ONLINE ACCOMMODATION OF REGIONAL ACCENTS

BY

ALISON M. TRUDE

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Arts in Psychology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2010

Urbana, Illinois

Master's Committee:

Assistant Professor Sarah Brown-Schmidt, Chair
Professor Gary S. Dell

ABSTRACT

Despite the ubiquity of between-talker differences in accent and dialect, little is known about how listeners accommodate this source of variability in online language comprehension. Here we sought to identify constraints on this process, in order to inform candidate theories. Three experiments used the visual world paradigm to examine the roles of memory and contextual cues to talker identity in the accommodation process. Listeners interpreted the speech of a male talker with an unfamiliar regional dialect of American English, in which the /æ/ vowel is raised to /eɪ/ only before /g/ (e.g., *bag* is pronounced /beɪg/), and a female talker without the dialect. We examined interpretation of words like *back* in the context of a competitor that has the same vowel in the familiar dialect only, as well as words like *bake*, which share a vowel with the competitor (*bag*) in the unfamiliar dialect only. In all three experiments, listeners rapidly used their knowledge of how the talker would have pronounced *bag* to either rule out or include *bag* as a temporary cohort competitor, in a talker-specific manner. Even though talkers randomly alternated across trials, providing an early cue to talker identity in the form of a preamble (Exp.1) or a portrait (Exp.2) did not overwhelmingly improve performance compared to performance in the absence of a cue (Exp.3). These results suggest that talker adaptation is rapid, even in multi-talker contexts, and that on-line adaptation processes access and use information learned during previous experiences with a talker based on minimal acoustic information.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

During the course of a typical day, we receive speech input from many different people, sometimes in rapid succession. These speakers may have very different accents or dialects, speak at different rates, and have different pitch rages, yet we can usually understand their speech quickly and easily. This phenomenon is a type of many-to-many mapping problem, or a lack of invariance problem, in that depending on the talker and context, a particular acoustic signal can correspond to different phonemes (or words, etc.), and at the same time, a particular phoneme can be conveyed with very different acoustic signals (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Peterson & Barney, 1952; Gordon, 1988). Despite this variability, listeners seem to quickly and effortlessly generate stable representations of speech. While this puzzle has attracted a tremendous amount of attention, theorizing, experimentation, and modeling, the mechanisms by which we are able to accommodate accents, as well as other sources of inter-talker variability in speech, remain poorly understood.

Numerous accounts of how listeners accommodate variability have been proposed. Here we focus on two very different types of proposed mechanisms for how talker variability in speech processing is accommodated: a normalization mechanism and an episodic mechanism. The normalization approach proposes that the listener mentally transforms the speech input so that it conforms to his or her set of prototypical speech sounds (Miller & Liberman, 1979; Nearey, 1989; Miller, 1989; see Pisoni, 1997). Conversely, the episodic approach does not require the speech input to conform to the listener's pre-set standards. Instead, it proposes that listeners activate memory traces of similar sounding, previously-heard speech to create an interpretation of a particular talker's speech (Goldinger, 1998; Goldinger & Azuma, 2003).

Currently, little research exists that shows clear evidence for one of these approaches over another, in part because a diagnostic data pattern which would clearly support one view over the other is lacking. Additionally, due to a prevalence in the literature of experimental methodologies using offline paradigms-that is, experimental techniques in which the probe is the ultimate interpretation of a word, as opposed to how that word is interpreted in real time-it is unclear how the mechanisms in the proposed models might operate during the online processing of speech sounds. Thus, the goal of the present research is not to present evidence which would unequivocally support a particular view of the accommodation process. Instead, the present research aimed to identify characteristics of accent accommodation during online speech processing that must be incorporated into future models of talker variability accommodation.

Due to our primary interest in typical language use, which often involves conversations among multiple familiar speakers with different speaking styles, the factors we focused on were the roles of memory for particular speakers and contextual cues to who might be speaking next. We examined the role of these factors in accommodating an unfamiliar regional accent of American English, and examined the interpretation process as it unfolded over time, using the visual world eye-tracking methodology (Tanenhaus, et al., 1995). In what follows, we report on the results of two eye tracking experiments that were conducted to shed light on these issues, with the ultimate goal of informing and improving models of online accommodation of variability in speech.

Normalization view

One school of thought for explaining how listeners deal with speaker variability can be referred to broadly as the *normalization* or *analytic* approach. Generally speaking, this view

presumes that listeners have a stored mental representation of an ideal set of speech sounds. Sources of talker variability, such as speech rate, speech style, and accent, are considered to be "noise" that must be filtered out of the speech stream in order to make the speech input match the listener's mental representations; word recognition is successful when the filtered input is matched against a stored representation of a word (Miller & Liberman, 1979; Nearey, 1989; Miller, 1989). According to one version of this proposal, the filtering process constitutes a transformational algorithm which changes each phoneme from the speech stream into one of the listener's standardized mental phonemic representations (Syrdal & Gopal, 1986; also see Pisoni, 1997). For example, if the listener stored the medial vowel in *tomato* as [eɪ], but heard a token pronounced with a medial [ɑ] vowel, to understand this word as *tomato*, she would apply a transformational algorithm that changed [ɑ] to [eɪ], and then proceed to interpret the word with the transformed vowel.

Normalization theories can be further divided based on how they propose that the listener arrives at the proper transformational algorithm. In one category, which we shall refer to as intrinsic normalization, decisions about when and how to apply a transformational algorithm depend only on acoustic information within a given syllable. For example, listeners may use the relationship between pairs of formants to estimate the speaker's vowel space and transform the input accordingly to make it map onto the listener's stored phonetic categories (Syrdal & Gopal, 1986; Nearey, 1989; Miller, 1989). Related proposals suggest that normalization might be accomplished based on estimations of the speaker's vocal tract length (Joos, 1948; Ladefoged & Broadbent, 1957; Fujisaki & Kawashima, 1968; Nearey, 1989).

However, it has been found that extra-syllabic information, such as acoustic information within a preceding linguistic context (Ladefoged & Broadbent, 1957; Nearey, 1989; Evans & Iverson, 2003), beliefs about speaker identity (Johnson, 1990; Johnson, Strand, & D'Imperio, 1999; Ladefoged & Broadbent, 1957), and knowledge of what words a given talker is likely to produce in a given context (Creel, Aslin, & Tanenhaus, 2008), affects perception. These findings pose challenges for the most extreme views of normalization. Likewise, evidence that participants perform more poorly in a variety of tasks when stimuli are produced by multiple speakers, as opposed to a single speaker (Nusbaum & Morin, 1992; Martin et al., 1989), suggests that normalization processes must be sensitive to information external to a single syllable or phrase. For example, Mullennix et al. (1989) presented participants with single-word stimuli with various signal-to-noise ratios and asked them to either type the word that they had heard or repeat it out loud. Across tasks and signal-to-noise ratios, the participants who heard multiple speakers were less accurate and slower than those who heard a single speaker. The drop in performance following a switch in speaker suggests that each speech sound is not handled independently, and that there is preservation of learning over time. These findings are inconsistent with versions of the intrinsic normalization view which predict that there should be no difference between single and multi-speaker contexts because only information from the current stimulus is used during its processing.

A different version of normalization, sometimes called extrinsic normalization, can account for the processing cost associated with adjusting to multiple speakers. It combines intrinsic strategies with a mechanism that retains speaker-specific normalization algorithms over time and adjusts these algorithms continuously as new stimuli are heard (Nearey, 1989;

Nusbaum & Morin, 1992; also see Joos, 1948). As long as the same speaker continues talking, the same algorithm continues to be refined. When a new speaker begins to talk, so long as the two talker's vowel spaces are sufficiently different, the process of determining a representation begins all over again, resulting in increased processing costs (Nusbaum & Morin, 1992; Magnuson & Nusbaum, 2007). On Nusbaum and Morin's (1992) account, two distinct processes operate: in mixed-talker conditions, a slow structural estimation process uses information within the speech to normalize itself, but in single-talker contexts, contextual tuning mechanisms combine information across utterances to map out the particular speaker's acoustic-phonetic space, thus reducing attentional and processing demands in subsequent interpretation.

Potential evidence in support of an extrinsic normalization view comes from a phoneme categorization task (Kraljic & Samuel 2007) in which participants are asked to identify phonemes presented in isolation (e.g., d and t). In this experiment, participants first completed two, single-speaker blocks of training in which they heard each speaker produce an ambiguous phoneme from a continuum (e.g., /d/-/t/) in the context of one of the two endpoints (e.g., croco?ile or cafe?eria). Later, the participants were asked to classify sounds on the continuum produced by the same two speakers that were heard in training, again in two single-speaker blocks.

It was predicted that if listeners can maintain multiple, speaker-specific phonemic representations, they should be more likely to classify the ambiguous phonemes as /d/ for the speaker who produced these phonemes in the "d" context during training, and more likely to classify the ambiguous phonemes as /t/ for the speaker who pronounced the phonemes in the "t" context. Conversely, if listeners must readjust the same phonemic representation each time a new

speaker begins talking, perceptual learning would not be demonstrated for more than one speaker.

The results indicated that participants shifted their phonemic representations toward those of the test speaker only when one speaker's training and testing sessions were presented in consecutive blocks. When a speaker's training and testing sessions were separated by a block of another speaker, participants had shifted their categories away from the test speaker's (i.e., toward the categories of the immediately *previous* speaker). The authors took this as evidence that speaker-specific representations were maintained only until a new speaker began speaking, at which point, the listener needed to shift his or her representations back to baseline before adjusting to the new speaker. This explanation is potentially consistent with an extrinsic normalization view, in which a listener must restart the process of developing a transformational algorithm every time a new speaker begins to talk.

A more recent view characterizes normalization as a hypothesis testing process in which listeners entertain multiple, simultaneous interpretations of an acoustic signal (Magnuson & Nusbaum, 2007; Nusbaum & Magnuson, 1997). By using active control mechanisms to shift attention to various characteristics of the acoustic signal and by using other information (e.g., linguistic knowledge, previous utterances) to constrain the list of possible interpretations, listeners identify the possible interpretation that most closely maps onto the speech sound that they are hearing, allowing them to successfully interpret the input. For example, Magnuson and Nusbaum (2007) found that listeners' expectations about what they were going to hear determined how they processed the speech: when listeners expected to hear multiple talkers, they showed slower interpretation times compared to listeners who heard the exact same acoustic

stimuli but expected to hear only one talker (also see Remez, Rubin, Pisoni, & Carell, 1981). The hypothesis testing account allows for more flexible processing of speech than previous versions of the normalization view because it does not presume that listeners are storing one-to-one mappings of speech sounds to stored phonemes. Instead, it allows listeners to consider multiple sources of evidence before settling on an interpretation of the acoustic signal.

In summary, the normalization view of talker adaptation posits that listeners map speech input onto invariant mental representations of phonemes. Although early versions of the theory proposed that only characteristics of the stimulus currently being processed were used to achieve this mapping, more recent experimental findings suggest that some information is retained and applied during processing of subsequent stimuli as long as the speaker's identity remains constant. A newer, hypothesis testing version of normalization characterizes the process as a sort of "decision tree," but the basic principles of variable speech input being categorized based on existing information about the language's phonological repertoire remain.

