

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 24 (2000)
Indiana University

Perceptual Adjustments to Foreign Accented English¹

Constance M. Clarke²

*Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405*

¹ This work was supported by a National Science Foundation Graduate Research Fellowship to the author and was conducted at the Speech Research Laboratory, Indiana University. The author would like to thank Dr. David Pisoni for the opportunity to conduct this research in the SRL and for his assistance and guidance. Much gratitude is also given to Luis Hernandez for his technical support, to Darla Sallee for office support and to Dr. Rebecca Herman for assistance with acoustical analyses.

² University of Arizona, Department of Psychology, Tucson, AZ.

Perceptual Adjustments to Foreign Accented English

Abstract. Two training tasks were evaluated for a study of perceptual learning of foreign-accented speech. The purpose of the planned perceptual learning study is to examine whether exposure to the speech of several foreign-accented voices will improve perception of a new voice with the same accent. Two important characteristics of the listeners' task during training with the voices are emphasis on similarities among accented voices and the availability of a method for evaluating listener performance during training. The first task examined was a similarity judgment task for pairs of voices. Multidimensional scaling was used to assess changes in perception through the course of the exposure, but this technique proved not to be sensitive enough to subtle changes in perception for these stimuli. The second task examined was the combination of a similarity judgment task and a transcription task. This dual task method was more successful in satisfying the goals for the training task and is a promising technique for use in the perceptual learning study.

Introduction

An important and still unanswered question in the study of speech perception is how the human speech processing system achieves perceptual constancy in the face of enormous variability in the acoustic signal. Productions of the same speech sound by different speakers are acoustically different. Even different productions of the same sound by one talker are not identical. Yet listeners still perceive the same speech sound across such variation. How does the perceptual system so successfully extract a single phoneme when there are few, if any, truly invariant features across different productions of that phoneme?

In the traditional approach to solving this problem, the perceptual system was thought to engage in a process of normalization when processing speech (Shankweiler, Strange, & Verbrugge, 1977). It was believed that this process stripped away and discarded variability that did not directly specify the intended speech segment (e.g., acoustic consequences of vocal tract characteristics, phonetic context, or speaking rate). What remained was invariant information of some kind that would unambiguously specify an abstract linguistic category (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). However, there is evidence that the features of speech that vary from token to token are not simply noise. They may actually be useful in the processing of speech.

Over the past ten years, research by Pisoni and his colleagues has shown that variable features of speech, such as those caused by the differences between talkers, are not lost or discarded but are stored and used later in perception. For example, using a continuous recognition paradigm, Palmeri, Goldinger, and Pisoni (1993) found that subjects were faster at recognizing that a word had been presented previously in a list when it was presented in the same voice, compared to when it was presented in a new voice. In addition, a study by Nygaard, Sommers, and Pisoni (1994) established that familiarity with voices improves intelligibility for those voices when speaking novel words. These pieces of evidence and others (e.g., Greenspan, Nusbaum, & Pisoni, 1988; McGarr, 1983) call into question the traditional view that the characteristics of specific voices are discarded by the speech perception system as it decodes a speaker's intended message. It seems that these "irrelevant" details are, in fact, learned and put to use in processing new input.

If perceptual constancy for speech cannot be explained by normalization, what other mechanism might be responsible? Previous study of the problem of talker variability has yielded two distinct hypotheses about how phonetic categories are structured and how the system deals with variability. The classic model holds that a phonetic category is based on a single, abstract prototype (Posner & Keele, 1968). “Perceptual operations” (Kolers, 1976; Nygaard et al., 1994; Pisoni, 1997) analyze each incoming speech token and match it with the correct category. Experience with a particular type of speech (e.g., with a particular person’s voice), allows the system to learn the unique perceptual operations needed to process that speech. Those perceptual operations can be stored and used for later perception. An alternative approach is based on exemplar-based models of categorization (Jacoby & Brooks, 1984; Medin & Shaffer, 1978; Nosofsky, 1986). Applied to speech perception, exemplar-based approaches claim that phonetic categories are made up of episodic traces of speech segments. In a simple version of this model, the token currently being perceived is matched with the most similar acoustic trace held in memory and is assigned the corresponding phonetic category (Goldinger, 1998; Pisoni, 1997).

Thus far, the studies demonstrating retention of variable features of speech do not distinguish between the prototype-with-perceptual-operations model and the exemplar-based model; the results can be explained by both models. For example, the continuous recognition study by Palmeri et al. (1993) described above is intuitively consistent with the exemplar-based model since improved perception was found for the same word produced by the same voice. But perceptual operations used for processing a particular word in a particular voice could also have been responsible for the results. For example, when a listener first hears a word in the experiment, she retains the perceptual operations used to process and identify that word. When the word, spoken by the same voice, is later presented for recognition, the procedures for processing the item will match the stored procedures. Consequently, the overall familiarity will be greater, and the listener will be more likely to correctly report that the word is old. In comparison to the Palmeri et al. (1993) study, the Nygaard et al. (1994) (see also Nygaard & Pisoni, 1998) study showed the perceptual system capable of a higher level of abstraction. There, the words at training and test were different, yet performance was still better with the familiar voices. This suggests that it is not just the particular episodic word tokens that are held in memory, but rather something more general about a talker’s speech. This knowledge might be stored in the form of perceptual procedures for matching the acoustic-phonetic input with phonetic categories. However, an exemplar-based model can also explain the results if it is argued that although the words were different at training and test, exemplars of the individual phonetic categories could have been stored.

