UNIVERSITY OF CALIFORNIA
RIVERSIDE

Do you Hear What I See?
The Voice and Face of a Talker Similarly Influence the Speech of Multiple Listeners

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Psychology

by

Kauyumari Sanchez

March 2011

Dissertation Committee:

Dr. Lawrence Rosenblum, Chairperson
Dr. Chris Chiarello
Dr. Steven Clark

The Dissertation of Kauyumari Sanchez is approved:

_____

_____

_____

Committee Chairperson

University of California, Riverside

Acknowledgements

There are many people I would like to thank for their role in helping me attain many wonderful accomplishments, including this Dissertation. First, I would like to express a great deal of gratitude to my advisor, Lawrence Rosenblum. His patience, insight, and encouragement are truly invaluable and will continue to shape me as a scientist. I would also like to give a special thanks to my Dissertation committee, Chris Chiarello and Steven Clark, who have provided support and insight in past committee roles. In addition, appreciation should be bestowed upon my colleges in the department, especially Patrick LaShell for his incalculable help in everyway imaginable. Recognition should also be given to the lab mates that have come and gone, especially Rachel Miller. Within this vein, I would also like to acknowledge past and present research assistants who have helped in the many experiments that have failed and succeeded. Appreciation is also given to the Psychology Department Staff, Faye Harmer, Dianne Fewkes, and Conrad Colindres who have assisted me in numerous ways over the years. Also, special thanks is given to my undergraduate mentors, Lorin Lachs and Carl Oswald, who helped spark the fire of the scientist within me and encouraged me to pursue my current path. And finally, last, but not least, I would like to give immense gratitude to my close friends and family, especially my aunt Maria. Her impact on my life is without comparison and has made me who I am today. She is an inspiration and embodies the notions of love and sacrifice.

Thank you all so very much!

To lovers, dreamers, and scientists.

ABSTRACT OF THE DISSERTATION

Do you Hear What I See?
The Voice and Face of a Talker Similarly Influence the Speech of Multiple Listeners

by

Kauyumari Sanchez

Doctor of Philosophy, Graduate Program in Psychology
University of California, Riverside, March 2011
Dr. Lawrence Rosenblum, Chairperson

Speech alignment occurs when interlocutors shift their speech to become more similar to each other. Alignment can also be found when one is asked to shadow (quickly say out-loud) perceived words recorded from a model. Prior investigations on alignment have addressed whether shadowers of auditory (e.g. Goldinger, 1998) or visual (e.g. Miller, Sanchez, & Rosenblum, 2010) speech would shift in the direction of a model. However, it is unknown whether multiple shadowers align to a specific model in the same ways or uniquely. This Dissertation addressed two questions: *Are utterances of shadowers of the same model more similar to each other than they are to the utterances of shadowers of a different model? Does the sensory modality of the shadowed speech affect the perceptual similarity between the shadowers of the same model?* In Experiment Series 1, evidence that shadowers similarly aligned to the auditory speech of a model was obtained. In Experiment 1a perceptual raters judged the utterances of

shadowers of the same heard model as being more similar than utterances from shadowers of another heard model. In Experiment 1b it was found that the results from Experiment 1a were due to speech style shifts towards those of the shadowed model and that the shadowers were not similar before exposure to the model. Acoustical analyses of the shadowed words also revealed that shadowers of the same model were more similar along some acoustic dimensions to each other than words from shadowers of a different model. The articulatory dimensions behind these similar acoustic dimensions could also potentially be perceived in visible articulation, suggesting that the results from Experiment 1a might also be found for shadowers of *visual* speech (lip-reading). In Experiment Series 2, evidence that shadowers similarly aligned to the visual speech of a specific model was obtained. In Experiment 2a perceptual raters judged the utterances of shadowers of the same lip-read model as being more similar than the shadowed utterances of the other lip-read model. Experiment 2b compared auditory and visually shadowed speech of shadowers of the same or a different model. Utterances of multiple shadowers of the same model were judged as being more similar than those of shadowers of another model, regardless of whether the model's speech was shadowed auditorily or visually. These results suggest that shadowers align to similar properties of a specific model's speech even when doing so based on different modalities. Implications for episodic encoding and gestural theories are discussed.

