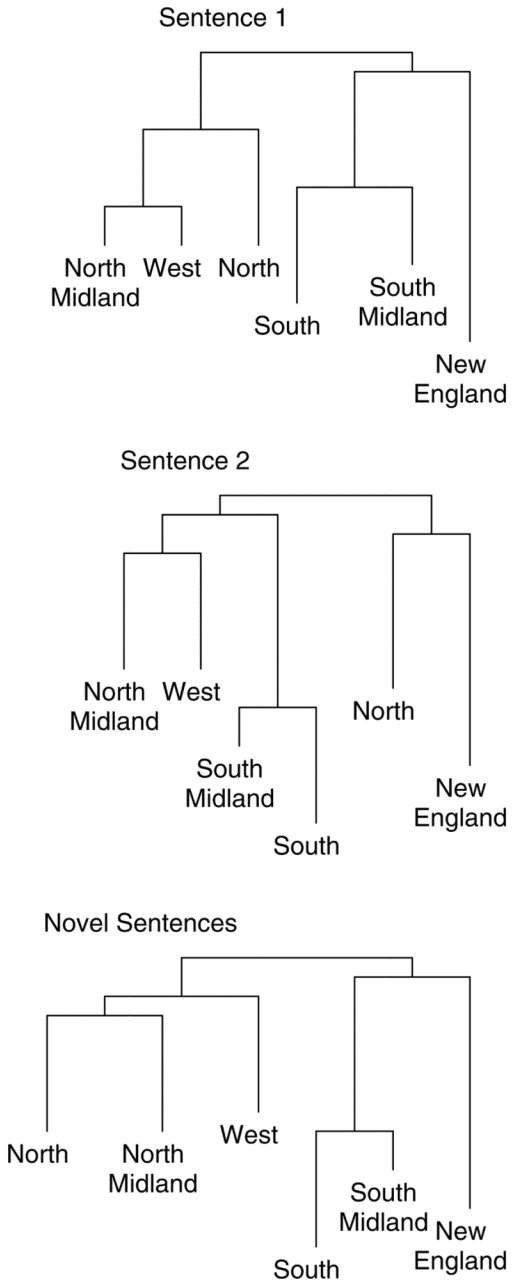

Figure 1. Clustering Analysis Results for the Male Talker Condition for Sentence 1, Sentence 2, and Novel Sentences

Source: Clopper and Pisoni, 2004b.

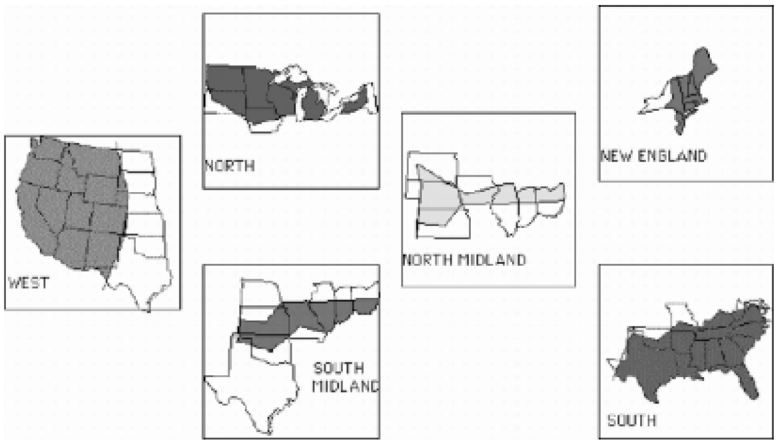

Figure 2. Response Alternatives in the Six-Alternative Categorization Task
Source: Clopper and Pisoni, 2004b.

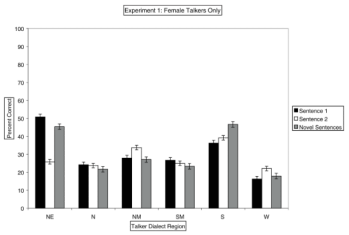

Figure 3.
Percentage Correct Categorization Performance in the Female Talker Condition for Each Talker Dialect Region in Each of the Three Experimental Blocks.

