Published in final edited form as:

Lang Speech. 2004 ; 47(Pt 3): 207–239.

Effects of Talker Variability on Perceptual Learning of Dialects

Cynthia G. Clopper and David B. Pisoni

Indiana University, Bloomington

Abstract

Two groups of listeners learned to categorize a set of unfamiliar talkers by dialect region using sentences selected from the TIMIT speech corpus. One group learned to categorize a single talker from each of six American English dialect regions. A second group learned to categorize three talkers from each dialect region. Following training, both groups were asked to categorize new talkers using the same categorization task. While the single-talker group was more accurate during initial training and test phases when familiar talkers produced the sentences, the three-talker group performed better on the generalization task with unfamiliar talkers. This cross-over effect in dialect categorization suggests that while talker variation during initial perceptual learning leads to more difficult learning of specific exemplars, exposure to intertalker variability facilitates robust perceptual learning and promotes better categorization performance of unfamiliar talkers. The results suggest that listeners encode and use acoustic-phonetic variability in speech to reliably perceive the dialect of unfamiliar talkers.

Keywords

dialect categorization; indexical properties; perceptual learning; speech perception; talker variability

1 Introduction

1.1 Sources of variation in the speech signal

Recent studies suggest that speech perception is a talker-contingent process that exploits the close coupling observed between the indexical and linguistic attributes of speech. Abercrombie (1967) described indexical properties of the speech signal as those which provide an index to some attribute of the talker, such as membership in a group (e.g., gender or regional dialect), talker-specific idiosyncrasies that allow us to identify familiar talkers, and context-specific attributes of the talker, such as current emotional state. While some of the talker-specific information carried by the linguistic signal can be attributed to biological and physiological differences between humans (e.g., gross differences in fundamental frequency between adult males and females), many of these attributes are socially learned (e.g., phonological variation due to gender, ethnicity, region of origin, or socioeconomic status). Thus, both physiological and sociolinguistic sources of variation are included in Abercrombie's set of indexical properties of speech.

The indexical properties of speech can be described independently of those properties of the speech signal that contain linguistic information (Joos, 1948; Ladefoged & Broadbent, 1957). However, both sources of information appear to be perceived and encoded in

memory by naïve listeners (Pisoni, 1997). Much of the previous research examining the interaction of indexical and linguistic information in the speech signal has focused on talker-specific properties related to the identity of a given talker using a variety of behavioral tasks involving spoken word recognition.

Several recent studies have shown that indexical information can interfere with linguistic processing. In one set of studies, Mullennix, Pisoni, and Martin (1989) and Sommers, Kirk, and Pisoni (1997) found that word recognition performance in noise was better when the talker remained the same across a block of trials than when a set of different talkers was used in the same block of test trials. Their results suggest that talker-specific properties of the speech signal may provide useful information for recognizing novel spoken words in noise because the talker becomes predictable across trials and the amount of uncertainty about the source of the signal is reduced. Conversely, when the talker is unpredictable across trials, performance is worse because of the interaction between talker-specific and lexical information.

Similarly, talker variability has been shown to affect performance on selective attention tasks. In one study, Mullennix and Pisoni (1990) asked participants to listen to isolated words that varied in the initial consonant (/b/or/p/) and talker gender (male or female). The listeners were instructed to ignore the talker and respond only to the initial consonant. Like Mullennix et al. (1989), Mullennix and Pisoni (1990) found that listeners performed more poorly when the talker was different across trials than when a single talker was used. More recently, Green, Tomiak, and Kuhl (1997) replicated these effects in a speeded classification task in which they manipulated initial consonant, talker gender, and speaking rate. Green et al. (1997) also found interference due to talker variability on initial consonant identification with poorer performance when different talkers were used in a block of experimental trials than when just one talker was used. The results of these two studies suggest that lexical characteristics of the speech signal are not processed independently of the indexical properties of the talker's voice. Talker-specific information cannot be ignored and appears to be processed automatically and simultaneously with the phonological and lexical information in the signal.

Just as a single talker leads to more accurate speech processing, familiar talkers can also facilitate spoken word recognition. Nygaard, Sommers, and Pisoni (1994) explored the role of talker familiarity on speech intelligibility. They trained naïve listeners to identify 10 different talkers by first name. The listeners then completed a spoken word recognition task in noise to assess the effects of talker familiarity on speech intelligibility. Half of the listeners heard novel words spoken by the same talkers used in the earlier training phase and half of the listeners heard novel words spoken by unfamiliar talkers. The listeners who heard the familiar talkers performed better on this speech intelligibility task than the listeners who heard novel talkers, suggesting that experience with a specific talker's voice facilitates extraction of the linguistic message from degraded speech signals even when the test materials are novel utterances (see also Nygaard & Pisoni, 1998).

Research on the perception of sinewave replicas of speech has also demonstrated a close coupling between indexical and linguistic information in the speech signal, even under degraded listening conditions. Sinewave speech replicas are signals consisting of three sinewaves that follow the frequency peaks of the first three formants of speech (Remez, Rubin, Pisoni, & Carrell, 1981). Although listeners often report that these signals sound like new-age music or a series of tones, they can also hear and understand the signals as speech with a small amount of exposure. In several recent studies, Remez and his colleagues (Remez, Fellowes, & Rubin, 1997; Sheffert, Pisoni, Fellowes, & Remez, 2002) have reported that some talker-specific information is also retained in sinewave speech. Listeners

can identify familiar talkers by name from sinewave replicas of natural speech (Remez et al., 1997) and naïve listeners can be trained to identify talkers from sinewave speech and then transfer that knowledge to the original unprocessed speech signals (Sheffert et al., 2002). Thus, not only is indexical information in speech processed simultaneously with linguistic information, but some talker-specific information is also present in the phonetic properties of the speech signal that carry phonological information (i.e., the talker's vocal tract transfer function and formant frequency trajectories).

Taken together, the results of these studies examining talker-specific information in spoken language processing suggest that the normal process of speech perception not only involves extraction of the abstract phonetic information needed to specify the symbolic linguistic content of the message but also entails perception and encoding of the indexical properties of the talker's voice. Further evidence that naïve listeners perceive and encode indexical information can be found in studies exploring the explicit identification of talkers and their membership in social groups.

In one early study, Pollack, Pickett, and Sumby (1954) explored the identification of familiar talkers as a function of set size, duration of utterance, and voicing. They found that the listeners were fairly accurate in identifying their colleagues, but that performance decreased as the number of talkers increased. In addition, performance improved as the length of the utterance increased and the listeners were more accurate on voiced speech samples than on whispered utterances. In general, however, naïve listeners are able to identify talkers whom they know personally, even when very short speech samples are presented.

Listeners can also make reliable categorization judgments about unfamiliar talkers based on indexical information. For example, adult listeners are quite accurate in identifying the gender of unfamiliar talkers. In fact, Lass, Hughes, Bowyer, Waters, and Bourne (1976) reported that listeners were 96% accurate in categorizing unfamiliar talkers by gender based on isolated vowels. In addition, performance was still above chance when the vowels were presented with degradation due to low-pass filtering (91% correct) and in whispered speech (73%). The results of this study suggest that gender-specific information is conveyed through both prosodic (i.e., fundamental frequency) and segmental (i.e., vowel formant frequency) information and that naïve listeners are able to use both sources of information in an explicit identification task.

Mullennix and Pisoni (1990) also manipulated talker gender in their speeded classification task described above. In particular, a second group of listeners was asked to ignore the segmental linguistic content of the stimuli and to categorize each talker by gender as either male or female. As in the consonant classification task described above, Mullennix and Pisoni (1990) found interference of the lexical variation in the gender classification task. However, the interference from the lexical information in the gender identification task was less pronounced than the interference due to talker variability in the consonant identification task. Green et al. (1997) also replicated this finding. Taken together, these results suggest that in addition to automatically processing indexical information in a selective attention task, listeners also have rapid access to these sources of information when the task requires it. Gender classification in particular seems to be a relatively easy task for naïve listeners.

Categorization of unfamiliar talkers based on other social dimensions, however, can be more difficult for naïve listeners. Regional dialect variation in particular seems to be more difficult for naïve listeners to identify than gender. In one recent study, Clopper and Pisoni (2004b) carried out a perceptual categorization experiment to assess naïve listeners' ability to explicitly use dialect variation in speech. Using a six-alternative forced-choice perceptual

categorization task, Indiana University undergraduates were asked to identify which region of the United States unfamiliar talkers were from based on short, meaningful English sentences. We found that naïve listeners could perform this task above chance, but they were only about 31% accurate in making this kind of perceptual categorization judgment without any prior training or feedback.