Episodic view

An alternative theory of speech perception can account for many of the same phenomena as extrinsic normalization. In this alternative, *episodic*, view of speech perception, listeners store specific episodes of speech input rather than having a set of idealized sound representations stored in long-term memory (Goldinger, 1998; Goldinger & Azuma, 2003; Hawkins, 2003; Johnson, 1997; Pisoni, 1997). Each time a word is heard, a new episodic memory trace is created. Along with an acoustic record of the word, these traces include information such as the identity of the speaker and the context in which the word was spoken. When a listener hears a word, previously stored traces that share characteristics with the current speech input are

activated. The simultaneous activation of many partially redundant traces creates a single generalized representation of the input (Goldinger, 1998). The episodic theory can also account for the previously mentioned processing cost associated with switching speakers. According to this theory, when speakers switch, the activated traces from that speaker must at first compete with the still-active traces from the previous speaker, leading to a processing cost such as the one found by Kraljic & Samuel (2007).

Goldinger and Azuma (2003) proposed an explanation of how an exemplar-based theory of speech recognition would work using Adaptive Resonance Theory (ART) (Grossberg, 1980), in which bottom-up and top-down information create a feedback loop to facilitate speech recognition. According to ART, speech input activates clusters of features in working memory, which can combine or interact in order to activate chunks from long-term memory. These chunks are prototypes created through prior experience and can be the size of any speech unit, from phonemes to whole words. Activated chunks send activation back to the feature clusters, creating a resonance. The achievement of resonance draws attention, creating a conscious experience for the listener of having heard an entire word, rather than a collection of smaller units. Although smaller (e.g., phoneme-sized) and larger (e.g., word-sized) resonances can exist simultaneously, various processes, such as integration of units over time, mask the smaller units, creating this cohesive perceptual experience. For example, upon hearing the word *jigsaw*, smaller resonances may occur for the phoneme /dʒ/ or the syllable /dʒɪg/ (itself a word); however, we perceive having heard a single word, and not a number of disjointed components.

In ART, circumstantial constraints, such as speaker identity or context, can activate top-down information, which can speed resonance, leading to faster recognition of words from the

same speaker or context (Goldinger & Azuma, 2003). In one experiment, participants were asked to listen to recordings of single-word stimuli and repeat the words as they heard them (Goldinger, 1998). They were also recorded reading the same words from a list. Afterwards, a new set of participants completed an AXB task, in which they heard recordings of the previous participants saying each word both in the reading (baseline) and repetition (shadow) contexts. The listeners were asked to compare these recordings to the initial recording of the stimulus word and determine which of the participant's recordings was the shadowed version.

Among other findings, the results showed that participants' shadowing RT's were faster for targets that had been repeated more often. Additionally, listeners were more likely to correctly identify the shadowed token in contexts where the participant had heard more repetitions of the stimulus (all produced by the same speaker). The explanation for this finding was that each repetition of the word created a trace in long-term memory. These identical traces converged, creating a stronger representation of the word and making the particular characteristics of the audio token more prominent, prompting more imitation of the target audio (Goldinger, 1998). Based on these findings, the episodic theory would predict that increased exposure to a particular speaker would lead to faster processing of that speaker's speech and possibly more complete representations of talker-specific information.

The episodic view of talker adaptation provides an alternative to normalization accounts that does not depend on the use of a stored bank of phonemic representations. Instead, traces from previous speech input are stored in episodic memory and activated upon hearing a new, similar speech event. The episodic account also allows for the use of top-down information, such as speaker identity, to aid in limiting the active traces to just the most relevant ones. This theory

would predict a strong role for long-term memory in the processing of speech and the accommodation of talker variability.

Evaluating the theories

The extrinsic normalization and episodic views of speech perception make very different claims about how listeners accommodate variability in speech; however, it is not immediately clear how to distinguish the theories. For example, while the episodic account makes a clear prediction that increased exposure to a talker should facilitate processing of his or her speech, the extrinsic normalization view could also predict these findings by allowing for the storage of talker-specific algorithms that can be refined with experience. Similarly, while the extrinsic normalization view makes a strong prediction that there is a cost when a new talker begins speaking, the episodic view would also predict this phenomenon because traces from the previous talker may remain active as the next talker begins speaking, leading to interference and slowed processing.

In the absence of a diagnostic data pattern, the goal of the current research was to refine both theories by identifying key features of the accommodation process that should be accounted for by models of talker variability accommodation. We chose to focus on two relevant features of the accommodation process: the roles of long-term memory and contextual information. Specifically, we examined the listener's ability to represent two different talkers' accents as talkers alternated, as well as the contextual information that a listener might use to prepare to access information about a particular talker's accent. We chose to focus on these two aspects of talker variability accommodation because they are highly relevant to the way we process speech in everyday situations. We often engage with more than one talker at a time, so understanding

how participants keep track of talker-specific information for multiple conversation partners is a critical feature for any theory of talker accommodation. Additionally, one could imagine that contextual information, such as an indication of who was going to talk next (e.g., a particular talker raising her hand in class before speaking) could provide useful cues that listeners could use to prepare themselves for speech that they are about to hear. Thus, contextual information may facilitate processing if the relevant information about a particular talker is in place before speech input begins. Understanding the constraints on these preparatory and memory-retention processes would then provide a variety of insights into the accommodation process as well as provide constraints on models of accommodation.

Although the normalization and episodic accounts are intended to explain accommodation of all kinds of talker variability, our research focuses on the processing of unfamiliar regional accents. Accents are a source of variation that is often encountered in daily life, especially as our society becomes increasingly more globalized and we are more likely to live and work among people from a variety of geographical locations. Thus accent accommodation represents not only a topical, but a very common type of speech accommodation.

In addition to its practical relevance for accommodating the speech of familiar interlocutors, long-term memory is a central component of the episodic theory, which relies upon the access of stored episodic traces as a basis for processing accents and other sources of talker variability (Goldinger, 1998). However, the role of long-term memory is less clear for the normalization account. In this account, it is proposed that a set of prototypical speech sounds is stored in long-term memory, but different versions of the theory would predict that different

elements of an individual's speech would be stored. Current extrinsic normalization accounts indicate that listeners can store information about an individual's speech in long-term memory, allowing them to recognize a familiar speaker's voice (Nusbaum & Magnuson, 1997). However, it is unclear whether this information can be used in the creation of transformational algorithms, or whether the algorithms themselves can be stored in long-term memory. For example, on one version of extrinsic normalization (Nusbaum & Morin, 1992), following a switch in talker, the listener embarks on a slow, attention-demanding structural estimation process which "self-normalizes" the speech sounds; when the same talker continues, listeners use contextual tuning mechanisms to learn vocal characteristics of the talker, based on multiple utterances. Whether this information is thought to be stored is unclear[1].

Another area of exploration is what types of speaker-specific information are stored in long-term memory and whether they can be used during online speech processing. On the episodic view, contextual information, such as talker identity and location, is stored along with acoustic information in the episodic traces (Goldinger & Azuma, 2003). However, it is unclear at precisely what stage of processing and at what speed this information could be applied during speech processing. Traditionally, while intrinsic normalization accounts focused mainly on the use of acoustic information in determining and applying transformational algorithms, and not on the application of non-linguistic information (e.g., Syrdal & Gopal, 1986; Nearey, 1989; Miller, 1989), revised versions of this theory have addressed the potential use of linguistic knowledge

[1] The fact that listeners must do the slow structural-estimation process following a change in talker suggests the model assumes that what was learned during contextual tuning is discarded. The advantage of this structure is that it accounts for the drop in performance in mixed-talker conditions.

and contextual information (Nusbaum & Magnuson, 1997). Thus, it could be possible that non-acoustic or extralinguistic information is being utilized during accent accommodation.

A final area of inquiry is how these views can be extended to account for the real-time processing of speech. After all, words unfold at roughly 2-3 per second (Levelt, 1989), and listeners begin making provisional commitments immediately on the basis of sublexical acoustic information (Allopenna, Magnuson, & Tanenhaus, 1998; Dahan, Magnuson, Tanenhaus, & Hogan, 2001; Salverda, Dahan, & McQueen, 2003); thus, any mechanism would have to perform quickly. For example, Allopenna, et al. (1998) monitored participants' eye movements as they viewed a display that contained pictures of a target word (e.g., *beaker*), a cohort competitor that shared an onset with the target (e.g., *beetle*), a competitor that rhymed with the target (e.g., *speaker*), and an unrelated distractor (e.g., *dolphin*). On critical trials, the participant heard the target word and was instructed to click on it. The results indicated that early in these trials, the cohort competitor competed more strongly with the target word than the rhyme competitor, due to their shared onset (i.e., *be-*). However, as the trial progressed and participants heard the portion of the target word that is shared with the rhyme competitor (i.e., *-eaker*), participants showed increased fixations to the rhyme competitor. These results suggest that not only do listeners entertain multiple interpretations of a speech signal, but that these interpretations can change over time, as more of the speech signal is realized.

In Allopenna, et al.'s (1998) experiment, all of the stimuli were produced by the same talker; therefore, the effects of talker-specific variability were not examined. Thus, a key open question is how listeners accommodate variability in on-line processing. The present research focuses on one specific type of variability accommodation: the accommodation of regional

accents. While these accents can vary from being quite similar to one's own accent to quite distinct, even the more subtle accents, such as the one we examined here, can result in surprising changes in the interpretation of words, including the elimination and addition of temporary competitors (e.g., cohort competitors), based on pronunciation variations in these different accents. We used a visual word paradigm similar to the one used by Allopenna et al. in order to test how accent information is incorporated during online speech processing.

Eye tracking as a measure of online accommodation processes

A limited amount of research has examined the on-line interpretation of regional accented speech. Dahan, Drucker, and Scarborough (2008) used a variant of the visual-world eye tracking technique (Tanenhaus, et al., 1995) to test whether a normalization or episodic mechanism could better account for accent accommodation. Participants heard the speech of an American English speaker with a regional accent in which the /æ/ vowel is raised to [ɛ] before /g/ (e.g., *bag* [bɛ̲g]) but not before /k/ (e.g., *back* [bæ̲k]). On critical trials, participants viewed a screen with four words presented orthographically: an *-ag* word, an *-ack* word beginning with the same consonant (e.g., *bag* and *back*), and two unrelated fillers, and were asked to click on an auditorily-presented word. Before exposure to words containing the accented vowel, participants exhibited a cohort competition effect: when hearing *back*, they initially fixated *bag* and *back* equally until disambiguating information (e.g., [k]) was heard (Allopenna, et al., 1998). However, after exposure to accented words, upon hearing the [bæ] in *back* words, participants quickly identified the word as *back* with little consideration of *bag* (because *bag* would have been pronounced by this speaker as [bɛ̲g], and thus is not a cohort competitor with *back*). They argued that this result is inconsistent with normalization approaches because listeners used their knowledge of the

accent even when interpreting non-accented words, that is, in circumstances where a transformation would not have been generated or applied. The authors concluded that the results support an episodic account of accent accommodation, which proposes that listeners can use contextual, top-down information about a speaker in order to adjust to phonemic representations and interpret speech accordingly. This process results in an expectation that if the speaker wanted to refer to *bag*, that s/he would have said [bɛg]; therefore, on *back* trials, *bag* is more quickly ruled out as the target word.