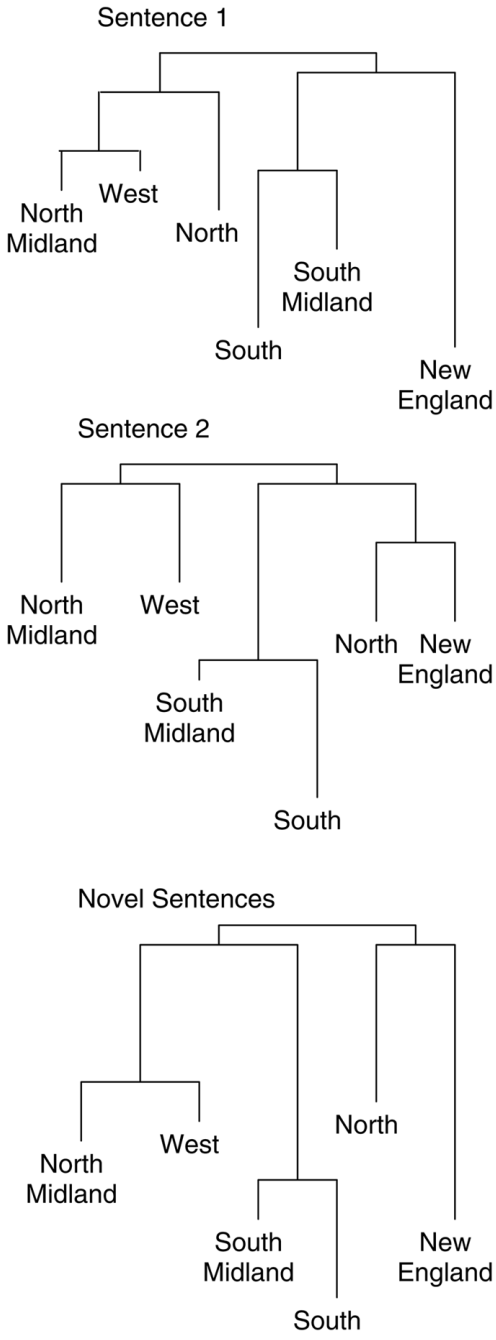

Figure 4.
Clustering Analysis Results for the Female Talker Condition for Sentence 1, Sentence 2, and Novel Sentences.

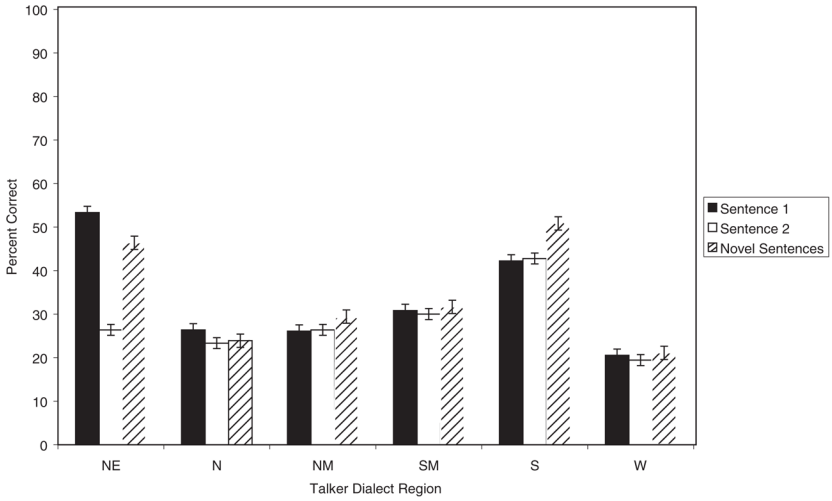

Figure 5.
Percentage Correct Categorization Performance in the Mixed Gender Talker Condition for Each Talker Dialect Region in Each of the Three Experimental Blocks.

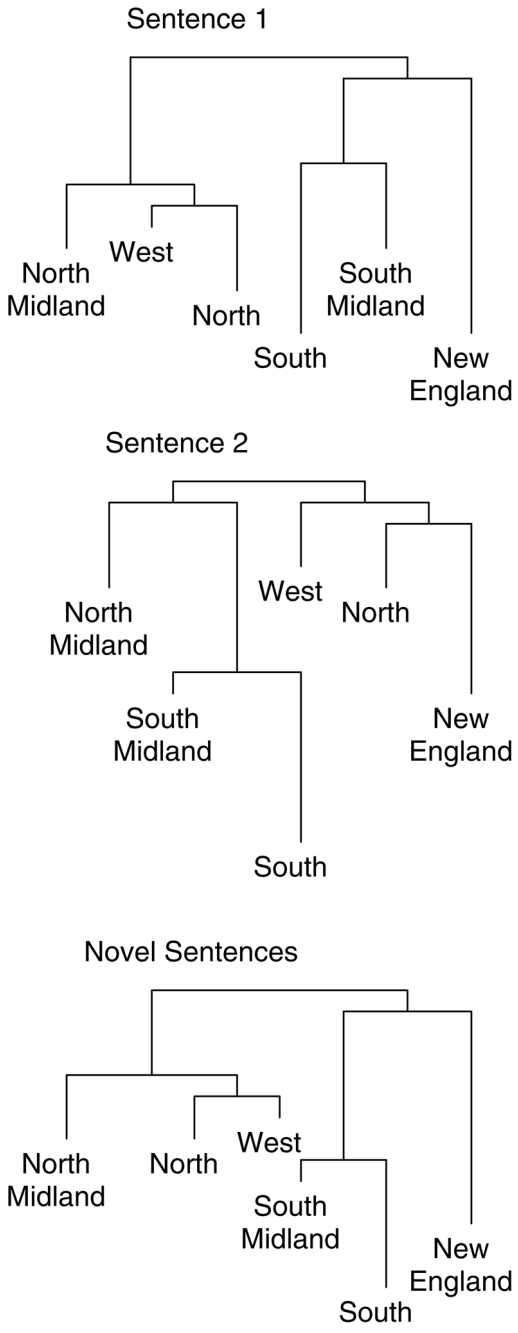

Figure 6.
Clustering Analysis Results for the Mixed Talker Condition for Sentence 1, Sentence 2, and Novel Sentences.

Table 1

Mean Percentage Correct Categorization Scores for Each of the Three Experimental Blocks for the Male Talker Condition (Clopper & Pisoni, 2004b), the Female Talker Condition (Experiment 1), and the Mixed Talker Condition (Experiment 2)

	% Male Talkers	% Female Talkers	% Mixed Male and Female Talkers
Sentence 1	33 (5)	31 (7)	33 (8)
Sentence 2	28 (5)	28 (6)	28 (7)
Novel sentences	33 (7)	31 (8)	34 (8)

Note: Standard deviations are shown in parentheses.