Table of Contents

List of Tables

Chapter 1

Introduction

A multitude of factors, including life experiences and biology, make each individual unique. The effects of these factors are evident in one's personality and even in how one speaks: one's *talker-specific characteristics*. People are influenced not only by the thoughts and ideas of others, but they are also influenced by the *way* others speak (Goldinger, 1998; Shockley, Sabadini, & Fowler, 2004; Namy, Nygaard, & Sauerteig, 2002; Pardo, 2006; Nielsen, 2008; Miller, Sanchez, & Rosenblum, 2010; Sanchez, Miller, & Rosenblum, 2010). The influential effect of others may be found in the words people say and how these words are used. However, the influencing effect of others can also be found in how one *articulates* words, a phenomenon commonly referred to as *speech alignment*.

Speech alignment research has revealed that when listeners hear the speech of a talker—or *model*—the listeners produce speech that is more similar to that model (Goldinger, 1998; Goldinger & Azuma, 2004; Shockley et al., 2004; Namy et al., 2002; Pardo, 2006; Nielsen, 2008; Miller et al., 2010; Sanchez et al., 2010). However, it is not yet known whether multiple listeners are influenced in the same way when perceiving the

same talker. The nature of the speech information that influences alignment is also unknown. These two unresolved issues are the impetus of this Dissertation.

This Dissertation addresses speech alignment and the nature of talker-specific characteristics that induce speech alignment. This Dissertation attempts to answer the following questions: *Are utterances of shadowers of the same model more similar to each other than they are to the utterances of shadowers of a different model? Does the sensory modality of the shadowed speech affect the perceptual similarity between the shadowers of the same model?*

These questions were investigated through the use of perceptual ratings and some acoustical analyses. The speech of shadowers influenced by a model perceived auditorily or visually (lip-read) was compared to the speech of shadowers who perceived a different model in these ways. The outcomes of the experiments within this Dissertation are relevant to the episodic encoding (Goldinger, 1998) and gestural theories (Liberman & Mattingly, 1985, 1989; Fowler, 1986) of speech.

1.1 Dissertation Organization

The Dissertation is organized in the following way. Chapter 1 is comprised of a brief introduction and the outline of the Dissertation. Chapter 2 provides an in-depth review of the speech alignment literature. Chapter 3 discusses other relevant concepts in the speech literature and presents theoretical perspectives. Chapter 4 introduces the experiments of the Dissertation and rationale. Chapter 5 is comprised of Experiment Series 1. Chapter 6 is comprised of Experimental Series 2. Chapter 7 presents a general

discussion of the results of the experiments, expands on theoretical implications, and concludes the Dissertation.

Chapter 2

Alignment

2.1 Speech Alignment

Speech alignment has been found to occur unconsciously and spontaneously in the laboratory setting for both socially-interactive and socially-isolated experiments. For example, while engaging in a shared task, participants have been found to shift their speech in the direction of their fellow interlocutor on speech tempo, intonational contour, voice-onset time, as well as to more phonetically-relevant dimensions such as vowel spectra (e.g., Giles, Coupland, & Coupland, 1991; Gregory, 1990; Natale, 1975; Sancier & Fowler, 1997; Pardo, 2006). Speech alignment has also been found in the absence of social interaction. Participants have been observed to shift their speech in the direction of a voice (of a model) heard over headphones in word-identification experiments (e.g. shadowing) (Goldinger, 1998; Shockley et al., 2004; Namy et al., 2002; Sanchez et al., 2010; Miller et al., 2010).