This interpretation of these results is largely based on the fact that only an episodic mechanism could explicitly allow for the use of top-down evidence in speech processing (Dahan et al., 2008). However, we would like to suggest that a version of a hypothesis testing account of normalization could potentially explain these results as well. One way a hypothesis-testing account could explain this data pattern is if we assume that listeners, upon hearing the temporarily ambiguous phoneme string /bæ/, generate a decision tree with all possible continuations (e.g., bathrobe, batter, back, bag), and then use contextual information, such as speaker identity, to prune those contextually-inconsistent branches, thus eliminating *bag* as a contender. Another way that the hypothesis-testing mechanism could account for the results is if it includes not only simple matching rules, such as a /b/ onset predicting words with b-onsets, but also contextually constrained counterfactual rules, such as /g/ never following /æ/ for a particular speaker. Thus, upon hearing /bæ/, the listener could employ the counterfactual rule to eliminate *bag* as a potential referent, because a /g/ cannot follow an /æ/ vowel for this speaker.

While these explanations would suggest how the two views could account for these results, there may be an even simpler explanation. Because the pairs of target stimuli in this

experiment (Dahan et al., 2008) always consisted of an unaccented *–ack* word and an accented *–ag* word, an alternative explanation is that participants simply learned that anytime they heard [æ], the target word would be the one ending in "k," resulting in fewer fixations to the competing *–ag* word. In the experiment, the target words in the filler trials all contained vowels other than /æ/ and /ɛ/; hence, they would not have kept participants from adopting this heuristic. This strategy could generate the observed results without requiring participants to create any representations of the speaker's speech or normalize the speech in any way.

Here we present the results of three experiments investigating how listeners accommodate accented speech on-line, specifically examining the roles of long-term memory and non-linguistic cues during accent accommodation processes. Experiment 1 uses multi-speaker contexts as a test case to determine whether speaker-specific representations are stored in long-term memory. Experiment 2 tests whether these representations can be applied during online speech processing without a preceding auditory cue to the talker's identity. Experiment 3 tests whether participants are able to accommodate talker-specific variability with only the initial consonant of the target word as a cue to talker identity. The designs of Experiments 1-3 were largely based on Dahan et al.'s experiments. Half of our critical trials were modeled after those in Dahan, et al.'s experiments, in which participants heard words like *back* in the context of *–ag* competitors such as *bag*, as well as the target *back*. In a second (novel) type of critical trial, participants heard words like *bake* in the context of *bag* and *bake*. Note that our speaker raised the /æ/ vowel before /g/ to [eɪ], rather than [ɛ]. Critically, the inclusion of *bake* trials means that the [æ] vowel is not automatically associated with *bag*-type (i.e., accented or g-final) targets. We predict that if listeners are able to learn the speaker's accent, and apply this knowledge to guide

online processing, they should show increased fixations to the target -*ack* word when hearing the accented speaker, as compared to an unaccented speaker, due to the reduced competition between the –*ack* and –*ag* words. Conversely, we expect to see the opposite effect on –*ake* trials.

CHAPTER 2: EXPERIMENT 1

The goal of Experiment 1 was to test whether talker-specific accent information is stored in long-term memory and then used to guide the online perception of speech stimuli. In order to do this, a two-talker test design was adopted. A multiple-talker paradigm is valuable not only because it approximates real-life experiences, in which we may be simultaneously conversing with multiple people with different accents, but also because it has the potential to help determine whether listeners store information about a talker in long-term memory and whether listeners are able to quickly retrieve that information even after hearing a different, intervening talker. We predicted that if listeners do store talker-specific information in long-term memory, they should be able to access that information quickly, leading to a successful interpretation of the speech signal based on that talker's accent.

In Experiment 1, native English-speaking participants sat in front of a computer screen with pictures and followed pre-recorded instructions to click on one of the images while their eye movements were monitored. The instructions were produced by one of two different native English speakers, one of whom had a regional American English accent different from the typical regional accent of the participant population, and one of whom had a typical local accent. The talkers randomly alternated from trial to trial in order to evaluate the listeners' ability to accommodate one talker's accent after hearing another talker with a different accent.

We hypothesized that if listeners store talker-specific information in long-term memory, either as transformational, hypothesis-testing algorithms or episodic information, then we should observe different patterns of interference on *-ack* and *-ake* trials. Specifically, on *-ack* trials, participants should make more fixations to the target when hearing the accented talker compared

to the unaccented talker, because for the accented talker, the target (e.g., *back*) and the competitor (e.g., *bag*) do not share a vowel, and are thus less similar than they are for the unaccented talker, for whom the two words do share a vowel. Conversely, on *-ake* trials, they should make more fixations to the target when hearing the unaccented talker compared to the accented talker because the unaccented talker produces these two words with different vowels, while the accented talker produces them with the same vowel.

According to an episodic account, listeners should be able to use top-down cues to quickly determine the talker and constrain potentially activated traces. Limiting the active traces to just those of the current talker should ensure that the average of these traces will converge on the pronunciation of the word that is particular to that talker, enabling the listener to process that talker's particular pronunciation more quickly and easily than other pronunciations of the word. In an extrinsic normalization account that includes long-term memory for algorithms, listeners could quickly access the previously stored transformational algorithm for the talker rather than creating one from scratch, speeding the accent accommodation process by eliminating the time needed to construct an algorithm.

Alternatively, if listeners do not store talker-specific information in long-term memory, they should have difficulty switching between talkers from trial to trial because they cannot use their previous experience with that speaker to guide processing in subsequent trials. This hypothesis is consistent with some accounts of extrinsic normalization which indicate that transformational algorithms are not stored and that each time a new talker is heard, the process of building a transformational algorithm must begin again from scratch (see Nusbaum & Morin, 1992). On this account, there should be no difference in fixations to target images between the

accented and non-accented talker on either *–ack* or *–ake* trials because an accurate transformational algorithm could not be created in time to process the word online.

In Experiment 1, listeners heard the talker say the phrase "*Click on*," followed by a 200 ms pause before hearing the target word. The preamble was the first indication of the talker's identity, so in order for speaker-specific information to be used in the processing of the word, it would have to be activated during the course of the preamble or subsequent 200 ms of silence. Additionally, it is important to point out that the preamble did not contain any words with /æ/ or /eɪ/ in them, so participants were not explicitly "reminded" of the critical contrast at the start of each trial, and any transformational algorithms constructed based on the preamble should not contain information about the accented vowel.

Method

Participants

38 members of the University of Illinois community participated in Experiment 1. Seven additional participants were excluded from analysis because of technical difficulties, and one participant was excluded because he did not complete the experiment. Participants received either payment ($16) or partial course credit for their participation. All participants were native speakers of North American English and had normal or corrected-to-normal hearing and vision. Most of the participants had a suburban Chicago accent; given the ubiquity of this accent on the University campus, those who did not exhibit this accent were certainly familiar with it. Crucially, it was established using a written survey that none of the participants shared an accent with the "accented" talker (see below).

Stimuli

The acoustic stimuli were produced by a male and a female talker who did not interact with the participants. The male talker was a native English speaker from Oregon with an accent similar to the speaker in Dahan et al. (2008). In this case, the talker raised the /æ/ vowel before /g/ to [eɪ], rather than [ɛ]. The female talker was a native English speaker from the Chicago area. Critically, although the female talker exhibited traits that would be considered characteristic of a Chicago-area accent, she did not exhibit the raised [eɪ] vowel before /g/. Talkers of different gender were used to ensure that the two voices in the experiment were perceptually dissimilar enough to be recognized as belonging to two different people. For convenience, we will refer to the female Chicago-area speaker as the "unaccented talker" and her recordings as the "unaccented" words because although she (like everyone) spoke with an accent, her accent was familiar to the participants. Likewise for expository purposes, we will refer to the male talker as the "accented" talker, his *–ag* tokens as the "accented" words, and his tokens which did not display the vowel raising as "unaccented."

Participants listened to the speech of the two talkers during a training phase and a testing phase. The acoustic stimuli for the testing phase consisted of eleven sets of six monosyllabic English words. Each set contained three critical words, ending in /æg/, /æk/, and/eɪk/, which shared the same onset (e.g., *bag*, *back*, and *bake*). Each set also contained three filler words, one ending in /g/ and two ending in /k/. All of the filler words in a set had the same onset, and they all contained vowels other than /æ/ and /eɪ/ (e.g., *league*, *leak*, and *lock*). The stimuli were adapted from Dahan et al. (2008) (see Appendix A for the complete list of stimuli). The acoustic stimuli for the training phase was a dialogue containing at least four instances of each of the 11 -

ag words (two instances per speaker), as well as four instances of both *–ack* words and *–ake* words being pronounced by each talker.

All acoustic stimuli were recorded to a computer in the open sound field using a headset microphone. For the training story, the two talkers were recorded together reading a dialogue. For the testing phases of the experiment, each critical and filler word was recorded in isolation.

The visual stimuli were color drawings taken from an online clip art database, and were selected to provide the clearest possible depiction of each associated word.

Equipment and Procedure

Experiment 1 consisted of a training phase followed by a testing phase. The entire experiment lasted approximately 2 hours. The experiment was programmed in Matlab using the Psychophysics toolbox (PTB-3, Brainard, 1997; Pelli, 1997).

Training. Due to the tight restrictions on the characteristics of the auditory stimuli, some of the critical pictures may not have been easily identifiable (due to low imageability of some of the target words, e.g., *flack*). In order to assure that all participants could identify the images using the critical word, participants first completed a picture training session. Over the course of 66 trials, each experimental picture was displayed on the screen in isolation with the target word written above it. Following this, participant were tested for their understanding of each word-picture pair by viewing 4 pictures on the screen, along with the written name of one of the pictures (e.g., "flack"), and clicking on the target. All participants successfully completed the test on their first attempt.

The participants then listened to a dialogue between the two talkers. The dialogue was intended to familiarize the participants to the two talkers' voices and expose them to each talker's pronunciation of *–ag*, *-ake*, and *–ack* words in a naturalistic conversation setting.

Test. During the testing phase of the experiment, participants' eye movements were recorded using an Eyelink 1000 desktop-mounted eye tracker which sampled eye position monocularly at 1000hz. Participants first viewed a fixation cross for 1000 ms. Participants then viewed a display containing pictures of four of the six words from one of the word sets: an *-ag* word, an *-ack* or *-ake* word, a filler word ending in /g/ (e.g., *wig*), and one of two possible filler words ending in /k/ (e.g., *wick* or *weak*) (Figure 1). After 2000 ms, the participants heard the preamble "*Click on*," followed by the target word, played through speakers, and were instructed to click on the word that they had heard. The preamble and target word were both spoken by the same talker. The test consisted of 352 trials per talker, for a total of 704 trials. For each talker, there were 88 trials each of *-ag* targets and filler targets ending in /g/, and 44 trials each of *–ack* targets, *-ake* targets, the first group of filler targets ending in /k/, and the second group of filler targets ending in /k/. The order of trials was completely random, with a different random order for each participant.