One way to distinguish between these two models is to investigate whether the perceptual system can learn even more abstract characteristics of the non-linguistic aspects of speech. A certain level of abstraction cannot be explained with an exemplar-based model. A foreign accent, for example, is a source of non-linguistic variability in speech that causes the speech to differ from native speech in abstract, phonetic rule-governed ways. It has been shown that experience with an accented voice helps perception of new words spoken by that voice (Clarke, 2000; Wingstedt & Schulman, 1987). However, if it could be shown that learning an accent aids in the perception of new words spoken by a *new* talker with the same accent, another level of abstraction would be introduced. Perceptual learning of the accent itself would be demonstrated. This would indicate that the speech perception system can learn at a more abstract level than has so far been established through voice training studies. The characteristics of a voice are largely based on a particular vocal tract and glottal source. However, a foreign accent is based on the structure and content of the native language’s phonetic system, its phonological rules, and the way it interacts with the target language (Tarone, 1987). These characteristics are overlaid on a voice and are assumed to be more or less consistent across different speakers with the same accent. This consistency is the basis of a listener’s ability to identify “what kind of accent” a non-native speaker has. Yet these consistencies across speakers are abstract. They are phonetic, not acoustic, in nature.

Generalization of accent learning to a new talker would call into question a strict exemplar-based view of speech perception. If voice learning is due to the referencing of previously stored acoustic tokens of each voice, as claimed by exemplar-based models, transfer of learning to a new voice with the same accent would not be expected. This is because the acoustic characteristics of the new voice may be quite different from the stored tokens; the only similarities would be abstract, phonetic ones. The type of perceptual learning that transfers to new voices may be better explained by the storing of perceptual operations. Perceptual operations may be more flexible than episodic traces because different levels of analytical rules could be retained, from very specific (i.e., at the level of acoustic characteristics) to very abstract (i.e., at the level of phonological regularities). The abstract rules could be applied to a larger variety of tokens of the original type of speech (e.g., Spanish-accented speech).

In a recent study, Clarke (2000) investigated whether experience with foreign-accented voices improves perception of the speech of a new talker with the same accent, that is, whether an accent itself can be learned. Borrowing the experimental methodology used by Nygaard et al. (1994), Clarke investigated accent learning by giving two groups of listeners three days of accent training. One group was trained with four Spanish-accented voices and four non-accented voices. The other group was trained with four Chinese-accented voices and four non-accented voices. All voices were female speakers producing American English. The listeners' task during training was to learn the name that went with each of the voices. After three days of recognition training, subjects were given a word intelligibility test with new sentences presented in noise. Test sentences included Spanish- and Chinese-accented voices used in training, as well as new Spanish- and Chinese-accented voices. Test sentences also included one new and one old non-accented voice. Clarke found that the Spanish-trained group had an advantage with the old Spanish-accented voice (one of the voices from training), and the Chinese-trained group had an advantage with the old Chinese-accented voice. This replicated the Nygaard et al. findings and showed that voice learning also occurs with foreign-accented voices. However, the listeners' experience with the accented voices did *not* improve their perception of the new accented voices: the Spanish-trained group showed no advantage for the new Spanish-accented voice, nor did the Chinese-trained group for the new Chinese-accented voice. The results suggested that the perceptual learning of speech is voice specific.

It may be, however, that the lack of transfer to new voices was due to a particular aspect of the training methodology. The training task itself may have interfered with the listener's "motivation" for finding similarities among voices of the same accent. For the three days of training, the listeners' goal was to discriminate the voices and match them with the correct name. This is the same procedure Nygaard et al. (1994) used. While the task encouraged close attention to the acoustic and phonetic characteristics of each voice, it effectively required listeners to look for differences among the voices, not similarities. Perhaps a task that emphasized the commonalities among the voices in each accent group would better support accent learning in addition to individual voice learning. An experiment using such a task would be better able to demonstrate whether the perceptual system can learn the abstract phonological characteristics common to all the accented voices and apply them to a new voice.

The purpose of the following two experiments was to find a new training task that can be used in a replication and extension of Clarke (2000). The new training task had to fulfill two goals: first, emphasize the similarities among the voices with the same accent (e.g., among the four Spanish-accented voices); and second, allow for a way to measure the success of training. This second requirement is necessary in order to, for example, determine which subjects were attending to and benefiting from the task. The first experiment assessed the use of a similarity judgment task. The second investigated a task that included both similarity judgments and sentence transcription. The second task was found to be more successful in meeting the goals stated above.

Experiment 1: Similarity Judgments

In the first task investigated, listeners were asked to make similarity judgments between pairs of voices on a seven-point scale from Very Similar to Not Similar At All. Multidimensional scaling³ was then used to examine whether their similarity spaces for the voices changed from the beginning of the experiment to the end. It was hoped that the similarity judgment task would serve the purpose of encouraging listeners to focus on the similarities among the accented voices, rather than the differences. The multidimensional scaling technique provided a way of measuring whether listeners' perception of the voices was affected by the task demands. For example, one possible change could be a shift from making similarity judgments based solely on the presence or absence of an accent, to judgments based on perceiving and encoding more fine-grained characteristics of the voices.

Method

Subjects

Twenty-four Indiana University undergraduates (20 female, 4 male) participated as listeners in the experiment for partial fulfillment of a course requirement. Eight participants were excluded from the final analysis: one because of a history of hearing disorder, one because of an error in the experimental program, two because of a failure to follow instructions, and four so that the correct counterbalancing of conditions was maintained⁴. The remaining 16 participants (13 female, 3 male) were monolingual, native speakers of American English who reported no history of speech or hearing disorders at the time of testing.