Similar results have been reported for dialect categorization studies in Wales by Williams, Garrett, and Coupland (1999) and Great Britain and the Netherlands by Van Bezooijen and Gooskens (1999). Williams et al. (1999) asked adolescent male listeners to categorize other adolescent male talkers by dialect in an eight-alternative forced-choice task. The boys performed the task with about 30% accuracy and were only able to correctly identify talkers from their own region about 45% of the time. Van Bezooijen and Gooskens (1999) reported higher levels of performance in their investigation of adult dialect categorization performance. Dutch listeners could accurately identify the province of origin of 35% of the talkers in the study in the Netherlands. British listeners, however, were 52% accurate in identifying the area of origin of the talkers from Great Britain. Van Bezooijen and Gooskens (1999) also reported that naïve listeners could still perform the task when presented with only segmental or only prosodic information, again suggesting close links between indexical information in speech such as dialect and phonetic properties of the symbolic linguistic information about the intended message.

1.2 Perceptual learning and stimulus variability

Just as experience with an individual talker leads to improved perceptual processing of that talker's speech, prior experience and familiarity with a given dialect should also lead to improved categorization of that dialect. Thus, explicit exposure to dialect variation should provide naïve listeners with the opportunity to robustly encode and represent the detailed properties of dialect-specific variation in memory and improve categorization performance on a set of novel utterances produced by unfamiliar talkers.

In their seminal paper on perceptual categorization, Posner and Keele (1968) found that experience with highly variable stimulus materials during perceptual learning of abstract dot patterns led to better performance on a transfer task involving novel stimuli than exposure to less variable stimuli during perceptual learning. Their dot pattern results were in direct contrast to traditional views of perceptual learning in which it was assumed that learning would proceed more quickly for invariant as opposed to highly variable categories.

Posner and Keele's (1968) basic findings on variability in perceptual learning have been replicated in a number of different fields, including the domain of speech perception and production. Logan, Lively, and Pisoni (1991) found similar effects of stimulus variability in their study of perceptual learning of English /r/ and /l/ by Japanese listeners. They showed that variability in talker and phonetic context in the stimulus materials during initial training led to better discrimination performance at testing. In a later study, Bradlow, Pisoni, Akahane-Yamada, and Tohkura (1997) found that high-variability training materials in perceptual learning not only affected speech perception, but also improved speech production skills. In particular, the productions of /r/ and /l/ by the Japanese speakers were more accurately identified by English-speaking listeners after the high stimulus variability perceptual training sessions than before. In both studies, the Japanese listeners were able to form robust representations of the phonemes /r/ and /l/, while ignoring the context- and talker-specific information in each token that did not provide useful information about the identity of the target phone.

More recently, Bradlow and Bent (2003) reported a study in which they trained native English-speaking listeners in a speech intelligibility task using Chinese-accented speech.

Listeners were trained over the course of two sessions to transcribe sentence-length utterances in noise. Listeners who were exposed to multiple Chinese-accented talkers during the perceptual learning phase performed better on a subsequent speech intelligibility task with an unfamiliar Chinese-accented talker than listeners who were exposed to only a single talker during training. Bradlow and Bent's (2003) results suggest that when the stimulus materials were highly variable, the listeners were able to extract talker-independent accent-specific information during perceptual learning. Given these earlier findings on perceptual learning, a similar pattern of results would be anticipated in a dialect-learning task such as the one conducted in the present study.

To investigate the effects of perceptual learning on dialect categorization, two groups of listeners were recruited for the current study. The first group listened to only one talker from each of six dialect regions in the United States during training. A second group listened to three different talkers from each of the six dialect regions during training. Both groups received feedback after each trial. A between-subjects design was used to assess the effects of talker variability on dialect categorization performance with novel sentences produced by unfamiliar talkers. We predicted that although the three-talker group might initially perform more poorly than the one-talker group in the training phases due to greater stimulus variability and uncertainty in the stimulus materials, the three-talker group would actually perform better on the more difficult generalization task using unfamiliar talkers than the one-talker group. The three-talker group was exposed to more intertalker variability during perceptual learning and would therefore have a better opportunity to extract dialect-specific properties of the speech signal and form more robust and phonetically detailed representations of the six regional dialects. For the one-talker group, however, dialect-specific and talker-specific properties in the speech signal were perfectly correlated. We therefore predicted that these listeners would perform more poorly on the final generalization task with novel sentences produced by unfamiliar talkers.

Our predictions contrast with two other potential outcomes. First, traditional accounts of perceptual learning would predict that the one-talker group would perform better on the generalization task because they were provided with a single good exemplar from each region and learning is traditionally assumed to be easier when less variability is present (Reber, 1985). Second, performance in the generalization task could be equivalent for the two groups. In some current prototype models, categorization of novel stimuli is achieved through comparison to category prototypes stored in memory (Murphy, 2002). If listeners were trained on stimulus materials approximating the prototype from each category in the one-talker condition, generalization performance should be equivalent to performance by the three-talker group who had to construct prototypes based on the speech of three different talkers from each group.

2 Methods

2.1 Talkers

Sixty-six 20- to 29-year old white male talkers were selected from the TIMIT Acoustic-Phonetic Continuous Speech Corpus (Fisher, Doddington, & Goudie-Marshall, 1986). The TIMIT corpus contains recordings of 630 different talkers reading 10 isolated sentences each. Documentation accompanying the corpus provided information about each talker including his or her age, gender, ethnicity, and region of origin. The eight regional labels used in constructing the TIMIT corpus were: New England, New York City, North, North Midland, South Midland, South, West, and Army Brat. The regional labels were assigned by the TIMIT authors based on where the talkers grew up and more specific information about how long they had lived in that region or where they lived at the time of recording was not provided (Fisher et al., 1986). For the current study, the talkers represented six regional

varieties of American English: New England, North, North Midland, South Midland, South, and West. Eleven talkers were selected from each dialect region.

For the one-talker listener group, a single talker was selected from each dialect region to serve as the training talker for that region. The training talkers were selected by the first author, a trained linguist, after repeated listening to all 10 sentences spoken by each talker. The training talker was impressionistically the “best representative” of his dialect region. The remaining 10 talkers from each of the six dialects were used in the final generalization task for the one-talker group.

For the three-talker listener group, three talkers were selected from each dialect region to serve as the training talkers. These talkers were selected using the earlier categorization data collected by Clopper and Pisoni (2004b). We selected talkers that represented a high degree of variability within each dialect. Thus, the best- and worst-categorized talkers were selected as well as a talker in the middle for each dialect region based on the perceptual categorization of the same talkers. The remaining eight talkers from each dialect were used in the final generalization test for the three-talker group.¹

2.2 Stimulus materials

All of the talkers recorded for the TIMIT corpus read two calibration sentences that were intended to elicit dialect variation through the explicit inclusion of certain lexical items known to reveal reliable phonological differences between regional varieties of American English (Fisher et al., 1986). These two calibration sentences are shown in (1). Each talker in the TIMIT corpus also recorded eight additional novel sentences. The complete list of novel sentences used in this experiment is listed in the Appendix. Of the eight novel sentences recorded by each talker on the TIMIT corpus, five were recorded by six other talkers and three were recorded only by a single talker. Talkers and sentences were selected from the TIMIT corpus for use in this experiment such that none of the novel sentences were ever repeated in the experimental session.

(1) Calibration sentences:

- a. She had your dark suit in greasy wash water all year.
- b. Don't ask me to carry an oily rag like that.

All 10 sentences from each of the training talkers were used in the training and test phases of the experiment. The calibration sentences were used in the first two training phases and the novel sentences were used in the final training and test phases. No sentence was ever repeated during the course of training and testing, with the exception of the two calibration sentences.

For the unfamiliar talkers used in the generalization task, a single sentence was selected for each talker. These novel sentences were chosen so that no sentence was ever repeated within the generalization phase. In addition, none of the sentences that were used in the training or test phases were used in the generalization phase. All of the sentence materials were reproduced into individual digital sound files for playback to the listeners and were leveled to 55 dB using Level16 (Tice & Carrell, 1998). The specific talkers and sentences used in the current study are shown in the Appendix.

¹Different methods were used to select the training talkers for the two groups of listeners due to the relative timing of data collection. Data collection for the one-talker group was begun prior to the collection of the data reported by Clopper and Pisoni (2004b). Data collection for the three-talker group, however, did not start until after the Clopper and Pisoni (2004b) data had been collected.