Results

-ack *word trials*

We predicted that if participants were able to store multiple representations in long-term memory, then they should show a greater proportion of fixations to the target when listening to the accented talker than when listening to the unaccented talker. The proportion of fixations to the target word was calculated by subject and item in 100 millisecond intervals beginning at the

onset of the critical word (e.g. *back*), continuing until 1000 ms after word onset. A baseline analysis region from -100 ms to 200 ms revealed no difference between talker conditions (t = 1.03). Repeated measures ANOVAs by subject and by item were performed, with talker (accented male vs. unaccented female) and time (200-1000 ms, in 100 ms intervals) as within-subject factors. The results generally replicated the findings by Dahan et al. (2008). The ANOVA revealed a main effect of time, $F_1(8, 296)$ [2]= 751.8, $p < .001$, ($\varepsilon = .281$); $F_2(8, 168) = 247.1$, $p < .001$, ($\varepsilon = .225$), due to increasing target fixations during the trial. The main effect of talker was also significant, $F_1(1, 37) = 43.1$, $p < .001$; $F_2(1, 21) = 23.6$, $p < .001$, with a greater proportion of fixations to the target when hearing the accented male talker than when hearing the unaccented female talker (.50 and .43, respectively, see Figure 2). These main effects were qualified by a significant interaction between talker and time, $F_1(8, 296) = 18.2$, $p < .001$, ($\varepsilon = .418$); $F_2(8, 168) = 11.7$, $p < .001$, ($\varepsilon = .351$). A series of planned comparisons indicated that this difference was significant from 400 to 1000 ms by subject ($ts > 2.32$) and from 500 to 1000 ms by item ($ts > 2.55$).

-ake *word trials*

On *–ake* trials, we predicted that if participants were able to store information about multiple talkers, they should show fewer fixations to the target when listening to the accented talker than when listening to the unaccented talker. An analysis of the baseline region (-100 ms to 200 ms) indicated that there was no baseline effect of talker ($t = 0.015$). Repeated measures ANOVAs by subject and by item were performed, with talker (accented male vs. unaccented

[2] All p-values were corrected for violations of sphericity using the Greenhouse-Geisser correction. For cases in which sphericity was violated, the Greenhouse-Geisser ε is reported. Where no ε value is reported, sphericity was not violated. For clarity, the uncorrected degrees of freedom are reported in all cases.

female) and time (200-1000 ms, in 100 ms intervals) as within-subject factors. The analyses revealed a main effect of time, $F_1(8, 296) = 493.0$, $p < .001$, ($\varepsilon = .260$); $F_2(8, 168) = 155.8$, $p < .001$, ($\varepsilon = .180$), due to an increase in target fixations during the trial. While the main effect of talker was not significant, the interaction between talker and time was, $F_1(8, 296) = 8.5$, $p < .001$, ($\varepsilon = .398$); $F_2(8, 168) = 4.8$, $p < .05$, ($\varepsilon = .285$). Inspection of the data (Figure 3), suggests that before 700 ms, there were significantly fewer target fixations when the accented talker was speaking, compared to the unaccented talker, and after 700 ms, the effect reversed. A series of paired, by-subjects comparisons at each 100 ms interval confirmed this observation: the predicted effect of talker was significant between 400-600 ms ($ts > 2.84$). At 700 ms, the effect was not significant ($t = 0.76$). At 800 ms, the reverse effect was marginal ($t = 1.73$), and from 900 to 1000 ms, it was significant ($ts > 2.56$). The comparisons by items were not as robust, and were significant only from 400-600 ms ($ts > 2.43$).

Unlike the earlier effect, which demonstrates differences in the processing of accented and unaccented speech during the unfolding of the target stimulus, the late reversal of the effect is most likely a result of the varying degree of difficulty in disambiguating the target and distracter words depending on talker. Because the target and distracter words in this condition were less similar for the unaccented talker, participants were more quickly able to settle on an interpretation of the target stimulus. The late reversal of the effect indicates that when hearing the unaccented talker, the participants finished processing the target stimulus more quickly and thus began looking around at the other pictures on the screen.

Discussion

In Experiment 1, in contexts containing a *bag*-type picture and a *back*-type picture, as listeners heard the word *back,* they were significantly more likely to fixate the target picture *back* with the accented talker, compared to the non-accented talker. This result replicated the Dahan et al. (2008) findings, using a standard picture-viewing paradigm rather than orthographic stimuli, and controlling for an alternative interpretation of the Dahan et al. findings based on the characteristics of their stimulus set. The effect that we observed occurred because listeners were able to use information about how the accented talker *would have* produced the *–ag*-type word (e.g. *bag)* to eliminate *bag* as a potential competitor upon hearing the initial phonemes /bæ/ in *back*, thus increasing the likelihood of a target fixation. We also observed the predicted, reverse effect on *-ake* trials: In contexts containing a *bag*-type picture and a *bake*-type picture, listeners were significantly less likely to fixate a target picture like *bake* upon hearing *bake* with the accented, compared to the non-accented, talker. Thus, as they interpreted *bake*, listeners were able to use information about how the accented talker *would have* produced the *–ag*-type word (e.g. *bag)* to temporarily consider *bag* as a potential competitor upon hearing the vowel in *bake,* thus temporarily reducing the likelihood of a target fixation.

Taken together, these findings suggest that participants were successful at accommodating the accents of both talkers online as the stimulus was being heard. The fact that listeners interpreted the words *bake* and *back* differently, depending on who was speaking, despite the fact that the talker switched randomly, suggests that listeners stored talker-specific information about accent in memory. This result is consistent with other findings that experience with multiple, specific talkers improves interpretation even in multi-talker contexts (Nygaard & Pisoni, 1998; Nygaard, Sommers, & Pisoni, 1994). The result is also consistent with findings that

listeners can use information about what words a particular talker does and does not produce in a given context, to eliminate potential competitors (Creel et al., 2008). The effects we observed took place relatively quickly and easily, in contrast to the phoneme categorization results found by Kraljic and Samuel (2007). The effects also occurred on non-accented trials; that is, listeners demonstrated sensitivity to knowledge of the accented talker's unfamiliar production of *–ag* type words on trials which they did not hear an *–ag* type word. Further, since the accented and unaccented talkers alternated randomly from trial to trial, and only 25% of trials contained the unfamiliar *–ag* vowel, this effect was clearly robust across changes in talker. Because the *Click on* preamble identified the talker, listeners may have prepared for interpretation of the critical word, based on previous findings that listeners can use information about the regional accent in a carrier phrase to modulate vowel categorization (Evans & Iverson, 2003; also see Johnson, 1990). How might normalization and episodic views account for these results?

Under an episodic approach, the results are expected and would suggest that listeners, upon hearing the initial portion of the instruction (*Click on…*) spoken by the accented talker, activated only those traces associated with the current talker, thus decreasing the relative activation of the standard pronunciation of the *–ag* words in the representations they created. Thus, when hearing *back*, the initial portion of the word, /bæ/, was consistent with *back*, as well as other non-present competitors, such as *backpack, battle, bad,* etc. (see Magnuson, Tanenhaus, Aslin, & Dahan, 2003), but not with *bag*. In this model, this occurred because traces associated the *–ack* word were activated more strongly than those associated with the *–ag* word, thus eliminating the *–ag* word as a competitor. Conversely, when hearing *–ake* words spoken by the accented speaker, both *-ake* and *–ag* word traces received high levels of activation because of

their similarity in the accented talker's dialect. Upon hearing a word like *bake*, the initial portion of the word, /beɪ/, was consistent with both *bake* and *bag* (as well as a number of non-present competitors like *bagel, baby,* etc.); thus, participants fixated both the target and the competitor, reducing overall fixations to the target.

Recall that on some versions of the extrinsic normalization view (Nusbaum & Morin, 1992), in mixed-talker conditions, listeners must re-create a transformational algorithm each time a new speaker begins talking. Thus, this view generates the hypothesis that in an experiment like Experiment 1, where the talker is randomly alternated on trial to trial, listeners should not be able to adjust to the different talkers. However, if extrinsic normalization could (a) take place rapidly (in this case, during a two-word preamble), and (b) on the basis of unaccented input alone (because the preamble lacked the accented vowel, and the effect appeared on unaccented words), this could account for our findings. Accommodation based on a preamble alone may be a distinct possibility, as characteristics of carrier phrases are known to affect interpretation of ambiguous vowels (Ladefoged & Broadbent, 1957; Johnson, 1990; Evans & Iverson, 2003). Another possibility would be a version of extrinsic normalization in which transformational algorithms can be stored in long-term memory and retrieved quickly when a familiar talker is heard. If this is the case, a stored algorithm could contain information based on previous experiences with that talker, allowing for a better representation of that person's speech than if only the current speech input could be used to construct the algorithm. Additionally, storage of algorithms is an attractive option because it potentially requires less of a burden on long-term memory than the episodic account. Rather than needing to store traces of every instance of speech, stored algorithms would allow listeners to represent a talker's speech characteristics in a more abstract form.

An outstanding issue is why our results differed from those of Kraljic and Samuel (2007). One difference between our Experiment 1 and Kraljic and Samuel's experiment is that our trials were presented randomly, causing the talkers to alternate frequently. By contrast, Kraljic and Samuel's stimuli were blocked by talker, so participants listened to the same talker for many trials before hearing the other talker. According to episodic theories, speech sounds activate traces stored in memory, with activation being stronger for traces that are most similar to the spoken input (e.g., words spoken by the same talker, phonetically similar words, etc., Goldinger & Azuma, 2003). After completing an entire block of trials where only one talker was heard and where the stimuli were all quite similar to one another, listeners should have many highly active traces from that talker. When a new block with a different talker begins, a large processing cost is incurred because many traces from the first talker are still active. Because the stimuli in Experiment 1 were not presented by block, but rather in pseudo-alternating order, there was less of a chance for traces from one talker to accumulate, leading to less competition when a switch occurred and, subsequently, less of a visible processing cost. One could imagine a similar cost occurring for a memory-based extrinsic normalization proposal: Perhaps when one transformational algorithm is repeatedly accessed, it becomes more highly activated, making it more difficult to access a different representation.

Experiment 1's results suggested that there is a long-term memory component to talker variability accommodation that allows for information about a talker to be accessed and applied quickly. However, it is still unclear what types of talker-specific information are stored and if different types of information can all be applied rapidly. Additionally, participants may have been prompted by the preamble to access a stored algorithm or activate talker-specific

information. In Experiment 2, we investigated whether listeners could process an accented talker's speech without a priori acoustic cues to talker identity.

CHAPTER 3: EXPERIMENT 2

In Experiment 1, it was unclear whether information about the accented talker's speech was being compiled or activated during the preamble in advance of the target word, or whether it was being activated as the target word was being processed. Experiment 2 tested whether listeners were using the preamble to create a transformational algorithm or otherwise represent the talker's accent. In Experiment 2, listeners were given a visual, rather than auditory, cue to the identity of the upcoming talker. Switching to a non-auditory cue ensured that the first auditory input received on each trial was the onset of the target word, thus eliminating the possibility that a transformational algorithm was being created or accessed based on acoustic information prior to the start of the target word.