Materials and Stimuli

Two groups of eight participants each listened to eight female voices. For each group, four of the voices were non-accented (NA) when speaking English (native speakers of English), and four had a noticeable accent (non-native speakers of English)⁵. For one group (Spanish/NA) the accented voices had a Spanish accent, and for the other group (Chinese/NA) the accent was Chinese⁶. These twelve voices (four non-accented, four Spanish-accented, and four Chinese-accented) were the same voices used in the training portion of the Clarke (2000) study. The non-accented speakers were native speakers of American English with no obvious regional accent, ranging in age from 19 to 31. The four Spanish-accented speakers were native speakers of Mexican Spanish, all from the region of Sonora, Mexico, who began learning English after the age of 25 (mean age of English acquisition: 33 years; mean age at time of recording: 38 years). The four Chinese-accented speakers were native speakers of Mandarin Chinese, all from Taiwan, ROC, who began learning English after the age of eleven (mean age of English acquisition: 12 years; mean age at time of recording: 24 years). All accented speakers reported using their native language at least thirty percent of the time in their current daily lives. The voices had originally been recorded in the Speech Perception Laboratory at the University of Arizona, Department of Psychology.

³ Multidimensional scaling (MDS) is a statistical technique for representing similarity among objects. Similarity data specify the location of objects in an n-dimensional space in which distance is inversely related to similarity.

⁴ Beyond the counterbalancing requirements, two of the participants who were excluded from further analyses were chosen because they showed the greatest trend toward a bias in their similarity judgments. The other two were excluded because they were the last to participate.

⁵ Although the main interest in these experiments was evidence of perceptual learning for the accented voices, the non-accented voices were included in order to keep the voice set identical to that used in the Clarke (2000) study and in the planned follow-up study. The inclusion of non-accented voices in the full studies is important for verifying that the basic voice learning effect can be obtained with the methodology used.

⁶ Different groups listened to the Spanish-accented and Chinese-accented voices because in the original study (Clarke, 2000) accent type was a between-subjects variable. There are no experimental comparisons between groups in the present study.

The voices were recorded onto tape and digitized onto a Macintosh PowerPC 8100 at a sampling rate of 22.05 kHz and a resolution of 16 bits. Amplitude was normalized to 90% of maximum for all sentences, and the individual sentence files were converted to WAVE format.

The sentences used in the experiment were taken from the Revised Speech Perception In Noise (SPIN) test (Bilger, 1984; Bilger, Nuetzel, Rabinowitz, & Rzeczkowski, 1984; Elliot, 1995; Kalikow, Stevens, & Elliott, 1977). The Revised SPIN Test comprises a set of phonetically and frequency balanced sentences designed to assess impairment of hearing for speech. It is made up of five- to eight-word sentences, each ending in a common, one-syllable word. All the sentences used in the current experiment were High Predictability (HP) sentences, meaning that the final word in each sentence was highly predictable from the semantic context of the sentence (e.g., Stir your coffee with a spoon). One hundred four of these sentences were used in the present study.

Among the eight voices that each group heard (four non-accented and four accented), each voice was paired with every other voice, for a total of 28 unique pairings (a voice was never paired with itself). Each of the 28 pairs of voices was presented once in each of six blocks, for a total of 168 trials. Each individual voice was heard seven times per block, 42 times total. The ordering of the voices in each pair was counterbalanced across blocks such that each voice was heard first and second an equal number of times across the experiment. The order of voices within a particular pair was identical for blocks 1, 3, and 5; the mirror order occurred in blocks 2, 4, and 6. In addition, voice order was counterbalanced across subjects. Within each block, the voice pairs were presented in random order using an on-line randomization program. Finally, in each trial, both voices produced the same sentence. Because of constraints stemming from which speakers had originally recorded which sentences, 64 of the 104 unique sentences had to be repeated once during the experiment in order to fill the 168 trials. However, a sentence was never repeated by the same voice and was never repeated in the same block. The sentences spoken by the non-accented voices were identical across the two groups; the sentences spoken by the Spanish-accented voices for the Spanish/NA group were identical to those spoken by the Chinese-accented voices for the Chinese/NA group.

Procedure

Participants were seated in a quiet room in front a computer keyboard and monitor. Up to six participants were run at a time, in separate booths, with separate computers, and at their own pace. Stimuli were presented to each participant over Beyer Dynamic DT100 headphones at approximately 71 dB SPL from a Pentium 133 MHz IBM compatible computer with a Soundblaster 16 AWE 32 sound card. After reading through an instruction sheet, listeners heard each of the eight voices say one sentence each. This phase was simply to familiarize them with the range and type of voices they would be exposed to; no response was required. Then the main portion of the experiment began. In each trial, listeners were alerted with the word "READY" displayed on the computer screen for 1000 ms. The listeners then heard two voices say the same sentence, with a 500 ms inter-stimulus interval (ISI). Five hundred milliseconds after the second voice, listeners were asked to rate how similar the two voices were to one another on a scale from 1 (labeled "Not Similar At All") to 7 (labeled "Very Similar") with the prompt, "PLEASE RATE SIMILARITY". They typed the rating response into a keyboard, and the response appeared on the screen. Participants were allowed to change the response if they wanted to before submitting it by pressing the ENTER key. In the instructions, listeners were asked to use the whole range of the scale during the experiment. After the response was submitted, a 1000 ms inter-trial interval (ITI) occurred before the next trial began. The entire experiment took approximately 25 minutes, and participants were given a break half way through.

Results

Within each block of the experiment, the similarity judgments for each pair of voices were averaged across all subjects in a group. This produced an 8 x 8 matrix of similarity data for each block in which the average similarity score for every combination of two voices was represented. The first block was considered warm-up and was not included in the analysis. For each group (Spanish/NA and Chinese/NA), Blocks 2 through 6 were submitted as separate matrices to a non-metric multidimensional scaling (MDS) analysis (Euclidean distance metric) using the INDSCAL model in the SPSS 10.0 ALSCAL program. This model takes several matrices and finds a multidimensional spatial solution that best fits the data in all the matrices. The model then determines dimension weights for each individual matrix that describe how much emphasis that matrix gives to each dimension relative to the overall solution. Our interest was in the change in these dimension weights from Block 2 (beginning of the experiment) to Block 6 (end of the experiment) for both groups. A change in the importance listeners placed on each dimension due to experience with the voices would indicate that the exposure was having an effect on perception.