As noted above, the calibration sentences were designed to include phonological and lexical information that would distinguish regional varieties of American English. Clopper and Pisoni (2004b) carried out an acoustic analysis of these two calibration sentences for each of the 66 talkers used in the current study. We examined 11 acoustic-phonetic properties in the two calibration sentences that were predicted to differentiate the six regional varieties and found significant differences for six of the measures. In particular, New England talkers were significantly more r-less than the South Midland and Western talkers in the word *dark*. Southern talkers had significantly more voicing in the fricative in *greasy* than New England talkers and the fricative in *greasy* was also significantly longer for Southern talkers than for Northern talkers. South Midland, Southern, and Western talkers had a significantly fronted /u/ in *suit* compared to New England talkers and the South Midland talkers also showed more /u/ fronting than the Northern talkers. Northern and Southern talkers also differed in the direction of /ow/ diphthongization in *don't*. Finally, Northern talkers had a significantly fronted /æ/ in *rag* compared to the New England talkers. North Midland talkers were the least linguistically marked among the six regional varieties and were not significantly different from the other dialect groups on any of the 11 acoustic-phonetic measures that we examined.

The results of a multiple regression analysis conducted by Clopper and Pisoni (2004b) revealed that the listeners in our six-alternative forced-choice dialect categorization study attended to four of these significant acoustic-phonetic variables in making their categorization responses: r-fulness, /ow/ diphthongization, /æ/ backness, and /u/ backness. Table 1 shows a comparison for each of these four acoustic-phonetic properties across all of the talkers as well as for each set of training talkers. In particular, the first column summarizes the significant differences across the different talker dialect groups described above. The remaining columns show how the training talkers compare for the one-talker group and the three-talker group. Separate comparisons were made for the best-, middle-, and worst-categorized talkers in the three-talker training condition.

As shown in Table 1, the talkers used in the one-talker condition and the best-categorized talkers used in the three-talker condition reflect the significant acoustic-phonetic differences between the six dialects for these four measures. The middle-categorized talkers in the three-talker group pattern similarly, but the /u/ fronting variable is less reliable in differentiating this set of talkers. Finally, the worst-categorized talkers show incorrect relationships for three of the four acoustic-phonetic variables. Thus, the training talkers selected for the one-talker group by the first author are both impressionistically and acoustically representative of their respective dialect regions. In addition, the best-, middle-, and worst-categorized talkers selected for the three-talker condition are also good, medium, and poor representatives of their dialect regions based on the acoustic-phonetic measures analyzed by Clopper and Pisoni (2004b).

2.3 Listeners

Fifty-nine listeners between the ages of 18 and 25 were recruited from the Indiana University community for this experiment. They were all monolingual native speakers of American English with native English-speaking parents who reported no history of hearing or speech disorders at the time of testing. Listeners either received partial course credit for an introductory psychology course or were paid \$6 for their participation. Thirty of the 59 listeners were assigned to the one-talker group. The remaining 29 listeners were assigned to the three-talker group.

The dialect of the listeners was not explicitly controlled. However, residential history information was obtained from each listener and both groups included a mix of listeners who had lived in just a single dialect region and listeners who had lived in multiple dialect

regions. The ratio of the former type to the latter was approximately two-to-one for both the one-talker group and the three-talker group. While the residential history of the listener has been shown to affect dialect categorization performance (Clopper & Pisoni, 2004a), the similarity of the two listener groups with respect to this variable should not result in significant group differences in categorization performance.

2.4 Procedure

The experimental procedures used for both groups of listeners in this study were identical and involved three phases: a training phase, a test phase, and a generalization phase. The initial training phase consisted of three blocks of trials with feedback provided after each trial. The test and generalization phases each consisted of a single block of trials without feedback. A schematic of the testing procedure is displayed in Table 2.

In the first two training blocks, the listeners heard the calibration sentences spoken by each of the training talkers. The two calibration sentences were used in the first and second blocks, respectively. In each training block, the one-talker group heard each talker reading the sentence 10 times each, while the three-talker group heard each talker reading the sentence five times each. In the third training block, both groups of listeners heard all of their training talkers reading four different sentences. After each trial in the training blocks, the correct response category was outlined in red immediately after the listener entered his or her choice (see Fig. 1). The presentation order of sentences and talkers was randomized in each block for each listener.

To ensure that the listeners had actually learned where the training talkers were from, the test phase of the experiment consisted of a set of trials that was identical to the last training block except that no feedback was given. To assess generalization performance after perceptual learning, the listeners heard a set of novel sentences spoken by unfamiliar talkers during the final block of the experiment. The generalization phase provided a measure of what the listeners had learned in the training phase about dialect variation that was independent of the specific talkers or the specific utterances they were exposed to during the training phase. As in the final test block, no feedback was given during the generalization task.

The listeners were seated at personal computers equipped with KeyTec Inc. pressure sensitive activation touchscreens (KTMT1315 ProE). The six response alternatives were represented on the screen as partial maps of the United States and were labeled with the name of the geographic region. Figure 1 shows the six dialect regions as they were displayed on the computer screen. The icons were roughly 5cm × 5cm in dimension and adequate space was left between the icons to minimize response errors. On each trial, the listeners heard a single talker reading one sentence over Beyerdynamic DT100 headphones at 70 dB SPL. The participants were asked to listen to each sentence carefully and then to select the dialect region that they thought the talker was from by pressing the icon on the computer touchscreen. The procedure was self-paced; each trial was initiated by the individual participant by clicking on a “Next Trial” button on the computer screen after the feedback was presented.

3 Results

A summary of the mean percent correct responses for both groups of listeners for each phase of the experiment is shown in Figure 2. The left panel of Figure 2 shows the performance for the two groups of listeners in the training phase. The right panel of Figure 2 shows their performance during the test and generalization phases. A repeated measures ANOVA with listener group (one-talker or three-talker) as the between-subjects factor and block (Train IA,

Train IB, Train II, Test, or Generalization) and talker dialect (New England, North, North Midland, South Midland, South, or West) as within-subject factors revealed a significant main effect of block, $F(4, 1769) = 107.1, p < .001$. Post hoc Tukey tests collapsed across listener group and talker dialect revealed that performance on each of the blocks in Train I (in which the sentence remained constant across all talkers) was significantly better than performance on any of the other three blocks (all $p < .05$). Performance on Train II and Test (in which novel sentences were read by the training talkers) was significantly better overall than performance on the final generalization phase in which unfamiliar talkers and novel sentences were used (both $p < .001$). Performance did not differ significantly between Train II and Test.

The repeated measures ANOVA also revealed a significant main effect of talker dialect, $F(5, 1769) = 55.9, p < .001$. Post hoc Tukey tests collapsed across listener group and experimental block revealed that performance on the New England talkers was better overall than performance on any of the other dialect regions (all $p < .001$). In addition, performance on the Northern, North Midland, and South Midland talkers was better than performance on the Western talkers (all $p < .005$) and performance on the North Midland talkers was significantly better than performance on the Southern talkers ($p = .012$). A summary of the mean percent correct responses for each experimental block for each talker dialect is shown in Table 3 for the one-talker group and in Table 4 for the three-talker group. The main effect of listener group was also significant, $F(1, 1769) = 96.5, p < .001$. Collapsed across all five experimental blocks, the one-talker group performed better overall than the three-talker group.

Each of the two-way interactions as well as the three-way interaction were significant in the repeated measures ANOVA analysis. The locus of the interaction between talker dialect and listener group, $F(5, 1769) = 4.2, p = .001$, can be seen by comparing the mean percent correct scores for each talker dialect across the two listener groups (see Tables 3 and 4). While the one-talker group performed better overall than the three-talker group, the difference in performance between the two groups varied from 13% for the Southern talkers to 29% for the North Midland talkers.

The significant listener group by experimental block interaction, $F(4, 1769) = 59.4, p < .001$, is shown in Figure 2. Because performance did not differ for the two calibration sentences in Train I, these data have been combined in Figure 2. An inspection of Figure 2 suggests two loci for the listener group by experimental block interaction. First, it appears that the one-talker group (striped bars) suffered a greater decline in performance than the three-talker group (dotted bars) between Train I (in which the same sentence was heard on every trial) and Train II (in which a different, novel sentence was heard on each trial). In order to quantitatively assess this aspect of the block by listener group interaction, difference scores were computed for each listener in each group from Train I to Train II. A histogram in Figure 3 shows the distribution of the difference scores for the participants in each listener group. This difference was statistically significant by a *t*-test, $t(57) = 5.0, p < .001$. As shown in Figure 3, the one-talker group displayed a greater decline in performance from Train I to Train II than the three-talker group. The pattern of results suggests that variation in the linguistic content of the utterance had a much greater effect on the categorization performance of the one-talker group than the three-talker group.