The use of a visual cue, rather than an auditory one, also served another purpose. Both the episodic view and newer iterations of the extrinsic normalization view allow for some use of contextual information in online speech processing. One could imagine that it would be useful if listeners could use this type of information as a signal to prepare for upcoming speech input. Indeed, it is well known that listeners integrate visual and acoustic information to arrive at a blended percept, as in the case of the McGurk effect (McGurk & MacDonald, 1976). More closely related to the current question, Johnson, Strand, and D'Imperio (1999) demonstrated that the gender of a concurrently-presented or imagined face shifted phoneme boundaries in a categorization task. Their results are consistent with our Experiment 1 findings, in that listeners brought learned expectations for how certain (categories of) talkers sound; the fact that these expectations were cued on the basis of a picture or by imagining the speaker suggests a very real contribution of non-linguistic, contextual information about talker identity on speech perception.

Here, we examine a similar issue for the case of accent accommodation, whether talker-specific expectations about accent can be cued on the basis of non-linguistic information alone. Crucially, unlike the McGurk effect, for our picture cue to have an effect, it would be giving the listener indexical information about the identity of the talker which could then be used to make inferences (though not necessarily explicitly) about how the talker would have pronounced the names of non-target pictures.

Experiment 2 tested whether non-linguistic information can be used to cue listeners to talker identity and allow them to activate talker-specific information in time to affect online speech processing. Experiment 2's method was identical to that of Experiment 1 except that, instead of hearing the preamble "*Click on*" before the target word, participants saw a picture of the talker. This paradigm made it possible to test whether listeners could prepare talker-specific information in the absence of acoustic cues.

On the normalization view, if participants do use the picture cue to guide comprehension (or, alternatively, if they simply do not need an acoustic cue), eliminating the *–ag* competitor on *–ack*-type trials, and considering *–ag* words to be competitors on *–ake*-type trials, this would not only suggest that non-linguistic information can be used within the framework of normalization, but also that transformational algorithms can either be retrieved from long-term memory or reconstructed based on previous instances of that talker's speech. Conversely, if participants do not use the cue, this would support the interpretation of the Experiment 1 findings that participants had constructed a transformational algorithm over the course of the preamble.

The episodic view makes strong predictions about the use of non-linguistic information. Specifically, it states that this type of information is used to limit the activation of traces to just

those that converge with the top-down information given. Therefore, if participants do use the picture cue to aid processing, it will confirm the current version of the episodic view. However, if participants do not use the picture cue, it would indicate either that this type of information is not stored within episodic traces or that it is stored but cannot be used to constrain online processing.

It was hypothesized that if participants were unable to use the picture cue to retrieve talker-specific information, they should have difficulty switching between talkers because they would not be able to apply any previously-stored information about the talker to aid in processing. On this account, there should be no difference in the proportion of fixations to target images between the accented and non-accented talker on either *–ack* or *–ake* trials. Alternatively, if participants were able to use the picture cue to retrieve talker-specific information, on *-ack* trials, participants should make more fixations to the target when hearing the accented talker, and on *-ake* trials, they should make more fixations to the target when hearing the unaccented talker.

Method

Participants

55 members of the University of Illinois at Urbana-Champaign community who did not participate in Experiment 1 participated. 17 additional participants were run but excluded from analysis due to technical difficulties (15), not meeting the participation criteria (1), and not completing the experiment (1). Participants received either payment ($16) or partial course credit for participation. All participants were native speakers of North American English and had normal or corrected-to-normal hearing and vision. The participants were surveyed to ensure that they did not share an accent with the accented talker.

Stimuli

The word list and recordings were the same as those used in Experiment 1.

Procedure

Like Experiment 1, Experiment 2 consisted of a training phase followed by a testing phase. The entire experiment lasted approximately 2 hours.

Training. Participants in Experiment 2 completed the same training as participants in Experiment 1, and all participants successfully completed the picture test. Participants then listened to the same dialogue between the accented and unaccented talkers as Experiment 1. During the dialogue, pictures of the two talkers were displayed on the computer monitor so that the participants would be able to associate the pictures with the talkers' voices.

Test. The procedure for the test was similar to that of Experiment 1. At the start of each trial, a fixation cross appeared for 1000 ms. Then, a picture of the talker appeared for 1000 ms, followed by a screen containing four stimulus pictures. The four pictures remained on the screen for 2000 or 4000 ms, and then the participant heard the target word produced by the talker who was pictured at the start of the trial. The latency between stimulus picture onset and speech onset was varied in order to test the hypothesis that participants were mentally rehearsing the picture names in the talker's accent before hearing the target audio, rather than processing the speech online. We hypothesized that if this was the case, the predicted effects would be larger in the 4000 ms condition because the participants would have more time to rehearse the picture names as if they were the talker (accented or unaccented, depending on the picture cue), thus emphasizing the similarities and differences between the target and competitor words created by the accented talker's accent. Preliminary analyses indicated two significant effects between delay

conditions. A main effect of delay was found on *–ake* word trials, $F_1(1, 54) = 7.0$, $p < .05$; $F_2(1, 21) = 8.5$, $p < .01$, such that that the overall proportion of target fixations was greater in the 4000 ms condition for both the accented and unaccented talker conditions (.49 vs .45 for the accented talker and .51 vs .48 for the unaccented talker). Additionally, a delay-by-time interaction was found on *–ack* word trials, but was only significant by items, $F_2(8, 168) = 3.4$, $p < .01$, ($\varepsilon = .381$). The nature of the interaction was such that at the start of the trial, participants had a higher proportion of target fixations in the 2000 ms condition, and at the end of the trial, participants had a higher proportion of target fixations in the 4000 ms condition, regardless of talker condition. Crucially, latency never interacted with talker, indicating that longer latencies did not exaggerate the differences between the two talkers' accents. The lack of a talker-by-latency interaction suggests that participants were not rehearsing the picture names prior to hearing the target audio, or that if they did, it did not improve accent accommodation. Due to our primary interest in talker effects, all subsequent analyses collapse across latency conditions.

Results

-ack *word trials*

We predicted that if participants used the picture cue to prepare talker-specific information, then they should show a greater proportion of fixations to the target when listening to the accented talker than when listening to the unaccented talker. The proportion of fixations to the target word was calculated by subject and item in 100 millisecond intervals beginning 200 ms before the onset of the critical word (e.g. *back*), continuing until 1000 ms after word onset. A baseline analysis from -200 ms to 200 ms revealed no difference between talker conditions ($t = 0.43$). Repeated measures ANOVAs by subject and by item were performed, with talker

(accented male vs. unaccented female) and time (200-1000 ms, in 100 ms intervals) as within-subject factors. The ANOVA revealed a significant main effect of time, $F_1(8, 432) = 494.1$, $p < .001$, ($\varepsilon = .286$); $F_2(8, 168) = 221.0$, $p < .001$, ($\varepsilon = .236$), due to increasing fixations to the target as the trial progressed. The main effect of talker was also significant, $F_1(1, 54) = 21.3$, $p < .001$; $F_2(1, 21) = 25.3$, $p < .001$, with a greater proportion of fixations to the target when hearing the accented male talker than when hearing the unaccented female talker (.45 and .40, respectively, see Figure 4). The main effects were qualified by a significant interaction between talker and time, $F_1(8, 432) = 22.5$, $p < .001$, ($\varepsilon = .370$); $F_2(8,168) = 13.6$, $p < .001$, ($\varepsilon = .278$). A series of planned comparisons indicated that this difference was significant from 500 to 1000 ms by both subjects and items ($ts > 2.22$).

-ake *word trials*

On *–ake* trials, we predicted that if participants use the picture cure, they should show fewer target fixations when listening to the accented talker than when listening to the unaccented talker. An analysis of the time window from -100 ms to 200 ms indicated that there was no baseline effect of talker ($t = 1.83$). Repeated measures ANOVAs by subject and by item were performed, with talker (accented male vs. unaccented female) and time (200-1000 ms, in 100 ms intervals) as within-subject factors. The main effect of time was significant, $F_1(8, 432) = 719.5$, $p < .001$, ($\varepsilon = .288$); $F_2(8, 168) = 126.1$, $p < .001$, ($\varepsilon = .171$), due to increased target fixations as the trial progressed. The main effect of talker was also significant, $F_1(1, 54) = 10.0$, $p < .01$; $F_2(1, 21) = 8.7$, $p < .01$, ($\varepsilon = .366$), with fewer target fixations in the accented talker condition than in the unaccented talker condition (.47 and .50, respectively, see Figure 5). A series of paired comparisons at each 100 ms interval indicated that the effect of speaker was significant between

500 and 900 ms by subject (*ts* > 2.09) and from 500-800 ms by item (*ts* > 2.36). The interaction between talker and time was not significant.

Discussion

In Experiment 2, we found that on-*ack* trials, listeners were significantly more likely to fixate the target when listening to the accented, compared to the non-accented talker. The effect of talker was also significant on -*ake* trials: listeners were significantly less likely to fixate the target when listening to the accented, compared to the non-accented talker.

These findings indicate that the effect found in Experiment 1 was not due to the creation of transformational algorithms during the preamble. Instead, listeners accessed information from memory about the talker's accent, possibly on the basis of the picture cue; this information was subsequently used to process the target word. There are several implications for the episodic and normalization accounts that follow from these results.

The episodic view currently accounts for the inclusion of non-linguistic contextual information in stored episodic traces, and our findings are consistent with this aspect of the theory. Our results also speak to the speed with which this information can be applied. In Experiment 2, listeners were able to accommodate the talkers' accents over the course of a single word, indicating either that the process of limiting traces based on contextual information begins very shortly after speech input begins, or that prior to speech perception, non-linguistic contextual information limits the potential pool of traces that can be activated to those that match the context (e.g., the talker's identity).

On a normalization account, the results of Experiment 2 suggest that non-linguistic information about talker identity may be used during the normalization process and quickly

applied during processing. This result is consistent with Magnuson and Nusbaum's (2007) view of extrinsic normalization as an active control process which makes use of contextual information.

However, another possibility is that participants did not use the visual cue to prepare talker-specific information in advance of the target word. Instead, they may have successfully accommodated the accented talker's vowel shift based solely on acoustic information contained in the onset of the target word. If so, this would suggest that simply hearing one consonant (or consonant cluster) is enough to activate specific information about the rest of that talker's phonemic inventory.

We explored this possibility in Experiment 3 by eliminating all pre-speech cues to talker identity.

CHAPTER 4: EXPERIMENT 3

Experiments 1 and 2 showed that listeners are able to quickly incorporate talker-specific information during speech processing, even when talkers are alternating rapidly and unpredictably. In each of these two experiments, participants were given talker-identifying information before the onset of the critical word (an auditorily presented preamble and a picture of the talker, respectively). We hypothesized that participants could have been using these cues to access either episodic traces or a transformational algorithm based on previous exposure to that talker. This information could then have been used during the processing of the target word.

However, it is possible that in one or both of these experiments, participants were not using the cue at all, and were instead accessing talker-specific information based only on the onset of the target stimulus word. Under a normalization account, this would mean that enough information must be available in the word onset to identify the talker. Once the talker has been identified, his or her transformational algorithm can be retrieved from memory or rebuilt from what has been remembered of previous experiences with that talker. Under an episodic view, hearing the onset of the target word could activate top-down constraints such that only traces from that talker could become active. In both cases, it must be assumed that hearing one phoneme produced by a particular talker is enough to activate knowledge about how the talker would produce other sounds. Additionally, this process would have to happen rapidly enough to immediately guide processing of the subsequent phoneme.

