FOREIGN ACCENTED SPEECH:

ADAPTATION AND GENERALIZATION

A Thesis

Presented in Partial Fulfillment of the Requirements for

the Degree Master of Arts in the

Graduate School of The Ohio State University

By

Shawn Aaron Weil, B.A.

* * * * *

The Ohio State University

2001

Master's Examination Committee:

Dr. Mark Pitt, Adviser

Dr. Keith Johnson

Dr. Mari Riess Jones

Approved by

_____

Adviser

Department of Psychology

ABSTRACT

Characteristics of a first language largely determine speech production in a second, non-native language. Foreign Accented Speech (FAS) is the result. Previous research regarding talker variability and normalization has been limited to non-accented talkers, and has found that non-phonetic talker characteristics are encoded into memory along with phonetic information, and that this information implicitly helps subsequent speech perception. The current study extends this research to FAS, and examines the ability to adapt to FAS and generalize from one talker to another. Listeners were exposed to one speaker for four experimental sessions via a battery of tests measuring speech intelligibility. Listeners were tested on either the talker they had been trained on, a similarly accented talker, or a talker with a different, unrelated accent. Control groups received the post-test only. It was hypothesized that training would improve performance in the same speaker/same accent condition to the greatest degree, in the different speaker/same accent condition to a lesser degree, and negligibly affect performance in the different speaker/ different accent condition. Instead, performance in the different speaker/same accent condition differed as a function of the task. Implications onto models of speech perception are discussed.

Dedication

To Beren Gayle - Thank you for everything.

ACKNOWLEDGMENTS

Many people helped be complete this investigation, more than I can mention in a simple page. Support came in several forms: intellectual, moral, and monetary.

First, I would like to thank my advisor, Mark Pitt, for his guidance, patience, and ideas. This was new territory for both of us, but he was willing to persevere.

Keith Johnson not only served on my master's committee, but also as my advisor during my Summer 2000 Cognitive Science Center. His perspective on adaptation and knowledge of the literature were invaluable, as was his generosity toward the expense of recording and recruiting speakers and subjects.

Thank you to Mari Jones, for agreeing to serve on my master's committee. I have a tremendous amount of respect for you, and I value your opinions and criticism highly. Thank you to the faculty of the Ohio State School of Education for pointing me towards some resources. Thank you to Anne Feldhaus. Peter Hook, and Madhav Deshpande for their knowledge of Marathi phonetics. Thank you to my accented talkers. Thank you to my friends and family members, who listened to me *kvetch* over the past several years. It finally worked out.

This research was supported in part by a Summer 2000 Ohio State University Center

for Cognitive Science Graduate Research Fellowship.

VITA

October 21, 1975 . . . . . Born - Boonton, NJ

1994 . . . . . B.A. Psychology/Music, Binghamton University.

1994 - present . . . . . Graduate Research and Teaching

Fellow, The Ohio State University

FIELDS OF STUDY

Major Field: Psychology

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

The human speech perception system is exceedingly flexible. Over the course of a lifetime, a speaker is bombarded with countless permutations of his native tongue, each deviating from his own speech on a multitude of dimensions. Variation of average pitch, speaking rate, intensity, and timbre interact to form a virtually limitless palette for speaker variability within a particular dialect. If variability due to articulatory dysfunction or environmental factors (i.e., echo or distortion) is considered, and the variation compounds. Despite all of this talker variability, listeners seem to have the surprising ability to comprehend spoken language without a great deal of effort.

Nygaard and Pisoni (1998) attribute our ability to recognize spoken language to "a period of perceptual adaptation in which listeners learn to differentiate the unique properties of each talker's speech patterns from the underlying intended linguistic message." This process is automatic, involuntary, and effortless. It is only when the idiosyncracies of the talker in speech deviate greatly from that of the listener's dialect, as happens when talking to an individual with a different dialect or accent, that this process becomes difficult and noticeable.

Errors in perception, however, do occur. At some point a talker's speech can deviate so far from a listener's that adaptation becomes difficult. The results are errors in word recognition that lead to miscommunication and slowed processing. Foreign accented speech (FAS) is the prototypical example of speech that deviates so much from the listeners' norm that it cannot be readily understood by the average native talker of a language..

Nygaard and Pisoni (1998; Nygaard, Sommers, & Pisoni, 1994) argue that adaptation to FAS is no more than an extension of the normal process that occurs when conversing with unaccented talkers. In several studies, they have examined the role of indexical talker properties (such as speaking rate, average F0) in word recognition, concluding that indexical properties of talkers are encoded into memory along with non-indexical information (i.e., phonetic), and that this information is helpful in subsequent lexical processing with familiar talkers.

Nygaard et al (1994) familiarized subjects with ten non-accented talkers during nine days of training. During the training sessions, subjects were first presented with single words, and given feedback regarding the name of the talker. They were then presented with a second set of words and asked to explicitly identify the talker. Subjects who were above an arbitrary criterion of 70% on the ninth day participated in a final testing session. Novel words were presented at different signal to noise (S/N) ratios, and subjects were instructed to transcribe these words. In a between subjects design, words were spoken either by the same ten talkers with whom they had been trained, or a different group of ten talkers. Subjects presented with new utterances spoken by familiar voices had consistently higher word recognition rates than

subjects presented with unfamiliar talkers (~51% versus ~42% over all S/N ratios). The authors concluded that with sufficient training, listeners are able to "attend to and modify the specific perceptual operations used to analyze and encode each talker's voice during perception." (p. 45). Somehow, indexical information is utilized in subsequent encounters with a particular talker's voice, helping recognition implicitly. Nygaard and Pisoni (1998) replicated these results, and extended them from the word-length to sentence-length stimuli.

The Nygaard et al studies have important implications for FAS perception. An accented talker would have a great deal of idiosyncratic speech qualities which would be encoded in addition to the phonetic information. In essence, an accented talker has two kinds of variation in his speech: the random individual variation unique to him, and an underlying pattern of phonetic production that is shared with others with a common linguistic background (e.g., dialect). According to Nygaard, this indexical information would aid subsequent recognition of speech spoken by that talker. This is accomplished through an exemplar model of speech perception (Goldinger, 1996; 1998). Each word is represented by many examples, or episodes. When a word is heard, it is first compared with other tokens of that word that an individual has heard. Processing time is directly related to the similarity of the stored episodes to that of the newly encountered token. It is then stored in memory as a distinct episode or exemplar within that word category.

For non-accented speech (NAS), speech perception may be effortless because there are already many examples of similar tokens in memory. However, the word recognition process takes longer for FAS because there are fewer stored examples that closely resemble

the output of the accented talker. The accent is a hindrance to word identification. After an individual is exposed to a large amount of speech from a FAS talker, a variety of speech tokens should be stored in memory. This allows subsequent speech perception to become less labored. Thus, within this theoretical framework, FAS word identification should not be qualitatively different from NAS identification.

A suitable analog to FAS is synthetic speech - speech produced by a computer using algorithms designed to emulate native English phonetic segments. Synthetic speech is designed to create more efficient human/computer interaction, by allowing for a more natural medium between computer and listener. It is analogous to FAS in several respects. First, it is internally consistent. Because the speech parameters are determined by a single algorithm, identical phonemes do not vary from one token to another. This is also true in FAS. Rogers (1997) indicated that phoneme production in accented speech is based on the relationship between the native language (L1) and the non-native language (L2). Consequently, phoneme production in L2 will not vary randomly, but instead remain consistent for a given language. In both the synthetic speech and FAS, listeners can rely on the internal consistency of phoneme production. Second, synthetic speech differs noticeably from natural speech, and takes longer to process. In a lexical decision task (Pisoni, 1981), subjects responded to synthetic tokens approximately 145 msec slower than to the same items in natural, unaccented speech. Similarly, Schmid and Yeno-Komshian (1999) found that reaction time in a mispronunciation task varied as a function of accentedness. The more severe an accent was judged, the lower the performance in detecting errors in pronunciation.

Schwab, Nusbaum, and Pisoni (1985) investigated the improvement in synthetic speech intelligibility over time using a battery of tests. The experiment lasted ten days; the first and last days constituted the testing portion, while the intermediate eight days were learning days. The testing days consisted of five tasks: the Phonetically Balanced (PB) word identification task, the Haskins semantically anomalous sentence task, a prose passage comprehension task, the Harvard sentence identification task, and the Modified Rhyme Test (MRT).