The second locus of the block by listener group interaction is the cross-over effect observed from Test to Generalization, shown in the right-hand panel of Figure 2. Although the one-talker group performed better than the three-talker group throughout the training and test blocks, the one-talker group performed more poorly than the three-talker group in Generalization, $t(57) = 2.2, p < .03$. To assess the strength of this cross-over effect,

difference scores for each listener in each listener group were computed from Test to Generalization. Because the number of trials for the two listener groups in the final generalization block differed, we computed two difference scores: one from Test to Generalization using the full data set and one from Test to Generalization using only the subset of shared trials across both groups. A histogram showing the distribution of difference scores for the two groups in the Full and Subset conditions is shown in Figure 4. A series of *t*-tests on the difference scores confirmed a greater decline in performance from Test to Generalization for the one-talker group than the three-talker group for both the full set of stimuli, $t(57) = 6.6$, $p < .001$, as well as the subset of stimuli that were shared by both groups, $t(57) = 5.9$, $p < .001$. Thus, the cross-over effect from Test to Generalization for the two listener groups is reliable and statistically significant.

Difference scores were also computed to assess the significant experimental block by talker dialect interaction, $F(20, 1769) = 5.4$, $p < .001$. In particular, the difference scores in performance for each listener on each talker dialect from Train I to Train II and from Test to Generalization were submitted to one-way ANOVAs with talker dialect as the factor to examine the effect of experimental block on performance on the different dialect regions. The ANOVA on the difference scores from Train I to Train II was not significant. However, the difference scores between Test and Generalization did reveal a significant main effect of dialect, $F(5, 353) = 7.3$, $p < .001$. Post hoc Tukey tests showed a significantly greater decline in performance from Test to Generalization for the New England, Northern, and North Midland talkers than for the Southern talkers (all $p < .001$). This result is likely due to the overall poorer performance on Southern talkers reported above.

Finally, post hoc analyses of the significant three-way interaction between listener group, talker dialect, and experimental block, $F(20, 1769) = 4.5$, $p < .001$, also confirmed significant differences due to listener group. To assess this three-way interaction, a one-way ANOVA on the difference scores from Test to Generalization for each dialect was computed for each talker condition. Figure 5 shows the difference scores for each of the six dialect regions from Test to Generalization for the two listener groups. The ANOVA for the one-talker group was significant, $F(5, 179) = 13.2$, $p < .001$. Post hoc Tukey tests revealed significantly greater difference scores for New England, North, and North Midland than for the South Midland (all $p < .05$) and a significantly smaller difference score for the South than for all of the other dialect regions (all $p < .05$). These results parallel the results of the talker dialect by experimental block interaction reported above. The critical finding from this analysis is that the ANOVA for the three-talker group was not significant. That is, there were no effects of talker dialect on the difference scores from Test to Generalization for the three-talker group.

4 Discussion

The effects of talker variability on perceptual learning of dialect were robust, as revealed by the significant cross-over effect in categorization performance for unfamiliar talkers during the generalization phase. The listeners in the one-talker group performed better than the listeners in the three-talker group in all phases of the experiment except the final generalization test. We will first discuss categorization performance during training and then consider categorization performance in the generalization test. The remainder of the Discussion will focus on the effects of the training on generalization performance.

The better categorization performance observed in the training sessions by the one-talker group was expected and reflects several experimental factors that have been shown to affect perceptual learning. First, the listeners in the one-talker group were only exposed to six different voices whereas the listeners in the three-talker group were exposed to 18 different voices. The difference in performance in training may simply reflect a generalized set size

effect which is frequently observed in studies of attention and memory. For example, Schneider and Shiffrin (1977) showed that performance on target detection tasks decreased as the number of possible targets increased. We would therefore expect to find a similar set size effect in perceptual learning; as the number of talkers to be learned increases, categorization performance decreases accordingly.

Second, in the one-talker condition, there was a consistent one-to-one mapping between talkers and dialects, whereas in the three-talker condition, a many-to-one mapping was used. White (1977) showed that identification of letters and numbers (e.g., responses such as “1,” “2,” “A,” or “B”) in a visual search paradigm was faster than categorization of the same stimuli as either “letter” or “number.” The difference in response latency between the two types of tasks suggests that identification of individual stimuli might be easier than categorization of the same stimuli (White, 1977). It is possible that the one-talker group who only had to learn one-to-one identification mappings was better than the three-talker group who had to learn three-to-one categorization mappings because identification is easier than categorization (Estes, 1994; Murphy, 2002).

Performance by both groups of listeners in the generalization phase was close to the 31% accuracy found in our earlier dialect categorization study (Clopper & Pisoni, 2004b) without any perceptual learning or feedback ($M = 30\%$, $SD = 7\%$ for the one-talker group, $M = 35\%$, $SD = 9\%$ for the three-talker group). It appears that the one-talker training had little, if any, direct effect on categorization performance, whereas the three-talker training did improve performance above the levels reported by Clopper and Pisoni (2004b) for the listeners who were not given any training or feedback.

As in Clopper and Pisoni (2004b), the present results reveal effects of talker dialect on categorization performance. In particular, performance on New England talkers was consistently better than performance on talkers from the other dialect regions. This result is likely due to the robust r-lessness of all of the training talkers (see Table 1). Performance was poorer on the Southern talkers than would be expected based on the perceptual and cultural salience of the American South (see Clopper & Pisoni, 2004b). However, it seems that the Southern talkers were only difficult to categorize for the listeners in the one-talker group during Train II and Test (see Table 3). As discussed in more detail below, the one-talker group may have developed less robust representations of the different dialects because they could rely on talker-specific properties of the speech signal to categorize the six training talkers. This talker-specific strategy may have led to greater confusion between the Southern and South Midland talkers during Train II and Test, when the linguistic content of the utterances changed from trial to trial compared to Train I when the sentences were invariant across trials.

4.1 Effects of training on generalization performance

The difference in generalization performance between the two listener groups can be attributed to the differences in the nature of the training materials used during the experiment and the perceptual strategies developed by each group of listeners to solve the categorization task. In particular, the three-talker group was exposed to more stimulus variability in the training materials than the one-talker group (Posner & Keele, 1968). Like the earlier Japanese /t/ ~ /l/ studies by Bradlow et al. (1997) and Logan et al. (1991) and the recent accent intelligibility study by Bradlow and Bent (2003), the listeners who were exposed to greater stimulus variability during initial perceptual learning performed better than the listeners who were exposed to less variability. The three-talker listener group had to learn to ignore between-talker variability and focus instead on the criterial properties that were common across the three talkers in each dialect. That is, they had to learn dialect-specific properties of the speech signal, instead of talker-specific properties.

The difference in variability in the stimulus materials may also have led to the development of different categorization strategies in the two listener groups. Listeners exposed to only a single talker from each dialect region could simply encode and represent all of the indexical characteristics of each individual talker during training and use any available indexical properties of the individual talkers' voices (e.g., speaking rate, pitch, or voice quality) to make the necessary one-to-one mapping between talker and dialect region. That is, the listeners in the one-talker group could use a shallow talker-specific encoding strategy based on surface properties in the training phases. In contrast, the listeners exposed to multiple talkers from each dialect region needed to first encode the indexical properties of each talker and then construct a more general phonological representation of each dialect region. Thus, the listeners in the three-talker group needed to discriminate and dissociate the dialect-specific information in each sentence from the talker-specific information in order to be able to correctly categorize the talkers.

The development of different dialect encoding strategies by the two listener groups is revealed by the post hoc analysis of the three-way interaction between talker dialect, experimental block, and listener group. In particular, the decline in performance from Test to Generalization for the one-talker group was affected by dialect region which suggests that the listeners in the one-talker group did not develop robust representations of the different dialects that were equally generalizable to novel sentences spoken by unfamiliar talkers. The decline in performance from Test to Generalization for the three-talker group, however, was not affected by talker dialect. Thus, the listeners in the three-talker group were able to form more robust representations of each of the six varieties that they could use to categorize unfamiliar talkers. The strategy of forming talker-independent, dialect-specific representations became more useful in the final generalization test because it is more robust to variability and less susceptible to talker-specific idiosyncrasies.

Additional evidence for the development of different perceptual learning strategies in the two listener groups can be found in the data from the training blocks. In the first two phases of the experiment, we found that performance of the one-talker group was better when the sentence remained invariant across all of the trials (Train I) than when the sentence changed from trial to trial (Train II and Test). The difference scores shown in Figure 3 reveal a general trend for the one-talker group to show a decline in performance from Train I to Train II ($M = 16\%$, $SD = 11\%$ for the one-talker group). This difference in performance across the two training blocks suggests that the one-talker listeners focused on specific properties of the sentence when the same sentence was used repeatedly within an experimental block. That is, the listeners in the one-talker group rapidly developed task-specific categorization strategies and selectively attended to certain keywords or segments in making their dialect categorization judgments. However, when the sentences varied from trial to trial, the listeners in the one-talker group did not know in advance which particular stimulus properties to focus on. Under these variable presentation conditions, they could not base their judgments on a set of a priori characteristic features, because they did not know which linguistic elements would be present in the stimulus pattern from trial to trial. Instead, they had to rely on more global representations of dialect variation in making their responses.