2.1.1 Auditory Speech Alignment

Speech alignment has primarily been investigated with auditory stimuli using word-identification (shadowing) methodology (Goldinger, 1998; Goldinger & Azuma, 2004; Shockley et al., 2004; Pardo, 2006; Namy et al., 2002; Sanchez et al., 2010; Miller et al., 2010). The current investigation also uses the word-identification, or shadowing, methodology. There are typically three phases to these experiments. First, participants, known as *shadowers*, engage in a baseline task, in which they are recorded saying words out-loud that they read off of a computer monitor. These words are considered a fair representation of how the shadower normally speaks and are called the shadowers' baseline utterances. In the second part of the shadowers' task, they are instructed to listen to spoken words said by a model and are asked to say each word out-loud quickly, but clearly (they are not asked to repeat or imitate the model). The words from this second task are referred to as the shadowers' shadowed utterances. The recordings of the shadowers' baseline and shadowed utterances are then presented to perceptual raters in the final task. Perceptual raters are asked to judge the relative similarity of the shadowers' baseline and shadowed words to the words spoken by the model. Experiments using this methodology have found that raters judge the shadowers' shadowed utterances as being more similar to those of the model than are the baseline utterances (Goldinger, 1998; Shockley et al., 2004; Namy et al., 2002; Miller et al., 2010).

2.1.1.1 Assessing Alignment: The Perceptual Rating Task

The perceptual rating procedure will be discussed in detail because it is the primary method used to assess alignment in this Dissertation. This section will be referred to when discussing the procedure for the experiments within Experimental Series 1 and 2.

In a typical perceptual rating task, two items (A and B) are compared to a third item (X). Within the speech alignment literature, the perceptual rating task is often referred to as an AXB task (Goldinger, 1998; Goldinger & Azuma, 2004; Namy et al., 2002; Pardo, 2006; Shockley et al., 2004; Miller et al., 2010). The AXB task is used to establish whether the shadowers' shadowed utterances (A) are perceptually more similar to the models' utterances (X) than are the shadowers' baseline utterances (B). The presented words on a trial consist of the same word (e.g. *turkey* – ***turkey*** – turkey). Participants serving as perceptual raters are asked to listen to the three utterances and then indicate whether the word in the A position or B position was pronounced more like the word in the X position. Alignment is determined to occur when the rater judges the shadower's shadowed utterance as more similar to the model's utterance than is the shadower's baseline utterance. The utterances in the A and B positions are counterbalanced.

In the speech alignment literature, the perceptual method of determining relative utterance similarity is preferred over acoustical analyses for several reasons. First, the perceptual ratings method serves as an ecologically valid way to assess similarity. That is, speech alignment occurs through the perception of a given talker, and thus occurs in a

perceptually relevant way. The perceptual rating method ensures that this is the case. As Goldinger has stated, "many acoustic properties can be cataloged and compared, but they may not reflect perceptual similarity between tokens--imitation is in the ear of the beholder" (pg. 257). This method acknowledges that the average human perceiver is sensitive to speech differences and can also reliably assess speech similarities.

Second, the perceptual rating method avoids the difficulty in determining to which of the many possible acoustical dimensions participants are aligning (Goldinger, 1998). Along these lines, it has also been acknowledged that the psychological validity of acoustical measurements is not fully understood (Goldinger, 1998). One final reason for selecting perceptual methods over acoustical measurements is that this method has been used to evaluate alignment in a majority of the studies in speech alignment (e.g., Goldinger, 1998; Goldinger & Azuma, Namy et al., 2002; Pardo, 2006; Shockley et al., 2004; Miller et al., 2010).

The aim of this Dissertation is to identify whether the utterances of shadowers of the same model are *perceptually* more similar to each other than the utterances of those who shadowed a different model. Therefore, in the experiments of this Dissertation, perceptual raters will be asked to judge the similarity of speech utterances in the experiments. However, acoustical measures will be used in one of the studies.

2.1.2 Visual Speech and Alignment

The results from speech alignment experiments may indicate that the shared influences of talker-specific characteristics serve a communicative function and are

useful in establishing a shared understanding between interlocutors (Pickering & Garrod, 2004). However, most communication involves a social function in addition to its linguistic function. In these circumstances interlocutors often have the *benefit* of *visual* information in addition to auditory information. Still, auditory speech tends to be the modality that most experimenters use to demonstrate the influences of talker-specific characteristics. However, recent evidence has emerged showing that the talker-specific information that is available in visual (including lip-read) speech can also be influential (Rosenblum, Miller, & Sanchez, 2007; Miller et al., 2010; Sanchez et al., 2010). Thus, the role of visual speech alignment will be investigated in addition to auditory alignment.