On Day 1and Day 10, the testing days, subjects were presented with blocked groups of the tasks described above. All stimuli were created using the Votrax Type-'N-Talk synthetic speech system. During the intervening training days, subjects were presented with all tests except the MRT. After each trial, subjects received feedback regarding the correctness of their response. In a between subjects design, one third of the subjects received training using Votrax Type-'N-Talk tokens, and one third received natural speech. A third group received no training.

When Day 1 and Day 10 were compared, the Synthetic speech group showed significant improvement for the PB (27% to 69%), MRT (63.4% - 80.3%), Haskins (26.7% - 76.9%), and Harvard tasks (42% - 77.8%). The control groups (natural speech and no training) showed modest gains that were significantly less than the Synthetic speech group. None of the groups exhibited any improvement over time for the prose passages. There are several possible explanations of this. In contrast to the other tasks presented, the prose passages had a much larger memory component than the other tasks. Subjects not only had to understand the speech, but also comprehend what the passages meant, and retain that

information over time. Additionally, because the questions asked were taken from college level achievement tests, the questions were difficult regardless of word intelligibility. However, it is possible that exposure to multi-sentence connected speech may have aided other tasks by providing a richer variety of phoneme combinations. Removing the task may have resulted in a smaller effect of adaptation in the other tasks.

Examining the training data, Schwab et al (1985) found consistent improvement in recognition for the Synthetic Speech group during the course of the two weeks, while the Natural group had consistent ceiling performance. For example, in the PB task, the Synthetic Speech group averaged 35.6% correct over days 2-3, 48.4% correct over days 4-5, 56.3% over days 6-7, and 66.9% averaged over days 8-9. On all days, the Natural group scored above 94%. The results of the other tasks are very similar. In fact, significant improvement could be seen even between the first and fifth days for the Synthetic Speech group in the PB (35.6% versus 48.4%), Haskins (53.7% versus 61.6%), and Harvard (59.7% - 70.0%) tasks. Because the nature of the stimuli did not change over the course of the week, the improvements may be attributed to the listeners themselves.

The benefits of exposure to speech are not solely limited to understanding the talker who produced the speech. The speech of other talkers who share similar qualities is also perceived more easily, and with less effort. Goldinger (1996) found not only that listeners perform better with talkers that they have heard before, but also to novel talkers that are similar to talkers that are familiar to the listener compared to novel talkers that not as similar to learned talkers. Similarity in multi-dimensional space among talkers was measured in a pilot study.

Generalization effects might be more pronounced with FAS because there would be phonetic similarities may be more salient to the listener. If accented talkers share many indexical qualities (i.e., awkward prosody, ambiguous consonant production, etc.), encoding these characteristics for one talker should generalize to others.

Research from several sources suggests that this is likely to be true. Accent is largely based on the relationship of L1 production with L2 production. Flege (1995) hypothesized that accentedness is related to faulty production of position-sensitive L2 vowel and consonant allophones, due to the talker's inability to recognize phonetically relevant contrasts in L2. When a phoneme in L2 is very similar to a phoneme in L1, accented talkers will often produce the L1 phoneme when speaking in both languages. The more dissimilar related L1 and L2 phonemes are, the more likely the critical differences will be discerned by the talker and subsequently produced correctly. For instance, if the voicing boundary between /b/ and /p/ is close to /b/ in an individual's native language, but close to /p/ in L2, the talker may produce the bi-labial stop consonants in L2 in the manner that is appropriate for L1. Individuals who have a common L1 will have similar confusions in L2 production, and consequently similar accents.

When native English speakers rated the pronunciation of talkers with a variety of accents, Sutter (1980) found that the most predictive determinant of accuracy was L1 itself. Arabic and Farsi talkers were consistently rated as less accurate than Japanese and Thai talkers who had similar L2 experience (years of instruction, age of learning). Consistent with Flege (1995), Sutter noted that each L1 causes different mistakes in L2 pronunciation and consequently causes varying degrees of difficulty in speech intelligibility in L2 as a function of

L1. As Rogers (1997) points out, accentedness is based on the relationship between L1 and L2. If the properties of the accent that deviate from the dialect of the listener are not linguistically relevant (i.e., don't affect critical features), accent will not impede intelligibility. It is only when an linguistically meaningful sound is ambiguous that errors occur.

Rogers (1997) acknowledged the impact of L1 on L2 production, and focused on the causes of accent for a single L1. If FAS is caused by the phonological relationship between L1 and L2, examining the relationship between phonemes in L1 and L2 should help to predict probable errors in L2 production. Careful phonetic analysis was conducted on the English speech of two native Mandarin talkers in order to ascertain which phonemes would likely cause errors for native talkers of English. Certain phonemes were consistently mispronounced, either by confusing the place of articulation (i.e., /d/ became /b/), manner of articulation (/p/ became /f/), or voicing (/b/ became /p/). Confusions in vowels were common and somewhat consistent as well. Examining the phonology of Mandarin explained at least some of the misperceptions exhibited by English-speaking listeners. For example, there were more errors for consonants in word-final position, perhaps due to the fact that Mandarin has a limited number of allowable word-final consonants. Subsequent tests on native English subjects found that mistakes occurred for phonemes in L2 that were related to phonemes in L1. The pattern of mistakes was similar over several Mandarin talkers, reinforcing the claim that listeners adapt to the idiosyncracies of an accent.

It is fairly clear from Rogers (1997), Sutter (1980), and Flege (1995) that phonetic production of FAS, being determined largely by L1, is consistent within an accent. Thus, it

should be the case that exposure to an individual with an accent will facilitate speech perception for individuals who share a common native language, and thus a similar accent. The current research addresses not only the questions of FAS adaptation, but also of the generality of that adaptation, by manipulating the amount of exposure to an accented talker, and then testing intelligibility of both similarly and differently accented talkers.

The current experiments borrow from the design of Schwab et al (1985). The same five tests of intelligibility - PB, MRT, Haskins, Harvard, and Prose - were used. The length of the experiment was five days, with two testing days and three intervening training days. Instead of the Votrax synthetic speech, an accented talker (M1) was recruited to record the stimuli. A similarly accented speaker (M2) and a differently accented speaker (R1) were also recruited.

If Nygaard et al (1994, 1998) are correct, training on M1 should result in the encoding of the indexical properties of M1's voice, both his dialect and his individual idiosyncracies. This exposure should improve subsequent intelligibility with this speaker. Furthermore, this encoding should also result in improved performance for M2, who shares many qualities in common with M1 due to the common native language. However, this benefit would not extend to R1 because R1's accent is unrelated phonetically; any benefits over time would be modest, and due to familiarity with the design.

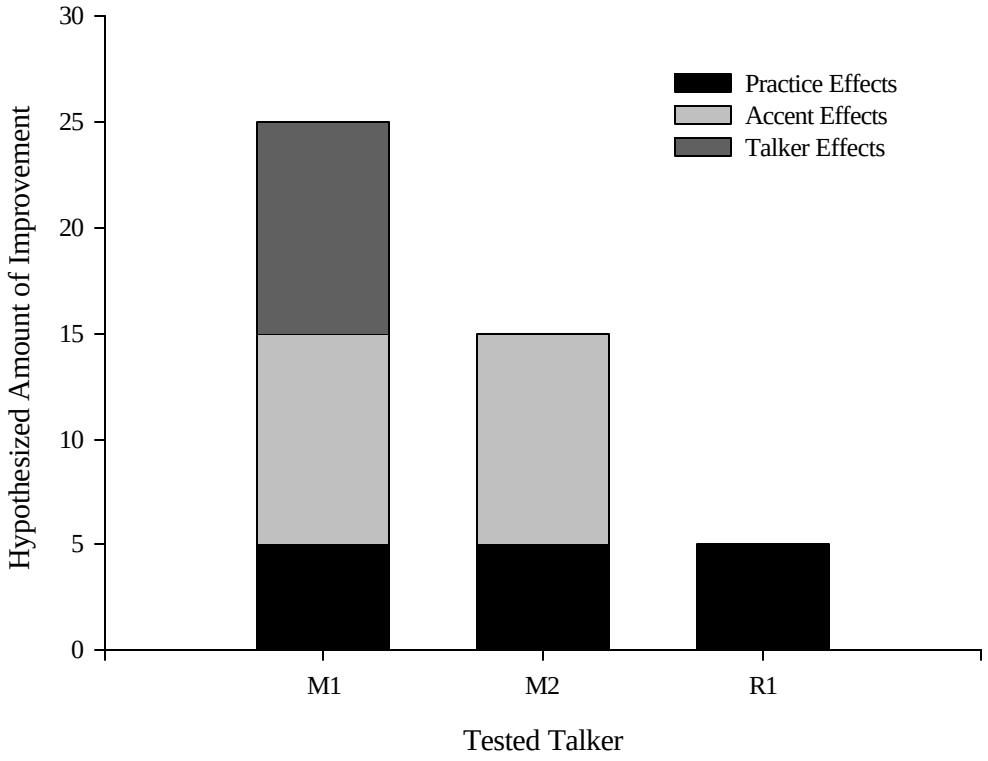

Figure 1.1: Benefits of M1 Training on Subsequent Testing

Figure 1.1 illustrates the predicted benefit of M1 training onto subsequent testing of M1, M2, and R1. Exposure to M1 will have the greatest effect on M1 testing because both talker and accent characteristics are the same from training to testing. Exposure to M1 will lead to some benefit in M2 testing because M1 and M2 share a common accent; the benefit of generalization should be a function of the similarity in accent. If two voices have different