The finding that variation in the linguistic content of an utterance affects recognition of talker-specific characteristics of an utterance is not surprising. In earlier spoken word recognition studies, it has been shown that talker variability reduces word recognition performance in noise (Mullennix et al., 1989; Sommers et al., 1997). The integral nature of linguistic and indexical information in the speech signal has been shown in selective attention tasks (Mullennix & Pisoni, 1990) as well as talker identification and word recognition studies using sinewave speech (Remez et al., 1997; Sheffert et al., 2002). The

interference effects of linguistic content on dialect categorization are consistent with these findings. In particular, the talker-specific encoding strategy adopted by the one-talker group may have initially focused on token-specific properties. Recall that the first training blocks included multiple repetitions of the same sentences with only six different stimulus items. The one-talker listener group could therefore perform the task by relying on token-specific information. This token-specific categorization strategy was effective for the two calibration sentences used in Train I, but proved less effective when the linguistic content varied randomly in Train II and led in turn to a significant decrease in performance.

The effect of linguistic content on the performance of the three-talker group was minimal and their performance did not differ significantly across the training and test phases ($M = 4\%$, $SD = 7\%$ for the three-talker group). As shown in Figure 3, more than two-thirds of the listeners in the three-talker group either improved from Train I to Train II or declined in performance by less than 10%. This result provides further support for the claim that the three-talker group developed a more robust coding strategy that permitted them to more accurately categorize not only novel utterances from unfamiliar talkers, but also novel utterances produced by familiar talkers.

We have argued that the cross-over effect observed from training to generalization is due to differences in perceptual learning strategies in the two listener groups due to differences in stimulus variability in the training materials. An alternative interpretation of the cross-over effect is that the listeners simply based their categorization judgments of the unfamiliar talkers on the similarity of these talkers to the training talkers they were exposed to during the initial phases of the experimental procedure. Given the range of within-dialect variability displayed by the training talkers for the three-talker group, one might expect that generalization would be easier for this group because the unfamiliar talkers might be more similar to the training talkers (Nosofsky, 1992). However, dialect categorization must make use of some kind of representation of the indexical properties of the talkers. In order to perform the generalization task, the listeners could not rely on global estimates of talker similarity based on just any talker-specific characteristics of the training talkers, but instead they needed to construct a similarity space to represent the *relevant* shared properties of the training talkers from each dialect region (i.e., dialect-specific properties of the speech signal).

A more detailed analysis of the performance of the three-talker listener group in the training and test phases for the three different talkers (best, middle, and worst) revealed a significant block by talker-difficulty interaction, $F(5, 503) = 3.45$, $p < .005$. Post hoc Tukey tests revealed significant differences between the best and worst talkers for the training phases (all $p < .001$), replicating the earlier findings of Clopper and Pisoni (2004b). This pattern of results demonstrates the acquisition of graded category knowledge and membership during the perceptual learning phases. However, post hoc Tukey tests failed to reveal any significant differences between those same talkers in the Test phase, suggesting that the listeners learned to make their categorization judgments using dialect-specific information that was independent of the individual talkers and the specific sentences used in this phase. Figure 6 shows the categorization performance of the three-talker listener group on the best-, middle-, and worst-categorized talkers (as determined by Clopper & Pisoni, 2004b) during training and test, collapsed across all six dialect regions. Thus, even if categorization is treated as a function of perceptual similarity, the underlying basis of similarity must be some sort of dialect-specific representations and not other variable talker-specific properties, such as voice quality or pitch which can be used to discriminate one talker's voice from another, but are not diagnostic in categorizing where the talker is from using novel sentences.

Several other differences in the training and testing procedures between the two groups of listeners may have affected the listeners' categorization performance during generalization. First, the three-talker listener group received more training trials than the one-talker group. If the total number of trials that each group received during the experiment affected performance, we would expect that the three-talker group, who received more trials per block, should perform better in later experimental blocks than the one-talker group because they simply had more experience with the task itself. However, the three-talker group only performed better than the one-talker group in the final generalization task of the experiment. If the number of trials and experience with the task were a relevant factor in this task, we would expect to see evidence of it operating earlier in the experiment. Instead, the three-talker group was consistently poorer than the one-talker group until the generalization phase with unfamiliar talkers. Given the evidence for robust dialect category development by the three-talker group, we conclude that the nature of the variability of the stimulus materials and not the total amount of experience with the categorization task was the factor responsible for the better performance of the three-talker group with unfamiliar talkers in the final generalization task.

Second, as discussed above, set size has been shown to affect performance in categorization tasks. One could argue that the three-talker group performed better than the one-talker group during generalization because these listeners had a smaller set of talkers to categorize. The results of the analysis on the subset of trials shared by both the one-talker group and the three-talker group suggest that the differences are not merely due to the set of talkers used in each generalization task. In addition, the effects of set size on speech perception and speech intelligibility grow vanishingly small as the size of the sets increases (Sumby & Pollack, 1954).

In the current experiment, the three-talker group categorized 48 different talkers in the generalization phase and the one-talker group categorized 60 talkers. Clopper, Conrey, and Pisoni (in press) reported the results of a six-alternative forced-choice categorization study comparing performance on male talkers, female talkers, and a set of mixed male and female talkers. We found that performance was not significantly different across the three conditions, although the number of talkers in each condition was different. In particular, the male talker condition included 66 talkers, the female talker condition included only 48 talkers, and the mixed talker condition included 72 different talkers. Performance actually increased (although not significantly) as the set size increased, with the best performance on the mixed group of talkers and the worst performance on the female talkers. Given that there is no a priori reason to think that female talkers should be harder to categorize than male talkers or a mixed group of talkers, Clopper et al.'s (in press) results suggest that set size effects are minimal when the sets are as large as 48 or 60 talkers. It therefore seems unlikely that the difference in set size in the present study between the two generalization tasks was responsible for the significant difference in categorization performance.

Finally, it may be the case that the listeners in the three-talker group were simply better at dialect categorization than the listeners in the one-talker group. We also find this possibility unlikely, however, for two reasons. First, all of the listeners were drawn from the same participant population, so we have no reason to assume that one group would be better than the other prior to training. Second, the two groups of listeners in the current study were also from the same population as the participants in our other work on dialect categorization. Clopper and Pisoni (2004b) reported similar categorization performance without training as that found here for the generalization task after training. In addition, Clopper et al. (in press) replicated the results of our initial study using female talkers and a mixed group of male and female talkers. Finally, while Clopper and Pisoni (2004a) found differences between listeners based on residential history, the overall performance was similar to that reported by

Clopper and Pisoni (2004b) and Clopper et al. (in press). In addition, the listeners in the two groups in the current study were well-matched for their experience and exposure to dialect variation.

We can therefore conclude from this study that stimulus variability during perceptual learning can contribute to the formation of robust categories of indexical properties such as regional dialect. Unlike gender identification, which is an easy task for naïve listeners (Lass et al., 1976), dialect categorization is more difficult but can be improved by exposure to highly variable stimulus materials, even without explicit instructions as to which features to attend to during learning. Logan et al. (1991) found that training with highly variable stimulus materials led to more robust encoding of the linguistic categories /r/ and /l/. The present findings extend the earlier results of Logan et al. (1991) and suggest that stimulus variability in perceptual learning of dialects can affect performance on generalization tasks that require categorization of indexical as well as linguistic information in the speech signal.

The current findings also provide several new insights into the nature of the encoding of indexical information. In particular, previous research on the interference of indexical information in spoken word recognition (Mullennix et al., 1989) and speeded classification tasks (Mullennix & Pisoni, 1990) is consistent with nonabstractionist accounts of language processing, such as that proposed by Goldinger (1996). In particular, Goldinger (1996) provided evidence from a range of lexical processing tasks that words are encoded in memory as individual exemplars and that abstraction across those exemplars is neither necessary nor desirable. The results of the present study, however, are more consistent with other recent proposals that both event-specific and abstract representations are represented in memory (Nosofsky, Palmeri, & McKinley, 1994; Pierrehumbert, 2001, 2002). Specifically, the participants in the three-talker group were exposed to multiple talkers from each regional dialect and were thus able to develop more robust abstract categories. On the other hand, the listeners in the one-talker group were not provided with enough variability during perceptual learning to differentiate dialect-specific properties from talker-specific idiosyncrasies. We predict that with additional training, the three-talker group would continue to improve, while the one-talker group would reach asymptotic performance quickly due to the limited range of stimulus materials with which they are presented and would be unable to show gain in performance even after additional training.