There is an ever increasing bed of knowledge supporting the importance of visual information in speech perception. Visual speech information is used by all sighted individuals, regardless of their level of hearing (see Rosenblum, 2005 for a review). For example, the ability to make use of both auditory and visual speech information has been found to play a role in speech development. It has been found that blind children have a difficult time distinguishing similar sounding phonemes (e.g. /m/ vs. /n/) in contrast to their sighted peers (Mills, 1987). In addition, when in a noisy environment, people tend to look more frequently at the mouth of the person speaking in order to better understand the words being said (e.g. Grant and Seitz, 2000; Kim and Davis, 2004). Also, it has been found that when conversing with someone who has a foreign accent, looking at the mouth aids intelligibility (Arnold & Hill, 2001). Similarly, research shows that when listening to a complicated message it is helpful to gaze at the mouth (Reisberg, McLean, & Goldfield, 1987). Finally, there is substantial evidence that the auditory and visual

speech signals automatically integrate for infants and adults, from all native language backgrounds (McGurk & MacDonald, 1976; and see Rosenblum, 2005, for a review).

Visual speech can also convey talker-specific information (Rosenblum, Niehus, & Smith, 2007; Rosenblum, Yakel, Baseer, Panchal, et al., 2002). For example, there is evidence that visual speech information can be used to identify a face even when the information for the face has been reduced to isolated speech movements. Isolating speech movements, which in effect reduces speech to kinematics (articulatory movements), can be achieved by using a point-light technique. In the point-light technique, reflective dots are placed in various places of articulation on a model. The model is then filmed, where only the moving dots on the model's face are seen against a dark background. Yet, even with this impoverished information, talkers can be recognized (Rosenblum et al., 2007; Rosenblum et al., 2002).

Recent speech alignment investigations have found evidence for the *influence* of talker-specific information present in visual only (lip-read) stimuli of a talker's articulating face. For example, Miller et al. (2010) found evidence for alignment to visual (lip-read) speech in a shadowing task. Participants engaged in a baseline task where they were recorded reading words from a monitor out-loud. They then engaged in a visual (via lip-reading) speech shadowing task. Accurate visual speech shadowing was achieved by implementing a two-alternative forced choice task. Participants first read two words on a computer screen ('tennis table') and then were immediately presented with the face of a model silently articulating one of those words ('tennis'). Participants were instructed to say out-loud the word they lip-read as quickly and as clearly as

possible. Participants' responses were audio recorded. Perceptual raters judged the shadowed lip-read words as sounding more similar to the model's utterances than were the baseline words. This suggests that visual speech information can induce alignment.

In addition, Sanchez et al. (2010) observed that the visual speech information of a talker influences speech productions. It was found that the produced voice-onset times (VOTs) of shadowers' utterances were different when a participant shadowed an auditory-only stimulus as compared to the utterances produced when presented with varying rates of a visual stimulus of an articulating face in conjunction with the VOT adjusted auditory stimuli. This too supports the notion of the influential effects of visual speech on alignment.

Thus, speech alignment has been observed for shadowers of auditory-only and visual-only speech. In both cases shadowers of a model shift their speech in the direction of the model perceived. However, it is not known whether shadowers of a model shift their speech in similar ways compared to shadowers of a different model. Also, it is not known whether the speech of shadowers of the same model, perceived either auditory-only or visual-only, is also more similar than the speech of shadowers of a different model. This Dissertation aims to address these issues. To understand why common alignment might occur, it is worth discussing social influences and the forming of dialects.