The data were analyzed with both a two-dimensional and a three-dimensional solution. The two-dimensional solution was the most appropriate for both groups' data because the fits were extremely good (Spanish/NA group: stress = .10, $R^2 = .96$; Chinese/NA group: stress = .06, $R^2 = .98$) and the dimensions were interpretable as 1) accentedness and 2) other voice characteristics. The two-dimensional MDS solutions (across all blocks) for both groups are shown in Figures 1A (Spanish/NA group) and 1B (Chinese/NA group). Each point represents a voice, and the points are labeled as non-accented (NA 1-4) or accented (Spanish/Chinese 1-4). Inspection of this figure shows that Dimension 1 clearly reflects accentedness: all accented voices have positive values on this dimension and all non-accented voices have negative values. Further support for this conclusion comes from the high correlation between rated accentedness (ratings obtained in the Clarke (2000) study) and Dimension 1 coordinate value ($r = +.99$, $p < .001$ for both groups). The source of Dimension 2 is less clear, but may reflect other general voice characteristics. One possible candidate is age of the speaker. There was a marginally significant positive correlation between speaker age and Dimension 2 value for the non-accented voices only (Spanish/NA group: $r = +.94$, $p = .06$; Chinese/NA group: $r = +.95$, $p = .05$; two-tailed; alpha set to .0125 for multiple correlations); the correlation was not significant for the accented voices. Another possible source of Dimension 2 is voice pitch. There was a trend toward a negative correlation between average minimum F0 and Dimension 2 value for non-accented voices only (Spanish/NA group: $r = -.93$, $p = .07$; Chinese/NA group: $r = -.97$, $p = .03$; two-tailed; alpha set to .0125 for multiple correlations); again, the correlation was not significant for the accented voices⁷. A definitive interpretation of Dimension 2 is not essential, however, for the objectives of this experiment. Of greatest interest is whether there was a systematic shift, over the course of exposure to the voices, in the relative weightings of the voice dimensions, whatever they may be.

The normalized dimension weights for Blocks 2 through 6 for both groups are shown in Figures 2A (Spanish/NA group) and 2B (Chinese/NA group). It can be seen from the dimension scales themselves that, for all blocks, similarity judgments were overwhelmingly based on accentedness. In all but one block across both groups, Dimension 1 (accentedness) commanded over 89% of the weight in similarity judgments. In terms of changes in dimension weightings from Block 2 to Block 6, however, the

⁷ Because the voices were not controlled for anything but accentedness, these analyses are post hoc, and the comments based on them are purely speculative. It is noted, however, that the finding that the accented voices are less separated in the similarity space is consistent with other studies of the perception of accented voices. Goggin, Thompson, Strube, and Simental (1991) and Thompson (1987) have found that listeners are worse at learning to discriminate foreign-accented voices than non-accented voices. These findings suggest that it is more difficult to distinguish subtle differences in the voice characteristics of accented voices compared to those of non-accented voices.