5 Conclusions

Short-term exposure to dialect variation in the laboratory affects dialect categorization of novel sentences produced by unfamiliar talkers. Increased variability during initial perceptual learning produced a significant gain in generalization performance, supporting recent theoretical proposals that encoding stimulus variability is a fundamental component of human speech perception that underlies the development of robust and stable phonological and lexical representations of spoken language in long term memory. In particular, exposure to dialect variability during training allowed the listeners to develop abstract, talker-independent representations of dialect variation. The present findings provide additional evidence for the claim that indexical properties of speech such as age, gender, and regional dialect are integral components of the speech signal. Both indexical and linguistic attributes are carried by the same acoustic signal and both sets of attributes are processed and encoded by the nervous system.

Acknowledgments

This work was supported by NIH-NIDCD R01 research grant DC00111 and NIH-NIDCD T32 training grant DC00012 to Indiana University. The authors would like to thank Caitlin Dillon for her assistance in selecting the talkers and Luis Hernandez for his technical advice and support.

References

- ABERCROMBIE, D. Elements of general phonetics. Chicago: Aldine; 1967.
- BRADLOW, AR.; BENT, T. Listener adaptation to foreign accented speech. Paper presented at the International Congress of Phonetic Sciences; Barcelona, Spain. August 3–9; 2003.
- BRADLOW AR, PISONI DB, AKAHANE-YAMADA R, TOHKURA Y. Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production. *Journal of the Acoustical Society of America*. 1997; 101:2299– 2310. [PubMed: 9104031]
- CLOPPER CG, CONREY BL, PISONI DB. Effects of talker gender on dialect categorization. *Journal of Language and Social Psychology*. (in press).
- 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. [PubMed: 21533011]
- CLOPPER CG, PISONI DB. Some acoustic cues for the perceptual categorization of American English regional dialects. *Journal of Phonetics*. 2004b; 32:110– 140.
- ESTES, WK. Classification and cognition. New York: Oxford University Press; 1994.
- 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.
- GOLDINGER SD. Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. 1996; 22:1166– 1183.
- GREEN KP, TOMIAK GR, KUHL PK. The encoding of rate and talker information during phonetic perception. *Perception and Psychophysics*. 1997; 59:675– 692. [PubMed: 9259636]
- JOOS M. Acoustic phonetics. *Language*. 1948; 24:5– 136.
- LADEFOGED P, BROADBENT DE. Information conveyed by vowels. *Journal of the Acoustical Society of America*. 1957; 29:98– 104.
- 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]
- LOGAN JS, LIVELY SE, PISONI DB. Training Japanese listeners to identify English /r/ and /l/: A first report. *Journal of the Acoustical Society of America*. 1991; 89:874– 886. [PubMed: 2016438]
- MULLENNIX JW, PISONI DB. Stimulus variability and processing dependencies in speech perception. *Perception and Psychophysics*. 1990; 47:379– 390. [PubMed: 2345691]
- MULLENNIX JW, PISONI DB, MARTIN CS. Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America*. 1989; 85:365– 378. [PubMed: 2921419]
- MURPHY, GL. The big book of concepts. Cambridge, MA: MIT Press; 2002.
- NOSOFSKY, RM. Exemplar-based approach to relating categorization, identification, and recognition. In: Ashby, FG., editor. *Multidimensional models of perception and cognition*. Hillsdale, NJ: Lawrence Erlbaum; 1992. p. 363-393.
- NOSOFSKY RM, PALMERI TJ, MCKINLEY SC. Rule-plus-exception model of classification learning. *Psychological Review*. 1994; 101:53– 79. [PubMed: 8121960]
- NYGAARD LC, PISONI DB. Talker-specific learning in speech perception. *Perception and Psychophysics*. 1998; 60:355– 376. [PubMed: 9599989]
- NYGAARD LC, SOMMERS MS, PISONI DB. Speech perception as a talker-contingent process. *Psychological Science*. 1994; 5:42– 46. [PubMed: 21526138]
- PIERREHUMBERT, JB. Exemplar dynamics: Word frequency, lenition and contrast. In: Bybee, J.; Hopper, P., editors. *Frequency effects and emergent grammar*. Amsterdam: John Benjamins; 2001. p. 137-158.
- PIERREHUMBERT, JB. Word-specific phonetics. In: Gussenhoven, C.; Warner, N., editors. *Laboratory phonology VII*. Berlin: Mouton de Gruyter; 2002. p. 101-140.

- PISONI, DB. Some thoughts on “normalization” in speech perception. In: Johnson, K.; Mullennix, JW., editors. *Talker variability in speech processing*. San Diego: Academic Press; 1997. p. 9-32.
- POLLACK I, PICKETT JM, SUMBY WH. On the identification of speakers by voice. *Journal of the Acoustical Society of America*. 1954; 26:403–406.
- POSNER MI, KEELE SW. On the genesis of abstract ideas. *Journal of Experimental Psychology*. 1968; 77:353–363. [PubMed: 5665566]
- REBER, AS. *The penguin dictionary of psychology*. New York: Penguin; 1985.
- REMEZ RE, FELLOWES JM, RUBIN PE. Talker identification based on phonetic information. *Journal of Experimental Psychology: Human Perception and Performance*. 1997; 23:651–666. [PubMed: 9180039]
- REMEZ RE, RUBIN PE, PISONI DB, CARRELL TD. Speech perception without traditional speech cues. *Science*. 1981; 212:947–950. [PubMed: 7233191]
- SCHNEIDER W, SHIFFRIN RM. Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review*. 1977; 84:1–66.
- SHEFFERT SM, PISONI DB, FELLOWES JM, REMEZ RE. Learning to recognize talkers from natural, sinewave, and reversed speech samples. *Journal of Experimental Psychology: Human Perception and Performance*. 2002; 28:1447–1469. [PubMed: 12542137]
- SOMMERS MS, KIRK KI, PISONI DB. Some considerations in evaluating spoken word recognition by normal-hearing, noise-masked normal-hearing, and cochlear implant listeners. I: The effects of response format. *Ear and Hearing*. 1997; 18:89–99. [PubMed: 9099558]
- SUMBY WH, POLLACK I. Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*. 1954; 26:212–215.
- TICE, R.; CARRELL, T. *Level16 (Version 2.0.3) [Computer Software]*. Lincoln, NE: University of Nebraska; 1998.
- 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.
- WHITE MJ. Identification and categorization in visual search. *Memory and Cognition*. 1977; 5:648–657.
- WILLIAMS, A.; GARRETT, P.; COUPLAND, N. Dialect recognition. In: Preston, DR., editor. *Handbook of perceptual dialectology*. Philadelphia: John Benjamins; 1999. p. 345-358.

Appendix

One-talker group training stimulus materials

New England	tjs0	When all else fails, use force.
		Tugboats are capable of hauling huge loads.
		How much allowance do you get?
		These exclusive documents must be locked up at all times.
		George is paranoid about a future gas shortage.
		Family loyalties and cooperative work have been unbroken for generations.
		So of course he stayed put.
North	jpm 0	But to the infuriation of scientists, for no known reason not all of them did.
		Alimony harms a divorced man's wealth.
		His scalp was blistered from today's hot sun.
		Norwegian sweaters are made of lamb's wool.
		Brush fires are common in the dry underbrush of Nevada.
		I just saw Jim near the new archeological museum.
		Don't plan meals that are too complicated.
		The batting average of one success out of seven increased to one out of three.