2.2 Social Influences on Speech Alignment

Speech alignment, as mentioned, is the unconscious and spontaneous tendency to be subtly influenced in how one speaks, based on the speech perceived. However, this is not to say that speech alignment is an *automatic* process. In fact, social factors have been found to alter one's likelihood or rate of alignment. These factors include the gender of the perceiver and model and the social role of the participants in an experiment (e.g., Namy et al., 2002; Pardo, 2006). In addition, social factors that affect speech productions are thought to not only affect interlocutors, but may also play a larger role in social relations and the formation of dialects. Although this Dissertation is primarily about alignment, the results of this Dissertation could have implications for the formation of dialects. Social influences on alignment will be discussed next followed by a discussion of dialect formation.

2.2.1 Gender

Gender is known to impact how a perceiver aligns and possibly the spread of dialects. For example, Namy et al. (2002) observed that female and male shadowers aligned differently. Overall, females were found to align more to their model than males overall. Also, the alignment ratings of female and male raters were found to be different. Female raters were more likely to detect alignment than male raters. These results were attributed to females, in general, as being more perceptually sensitive and thereby more likely not only to perceive alignment, but also more likely to align.

However, the effect of gender is unclear given conflicting evidence. For example, Pardo (2006), in contrast to the Namy et al (2002) results, found that males were more likely to align than females. This difference between the studies may be attributed to the methodology employed. In the Namy et al. experiment, participants engaged in a shadowing task, while the participants in the Pardo study engaged in an interactive task with another participant.

When comparing the studies, participants in the Namy et al. (2002) experiment served in a passive role, where they simply heard the speech of a talker and subsequently said the word that they heard. Whereas in the Pardo (2006) study, participants served in an active role, where they conversed with a fellow participant to achieve a common goal (e.g. navigating a map to arrive at a specific location). In the Pardo study, each participant pair were assigned an experiment role, where they either instructed the other participant how to arrive at the set location (e.g. the giver of information – dominant role), or asked questions about arriving to the location (e.g. receiver of information – less dominant role). Thus the differences in the studies may be due to several factors, including whether the participants engaged in an active or passive role, whether they heard recorded speech or speech from a live partner, and the role the participant played in the experiment.

In addition to finding that males were more likely to align than women, Pardo also found that gender interacted with the role the participants played in the experiment (giver or receiver; to be discussed next). Yet, despite the conflicting evidence for the

effect of gender on alignment, it seems that gender can play a role in the alignment story on the small scale, and possibly a role in dialect formation on a larger scale.

2.2.2 Role

Social role is a factor known to influence speech alignment and possibly dialect formation. As mentioned, Pardo (2006) found that the role played by participants was impacted by the person's gender, though overall, information givers aligned more than information receivers. In other words, those with a more dominant role were more likely to shift their speech toward their partner, who was less dominant. Regardless, it appears that one's conversational role seems to affect alignment.

Social class is a societal role that is known to affect general language alignment. It has been found that one can align in an *upward* or *downward* way, with respect to one's own societal class and the class of a fellow interlocutor (see Giles & Ogay, 2006, for a review). Upward alignment is the act of shifting one's language attributes to a perceived higher societal rank than one's own, as illustrated when communicating with a potential employer or person in authority. Downward alignment, on the other hand, is the act of shifting one's language characteristics to a perceived lower societal rank than one's own, as illustrated by an educator attempting to clarify class material by speaking the *lingo* of students of a younger generation. Researchers have found that when the social class of interlocutors differ, people will shift their speech on the pronunciation of words (Coupland, 1984), pitch (Gregory, & Webster, 1996), and word choice and meaning (Azuma, 1997).

2.2.3 Dialectical Groups

In addition to establishing a shared understanding between interlocutors or groups, speech alignment may also help serve an even more basic function, the need to belong; to be part of a group. On the small scale, the act of speech aligning may bring two individuals together in establishing rapport and a sense of closeness (Giles & Ogay, 2006). When this process is applied to a group of individuals, over a period of time, this may serve as a building block to group formation and cohesion. For instance, it has been found that the way high school students at an all-girls school pronounced certain words could be linked with whether they belonged to a particular group who shared their lunchtime meals together (Drager, 2006). In addition, the clothing worn by different schoolgirls has been found to covary with how these girls pronounced their speech (Eckert, 1996). These girls may be using several factors, including alignment to speech production mannerisms, to establish group identity and bonds to that group.