accents (i.e., M1 and R1), this effect should be negligible, and improvement due only to practice with the tasks employed. These effects may be additive; the more similarity between speakers, the more benefit. Task familiarity will cause improvement in all cases, and talker and accent effects should combine to aid perception.

In essence, the exposure to the accented talker results in two different effects. A talker effect, where exposure to a specific voice facilitates subsequent word recognition of that talker due to the encoding of idiosyncratic properties. A second effect is an accent effect that should occur because the listener is simultaneously encoding talker characteristics which are shared within a dialect. This should facilitate comprehension for all members of this dialect group.

CHAPTER 2

EXPERIMENTATION

Method

Listeners.

One hundred seventeen Ohio State University undergraduates participated in this experiment. All were native monolingual speakers of English with no history of speech or hearing dysfunction. The majority of participants reported being brought up in Central Ohio, while the remaining subjects were raised in adjacent areas. Participants did not report significant exposure to accented speech in general or to the specific accents utilized. Ten participants did not complete the full experiment, and will not be considered in any subsequent discussion. Forty-three participants were tested for five consecutive days ("Training" Group), while 64 participants were tested in a single session ("No Training" Group). All participants received course credit for their attendance.

Accented Talkers.

Three accented talkers were recruited from the Ohio State University community. Talkers were chosen by the similarities in their linguistic backgrounds. All three had spent relatively little time in the United States, and no time while they were first learning English. Two talkers (M1 and M2) share a common L1 (Marathi) as well as a common third language (Hindi). It is assumed that the accents of M1 and M2 will share many phonetic properties (Rogers, 1997). The third talker (R1) was a native Russian speaker. Table 2.1 summarizes the talkers' biographical information.

The two languages chosen, Marathi and Russian, are relatively unrelated from a historic standpoint; Marathi is a member of the Indo-Aryan language family, while Russian is a member of the Slavic language family. Marathi is characterized by a number of retroflex consonant minimal pairs (/š/ versus /t/, /`/ versus /d/, /}/ versus /l/, /"/ versus /s/), and aspirated consonants (/t^h/ versus /t/, /k^h/ versus /k/, /d^h/ versus /d/, /$š^h$/ versus /š/), as well as both tapped (/•/) and trilled (/r/) alveolar consonants (Jha, 1977). Russian does not contain retroflex or aspirated consonants, but does have several pairs of nonpalatalized/ palatalized consonants (/f/ versus /f^j/, /t/ versus /t^j/, /s/ versus /s^j/, /z/ versus /z^j/, /l/ versus /l^j/, and /k/ versus /k^j/; Halle & Jones, 1959). In consequence, the Marathi accent has more pronounced aspiration and trilled alveolar consonants (in place of the American English alveolar approximate, /<). The Russian accent has more palatalized consonants.

Recording and Materials.

All stimuli were recorded in a sound attenuated booth using a head mounted Crown CM 311A Differoid Condenser microphone. Stimuli were recorded onto Digital Audio Tape (DAT) using a Tascam DA-30 MKII recorder, and then transferred to an IBM compatible PC to be converted to .wav files (Channels: Mono; Frequency = 16,000 Hz; 16 bits) and edited. Speech tokens were normalized to achieve a comfortable average intensity level, and notch filtered at 60 and 120 Hz to remove extraneous ground noise.

Talker	Age	Native Language	Other Languages	Age of U.S.Arrival	Age of First English Instruction
M1	26	Marathi	Hindi	25	6
M2	26	Marathi	Hindi	24	3.5
R1	21	Russian	n/a	17	7

Table 2.1: Biographical information for accented talkers.

Sentence and paragraph tokens were edited to remove extraneous hesitations, stutters,

corrected mistakes, and environmental noise. All sound editing was accomplished using Cool Edit 2000 software (Johnston, 1999). Signal correlated noise (SCN) was added to stimuli used in the pre-test and post-test by randomly altering the sign of each sample in each .wav file. SCN was used because it provides a level of distortion that changes as a function of the intensity of the speech. Thus the signal to noise ratio is consistent throughout the signal.

Design.

Six between-subjects groups were tested in a 2 x 3 design (Experience x Day 5 Talker). There were two levels of Experience, "Training" and "No Training." Training groups received a pre-test (Day 1) and three days of training on M1 before completing a post-test on Day 5. No Training groups were only presented with the post-test. This post-test was identical to the Day 5 post-test that the Training groups received. The Day 5 talker was either the trained talker (M1), a similarly accented talker (M2), or a differently accented talker (R1). All Training groups received training on M1 tokens, regardless of the talker presented on Day 5. Table 2.2 illustrates these groups.

Procedure.

Stimuli were presented to participants via Sony MDR-V900 Dynamic Stereo Headphones at a comfortable volume level. Training groups were tested at the same time of day for five consecutive days, while No Training groups participated in only one session (Day 5). Within each task the stimuli used on successive days were completely different; no words

were repeated. This eliminated the participant's ability to remember the peculiarities of specific tokens while still presenting them with the full range of the characteristics of the accent. The stimuli were the same between groups of participants; every group heard the same stimuli on a particular day. Participants were tested in individual sound attenuated booths, and responded to all tasks in individual score books. Listeners were tested in groups of 3 or 4.

	Talkers		
Experience	Days 1-4	Day 5	n
Training	M1	M1	14
	M1	M2	15
	M1	R1	14
No Training	none	M1	22
	none	M2	21
	none	R1	19

Table 2.2: Conditions in current experiment.

Testing:

For the Training groups, testing occurred on Day 1 and Day 5. Testing procedures and order were the same for both the pre-test and the post-test (Table 2.3). Each testing day consisted of five tasks. Subjects were instructed to leave their booths and meet in an adjoining common room after each task was completed in order to receive instructions for the next task. The order was: PB, Haskins, prose passages, Harvard, and MRT.

The Phonetically Balanced (PB) lists (Egan, 1948) are each made up of fifty isolated mono-syllabic words. The frequency of phonemes in a single set is approximately proportional to the frequency of that phoneme in the language at large. The words were not related semantically. After hearing the word, participants wrote down the English word they heard. If they were uncertain, they were instructed to "make their best guess based on what they heard." One set (50 words) was presented on each testing day. Each trial lasted 10 seconds. In each session, words were presented randomly.

The Haskins Sentences (Nye & Gaitenby, 1974) are sets of ten semantically anomalous sentences, each containing four key words. All sentences were presented in the form "The [adjective] [noun] [past tense verb] the [noun]." The score sheets had ten lines which read "The _____ _____ _____ the _____." For each sentence, they were instructed to fill in the four blanks with the key words. Subjects were informed of the anomalous nature of the sentences beforehand, and instructed to "pay extra close attention." One set (10 sentences) was presented to the listeners on each testing day. Each trial lasted 25 seconds. Sentences were permuted randomly for each session.