Method

Participants

59 members of the University of Illinois at Urbana-Champaign community who had not participated in Experiments 1 or 2 participated. 34 additional participants were run but excluded from analysis due to technical difficulties (31), not completing the experiment (1), and making a comparatively low number of fixations during the experiment (i.e., fewer than half as many as any other participant) (2). Participants received either partial course credit or payment ($16) for their participation. All participants were native speakers of North American English and had normal or corrected-to-normal hearing and vision. The participants were surveyed to ensure that they did not share the accented talker's accent.

Stimuli

The word list and recordings were the same as those used in Experiments 1 and 2.

Procedure

Like Experiments 1 and 2, Experiment 3 consisted of a training phase followed by a testing phase. The entire experiment lasted approximately 2 hours.

Training. Participants in Experiment 3 completed the same training as participants in the previous two experiments. They were first trained on the names of the pictures that were used in the test phase of the experiment; however, the test on the picture names that was given in Experiments 1 and 2 was deemed not to be necessary in Experiment 3 because all participants in the previous two experiments had successfully passed the test on the first try. Participants then listened to the same training dialogue that was used in Experiments 1 and 2.

Test. The test phase was very similar to those of Experiments 1 and 2. Each trial began with a fixation cross, which appeared for 1000 ms, followed by a screen containing four pictures.

After 3000 ms, the target word was presented. The two talkers alternated randomly throughout the experiment.

Results

-ack *word trials*

We predicted that if listeners were able to use talker-specific information with no cue to talker identity preceding the target audio, they should fixate the target more when listening to the accented talker than when listening to the unaccented talker. The proportion of fixations to the target was calculated by subject and item in 100 millisecond intervals beginning 200 ms before the onset of the critical word and continuing until 1000 ms after word onset. A baseline analysis from -200 ms to 200 ms revealed no baseline difference between talker conditions ($t = 1.23$). Repeated measures ANOVAs by subject and by item were performed, with within-subject factors of talker (accented male vs. unaccented female) and time (200-1000 ms, in 100 ms intervals). A significant main effect of time, $F_1(8, 464) = 453.7$, $p < .001$, ($\varepsilon = .218$); $F_2(8, 168) = 177.7$, $p < .001$, ($\varepsilon = .200$), was due to increasing fixations to the target as the trial progressed. A significant main effect of talker, $F_1(1, 58) = 17.3$, $p < .001$; $F_2(1, 21) = 22.8$, $p < .001$, was due to a larger proportion of target fixations on trials with the accented male talker than trials with the unaccented female talker (.44 and .40, respectively, see Figure 6). The main effects were qualified by a significant interaction between talker and time, $F_1(8, 464) = 7.1$, $p < .001$, ($\varepsilon = .418$); $F_2(8,168) = 5.2$, $p < .001$, ($\varepsilon = .350$). A series of planned comparisons indicated that this difference was significant from 500-1000 ms by subject ($ts > 2.03$) and from 600-1000 ms by item ($ts > 2.82$).

-ake *word trials*

On *–ake* trials, we predicted that if listeners can use talker-specific information without cues to the talker's identity, they should fixate the target more on trials where they heard the unaccented talker than when they heard the accented talker. An analysis of the time window from -100 ms to 200 ms indicated that there was no baseline effect of talker (t = .44). Repeated measures ANOVAs by subject and by item were performed, with talker (accented male vs. unaccented female) and time (200-1000 ms, in 100 ms intervals) as within-subject factors. The main effect of time was significant, $F_1(8, 464) = 571.2$, $p < .001$, ($\varepsilon = .237$); $F_2(8, 168) = 156.8$, $p < .001$, ($\varepsilon = .170$), due to increased target fixations as the trial progressed. The main effect of talker was marginally significant in the by-subjects ANOVA only, $F_1(1, 58) = 3.4$, $p = .07$, with fewer target fixations in the accented talker condition than in the unaccented talker condition (.47 and .49, respectively, see Figure 7). The interaction between time and talker was significant in the by-subjects analysis, $F_1(8, 464) = 2.7$, $p < .01$. A series of paired comparisons at each 100 ms interval indicated that the effect of speaker was significant, by subject only, between 600 and 1000 ms ($ts > 2.14$).

Discussion

Experiment 3 demonstrated that participants were able to rapidly accommodate the speech of the two talkers even without prior knowledge of the talker's identity on a given trial. In order for the participants to be able to rule out or include potential competitors based on the identity of the talker, talker-specific information would have to be active before or during the processing of the target word's vowel. That participants were successful indicates that talker-specific information about the production of the /æ/ and /eɪ/ vowels was activated during the onset consonant of the target words. This suggests that hearing a very small sample of a talker's

speech (in this case, one consonant or a two-consonant cluster) may provide activation to all of a listener's representations of that talker's speech. This result is in line with other findings that listeners rapidly integrate fine-grained acoustic information, such as vowel duration and prosody, into the on-going interpretation of a word (Salverda et al., 2003; Dahan et al., 2001; Dahan, Tanenhaus, & Chambers, 2002; McMurray, Aslin, Tanenhaus, Spivey, & Subik, 2008).

The results of Experiment 3 are also consistent with evidence that listeners transfer knowledge of how a talker would produce one phoneme to perception of other, previously unheard phonemes. For example, Nielsen (2006) demonstrated that participants extended their knowledge about a talker's voice onset time from one voiceless stop to another. Our study goes beyond these findings by showing that listeners can transfer knowledge of a particular talker's vowel space across phoneme classes: The patterns of lexical competition in Experiment 3 demonstrated that upon hearing a single phoneme from a particular talker, listeners activated information about very different classes of phonemes (e.g., for "back," hearing a voiced stop would have to activate information about a vowel). Together, these findings provide evidence that listeners activate information not only about what they hear, but also about other sounds that they have previously heard that particular talker produce. The fact that listeners in both of these studies experienced only a small inventory of each talker's speech may have facilitated this transfer process. Thus, an open question is how the same processes might operate in relatively unconstrained contexts, such as natural conversation.

CHAPTER 5: GENERAL DISCUSSION

The results of Experiments 1-3 indicate that some form of talker-specific information is stored in long-term memory and can be quickly accessed, even in the presence of minimal acoustic input, to aid in accommodating talker variability during online speech processing. Non-linguistic information, such as the talker's appearance, may not overwhelmingly increase talker adaptation beyond this minimal acoustic information in all situations. Overall, the findings of these experiments demonstrate that listeners are able to easily interpret the speech of two talkers with different accents, even when the talkers alternate quickly and frequently.

A role for contextual information?

The fact that the results of Experiment 3 were strikingly similar to those of Experiments 1-2 suggests that having an acoustic or visual cue to the upcoming talker may not provide much additional processing benefit beyond what is obtained from the onset consonant. While small methodological differences between the experiments prevent us from making direct comparisons between the experiments, the time-course and size of the effects are generally comparable. As with all null effects, the lack of an obvious benefit from contextual cues does not mean this information is not used to facilitate accommodation processes, nor does it speak to the question of whether contextual information is stored along with talker-specific speech characteristics (e.g., Goldinger & Azuma, 2003; Magnuson & Nusbaum, 2007). Instead, our findings highlight the speed with which the language processing system integrates small bits of acoustic information with stored talker-specific representations, to arrive at a talker-specific interpretation of a word.

When might contextual information be the most useful? We suspect that the use of two talkers and a small lexical inventory may have limited the usefulness of the linguistic

(Experiment 1) and non-linguistic (Experiment 2) contextual cues we provided. With a total of 66 test words, it is unlikely that listeners stored tokens of each word from each talker in working memory. However, listeners may have stored some information about each talker's vocal characteristics in working memory, thus facilitating rapid accommodation on the basis of limited acoustic input (i.e., Experiment 3). In contexts where an upcoming talker's vocal characteristics are unlikely to be stored in working memory, an early cue to an upcoming talker might provide more of a benefit to accent accommodation. Candidate situations include multi-party conversations with a large number of talkers, or cases in which a listener has not heard a familiar talker's voice for a long period of time. Contextual cues to talker identity might also be more useful in cases where talker identity is not easily gleaned from the onset consonant. Because our talkers were of different gender and spoke with different pitch ranges, the onset consonant was generally a good cue to talker identity. The onset consonant may be a less useful cue in multi-talker situations with talkers of the same gender; in these cases, a pre-speech cue to talker identity may provide more of a benefit.

Asymmetry of results on –ake and –ack word trials

A puzzling but consistent finding was that the magnitude of the accommodation effect appears much larger for *–ack* word trials, compared to *–ake* word trials, across all three experiments. Why might this be the case? We suspect that the specific mechanisms of competition reduction may be at play. Talker-specific processing on *back* trials involved the elimination of a cohort competitor, *bag*, when the accented talker was speaking. In contrast, talker-specific processing on *bake* trials involved the inclusion of a new competitor, *bag*, when the accented talker was speaking. Perhaps exclusion of potential competitors is an easier or more

common task, as various linguistic and non-linguistic constraints routinely eliminate potential cohort competitors from consideration (Dahan & Tanenhaus, 2004; Brown-Schmidt & Tanenhaus, 2008). Fully understanding the mechanisms at play will undoubtedly require explicit models of the relevant processes.

Additionally, small differences in production may have driven this asymmetry. Although the accented talker's accented /æ/ approached /eɪ/, the two vowels were not produced identically. Thus, on *–ake* trials, upon hearing the /eɪ/ vowel of the target word, participants may have been able to eliminate the *–ag* competitor relatively easily due to their knowledge of the slight difference between the /eɪ/ vowel heard during the trial and the accented /æ/ vowel that would have been produced if the talker had been saying an *–ag* word. (Conversely, *-ack* trials did not depend on the similarity of the accented /æ/ to a standard /eɪ/ vowel, but rather on the much greater difference between the talker's unaccented /æ/ and accented /æ/ vowels.) The discrepancy between the effect sizes in the two conditions could therefore also indicate that listeners encoded and used detailed information about the accented vowel, rather than simply assimilating it into the closest pre-existing phonemic category.

Implications for models of accommodation

In this final section, we consider how the results of Experiments 1-3 can inform and constrain episodic and extrinsic normalization views of accommodation.

How might an episodic view account for our results? First, let us assume that when the listener hears the initial consonant of the target word, words that share that initial consonant are activated. Traces of the target stimuli that were heard in the experiment are likely the most highly activated, as they have been activated most recently (i.e., during the training dialogue and

previous trials during the test phase). As the initial consonant is interpreted, only those traces associated with the current talker remain activated, perhaps through a process in which the talker is first identified, and then active traces are limited to those tagged as being produced by the current talker. Alternatively, selective activation of traces associated with the current talker may be accomplished by a similarity-based mechanism whereby activation is limited to traces that match the initial acoustic input based on sub-phonemic acoustic properties, including the speaker's timbre and pitch. After the traces produced by the current talker have been identified and other traces ruled out, perception of the target vowel should result in an interpretation of the stimulus that is consistent with the current talker's accent. For instance, if the listener hears the accented talker say /b/, he should activate tokens of the accented talker producing *back*, *bake*, and *bag*. Then, when the accented talker says /æ/, *back* traces should receive the most activation because they are most similar to the acoustic input. Because *bake* and *bag* contain a dissimilar vowel, the listener should interpret the target stimulus as *back*.