Recently, Fagyal, Swarup, Escobar, Gasser, and Lakkaraju (2010) computationally modeled language change and maintenance in society. Within their model, they observed that (virtual) highly-connected charismatic individuals within a society can influence changes in language (in use and pronunciation). In their model, highly-connected charismatic individuals were those who had many connections within the network. The network was designed so that (virtual) people were weighted in favor of, but not determined to, imitating the behaviors of a (virtual) person who had many connections within his or her network. Fagyal et al. suggest that highly-connected charismatic people may be able to influence language directly and indirectly through their

network of connections. So for instance, if Larry is a highly-connected and charismatic person and he influences Chris and Steve's speech directly, the people who interact with Chris and Steve that are outside of Larry's network are indirectly influenced by Larry's speech. The influence of Larry's speech goes further and further along the chain of relationships. Thus the more ties Larry and his friends have, the more likely his speech is to lead to a shift in how society speaks. Less connected individuals (e.g. loners: those who have few connections in the network) on the other hand, serve to maintain the status quo, or the older, more traditional style of speaking given that they have little contact with others whose speech may be changing (Fagyal et al., 2010). Although the predictions from the model seem plausible, there is currently no behavioral data to support the model.

Yet, it is possible that multiple individuals might align to common speech properties of a particular individual. This process may play a pivotal role in the forming of dialectical communities as modeled by Fagyal et al. (2010). In their model, a highly-connected and charismatic figure was projected to be able to influence the speech of others. This is consistent with the speech alignment research (Goldinger, 1998; Shockley et al., 2004; Namy et al., 2002; Miller et al., 2010) which has found that people (e.g. shadowers) are influenced in the direction of a particular talker (e.g. a model). However, it is unknown if, when multiple talkers align to a model, alignment to the same aspects of the model's speech occurs. If so, then the speech alignment that has been observed with shadowing methods, could be the same mechanism that helps underlie the dialectic formation phenomena addressed by Fagyal et al. This Dissertation will serve as an initial

step in first identifying whether people are actually influenced in the same way by the speech of a particular talker.

Within the topic of changes in speech, Chambers (1992) outlines several aspects that may be involved in the adoption of a dialect. He notes that there are differences in the adoption of a dialect with respect to age. Generally, dialect adoption, with respect to words used and pronunciation are strongest for children, less so for young adults, and even less for older adults, a pattern that effectively separates early to late learners. According to Chambers, one of the first things to occur is changes in word use. For example, an American English talker who has relocated to England is likely to change the words used for things, like substituting the word "chips" for "fries". In addition, changes in word pronunciation follows changes in word use. The American English talker may start saying "to-mah-to" instead of "tomato". What is more, within the language change literature it has been found that words that occur rarely, such as the words "fries" and "tomato", are more susceptible to changes in word use and pronunciation than words that are commonly used (Diessel, 2007; Bod, Hay, & Jannedy, 2003; Bybee & Hopper, 2001). Thus, when adopting a dialect, aligning to the words used in the community and the way they are pronounced may help the newcomers assimilate, leading to better community relations. Yet, alignment is only one side of the process. There is a counter-process called divergence that will be discussed next.

2.3 Divergence

While the act of alignment results in reducing differences between people or groups, *divergence* is said to occur when people accentuate *differences* in actions or speech from each other. For example, persons or groups may diverge their actions and speech to accentuate a valued difference (Giles & Ogay, 2006). Such behavior can be observed not only physically (e.g. hair length and style of dress) but also in the words spoken and how they are articulated.

For instance, it has been found that when one feels offended or threatened by someone belonging to an out-group one will often speak (syntax, word choice, and pronunciation) in a way that highlights differences between them (Bourhis & Giles, 1977). This type of speech divergence is clearly illustrated in Bourhis and Giles's (1977) experiment where Welsh participants who interacted with a person with a Standard English accent (also called Received Pronunciation or the Queen's English) were found to diverge or align their speech depending on whether the Englishman behaved negatively or positively. Thus, speech can be used to divide or unite people based on whether one is perceived as a foe or friend.