Task	Response Type	Responses	Dependent Measure
1 PB List	Free Transcription	50	Correct Transcription
1 Haskins List	Free Transcription	40	Correct Transcription
4 Prose Passages	True/False	20	Correct Comprehension
1 Harvard List	Free Transcription	50	Correct Transcription
2 MRT Lists	6 AFC	100	Correct Identification

Table 2.3: Stimuli presented for pre-test (Day 1) and post-test (Day 5)

Short prose passages were taken from several sources (Ekwall & Shanker, 1993; Flynt & Cooter, 1998; Woods & Moe, 1989) designed for reading assessment for the middle school grades. After each passage, five true/false questions based on the passages were presented to the participants via a 15" Black and White monitor. Participants were instructed to respond by circling either “T” or “F” on the score sheet. They had ten seconds to answer each question before the next question was presented or the next passage began. The questions were written by the author, and were designed to be obvious if the passage was understood. Four passages were presented on each testing day. Each passage was approximately one minute long.

Listeners had ten seconds to answer each T/F question. The order of the story and question presentation was randomly permuted for each session.

The Harvard Sentences (IEEE, 1969) are sets of ten sentences, each balanced phonetically to reflect the frequencies of phonemes in the language. Each sentence contained five key words in addition to function words. Subjects transcribed the entire sentence. Sentences were meaningful, and had relatively complex syntactic structures. One set (ten sentences) was presented on each testing day. Each trial lasted 25 seconds. Sentences were permuted randomly for each session.

The Modified Rhyme Test (MRT; House, Williams, Hecker, & Kryter, 1965) consists of fifty groupings of six words. Participants were presented a single word aurally, and the six choices in the set were presented visually in a six-alternative forced choice task. Each set consists of words that deviate from each other by a single phoneme. The changes occurred in word initial and word final position equally often. For example late, lake, lay, lace, lane, and lame compose one complete set. Participants responded by circling the letter in their scoresheet corresponding to the word they had heard. Two sets (100 words total) were presented for the pre- and post- tests (Day 1 and Day 5). The order of the words was random for each session. The order of the six alternatives was random for each trial. Each trial lasted ten seconds.

On Day 5, a biographical questionnaire was presented to the participants after all tasks were completed.

For all tasks, the stimuli used on each day of testing and training were novel; within

each task, word and sentences were presented only once during the duration of the experiment. No feedback was given regarding the correctness of the listener's responses. After presentation of the stimulus, listener's had several seconds to transcribe the word or sentence before presentation of the next trial. The length of time to respond was 8 seconds for the PB and MRT task, 25 seconds for the Harvard and Haskins task, and 10 seconds for each True/False question in the prose passages. Listener's reported to difficulty responding in the allotted time.

Training:

Half of the groups participated in three days of training prior to the post-test. Procedures for the training sessions (Days 2-4) were similar to those of the testing sessions, with exceptions. To ensure exposure to the accent, stimuli were not presented with SCN. A pilot study showed that this improved intelligibility to ceiling levels. Because the full set of MRT groups is small, the MRT task was not presented, but saved for the testing phase. To enrich the listener's exposure, five prose passages were presented instead of four. All other aspects of the procedure were identical to the testing days.

Results

For all tasks in which the response was transcription, correct responses were those which exactly matched the intended utterance of the speaker. Responses that were homophones of the intended word ("earn" versus "urn") were judged to be correct even when they were semantically inappropriate (i.e., the Harvard sentences). In some cases, words that were near homophones ("hawk" versus "hock") were scored as correct because of their similarity in the dialect of the listener. Misspelled words were judged to be correct if the intention of the listener was clear (e.g., "knaw" for "gnaw"). Unless otherwise noted, all comparisons were significant at the $p < .05$ level.

The lists used for each task were created to be uniformly balanced. If this is true, the average score on each task should remain constant if the amount of experience and speaker are held constant. Comparing the Day 1 pre-test for the M1 Training group to the Day 5 post-test for the M1 No Training group for each task allows the necessary comparison because it is the same talker in both conditions, and listeners have no experience in either case. The mean proportion correct (PC) was not significantly different for the PB task (Day 1 = .32 versus Day 5 = .33), t = -.429 (ns), for the Prose task (.73 versus .78), t = -1.76 (ns), for the Harvard task (.74 versus .77), t = -1.01 (ns), or for the MRT (.78 versus .80), t = -1.03 (ns). There was a significant difference in the mean for the Haskins task (.60 versus .52), t = 2.504. The Haskins task did show a difference in baseline intelligibility. The effects of training on performance in the Haskins task must take this inequality into account. Overall, these results suggest that the lists did not differ in baseline intelligibility.

For each task, a 2 x 3 ANOVA was performed (Training x Day 5 Talker) comparing the PC for Day 5 to assess the overall effects of talker and training. A main effect for Training would indicate a difference in performance between groups with several days of experience versus no experience. It is hypothesized that having training would generally increase performance, regardless of talker. This would be caused by both practice effects and the talker/accent effects of interest. A main effect for Day 5 talker would indicate that words spoken by specific talkers are recognized with different degrees of success. Because talkers

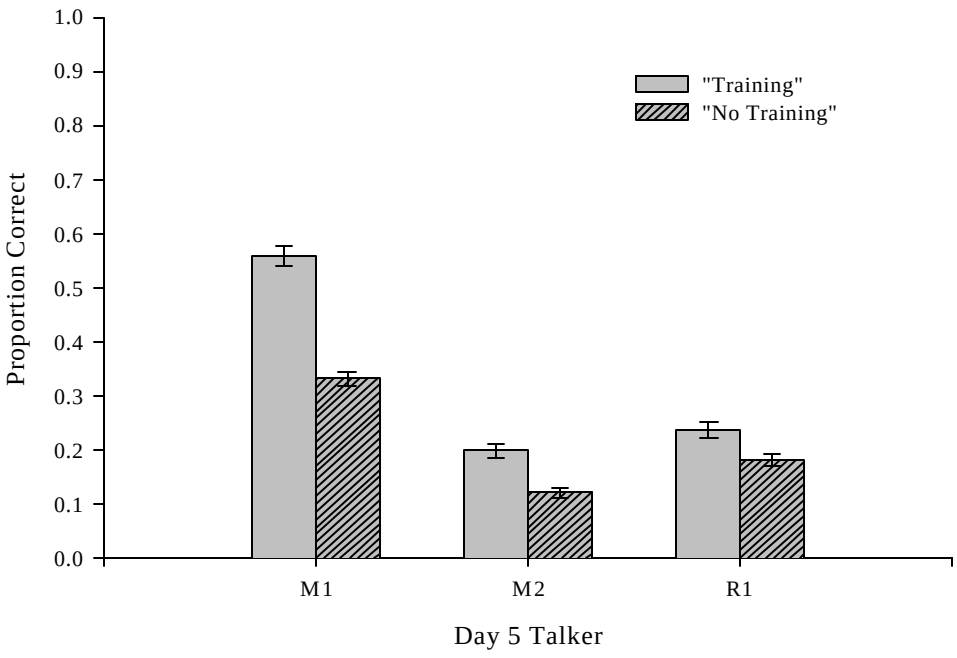

Figure 2.1: Day 5 PB task means.

differ from each other randomly, it is expected for the speech of some talkers to naturally be more intelligible than others independent of accent. The interaction of Training and Day 5 Talker is much more telling. For all tasks, listeners should have better performance on Day 5 when they are responding to the talker on which they had been trained (i.e., M1). If accents are generalizable, listeners should also have better performance on Day 5 when they are responding to a talker who has an accent similar to the one on which they had been trained (i.e., M2). The data from each task are discussed separately.

PB Task.