How might an extrinsic normalization view account for our results? The results from Experiments 1 through 3 appear inconsistent with some versions of the extrinsic normalization theory. On these accounts, listeners must start building a new transformation from scratch every time a new speaker begins talking (so long as their vowel spaces are sufficiently distinct, Nusbaum & Morin, 1992). If no information from previous experience can be utilized, adaptation to the speaker's accent should be slow and resource intensive in multi-talker contexts, a prediction which is inconsistent with our findings. However, our results are consistent with an extrinsic normalization view if it assumes either that: (a) transformational algorithms for familiar talkers are stored in long-term memory and can be accessed and applied during online speech

processing, or (b) information about a talker's speech is stored in long-term memory and can be used to re-create transformational algorithms when that talker is cued.

We suspect that the latter proposal may be untenable, given the speed with which a transformational algorithm would need to be reconstructed to account for our findings in Experiment 3 (although clearly this depends on the nature of the reconstruction process). We would like to suggest, then, a memory-based extrinsic normalization view in which transformational algorithms are stored in memory. When a listener hears the initial phoneme of a target stimulus, the identity of the talker is quickly determined, and his or her transformational algorithm is accessed. This algorithm can then be applied in order to correctly interpret the input. If, as some have suggested, that this process operates like a hypothesis-testing procedure (Magnuson & Nusbaum, 2007; Nusbaum & Magnuson, 1997), stored characteristics of a particular talker's speech are used to eliminate or introduce possible options for interpretation. For example, when the accented talker produces a /b/, the listener first identifies the talker. After the talker has been identified, stored information about that talker can be taken into consideration as the rest of the word is interpreted. So, when the talker produces the /æ/ vowel, the listener can use the talker-specific information that he retrieved to eliminate the possibility that a /g/ is coming up. Therefore, he can predict that *back* is likely to be the target word.

Conclusions

Results from three experiments on the online interpretation of accented and unaccented words demonstrate that in multi-talker environments, listeners store information in memory about the vocal characteristics of the talkers and use this information to rapidly accommodate the current talker's accent. These accommodation effects were observed on the perception of

unaccented words, demonstrating that listeners used their knowledge of each talker's full phonemic inventory when interpreting a given word. These findings add to the evidence against self-normalizing views of normalization in which each segment of speech provides information necessary for the normalization process (e.g., Syrdal & Gopal, 1986; see discussion in Nusbaum & Morin, 1992), as well as versions of extrinsic normalization which propose that the normalization process begins anew when a talker with a sufficiently different vowel space begins speaking (Nusbaum & Morin, 1992). Instead, the results are consistent with episodic theories of accommodation, with the caveat that the dynamics of trace activation must be prompt enough to support talker-specific activation of traces based on perception of an onset consonant. A memory-based extrinsic normalization view could account for our findings as well, with the caveats that talker-specific normalization must be stored in memory and rapidly accessible. The similarity of results across the three experiments suggests that in certain circumstances, contextual cuing of an upcoming talker does not overwhelmingly speed the accommodation process; identifying the cases in which this information can speed processing is an important goal for future research. More importantly, and finally, the results of our experiments show that listeners rapidly access information about a talker's vowel space on the basis of minimal acoustic input—in this case a single consonant or consonant cluster—and integrate this information into the ongoing interpretation of the word at a speed fast enough to eliminate, or induce, competition effects that are hallmarks of the real-time mapping of the unfolding acoustic signal onto lexical candidates (Allopenna, et al., 1998; McMurray, et al., 2008).

FIGURES

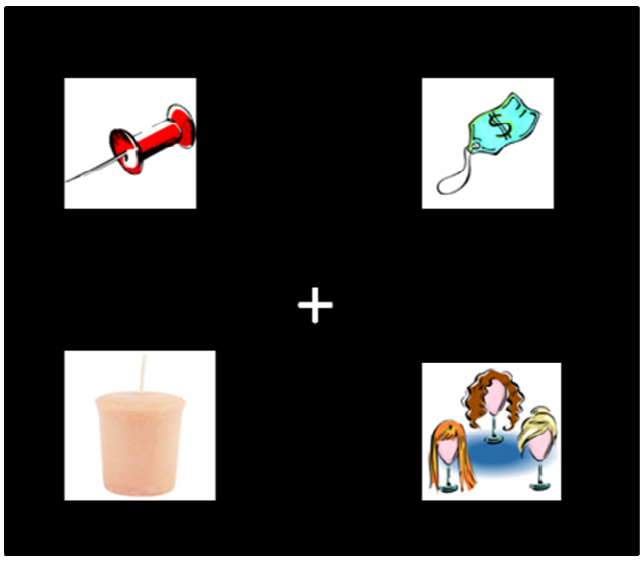

Figure 1. Example of a display for an *–ack* target trial. The target word is *tack* (upper left), and the competitor word is *tag* (upper right). Participants heard "*Click on tack*."

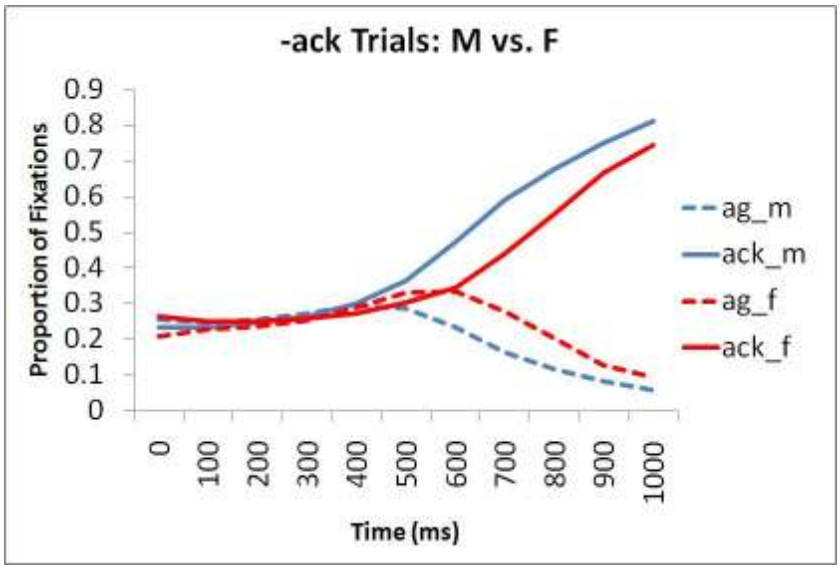

Figure 2. Proportion of fixations to the target and *–ag* competitor on *–ack* trials for male and female speakers as a function of time in Experiment 1.

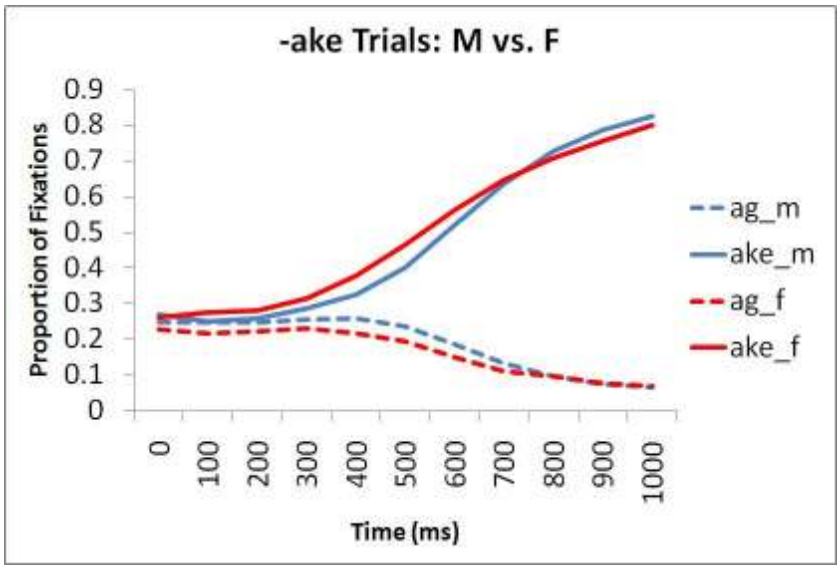

Figure 3. Proportion of fixations to the target and *–ag* competitor on *–ake* trials for male and female speakers as a function of time in Experiment 1.

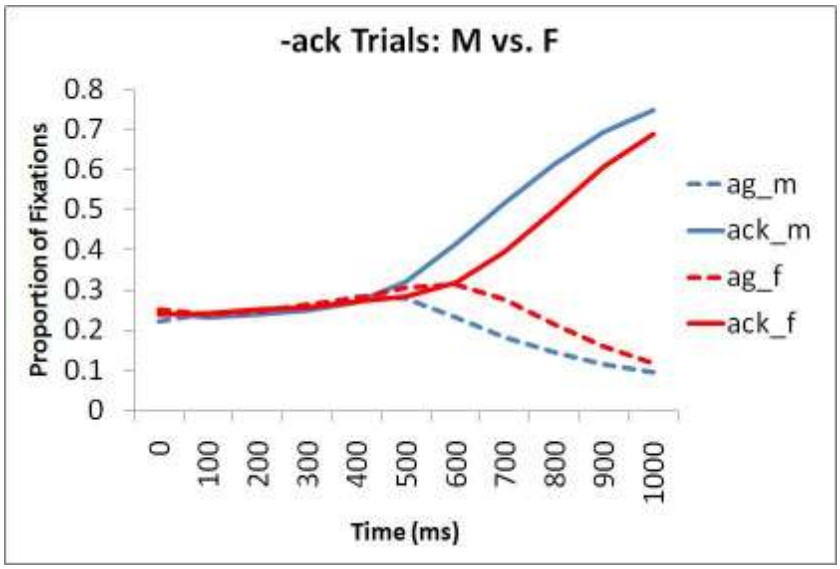

Figure 4. Proportion of fixations to the target and *–ag* competitor on *–ack* trials for male and female speakers as a function of time in Experiment 2.

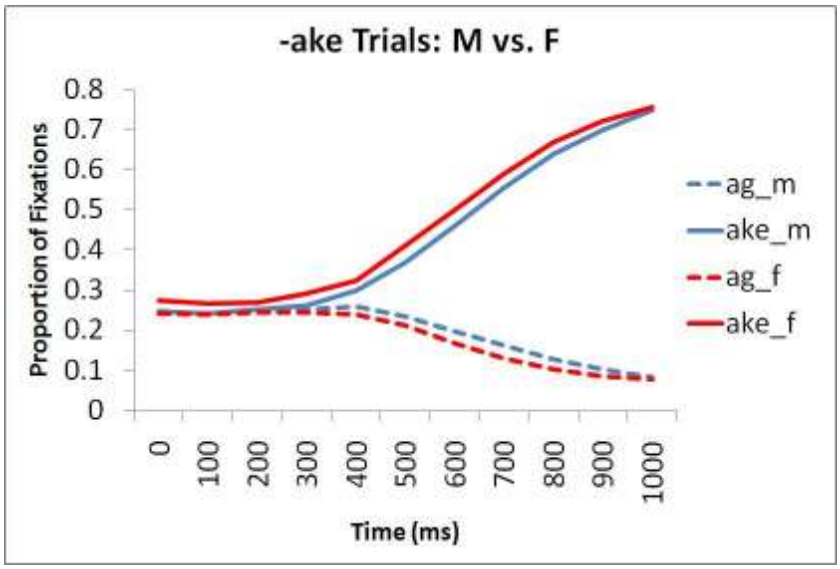

Figure 5. Proportion of fixations to the target and *–ag* competitor on *–ake* trials for male and female speakers as a function of time in Experiment 2.