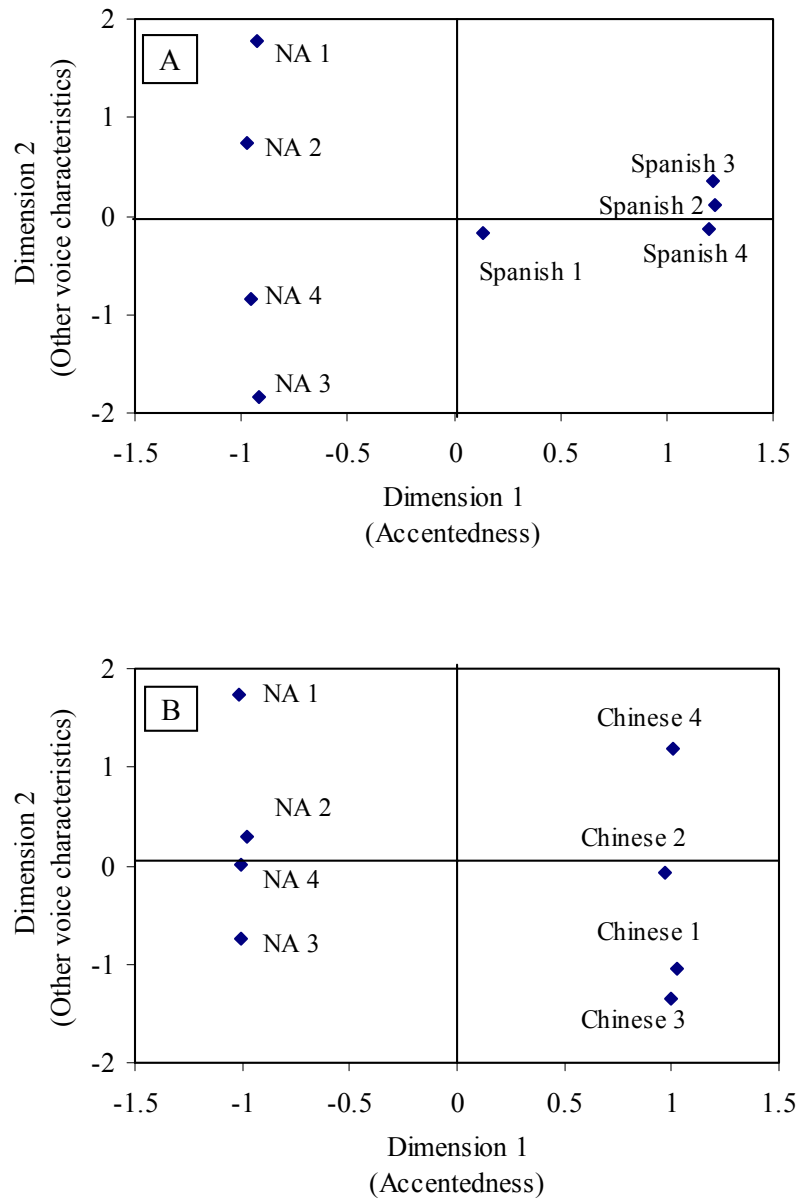


Figure 1. Multidimensional scaling solutions for A) the group listening to four non-accented (NA) and four Spanish-accented voices, and B) the group listening to four non-accented (NA) and four Chinese-accented voices. The solutions are based on similarity judgments between pairs of voices. Each point represents one voice (NA 1-4 were the same for both groups). For both solutions, Dimension 1 was interpreted as Accentedness and Dimension 2 as Other voices characteristics. The dimension scales are arbitrary.

two groups showed different patterns. For the Spanish/NA group, slightly more weight was given to Dimension 2 (other voice characteristics) as the experiment progressed over time. However, for the Chinese/NA group, after Block 2, in which only about 78% of the weight went to the accentedness dimension, almost 100% of the weight went to accentedness. This seems to indicate that after Block 2, the listeners in the Chinese/NA group shifted to a strategy of judging voice similarity almost entirely by

whether the voices matched on accentedness. That is, all accented/accented pairs and non-accented/non-accented pairs were judged as equally similar, and all accented/non-accented pairs were judged as equally dissimilar. This is in contrast with the listeners in the Spanish/NA group, who on the whole seemed to maintain a consistent strategy but gradually became slightly more influenced by the individual voice characteristics.

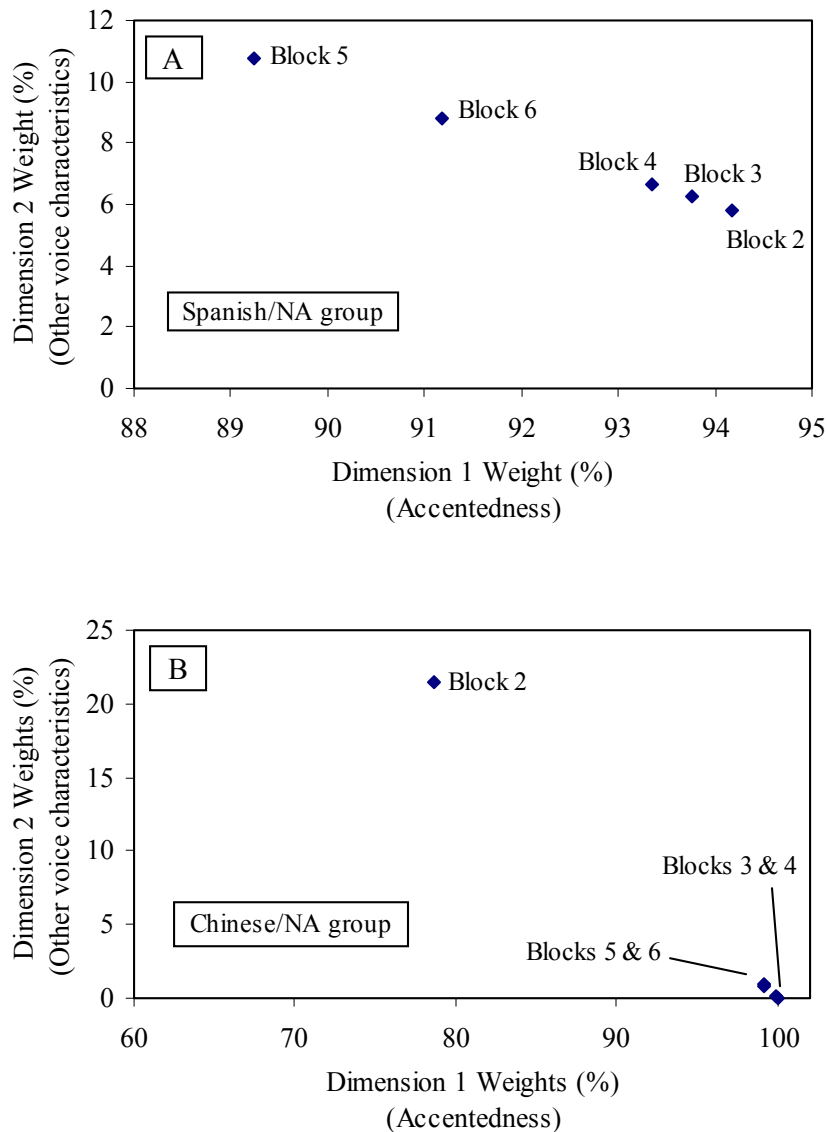


Figure 2. Normalized dimension weights for A) the group listening to non-accented and Spanish-accented voices, and B) the group listening to non-accented and Chinese-accented voices. Each point represents an experimental block of 28 trials. Each block's value on each scale indicates the percentage of weight or importance given to that dimension during that block. Dimension 1 was interpreted as Accentedness and Dimension 2 as Other voices characteristics.