For the PB task (Figure 2.1), a 2 x 3 ANOVA was performed (Training x Day 5 Talker). There was a main effect of Training, $F(1, 101) = 129.15$ and of Talker, $F(2, 101) = 274.75$, as well as a significant Training x Talker interaction $F(2, 101) = 24.871$. However, these main effects and interactions do not adequately describe the effects because they average over simple comparisons. The most informative measure can be shown graphically by noticing the difference between each pair of bars in Figure 2.1. This is the disparity in performance between Training and No Training listeners for each speaker. The difference between the M1 groups (.22 difference in performance) was much larger than the other two sets of groups (.08 for the M2 groups and .06 for the R1 groups). This may imply that while performance benefitted due to practice regardless of speaker, there was additional benefit when the training talker was the same as the testing talker. This is fundamentally the talker effects described by

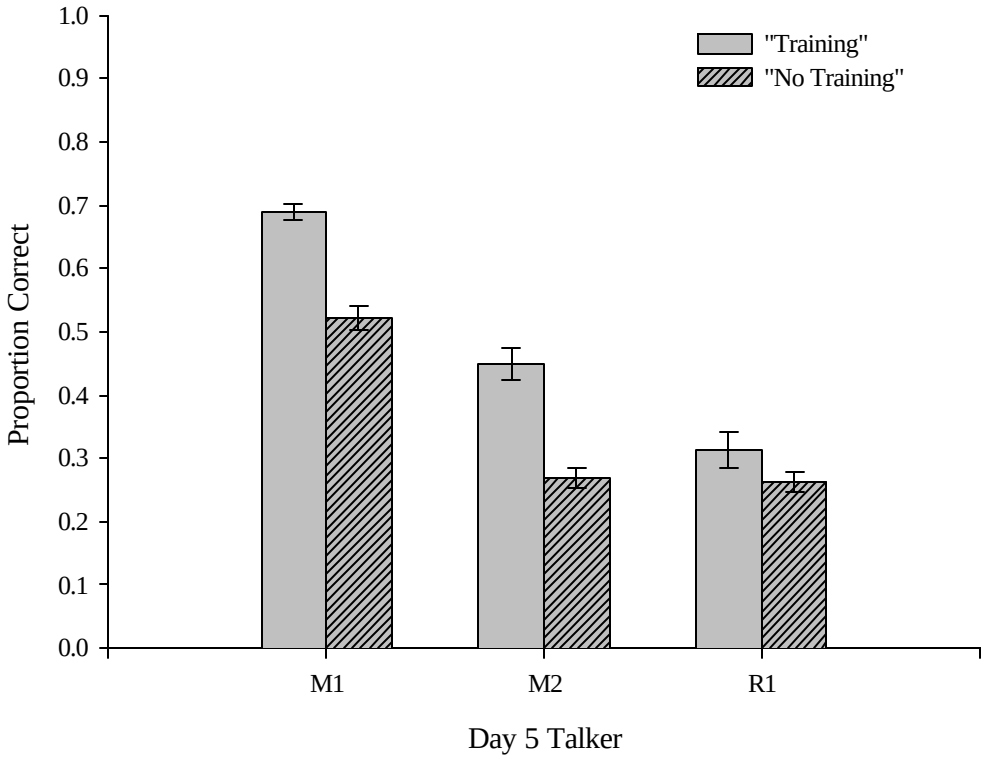

Figure 2.2: Day 5 Haskins task means.

Nygaard, Sommers, and Pisoni (1994); higher relative performance for familiar talkers. However, no generalized accent effects were found, as would have been indicated by a larger difference between the groups tested on M2 relative to the those tested on R1. This indicates that no significant benefits accrue to M2 intelligibility due to M1 Training.

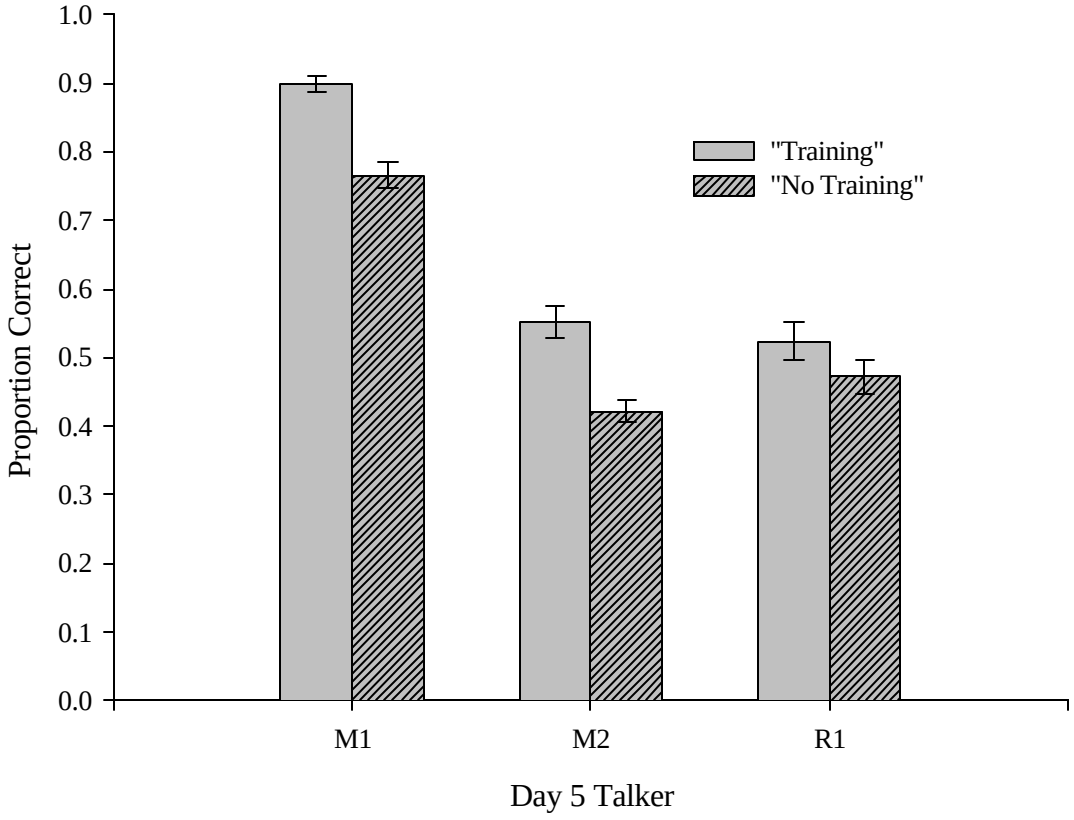

Figure 2.3: Day 5 Harvard Sentence task means.

Haskins Task.

A 2 x 3 ANOVA was performed (Training x Day 5 Talker) for the Haskins task (Figure 2.2). There was a significant main effect of Training, $F(1, 101) = 65.16$, of Talker, $F(2, 101) = 135.72$, and a significant interaction $F(2, 101) = 6.25$. As with the PB task, a

significant overall interaction emerges, but in this case it revealed a different underlying pattern. The differences for the M1 and M2 groups were both .17, while it was only .05 for R1 groups. In this task, training on M1 seemed to affect M2 as well as M1 intelligibility, an indication of generalization from one speaker to another. The modest improvements in the R1 condition can be attributed to practice effects.

Harvard Task

A 2 x 3 ANOVA (Training x Day 5 Talker) was performed (Figure 2.3). There was a significant main effect of Training, $F(1, 101) = 35.31$, and of Talker, $F(2,101) = 163.06$, although there was no significant interaction, $F(2, 101) = 2.3$. Although no significant interaction was found, an examination of the differences reveals a pattern which is very close to the pattern in the Haskins task. Differences for the M1 and M2 groups were similar (both ~ .13), while the R1 difference was lower (.05).

MRT Task.

A 2 x 3 ANOVA (Training x Day 5 Talker) was performed for the MRT[1] (Figure 2.4) task. There was a significant main effect of Training, $F(1, 99) = 5.8$, and of Talker, $F(2, 99) = 173.118$, but no significant interaction, $F(1, 99) = 1.00$. Differences in the MRT groups were very low (M1 = .05; M2 = .02; R1 = .01) while performance was high (.67 over all conditions). However, because the response set was limited, chance performance was

[1]Two participants were disqualified from this analysis because they had outlying means (PC > 80%)