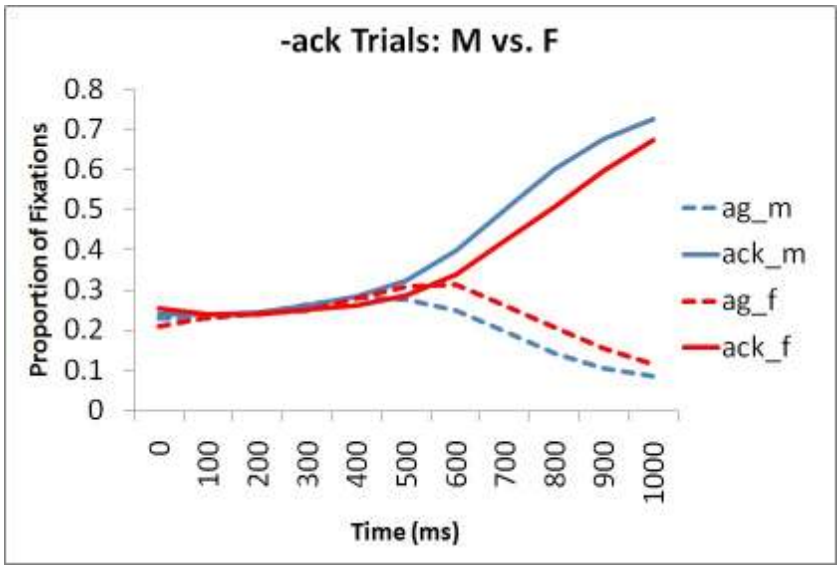

Figure 6. Proportion of fixations to the target and *–ag* competitor on *–ack* trials for male and female speakers as a function of time in Experiment 3.

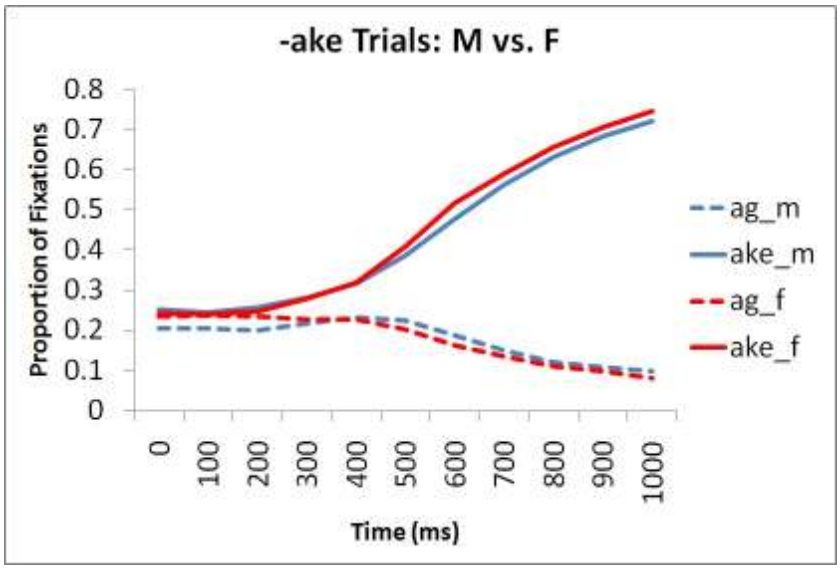

Figure 7. Proportion of fixations to the target and *–ag* competitor on *–ake* trials for male and female speakers as a function of time in Experiment 3.

REFERENCES

Allopenna, P.D., Magnuson, J.S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, *38*, 419-439.

Brainard, D. H. (1997) The Psychophysics Toolbox, Spatial Vision 10:433-436.

Brown-Schmidt, S., & Tanenhaus, M. K. (2008). Real-time investigation of referential domains in unscripted conversation: A targeted language game approach. *Cognitive Science, 32*, 643-684.

Creel, S. C., Aslin, R. N., & Tanenhaus, M. K. (2008). Heeding the voice of experience: The role of talker variation in lexical access. *Cognition, 106,* 633-664.

Dahan, D. (2010). The time course of interpretation in speech comprehension. *Current Directions in Psychological Science, 19*, 121-126.

Dahan, D., Drucker, S. J., & Scarborough, R. A. (2008). Talker adaptation in speech perception: Adjusting the signal or the representations? *Cognition*, *108*, 710-718.

Dahan, D., Magnuson, J. S., Tanenhaus, M. K., & Hogan, E. M. (2001). Subcategorical mismatches and the time course of lexical access: Evidence for lexical competition. *Language and Cognitive Processes, 16*, 507-534.

Dahan, D., Tanenhaus, M. K., & Chambers, C. G. (2002). Accent and reference resolution in spoken-language comprehension. *Journal of Memory and Language, 47*, 292-314.

Dahan, D., & Tanenhaus, M. K. (2004). Continuous mapping from sound to meaning in spoken-language comprehension: Immediate effects of verb-based thematic constraints. *Journal of Experimental Psychology, Learning, Memory and Cognition, 30*, 498-513.

Evans, B.G., & Iverson, P. (2003). Vowel normalization for accent: An investigation of best exemplar locations in northern and southern British English sentences. *Journal of the Acoustical Society of America, 115*, 352-361.

Fujisaki, H. & Kawashima, K. (1968). The roles of pitch and higher formants in the perception of vowels. *IEEE Transactions on Audio and Electroacoustics*, *AU-16*, 73-77.

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

Goldinger, S. D. & Azuma, T. (2003). Puzzle-solving science: The quixotic quest for units in speech perception. *Journal of Phonetics*, *31*, 305-320.

Gordon, P. C. (1988). Induction of rate-dependent processing by coarse-grained aspects of speech. *Perception & Psychophysics, 43*, 137–146.

Grossberg, S. (1980). How does the brain build a cognitive code? *Psychological Review*, *87*, 1-51.

Hawkins, S. (2003). Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of Phonetics*, *31*, 373-405.

Johnson, K. (1990). The role of perceived speaker identity in F0 normalization of vowels. *Journal of the Acoustical Society of America, 88*, 642-654.

Johnson, K. (1997) Speech perception without speaker normalization: an exemplar model. In: *Talker variability in speech processing* (K. Johnson, & J. Mullennix, editors), pp. 145-166. New York: Academic Press.

Johnson, K., Strand, E. A., & D'Imperio, M. (1999). Auditory–visual integration of talker gender in vowel perception. *Journal of Phonetics*, *27*, 359 –384.

Joos, M. (1948). Acoustic phonetics. Baltimore: Linguistic Society of America.

Kraljic, T. & Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. *Journal of Memory and Language*, *56*, 1-15.

Ladefoged. P., & Broadbent, D. E. (1957). Information conveyed by vowels. *Journal of the Acoustical Society*, *29*, 98-104.

Levelt, W. J. M. (1989). *Speaking.* Cambridge: MIT Press.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review, 74*, 431– 461.

Magnuson, J. S., Tanenhaus, M. K., Aslin, R. N., & Dahan, D. (2003). The time course of spoken word learning and recognition: Studies with artificial lexicons. *Journal of Experimental Psychology: General, 132*, 202-227.

Magnuson, J. S. & Nusbaum, H. C. (2007). Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. *Journal of Experimental Psychology: Human Perception and Performance*, *33*, 391-409.

Martin, C. S., Mullennix, J. W., Pisoni, D. B., & Summers, W. V. (1989). Effects of talker variability on recall of spoken word lists. *Journal of Experimental Psychology: Language, Memory, and Cognition*, *15*, 676-684.

Maye, J., Aslin, R. N., & Tanenhaus, M. K. (2008). The Weckud Wetch of the Wast: Lexical adaptation to a novel accent. *Cognitive Science*, *32*, 543-562.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*, 746-748.

McMurray, B., Aslin, R., Tanenhaus, M., Spivey, M., and Subik, D. (2008). Gradient sensitivity to within-category variation in speech: Implications for categorical perception. *Journal of Experimental Psychology, Human Perception and Performance, 34*, 1609-1631.

Miller, J. D. (1989). Auditory-perceptual interpretation of the vowel, *Journal of the Acoustical Society of America, 85*, 2114-2134.

Miller, J. L., & Liberman, A. M. (1979). Some effects of later-occurring information on the perception of stop consonant and semivowel. *Perception & Psychophysics, 25*, 457– 465.

Mullennix, J. W., Pisoni, D. B., & Martin, C. S. (1989). Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America*, *85*, 365-378.

Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. *Journal of the Acoustical Society of America*, *85*, 2088-2113.

Nielsen, K. Y. (2006). Specificity and generalizability of spontaneous phonetic imitation. In Proceedings of the 9th International Conference on Spoken Language Processing, 1–4.

Nusbaum, H.C., & Magnuson, J. S. (1997). Talker normalization: Phonetic constancy as a cognitive process. In K. Johnson & J. W. Mullennix (Eds.), *Talker variability in speech processing* (pp. 109-132). San Diego, CA: Academic Press.

Nusbaum, H. C., & Morin, T. M. (1992). Paying attention to differences among talkers. In Y. Tohkura, Y. Sagisaka, & E. Vatikiotis-Bateson (Eds.), Speech perception, speech production, and linguistic structure (pp. 113–134). Tokyo: OHM.

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

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

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies, *Spatial Vision 10*, 437-442.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of vowels. *Journal of the Acoustical Society of America*, *24*, 175–184.

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.

Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. (1981). Speech perception without traditional speech cues. *Science*, *212*, 947-950.

Salverda, A. P., Dahan, D., & McQueen, J. M. (2003). The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. *Cognition*, *90*, 51-89.

Syrdal, A. K. & Gopal, H. S. (1986). A perceptual model of vowel recognition based on the auditory representation of American English vowels. *Journal of the Acoustical Society of America*, *79*, 1086-1100.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science, 268,* 1632-1634.

APPENDIX A: LIST OF STIMULI FOR TESTING PHASES OF EXPERIMENTS 1-3

-ag word	*-ack* word	*-ake* word	/g/-end filler	/k/-end filler	/k/-end filler
bag	back	bake	League	leak	luck
flag	flack	flake	Chug	chuck	check
jag	jack	Jake	Dog	dock	duck
lag	lack	lake	Smog	smock	smoke
rag	rack	rake	Lug	luck	lick
sag	sack	sake	Plug	pluck	peak
shag	shack	shake	Pug	puck	pick
snag	snack	snake	bug	buck	beak
stag	stack	stake	tug	tuck	took
tag	tack	take	wig	wick	week
wag	whack	wake	jog	jock	joke