The factors that influence alignment and divergence are varied and complex. People might alter their speech articulations by virtue of hearing or seeing another person. These shifts in behavior may be affected by social factors such as gender and societal role. Yet, when alignment does occur, in the case of speech articulations, it is not known whether people are influenced in a similar manner, and is investigated in this Dissertation. The questions posed in this Dissertation may be relevant to the formation

and spread of dialects. Returning to the example of Larry, the well connected charismatic figure, this Dissertation will address the first step of dialect change, the change in Chris and Steve's speech due to Larry's direct influence. In this case, Chris and Steve's speech will be compared to identify whether they are influenced in the same ways by Larry's speech. This Dissertation will go further and address whether Larry's voice and face similarly influence Chris's speech (who only heard Larry's speech) and Steve's speech (who only lip-read Larry's speaking face).

Chapter 3

Speech Literature

3.1 Talker-Specific Characteristics

The way we speak is unique, like a fingerprint, and is the result of our own talker-specific characteristics. Talker-specific characteristics are the unique qualities of one's speech that are formed through a combination of factors such as dialect, and biological factors (e.g., gender and vocal tract size) (Abercrombie, 1967). Although the relationship between talker-specific characteristics and speech alignment will be the focus of this Dissertation, talker-specific characteristics also play a role in other speech phenomena. Talker specific-characteristics are influential in both speech perception and memory, in addition to speech production (alignment).

Talker-specific characteristics have been shown to influence speech perception via *talker familiarity* investigations (e.g., Goldinger, Kleider, & Shelley, 1999; Nygaard & Pisoni, 1998; Nygaard, Sommers, & Pisoni, 1994; Rosenblum, Miller & Sanchez, 2007). In these experiments, participants are first trained, or *familiarized*, with the speech of a talker or talkers. Participants subsequently engage in an auditory word

identification task in noise. It has been found that the auditory speech of the familiarized talker(s) is better identified than that of a novel talker even when the auditory signal has been degraded with white noise. This suggests that familiarity with a person's talker-specific characteristics facilitates perception of that person's speech.

Evidence for the presence of talker-specific information in memory has also been found in word recognition experiments (Craik & Kirsner, 1974; Palmeri, Goldinger, & Pisoni, 1993; Goldinger, 1996). In these experiments, participants listen to words during a study phase and are then given a recognition memory test on these words. Half of the words in the test phase are composed of old words (words from the study list). These old words are either stated in the same voice as they were originally presented in the study phase, or stated by a different voice that was also heard in the study phase. The other half of the list contains new words (words that were not previously presented). The results of these experiments find that participants are better at recognizing old words when stated by the original talker then when stated by a different talker.

To provide some insight as to the processes involved in alignment, speech theories that have addressed why listeners shift their speech productions will be discussed in the following section.

3.2 Alignment and Speech Theories

The utility of talker-specific characteristics have not always been accepted within the speech literature. For example, one traditional theory of speech, the abstractionist view (Joos, 1948; Gerstman, 1968; Summerfield & Haggard, 1973) considers the unique

qualities of one's voice to be extraneous features to lexical information. From the perspective of this theory, speech undergoes a normalization process when perceived, in which all of these superficial features (e.g., talker-specific characteristics) are removed.

However, the findings from experiments on alignment in speech production, as well as investigations on speech and memory discussed above, have turned many researchers away from the abstractionist approach to theories that include talker-specific characteristics as relevant speech information. The speech theories that account for talker-specific characteristics in speech alignment are the episodic encoding theory and the gestural theories (Goldinger, 1998; Liberman & Mattingly, 1985, 1989; Fowler, 1986). Although these theories traditionally have had different ideas about the nature of talker-specific characteristics, both theories can adequately account for the influence of talker-specific characteristics in speech alignment.