Discussion

The goal of this first experiment was to determine whether a similarity-rating task would be appropriate for use in a study of accent learning. At first glance, the task did meet the initial goal of emphasizing the similarities among the voices. However, the task did not robustly satisfy the second goal: to provide a method of measuring the task's success in affecting perception. A multidimensional scaling analysis was used to look for a change in the subjective similarity space of the voices from the beginning of the experiment, when the voices were unfamiliar to the listeners, to the end, when they were more familiar. To the author's knowledge, the use of MDS in measuring perceptual changes in voice familiarity has not been reported in the literature before. Although MDS seems to be a promising technique for this purpose, it was unsuccessful in the present experiment. First, due to the nature of the voice stimuli, the accentedness dimension of the voice set had almost complete influence on the similarity judgments. This fact likely rendered the similarity measure insensitive to any subtle changes in similarity space that might have been present. Second, the changes that were seen, that is, the change in dimension weightings from the beginning to the end of the experiment, were in opposite directions for the group listening to Spanish and NA voices and the group listening to Chinese and NA voices. The listeners assigned to the Spanish/NA group gave more weight to the "other voice characteristics" as the experiment went on, while the listeners assigned to the Chinese/NA group gave less (and, for most of the experiment, judged solely based on accentedness). Finally, we found that this similarity judgment task was probably too monotonous for a full, three-day training experiment. Therefore, a new task was used in Experiment 2.

Experiment 2: Similarity Judgment and Transcription

The main problem with the first task was that it was difficult to evaluate whether the training was having an effect on listeners' perception of the voices. Therefore, experiment two retained the similarity judgment task, since it was still the best candidate for emphasizing similarities among voices, but added a subsidiary activity: a transcription task. It has been well established that transcription of words or sentences improves with increased voice familiarity (e.g., Greenspan, Nusbaum, & Pisoni, 1988; Nygaard & Pisoni, 1998; Nygaard, Sommers, & Pisoni, 1994; Schwab, Nusbaum, & Pisoni, 1985). Transcription is therefore a more reliable and direct way of evaluating whether the experience with the voices is affecting listeners' perception. In this new task, sentence transcription trials were interspersed throughout the experiment along with the similarity rating trials. With this methodology, separate activities served the two objectives of this training task. The similarity trials encouraged listeners to attend to similarities among the voices, and the periodic transcription trials provided a way to track the listeners' ability to understand the voices. We expected that perception would improve with exposure to the voices over time and thus transcription would become more accurate from the beginning of the experiment to the end. As a final note, we also hoped that the inclusion of two different activities would make the task more interesting for the listeners. Moreover, a task that allows for a chance at improvement (the transcription task) might increase participant motivation during the entire procedure.

Method

Subjects

Twenty-eight Indiana University undergraduates (17 female, 11 male) participated as listeners in the experiment for partial fulfillment of a course requirement. Four participants were excluded from the final analysis: two because of exposure to a language other than English at an early age, and two in order

to maintain the correct counterbalancing of conditions⁸. The remaining 24 participants (14 female, 10 male) were monolingual, native speakers of American English who reported no history of speech or hearing disorders at the time of testing.

Materials and Stimuli

Participants were again assigned to two groups. One group (Spanish/NA) listened to four non-accented and four Spanish-accented voices, and the other group (Chinese/NA) listened to four non-accented and four Chinese-accented voices. The voices were the same twelve used in Experiment 1. A new set of the recorded SPIN sentences were used, 64 HP (High Predictability) sentences for the similarity judgment trials, and 24 LP (Low Predictability) sentences for the transcription trials (LP: the final word was not predictable from the semantic context of the sentence, e.g., We spoke about the knob).

Similarity Judgment Trials. The similarity judgment task was modified slightly from Experiment 1. Instead of hearing two voices and rating how similar they were, listeners heard three voices: a reference voice and two comparison voices. The three voices in a trial were always either all accented or all non-accented. Listeners were asked to choose which comparison voice was more similar to the reference voice. (This is a variant of an XAB task.)

All possible combinations of each reference voice with two other voices from the same accent category were presented twice, once in the first half of the experiment and once in the second half. Thus, there were 96 similarity judgment trials, consisting of 48 non-accented trials and 48 accented trials. Each of the eight voices was the reference voice 12 times total. Within each half of the experiment, the trials were presented in a random order, using on-line randomization. Finally, on a given trial, all three voices always said the same sentence. Because of constraints due to which speakers had originally recorded which sentences, 32 of the 64 unique sentences had to be repeated once during the experiment in order to fill the 96 trials. However, a sentence was never repeated by the same voice and was never repeated in the same half of the experiment. For each reference voice, four of the trials contained non-repeated sentences and eight contained repeated sentences.

Transcription Trials. The transcription trials consisted of one voice saying one sentence (in the clear), followed by the listener typing the entire sentence into the computer keyboard as accurately as possible. There were 24 transcription trials, three trials for each of the eight voices, interspersed among the 96 similarity judgment trials. Only new, LP sentences were presented for transcription. The 24 trials were divided into three blocks, with each block containing one sentence from each of the eight voices. To guard against item effects on transcription accuracy, the presentation order of the blocks was counterbalanced across three groups. The block orders were as follows: Group 1—1, 2, 3; Group 2—2, 3, 1; Group 3—3, 1, 2. Within each block of trials, the eight sentences were presented in random order, using on-line randomization. One transcription trial was presented after every two to six similarity judgment trials (the number between two and six, inclusive, was randomly chosen on-line); hence, the transcription trials were dispersed evenly throughout the experiment, but their occurrence was not predictable.

Procedure

Participants were seated in a quiet room in front of a computer keyboard and monitor. Up to five participants were run at a time, in separate booths, with separate computers, and at their own pace. Stimuli were presented to each participant over Beyer Dynamic DT100 headphones at approximately 71

⁸ The two participants who were excluded from further analyses were chosen because they were the last to participate.

dB SPL from a Pentium 133 MHz IBM compatible computer with a Soundblaster 16 AWE 32 sound card. After reading through an instruction sheet, listeners heard each of the eight voices say one sentence each. This phase was simply to familiarize them with the range and type of voices they would be exposed to during the full experiment; no response was required. The main portion of the experiment followed.