North Midland	cef0	She took it grudgingly, her dark eyes baleful as they met his.
		The lack of heat compounded the tenant's grievances.
		Too much curiosity can get you into trouble.
		I itemize all accounts in my agency.
		The full moon shone brightly that night.
		Move the garbage nearer to the large window.
		Are planning and strategy development emphasized sufficiently in your company?
		Have we not actually developed idea worship?
South Midland	lel0	No other visitor inquired for her that evening.
		Gently place Jim's foam sculpture in the box.
		Biologists use radioactive isotopes to study microorganisms.
		The irate actor stomped away idiotically.
		The barracuda recoiled from the serpent's poisonous fangs.
		Please sing just the club theme.
		He injected more vitality into the score than it has revealed in many years.
		He was, thus, an early and spectacular victim.
South	ctt0	A voice spoke near-at-hand.
		Beg that guard for one gallon of gas.
		Combine all the ingredients in a large bowl.
		Rob sat by the pond and sketched the stray geese.
		I'd rather not buy these shoes than be overcharged.
		Her right hand aches whenever the barometric pressure changes.
		Asked why, he replied primly: because that's no activity for a gentleman.
		We would establish no censorship.
West	dlr0	He says he'll be here on the one-o'clock plane.
		Barb's gold bracelet was a graduation present.
		Angora cats are furrier than Siamese.
		Nonprofit organizations have frequent fund raisers.
		The proof that you are seeking is not available in books.
		They assume no burglar will ever enter here.
		Later, we shall see what happened when an emperor took this idea too literally.
		How well do faculty members govern themselves?
		Oil-field workers were a rough-tough lot.
<i>One-talker group generalization stimulus materials</i>		
New England	cpm 0	Please shorten this skirt for Joyce.
	dab0	The bungalow was pleasantly situated near the shore.
	dac1	The thick elm forest was nearly overwhelmed by Dutch Elm Disease.
	jeb1	Computers are being used to keep branch inventories at more workable levels.
	pgh0	They consider it simply a sign of our times.
	pgr0	Seamstresses attach zippers with a thimble, needle, and thread.
	psw0	We are open every Monday evening.
	stk0	Don't do Charlie's dirty dishes.
	tpf 0	We would lose our export markets and deny ourselves the imports we need.

North	trr0	We know that actors can learn to portray a wide variety of character roles.
	dbp0	Who took the kayak down the bayou?
	jar0	Poach the apples in this syrup for 12minutes, drain them, and cool.
	jrj0	Our entire economy will have a terrific uplift.
	pgl0	In fact our whole defensive unit did a good job.
	ppc0	Mosquitoes exist in warm, humid climates.
	rcw0	His sudden departure shocked the cast.
	rjm0	She drank greedily, and murmured, thank you, as he lowered her head.
	rlr0	Contrast trim provides other touches of color.
North Midland	rms0	Jeff's toy go-cart never worked!
	wew0	Kindergarten children decorate their classrooms for all holidays.
	dwm0	A huge tapestry hung in her hallway.
	hmr0	The record teems with romance and adventure.
	jjb0	The football team coach has a watch thin as a dime.
	mjb1	Iris thinks this zoo has 11 Spanish zebras.
	msm0	When peeling an orange, it is hard not to spray juice.
	ree0	Should giraffes be kept in small zoos?
	rjb1	Amoebas change shape constantly.
South Midland	rwa0	By eating yogurt, you may live longer.
	tlb0	The paper boy bought two apples and three ices.
	tpg 0	She uses both names interchangeably.
	css0	Add a few caraway seeds, too, if you'd like.
	dls0	Who authorized the unlimited expense account?
	esg0	The gorgeous butterfly ate a lot of nectar.
	gag0	He ignores guidebook facts.
	jee0	Last year's gas shortage caused steep price increases.
	jmm0	What is this large thing by the ironing board?
South	jrj0	Only then did he decide he didn't want one.
	pcs0	Destroy every file related to my audits.
	trc0	One of the problems associated with the expressway stems from the basic idea.
	trt0	The prowler wore a ski mask for disguise.
	chl0	Draw every outer line first, then fill in the interior.
	gsh0	Bright sunshine shimmers on the ocean.
	hmg0	You must explicitly delete files.
	jpg 0	The avalanche triggered a minor earthquake.
	jwg0	He was kneeling to tie his shoelaces.
West	ram0	Growing well-kept gardens is very time consuming.
	rew1	It will accommodate firing rates as low as a half gallon an hour.
	sas0	The two artists exchanged autographs.
	srr0	But this doesn't detract from its merit as an interesting, if not great, film.
	wch0	The annoying raccoons slipped into Phil's garden every night.
	bar0	Puree some fruit before preparing the skewers.
	bbr0	The water contained too much chlorine and stung his eyes.

bml0 The system may break down soon, so save your files frequently.
 cth0 Usually, they titter loudly after they have passed by.
 dlfo The sound of Jennifer's bugle scared the antelope.
 dlr1 An adult male baboon's teeth are not suitable for eating shellfish.
 hbs0 Of course you can have another tunafish sandwich.
 jai0 Curiosity and mediocrity seldom coexist.
 klr0 Did Shawn catch that big goose without help?
 ntwo I'll have a scoop of that exotic purple and turquoise sherbet.

Three-talker group training stimulus materials

New England	dac1	The misprint provoked an immediate disclaimer. Rich looked for spotted hyenas and jaguars on the safari. Be careful not to plow over the flower beds. The speech symposium might begin Monday. The thick elm forest was nearly overwhelmed by Dutch Elm Disease. Program note reads as follows: take hands; this urgent visage beckons us. In two cases, airplanes only were indicated. Her hum became a gurgle of surprise.
	stk0	Don't do Charlie's dirty dishes. Objects made of pewter are beautiful. The morning dew on the spider web glistened in the sun. Cheap stockings run the first time they're worn. Calcium makes bones and teeth strong. At the base of the rocky hillside, they left their horses and climbed on foot. You young men get to be my age, you won't take flu so lightly. A woman met a famous author at a literary tea.
	tpf0	Challenge each general's intelligence. The water contained too much chlorine and stung his eyes. Does Hindu ideology honor cows? Gus saw pine trees and redwoods on his walk through Sequoia National Forest. Movies never have enough villains. We would lose our export markets and deny ourselves the imports we need. We often say of a person that he looks young for his age or old for his age. Now there's nothin left of me.
North	jar0	Why yell or worry over silly items? Guess the question from the answer. Who authorized the unlimited expense account? To further his prestige, he occasionally reads the Wall Street Journal. Lori's costume needed black gloves to be completely elegant. Poach the apples in this syrup for 12minutes, drain them, and cool. How's it strike you, foul or fair? Weakness in leadership.
	pgl0	Aluminum silverware can often be flimsy. She slipped and sprained her ankle on the steep slope.

Young children should avoid exposure to contagious diseases.
 Weatherproof galoshes are very useful in Seattle.
 Pam gives driving lessons on Thursdays.
 In fact our whole defensive unit did a good job.
 This does not allow the mystery to invade us.
 Shrugs met that, from room clerks, from bellhops.
 rjm0 We experience distress and frustration obtaining our degrees.
 Most precincts had a third of the votes counted.
 A doctor was in the ambulance with the patient.
 Gregory and Tom chose to watch cartoons in the afternoon.
 Iris thinks this zoo has 11 Spanish zebras.
 They find deep pessimism in them.
 Tragedy presumes such a configuration.
 She drank greedily, and murmured, thank you, as he lowered her head.
 North Midland jjb0 The singer's finger had a splinter.
 While waiting for Chipper she crisscrossed the square many times.
 Puree some fruit before preparing the skewers.
 Chip postponed alimony payments until the latest possible date.
 The football team coach has a watch thin as a dime.
 In time she presents her aristocratic husband with a coal-black child.
 So would radar picket ships.
 Suppose you tell me the real reason, he drawled.
 rjb1 Get a calico cat to keep.
 The coyote, bobcat, and hyena are wild animals.
 Jeff's toy go-cart never worked!
 Amoebas change shape constantly.
 He picked up nine pairs of socks for each brother.
 The cowboy called this breed of cattle magpies.
 In most discussions of this phenomenon, the figures are substantially inflated.
 Then he would realize they were really things that only he himself could think.
 tpg 0 Coconut cream pie makes a nice dessert.
 A screwdriver is made from vodka and orange juice.
 The news agency hired a great journalist.
 The bluejay flew over the high building.
 She uses both names interchangeably.
 Make a paste of brown sugar and mustard and spread lightly over scored surface.
 How much and how many profits could a majority take out of the losses of a few?
 They were already swollen to bursting.
 South Midland esg0 Spring Street is straight ahead.
 Her auburn hair reminded him of autumn leaves.
 The gorgeous butterfly ate a lot of nectar.
 Take charge of choosing her bride's maids' gowns.
 Coffee is grown on steep, jungle-like slopes in temperate zones.