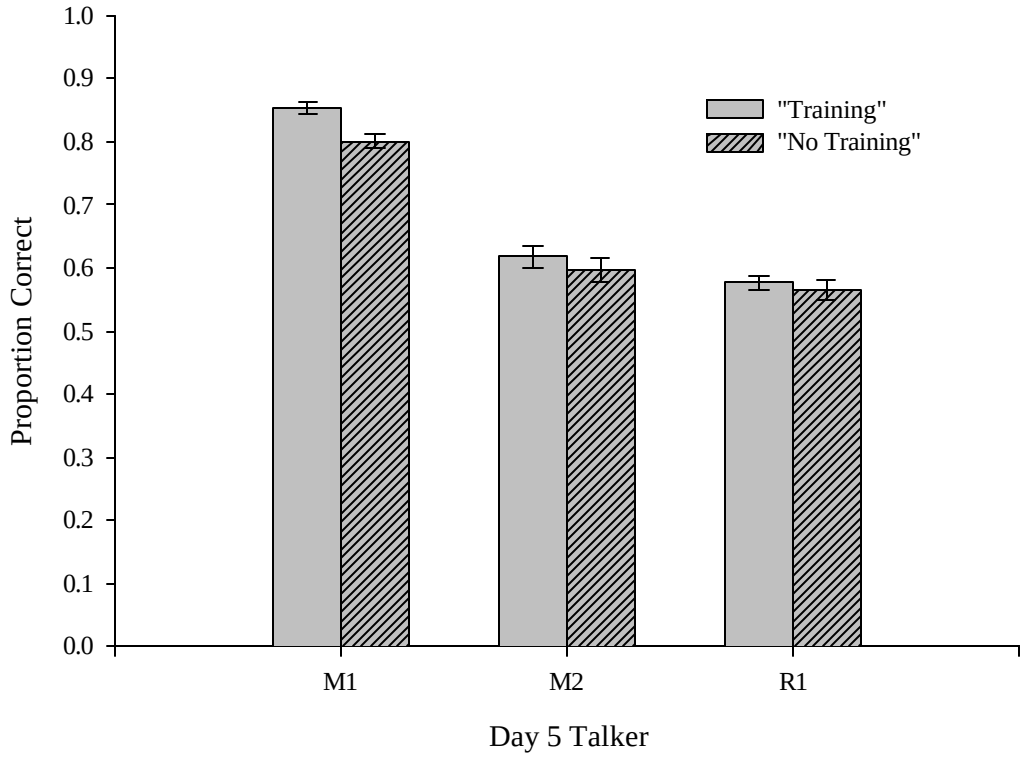

Figure 2.4: Day 5 MRT task means.

16.66%. This had the effect of increasing performance, lessening the effects of training. Nonetheless, the trend - higher M1 difference compared to similar M2 and R1 differences - is similar to the PB task.

Prose Passages.

For the Prose passages, there was a significant main effect for Talker, $F(2, 101)$ =

16.69, but not for Training, $F(1, 101) = .04$. There was no significant interaction, $F(2, 101) = 2.00$. Examination of the training data revealed that performance in this task varied as a function of the stories and questions used rather than talker factors. In fact, performance was similar between testing and training days, despite the addition of SCN. However, as in Schwab, Nusbaum, and Pisoni (1985), the exposure during training may have helped intelligibility in other tasks. Differences in performances on Day 5 were negligible.

There is evidence to suggest that listeners can adapt to an accented talker with experience, and that the adaptation generalizes to similarly accented talkers in some cases. All tasks except the prose passages showed a significant main effect of training. Listeners who participated in the three days of training had higher proportion correct in all conditions compared to the No Training groups. This was true regardless of the Day 5 talker, although the degree of the difference was not always the same among M1, M2, and R1 Training groups. Practice effects, adaptation to the talker, and adaptation to the accent all contributed to this improvement.

A significant main effect for talker was found in all tasks. Talker M1 yielded higher performance than M2 or R1 in every task, for both levels of training. The other two talkers, M2 and R1, were generally equivalent in terms of intelligibility between tasks. This disparity reflects the natural variability in talker intelligibility that is independent of accent. Because the talkers were not chosen on the basis of subjective intelligibility ratings, but instead by their linguistic experience, it is not surprising that some variability occurred in this study. Examining the talkers biographical information, talker M1 had some training in public speaking in his native

tongue, while the other talkers did not. This could have contributed to the main effect of talker.

The difference between the M1 Training and No Training groups is a between subjects measure of the effects of training on performance. It is a test of talker adaptation: exposure to an individual should improve subsequent performance. However, a within subjects comparison was available as well. To assess the degree in which experience (Day 2 - 4) changed performance on the test in noise, performance on each task from the pre-test (Day 1) and post-test (Day 5) were compared for the Training group which was both trained and tested on M1.

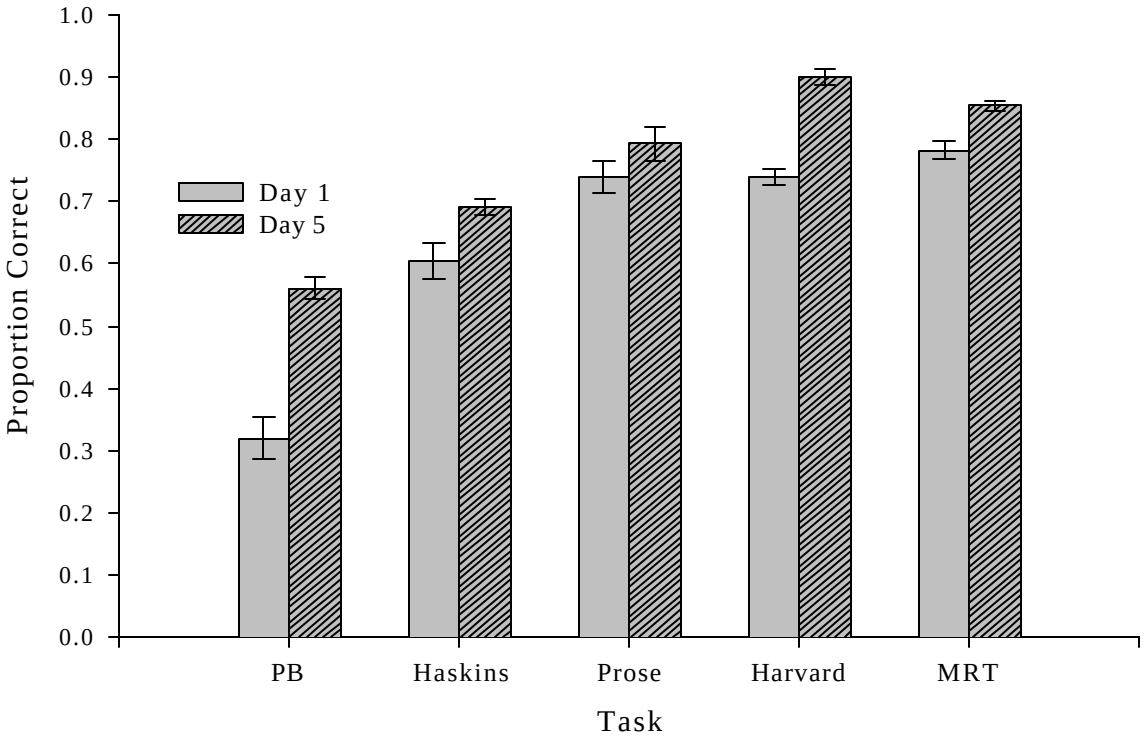

Figure 2.5: Day 1 Versus Day 5 for M1 "Training" Group
All Tasks

They are the only group who participated in the entire experiment, and had the same talker for both pre- and post-tests. Figure 2.5 illustrates the results for each task. A repeated measures analysis of variance (ANOVA) was performed for each of the 5 tasks. For the PB task, mean proportion correct (PC) changed from .32 on Day 1 to .56 on Day 5, $F(1, 13) = 46.63$. For the Harvard task, mean PC changed from .73 to .90, $F(1, 13) = 201.21$. For the MRT, mean PC changed from .78 to .85, $F(1, 13) = 32.51$. For the Haskins task, mean PC changed from .60 to .69, $F(1, 13) = 12.72$. The mean PC did not change considerably for the prose passages, $F(1, 13) = 3.8$ (ns). These results suggest that subjects were able to improve with experience with a particular accented speaker.

CHAPTER 3

GENERAL DISCUSSION AND CONCLUSION

Anecdotal evidence suggests that perception of FAS improves over successive encounters with an accented speaker. The first time an individual meets an accented speaker, speech perception is labored and faulty. After several sessions with the speaker, less mistakes in speech perception are made and the speaker is more readily understood. After large amounts of exposure, an accent can become unnoticed. Intelligibility of similarly accented speech also seems to benefits from this exposure.

This pattern repeats itself in many contexts; at school, in business, and in international relations. Miscommunication due to mistakes in accent perception can lead to minor annoyances, or have grave consequences in life or death situations where communication is vital (Scott, 2000). Although the effects of adaptation seem straightforward, accent adaptation has not been explained experimentally.

The results of the current investigation indicate that adaptation to FAS does occur. Listeners who were tested before and after three days of exposure to a single talker showed

significant increases in performance. Comparing Training and No Training groups on a single test provided similar results. Given exposure to a talker, subsequent speech perception improves. This extends previous research on talker variability (Nygaard & Pisoni, 1998; Nygaard, Sommers, & Pisoni, 1994; Goldinger, 1996; 1998) which has shown implicit memory for talker identity in word identification and recall tests. Words spoken by familiar talkers are identified more quickly than words spoken by unfamiliar talkers. This effect is magnified when the talkers are accented.