On each similarity judgment trial, listeners were alerted to the type of trial coming up with the prompt “SIMILARITY JUDGMENT” displayed on the computer screen for 1500 ms. The sentence they were about to hear was then displayed orthographically for 2000 ms. The words “Reference Voice” were then displayed in the middle of the screen while the reference voice was presented over the headphones. The first comparison voice began 1000 ms after the reference voice finished, and “Comparison 1” was displayed on the lower left side of the screen. “Comparison 1” remained on the screen, and after 500 ms the second comparison voice was played and “Comparison 2” was displayed on the lower right side of the screen. The screen cleared 500 ms after the second comparison voice ended, and a prompt for a response was displayed: “Which is more similar to the reference voice? 1 or 2?” Listeners responded by pressing one of two keys labeled “1” and “2” on the keyboard. After entering their response, participants pressed the ENTER key to move on to the next trial. The ITI was 1000 ms for all trials.

On each transcription trial, listeners were alerted to the type of trial with the prompt “TRANSCRIPTION” which remained on the screen throughout the presentation. After 1000 ms, the sentence to be transcribed was presented over the headphones. Following a pause of 500 ms, listeners were prompted for a response with the words, “Please type the sentence now.” They typed what they had heard into the keyboard, and the response appeared on the screen. Participants were allowed to correct mistakes as they typed. When they were finished they pressed the ENTER key to submit their answer. After a 500 ms pause, feedback was provided with the words, “The sentence was:” and the sentence text, displayed for 1500 ms. After the 1000 ms ITI, the next trial began. The entire experiment took approximately 40 minutes, and participants were given a break half way through.

Results

Transcription accuracy was evaluated by scoring predetermined keywords in each of the 24 sentences⁹. The keywords were content words only, including nouns, verbs, adjectives, and adverbs. A keyword was accepted as correct if: it matched the target word exactly, it was an obvious misspelling, it was a homophone, a plural had been added or deleted, or an inflectional affix had been added or deleted. One point was given for each correct keyword, and the score for each sentence was the percentage of correct keywords out of the total possible keywords. Trials in which no response was given were not counted in the total possible score. For each participant, the percent correct score was calculated separately for the first third of the sentences (first eight sentences) and the final third (last eight sentences). Each third included one sentence from each of the voices, and across subjects, each sentence appeared an equal number of times in the first third and in the final third of the experiment.

The mean scores for the first eight and final eight transcription sentences for both groups are shown in Figure 3. For the Spanish/NA group, a Sign Test revealed that a significant number of listeners, 10 out of 12, improved in their transcription accuracy from the first third to the final third of the experiment ($p < .05$). The average improvement for the Spanish/NA group from the first third, $M = 77.98\%$, $SD = 5.07$, to the final third, $M = 82.74\%$, $SD = 8.11$, was shown to be marginally significant with a paired t-test, $t(11) = 1.56$, $p = .07$. The Chinese/NA group did not show a significant improvement

⁹ Only the results of the transcription task will be reported here. In this experiment the similarity judgment results were of secondary interest and will be analyzed at a later time.

from the first third, $M = 81.33\%$, $SD = 10.88$, to the final third, $M = 85.06\%$, $SD = 5.16$, with either a paired t-test ($t(11) = 1.03$, $p = .16$), or a Sign Test (7 out of 12 improved, $p = .39$).

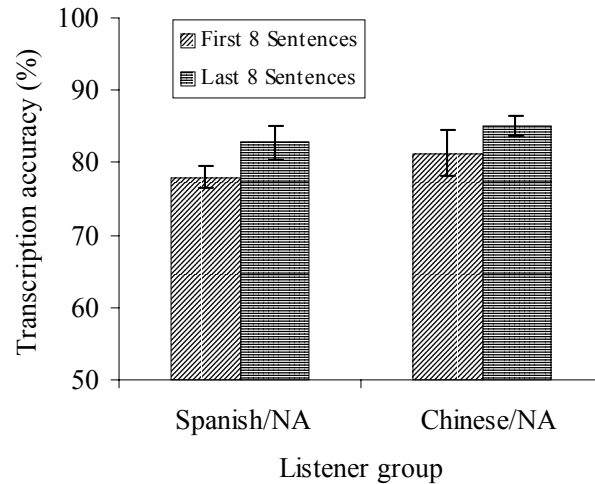


Figure 3. Mean percent accuracy of transcription of keywords in the first 8 and the last 8 sentences of Experiment 2 for the two listener groups. Spanish/NA group: group listening to non-accented and Spanish-accented voices; Chinese/NA group: group listening to non-accented and Chinese-accented voices. Error bars indicate standard error of each mean.

Discussion

The combination similarity judgment and transcription task seems to be promising as a voice-learning task for a full foreign accent learning study. Although the listeners had a relatively short exposure to the voices (40 minutes), the transcription task revealed an improvement in intelligibility of the voices, at least for the Spanish/NA group. The results for the Spanish/NA group are encouraging and suggest the possibility that a significant improvement will also be found for the Chinese/NA group when the training exposure is increased to the planned 2 ½ hours in the full experiment. Thus, we expect that this task will effectively fulfill the second objective of a good training task: to allow for a way to measure the success of training. In addition, the inclusion of similarity judgment trials, and the fact that the voices in each accent category are always grouped within a trial, are likely to fulfill the first goal: to emphasize the similarities among the accented voices. An additional benefit of this task is that the unpredictable transcription trials seem to make it more interesting for the participants and to provide motivation for improvement. This task is therefore an improvement over the similarity-rating task alone, as presented in Experiment 1, and will be the basis for developing the follow-up foreign accent training study.