		We have also seen the power of faith at work among us.
		Thyroid function tests yielded normal results.
		Another snarled close overhead.
jee0		Last year's gas shortage caused steep price increases.
		Remove the splinter with a pair of tweezers.
		The cigarettes in the clay ashtray overflowed onto the oak table.
		Bob bandaged both wounds with the skill of a doctor.
		Every cab needs repainting often.
		Mercenary: term of honor?
		We may say of some unfortunates that they were never young.
		Then at last the darkness began to dissolve.
pcs0		Where were you while we were away?
		Medieval society was based on hierarchies.
		Destroy every file related to my audits.
		Alice's ability to work without supervision is noteworthy.
		Bob papered over the living room murals.
		Best solution is to find an area that is predominantly sunlight or shade.
		In short, scientific sampling was introduced in place of subjective sampling.
		Accident, murder, suicide — take your pick.
South	chl0	Primitive tribes have an upbeat attitude.
		Does Creole cooking use curry?
		Draw every outer line first, then fill in the interior.
		The moisture in my eyes is from eyedrops, not from tears.
		First add milk to the shredded cheese.
		Again, the analyticity of the two curves guarantees that such intervals exist.
		Thinking the evidence insufficient to get a conviction, he later released him.
		He fought the panic of vertigo.
jpg 0		Alfalfa is healthy for you.
		His sudden departure shocked the cast.
		Approach your interview with statuesque composure.
		The avalanche triggered a minor earthquake.
		We welcome many new students each year.
		This is going to be a language lesson, and you can master it in a few minutes.
		All chance of fulfilling my destiny is over.
		These planets were much bigger, nearly all capable of holding an atmosphere.
sas0		Most young rise early every morning.
		Clasp the screw in your left hand.
		When peeling an orange, it is hard not to spray juice.
		The two artists exchanged autographs.
		Which church do the Smiths worship in?
		For sweet-sour sauce, cook onion in oil until soft.
		Sometimes strong stress serves to focus an important secondary relationship.
		His blue eyes sought the shimmering sea of haze ahead.

West	dlr0	Barb's gold bracelet was a graduation present.
		Angora cats are furrier than Siamese.
		Nonprofit organizations have frequent fund raisers.
		The proof that you are seeking is not available in books.
		They assume no burglar will ever enter here.
		Later, we shall see what happened when an emperor took this idea too literally.
		How well do faculty members govern themselves?
		Oil-field workers were a rough-tough lot.
	jai0	An official deadline cannot be postponed.
		Curiosity and mediocrity seldom coexist.
		Do they allow atheists in church?
		I know I didn't meet her early enough.
		Each stag surely finds a big fawn.
		Often they are able to get in only because the area is declining economically.
		However, the aircraft which we have today are tied to large, soft airfields.
		He believed that brave boys didn't cry.
	klr0	Eat your raisins outdoors on the porch steps.
		The government sought authorization of his citizenship.
		I ate every oyster on Nora's plate.
		Did Shawn catch that big goose without help?
		A toothpaste tube should be squeezed from the bottom.
		They know little about their machinery beyond mechanical details.
		Then he heard the outer door closing.
		Maybe they're delivering the desk now!
	Three-talker group generalization stimulus materials	
	New England	cpm 0 Please shorten this skirt for Joyce.
		dab0 The bungalow was pleasantly situated near the shore.
		jeb1 Computers are being used to keep branch inventories at more workable levels.
		pgh0 They consider it simply a sign of our times.
		pgr0 Seamstresses attach zippers with a thimble, needle, and thread.
		psw0 We are open every Monday evening.
		tjs0 But to the infuriation of scientists, for no known reason not all of them did.
		trr0 We know that actors can learn to portray a wide variety of character roles.
North	North Midland	dbp0 Who took the kayak down the bayou?
		jpm0 She took it grudgingly, her dark eyes baleful as they met his.
		jrp0 Our entire economy will have a terrific uplift.
		ppc0 Mosquitoes exist in warm, humid climates.
		rcw0 To use these new ways in daily life is the last step.
		rlr0 Contrast trim provides other touches of color.
		rms0 Whether historically a fact or not, the legend has a certain symbolic value.
		wew0 Kindergarten children decorate their classrooms for all holidays.
North Midland	North Midland	cef0 No other visitor inquired for her that evening.
		dwm0 A huge tapestry hung in her hallway.

	hmr0	The record teems with romance and adventure.
	mjb1	He murmured to himself, with firmness: no surrender.
	msm0	So, if anybody solicits by phone, make sure you mail the dough to the above.
	ree0	Should giraffes be kept in small zoos?
	rwa0	By eating yogurt, you may live longer.
	tlb0	The paper boy bought two apples and three ices.
South Midland	css0	Add a few caraway seeds, too, if you'd like.
	dls0	This is my hen ledger, he informed him in an absorbed way.
	gag0	He ignores guidebook facts.
	jmm0	What is this large thing by the ironing board?
	jrh0	Only then did he decide he didn't want one.
	lel0	A voice spoke near-at-hand.
	trc0	One of the problems associated with the expressway stems from the basic idea.
	trt0	The prowler wore a ski mask for disguise.
South	ctt0	He says he'll be here on the one-o'clock plane.
	gsh0	Bright sunshine shimmers on the ocean.
	hmg0	You must explicitly delete files.
	jwg0	He was kneeling to tie his shoelaces.
	ram0	Growing well-kept gardens is very time consuming.
	rew1	It will accommodate firing rates as low as a half gallon an hour.
	srr0	But this doesn't detract from its merit as an interesting, if not great, film.
	wch0	The annoying raccoons slipped into Phil's garden every night.
West	bar0	Men believed they could control nature by obeying a moral code.
	bbr0	Come on, let's hurry down before they lock up for the day.
	bml0	The system may break down soon, so save your files frequently.
	cth0	Usually, they titter loudly after they have passed by.
	dlf0	The sound of Jennifer's bugle scared the antelope.
	dlr1	An adult male baboon's teeth are not suitable for eating shellfish.
	hbs0	Of course you can have another tunafish sandwich.
	ntw0	I'll have a scoop of that exotic purple and turquoise sherbet.

Figure 1.
Six response alternatives in the perceptual learning experiment (adapted from Clopper & Pisoni, 2004b)

Figure 2. Percent correct categorization scores for the two groups of listeners in each of the four experimental phases in the perceptual learning experiment. Error bars are *SE*. The left panel shows the results for the Training phases when feedback was provided after each trial. The right panel shows the results for the Test and Generalization phases when no feedback was provided

Figure 3.
Histogram of difference scores from Train I to Train II for each participant in each listener group

Figure 4. Histogram of difference scores from test to Generalization for each participant in each listener group for both the Full (top) and the Subset (bottom) data sets

Figure 5.
Mean difference scores for each talker dialect group from Test to Generalization for each listener group

Figure 6. Percent correct categorization scores for the three-talker listener group in the training and test phases of the perceptual learning experiment as a function of talker difficulty. Error bars are *SE*

TABLE 1

Comparison of training talkers based on acoustic-phonetic measures

Acoustic-phonetic variable	Group means	One-talker	Three-talker "Best"	Three-talker "Middle"	Three-talker "Worst"
r-lessness	NE > SM, W	NE > SM, W	NE > SM, W	NE > SM, W	NE > SM, W
/æ/ fronting	N > NE	N > NE	N > NE	N > NE	*N < NE
/ow/-diphthongization	N < S	N < S	N < S	N < S	*N > S
/u/ fronting	SM, S, W > NE, N	SM, S, W > NE, N	SM, S, W > NE, N	SM, S, W > N	SM, S, W > NE *SM < N

Key: New England (NE), North (N), North Midland (NM), South Midland (SM), South (S), and West (W). Relative degree of production of each variable is indicated by '>' or '<'. Incorrect relationships relative to the group means are indicated by * *.

TABLE 2

Schematic of the experimental procedure for each of the two groups of listeners

	Train IA (Feedback)	Train IB (Feedback)	Train II (Feedback)	Test (No Feedback)	Generalization (No feedback)
One-Talker Group	6 talkers	6 talkers	6 talkers	6 talkers	60 talkers
	10 repetitions sentence #1	10 repetitions sentence #2	4 novel sentences	4 novel sentences	1 novel sentence
Three-Talker Group	18 talkers	18 talkers	18 talkers	18 talkers	48 talkers
	5 repetitions sentence #1	5 repetitions sentence #2	4 novel sentences	4 novel sentences	1 novel sentence

TABLE 3
Summary of percent correct performance by the one-talker group for each talker dialect in each experimental block

	New England	North	North Midland	South Midland	South	West	Mean
Train IA	98	80	86	75	67	80	81
Train IB	84	76	82	77	70	60	75
Train II	85	62	72	61	39	52	62
Test	88	70	72	58	34	43	61
Generalization	42	27	27	36	36	14	30
Mean	79	63	68	61	49	50	62

TABLE 4
Summary of percent correct performance by the three-talker group for each talker dialect in each experimental block

	New England	North	North Midland	South Midland	South	West	Mean
Train IA	66	39	42	46	41	31	44
Train IB	49	41	38	44	35	28	39
Train II	52	35	40	41	36	21	38
Test	65	33	41	43	36	25	41
Generalization	53	31	34	35	32	23	35
Mean	57	36	39	42	36	26	39