The generality of this adaptation differed from the hypothesized results. Recalling Figure 1.1, it was hypothesized that the benefit would be approximately additive; practice effects, talker effects, and accent effects were all expected to benefit intelligibility independently. Training on M1 should lead to encoding of both accent and individual characteristics for that speaker (M1), which will lead to a general benefit in M1 test intelligibility. Testing on M2 would only show the benefit of accent effects, not talker effects, because those are the characteristics that the two talkers share. Testing on R1 would only exhibit practice effects, the modest benefit due to exposure to the task.

The actual results did not exhibit this additive effect. Exposure to M1 did lead to improvement in M1 testing in all conditions (~13%), and to a very modest benefit for R1 (~5%). However, M2 testing changed as a function of the context provided by a given task. When the task was word transcription, M2 did not benefit from M1 training at all. The benefit of exposure to M1 for M2 perception was similar to the benefit of M1 exposure for R1 perception.

Conversely, when the task was sentence transcription, M2 benefitted from M1 training. In fact, the benefit of M1 training was as large in the M2 condition as in the M1 condition.

It is easy to see the benefit of M1 exposure by examining the difference in performance between the Training and No Training groups. Taking the difference negates the main effect of Talker, which varies randomly independently of accent. Figures 3.1 and 3.2 show these

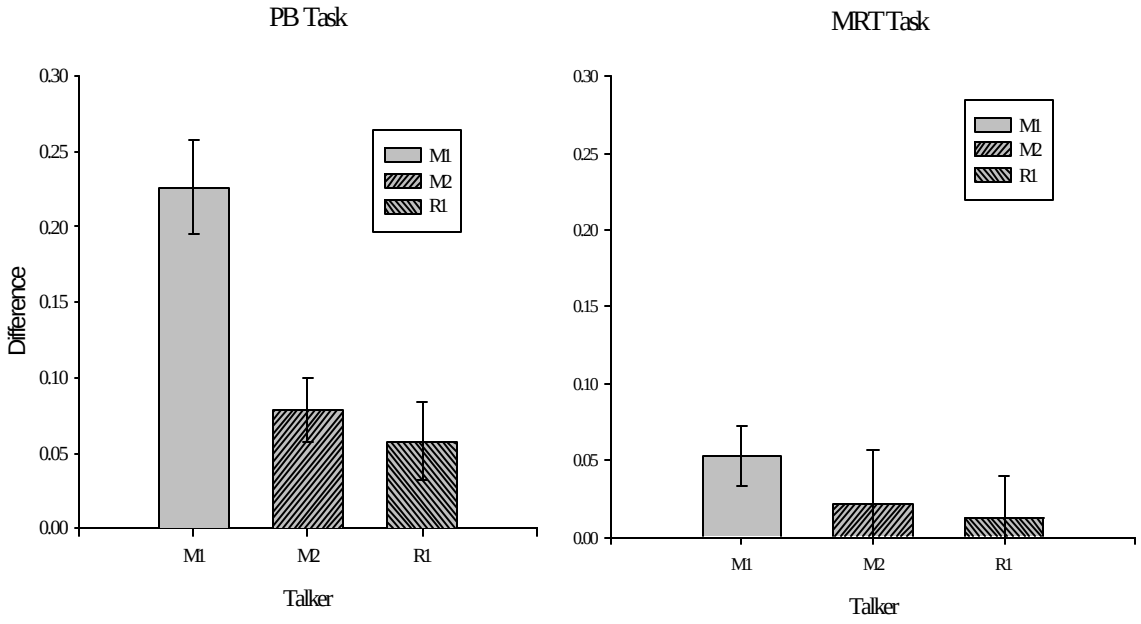

Figure 3.1: PB and MRT task differences.

differences for each Day 5 post-test talker in each task. In the PB and MRT tasks (Figure 3.1), training in M1 benefits only M1 testing. There is no advantage when listening to a similarly accented talker. These tasks both present the listener with a single word and then ask either for

transcription (PB) or for recognition (MRT). Although the differences between the groups in the MRT task is small, there does seem to be a general trend that is similar to the PB task; high M1 differences compared to similar M2 and R1 differences.

The Haskins and Harvard tasks show a different pattern (Figure 3.2). Exposure to M1 benefitted both M1 and M2 intelligibility equally. This is different than predicted; the difference in the M1 groups was expected to be larger than the difference for the M2 groups. Two

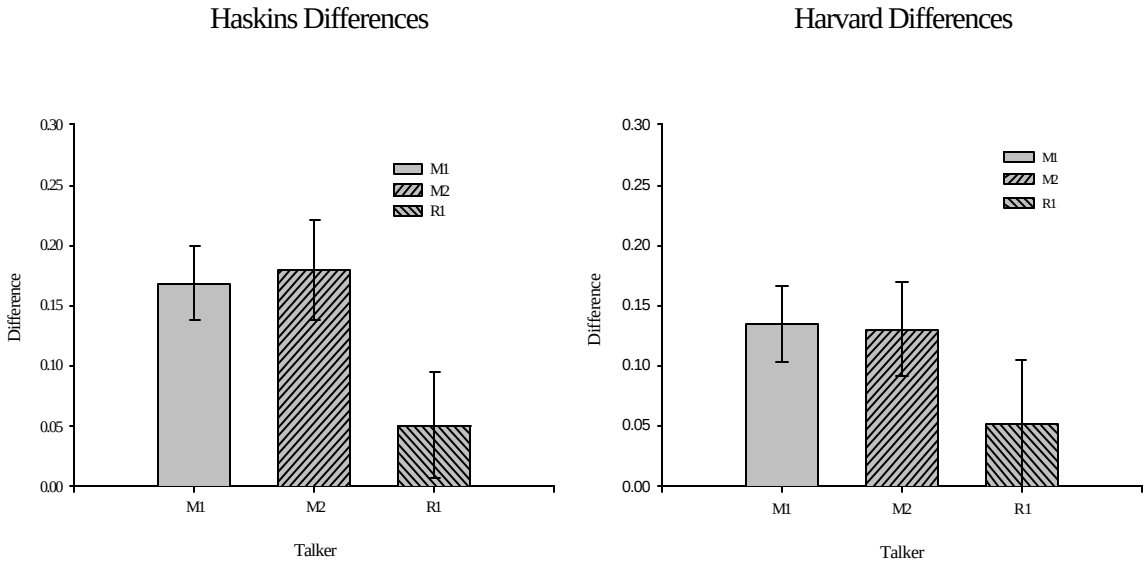

Figure 3.2: Haskins and Harvard task differences.

commonalities between these two tasks were the type of stimuli and type of response; both presented listeners with whole sentences, either anomalous or regular, and required transcription. Notice that the difference for R1 is comparatively smaller than

those of M1 and M2, and are consistently about 5% for the Harvard, Haskins, and PB tasks. This modest improvement is most likely related to simple practice effects.

The patterns of differences found in both the sentence and word level tasks are enigmatic in two respects. First, if the accent effect occurs, why would it occur only in the larger sentence context? Perhaps the level of processing (sentence versus word level) leads to the divergent results, due to memory demands in the longer sentence task. But if the underlying encoding is phonetic, the level should be immaterial. One could argue that the SCN may be the culprit; listeners may need a certain amount of time to adjust to the noise before they can attend to the signal. If this were true, performance in the word tasks should be much lower regardless of training because the initial phoneme would be difficult to recover. This is not the case. Perhaps contextual cues that come from the larger sentence context are all that is needed for improved performance. If this is true, the benefit is robust enough to overcome the semantic ambiguity of the Haskins task. Perhaps transcription is not sensitive enough to show the underlying pattern, and reaction time would reveal a different pattern.

Second, when the accent effect does occur, as indicated by high M2 benefit, why aren't the talker effects more obvious? The benefit of M1 exposure on M2 perception was as strong as the benefit of M1 exposure on M1 perception in the Haskins and Harvard tasks. Listeners seem to pick up on accent information exclusively, ignoring any additional talker information. This is similar to models of lexical access that propose that all indexical information is stripped away, leaving the pure linguistic form (Summerfield & Haggard, 1973). Does the

word recognition system choose which talker characteristics are important for the task? This seems unlikely.