Conclusion

The original motivation for these experiments was to find a new task that would draw listeners' attention to the similarities among accented voices presented during training, unlike the task used in Clarke (2000). This new training task was to be used in a follow-up study that would re-test whether experience with foreign-accented voices improves perception of a new voice with the same accent. It was argued that a task emphasizing the similarities among the accented voices would provide the best chance

for learning the abstract characteristics of a foreign accent. Transfer of that learning to a new accented voice might then be possible.

The two experiments described here reveal that finding a good voice-training task is not a simple undertaking. The procedure must satisfy several objectives at once, not the least of which is keeping participants interested and motivated so that the training has the desired effect on perception. Through these pilot experiments it was found that two different activities, interweaved throughout the training, provide the best solution for satisfying the two main goals desired for this study. The first goal, emphasizing similarities among the accented voices, is presumably satisfied by the use of a similarity judgment task as the main activity during training. The second goal, providing a way to measure training success, was fulfilled most successfully with a transcription accuracy measure. The transcription task was not completely successful, however, (i.e., for the Chinese/NA group) and may require some modification to make it a more robust measure of perceptual change. For example, transcription sentences could be presented in noise instead of in the clear, as done here. This change would lower overall performance, but may amplify the difference in the listeners' perceptual abilities from the beginning of training to the end. This amplification would be expected since the increased difficulty would demand the use of every perceptual advantage the listeners may have gained during training.

Future directions for this research include testing the similarity judgment/transcription task with the transcription trials in noise to see if this is a stronger measure of perceptual change. Finally, the new task will be applied to a replication and extension of the Clarke (2000) accent training study. It is hoped that this study will provide further insight into the mechanisms involved in the perceptual learning of novel voices and the larger issues of variation and variability in speech and spoken language processing.

References

- Bilger, R.C. (1984). Manual for the clinical use of the revised SPIN Test. Champaign, IL: The University of Illinois.
- Bilger, R.C., Nuetzel, J.M., Rabinowitz, W.M., & Rzeczkowski, C. (1984). Standardization of a test of speech perception in noise. *Journal of Speech and Hearing Research*, 27, 32-48.
- Clarke, C.M. (2000). Perceptual learning of foreign accented English. Unpublished masters thesis, Tucson, AZ: The University of Arizona.
- Elliot, L.L. (1995). Verbal auditory closure and the Speech Perception In Noise (SPIN) Test. *Journal of Speech and Hearing Research*, 38, 1363-1376.
- Goggin, J.P., Thompson, C.P., Strube, G., & Simental, L.R. (1991). The role of language familiarity in voice identification. *Memory & Cognition*, 19, 448-458.
- Goldinger, S.D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, 105, 251-279.
- Greenspan, S.L., Nusbaum, H.C., & Pisoni, D.B. (1988). Perceptual learning of synthetic speech produced by rule. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14, 421-433.
- Jacoby, L.L., & Brooks, L.R. (1984). Nonanalytic cognition: Memory, perception, and concept learning. In G. Bower (Ed.), *The Psychology of Learning and Motivation*, Vol. 18 (pp. 1-47). New York: Academic Press.
- Kalikow, D.N., Stevens, K.N., & Elliott, L.L. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *Journal of the Acoustical Society of America*, 61, 1337-1351.
- Kolers, P.A. (1976). Pattern analyzing memory. *Science*, 191, 1280-1281.

- Lieberman, A.M., Cooper, F.S., Shankweiler, D.P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*, 431-461.
- McGarr, N.S. (1983). The intelligibility of deaf speech to experienced and inexperienced listeners. *Journal of Speech and Hearing Research*, *26*, 451-458.
- Medin, D.L., & Schaffer, M.M. (1978). Context theory of classification learning. *Psychological Review*, *85*, 207-238.
- Nosofsky, R.M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39-57.
- Nygaard, L.C., Sommers, M.S., & Pisoni, D.B. (1994). Speech perception as a talker-contingent process. *Psychological Science*, *5*, 42-46.
- Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics*, *60*, 355-376.
- Palmeri, T.J., Goldinger, S.D., & Pisoni, D.B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*, 309-328.
- Pisoni, D.B. (1997). Some thoughts on "normalization" in speech perception. In K. Johnson, & J.W. Mullennix (Eds.), *Talker Variability in Speech Processing* (pp. 9-32). San Diego, CA: Academic Press.
- Posner, M.I., & Keele, S.W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, *77*, 353-363.
- Shankweiler, D., Strange, W. & Verbrugge, R. (1977). Speech and the problem of perceptual constancy. In R. Shaw, & J. Bransford (Eds.), *Perceiving, acting, and knowing: Toward an ecological psychology*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Schwab, E.C., Nusbaum, H.C., & Pisoni, D.B. (1985). Some effects of training on the perception of synthetic speech. *Human Factors*, *27*, 395-408.
- Tarone, E.E. (1987). The phonology of interlanguage. In G. Ioup, & S.H. Weinberger (Eds.), *Interlanguage Phonology: The Acquisition of a Second Language Sound System* (pp. 70-85). Cambridge, MA: Newbury House Publishers.
- Thompson, C.P. (1987). A language effect in voice identification. *Applied Cognitive Psychology*, *1*, 121-131.
- Wingstedt, M. & Schulman, R. (1987). Comprehension of foreign accents. In W. Dressler, H. Luschutzky, O. Pfeiffer, and J. Rennison (Eds.), *Phonologica 1984* (pp. 339-345). Cambridge: Cambridge, U.P.

This page left blank intentionally.