One possible partial explanation of these results is the role of prosody on speech processing. If prosodic aspects of speech (pitch contour, amplitude contour, speaking rate) are encoded into memory, they may affect tasks of different lengths differentially. This is due to the inherently temporal nature of prosodic characteristics. For example, if M1 and M2 had similar pitch contours when they speak, this similarity may be more salient over a longer presentation (i.e., the sentence tasks) compared to a relatively impoverished speech sample (i.e., the word tasks). Thus the benefits of M1 for M2 processing would be more pronounced for the sentence tasks.

These enigmatic patterns notwithstanding, the overall finding that learned talker characteristics improve intelligibility for similar talkers is important for models of speech perception. It confirms the hypothesis that indexical talker characteristics are utilized in perception, but also suggests that listeners use all available resources (i.e., similar speakers) to recognize unusual speech patterns, such as FAS. The contextual dependance of this effect may imply a fundamental difference in the way sentence and word length perception is achieved.

Implication for Models of Lexical Access.

The results of this study have implications for models of speech perception. The studies by Nygaard et al (1994; 1998) and Goldinger (1996, 1998) are based on an episodic model of

lexical access, a model that is specifically designed to account for talker effects. The viability of this model can be tested by predicting its reaction to FAS tokens.

Goldinger (1996, 1998) describes a model of word recognition that is based on Hintzman's (1986) MINERVA 2 model of episodic memory. The model is an exemplar model; Goldinger assumes that every time a known word is heard again, that new token is added in memory along side all previous tokens of that word. Many copies of each known word are retained in memory.

The model does not assume to capture the intricacies of human perception, but instead endeavors to capture the underlying manner of lexical access. For each word category in memory, there are many traces, or different versions of that word. When a new word token, the probe, is encountered, it is compared to each trace, by comparing many smaller features. Each trace is activated to some degree by relating the similarity of the probe to the trace. The similarity activations for each trace are themselves summed to determine the intensity of the echo - the measure of category membership. The more traces that are similar to the probe, the higher echo intensity will be. If the echo is intense enough, the probe is identified as being a member of that category, and is added to the traces in that category. Thus, "non-linguistic" information is not discarded, but encoded into memory as the features of an episode - a trace - to which subsequent word tokens are compared. When Goldinger (1998) compared experimental results to simulated results using MINERVA. The graphs were nearly identical. These results were replicated over various types of shadowing tasks, and multi-speaker tokens (Goldinger, 1996)

Theoretically, if an accented word is presented to this model, the probe will differ from all of the other traces on several features. Because of this difference, initial performance may be slow, and often incorrect. Assuming correct identification, when the accented probe is compared to the traces in memory, it may take several comparisons for echo intensity to rise to a critical level, identifying the word. The next presentation of the same word should lead to vastly improved performance, because new accented traces will be compared with the accented trace in memory, causing an increase in echo intensity. The more accented tokens with similar characteristics in memory, the faster and more reliable lexical access will be. When a new talker with a familiar accent is encountered, there are already traces with which to compare it. Thus, initial lexical access will not be disrupted to the same degree as the first speaker. Exposure to one accented speaker can generalize to others, if the words spoken share similar features.

Goldinger theorizes that episodes of words are stored, not smaller units such as syllables or phonemes. This is where the model is incompatible with the current data and with Nygaard, Sommers, and Pisoni (1994). If words are stored divorced from the phonetic properties, only words which have been heard before will benefit from exposure to an individual. In both studies, talker or accent effects were found for novel words as well as previously encountered words. Words that share phonemes must benefit from previous exposure to other words that contain those phonemes. This does not eliminate an episodic model as viable; if heard words are decomposed into their constituent parts - syllable or phoneme size objects - these units could be stored as exemplars. This would also reduce the

amount of space needed to retain the episodes; there are a finite quantity of syllables or phonemes, but an infinite number of words.

Future Research.

The current study found experimental evidence for adaptation and generalization of FAS. In the immediate future, the relative importance of talker and accent effects needs to be clarified. Are they functionally identical, or are they different processes? The nature of these effects also needs to be clarified: Can the results of single word recognition tests be generalized to larger contexts? How do these effects fit into models of word recognition and lexical access?

Transcription is not a very sensitive measure of perception. Once the listener hears an utterance, mistakes can occur both in perception and in response. Measuring reaction time may reveal more step-like graphs. Goldinger (1998) used reaction time as a meter of echo intensity. The longer the reaction time, the more passes needed for word recognition. If FAS at first requires many comparisons of novel words to stored episodes, reaction time should be slow. With many episodes of FAS, reaction time should decrease.

This research has important implications not only for models of speech perception, but also for practical communication. If this adaptation process is better understood, perhaps more efficient means of communication could be accomplished. Exposure to a variety of accents at a young age, for example, could lead to easier FAS intelligibility later in life. Speech recognition systems could be made more robust if they were able to utilize successful human strategies for overcoming speaker variability.

WORKS CITED

Egan, J.P. (1948). Articulation testing methods. *Laryngoscope,* 58, 955-991.

Ekwall, E.E., and Shanker, J.L. (1993). *Ekwall/Shanker reading inventory* (Third Edition). Boston: Allyn and Bacon.

Flege, J.E. (1995). Second language speech learning : theory, findings, and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: issues in cross-language research* (pp.245-277). Baltimore: York Press.

Flynt, E.S., and Cooter, R.B. (1998). *Flynt - Cooter reading inventory for the classroom* (Third Edition). Columbus, OH: Merrill.

Goldinger, S.D. (1996). Words and voices: episodic traces in spoken word identification and recognition memory. *Journal of Eperimental Psychology; Learning Memory and Cognition. 22,* 1166-1183.

Goldinger, S.D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105,* 251-279.

Halle, M., and Jones, L.G., (1959). *The Sound Pattern of Russian*. Mouton & Co.: The Hague, the Netherlands.

Hintzman, D.L. (1986). "Schema abstraction" in a multiple trace memory model. *Psychological Review, 93,* 411-428.

House, A.S., Williams, C.E. Hacker, M.H.L., and Kryter, K.D. (1965). Articulation-testing methods: consonantal differentiation with a closed-response set. *Journal of the Acoustical Society of America, 37,* 158-166.

IEEE. (1969). IEEE recommended practice for speech quality measurements (IEEE No. 297). New York: IEEE.

Jha, Aparna (1977). *An Outline of Marathi Phonetics*. Deccan College Press: Poona, India.

Johnston (1999). Cool Edit 2000 [Sound Editing Package]. Syntrillium Software Co: Author.

Nye, P.W., and Gaitenby, J.H. (1974). The intelligibility of synthetic monosyllabic words in short, syntactically normal sentences. Haskins Laboratories: *Status Report on Speech Research.* Sr-37/38, 169-190.

Nygaard, L.C., and Pisoni, D.B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics, 60*, 355-376.

Nygaard, L.C., Sommers, M.S., and Pisoni, D.B. (1994). Speech perception as a talker-contingent process. *Psychological Science, 5*, 42-46.

Pisoni, D.B. (1981). Speeded classification of natural and synthetic speech in a lexical decision task. *Journal of the Acoustical Society of America, 70,* S98.

Rogers, C.L. (1997). *Intelligibility of Chinese-accented English.* Unpublished doctoral dissertation, Indiana University.

Schmid, P.M., and Yeni-Komshian, G.H. (1999). The effects of speaker accent and target predictability on perception of mispronunciations. *Journal of Speech, Language, and Hearing Research, 42,* 56-64.

Schwab, E.C., Nusbaum, H.C., and Pisoni, D.B., (1985). Some effects of training on the perception of synthetic speech. Human Factors, 27(4), 395-408.

Scott, K.R. (2000, November). *The impact of accent, noise, and linguistic predictability on the intelligibility of non-native speakers of English.* Paper presented at the meeting of the Central Ohio Chapter of the Acoustical Society of America, Columbus, OH.

Sutter, R.W. (1976). Predictors of pronunciation accuracy in second language learning. *Language Learning, 26,* 233-253.

Thompson, I. (1984). *An experimental study of foreign accents.* Unpublished doctoral dissertation, The George Washington University.

Woods, M.L., and Moe, A. J. (1989). *Analytical reading inventory* (Fourth Edition). New York: Macmillian Publishing Company.