3.2.1 Episodic Encoding Theory

The theory of episodic encoding proposes that talker-specific characteristics are tied to lexical items in speech and that both lexical items and talker-specific characteristics are stored together in memory (Goldinger, 1998). Episodic encoding is supported by evidence from speech alignment (Goldinger, 1998; Goldinger & Azuma, 2004), perception (Goldinger et al., 1999), and memory (Palmeri, Goldinger, & Pisoni, 1993; Lachs, McMichael, & Pisoni, 2000; Goldinger & Azuma, 2004).

The episodic encoding theory proposes that speech alignment works in the following way: When one perceives a word, that word is stored in the perceiver's

memory along with information about the talker who said it and how it was said. This information unit is referred to as an *episode* (Goldinger, 1998). The words and voices in memory are then recalled when one speaks a given word, and are consequentially influential on the speech produced. However, it should be noted that within this theory, all words are not created equal. The more common the word, the less likely it is that a given talker's speech characteristics will be strongly influential on how a perceiver later produces that word. This is because many memory episodes are activated (i.e. other instances of that same word spoken by different voices) for common words, leading the influence of a particular talker's speech characteristics to be inconsequential. On the other hand, uncommon words are more influenced by a given talker's speech characteristics since there are fewer memory episodes that are activated, such that a given talker's speech information is given more weight. This then results in a higher likelihood that a perceiver will align to these words stored in memory episodes (Goldinger, 1998; Goldinger & Azuma, 2004).

From the episodic perspective, speech perception, memory, and production are related (Goldinger, 1998). Within this theory, speech perception begets the memory of speech episodes. These episodes facilitate the perception of similar items when a speech event activates them. Goldinger likens this process to Gibson's (1966) resonance metaphor. Resonance occurs between the speech event and the stored memory episodes much like a struck tuning fork vibrates another tuning fork within its proximity. Similarly, heard words from a talker activate, or cause a resonance with, the stored episodes in the listener, resulting in produced speech that is similar to what was heard.

To illustrate, Goldinger and Azuma (2004) combined recognition memory and alignment tasks in their investigation of talker-specific characteristics. They found that stored episodes of talker-specific information not only affects how a listener later produces items verbally (the degree of alignment), but also found that exposure to another's speech can cause long-term changes (at least over the course of two weeks) in how one speaks and remembers. In their experiment, words that were listened to prior (1 week) to reading the same text words out-loud, and prior (2 weeks) to a text based recognition memory task, activated stored memory episodes that contained information about the talker who stated that particular word in a listening task. This resulted in produced speech for rare words that sounded more like a particular talker from the listening task and resulted in higher recognition accuracy for those rare words. Rare words were found to be aligned to with greater fidelity and remembered better because there were fewer activated traces, allowing information from the heard talker from the listening task to carry more weight in memory, and thus lead to the observed effects. Common words, on the other hand, were not found to exhibit alignment and were not remembered as well because many episodes were activated, leaving the influence of the talker from the listening task to be non-existent. Thus, the perception of a talker's voice can influence how one speaks and remembers.

As it stands, the theory of episodic encoding is based on lexical episodes that contain only *auditory* talker-specific information. Yet it is, in principle, possible for episodes to be composed of *visual* talker-specific information (or, more generally, gestures, which will be discussed later). In fact, there is an ever increasing bed of

knowledge supporting the importance of visual speaker information, including in the context of speech alignment (Miller, Sanchez, & Rosenblum, 2010; Sanchez, Miller, & Rosenblum, 2010). While a visual speech examination of episodic encoding has not been conducted the Miller et al. (2010) examination of visual speech alignment can be interpreted as an initial test of episodic encoding. Sanchez et al. (2010) also entertain the idea that their observed alignment results may be due not only to visual speech episodes, but sub-lexical or even gestural speech episodes.

3.2.2 Gestural Theories

The gestural theories of speech (Liberman & Mattingly, 1985, 1989; Fowler, 1986) suggest that the objects of speech perception take a gestural (articulatory) rather than acoustic form. These gestures are thought to be available multi-modally, meaning that speech information is conveyed both auditorily and visually. From a gestural theory stance, the speech primitives (e.g. the fundamental units of speech) shared by perceptual events and actions in both the audio and visual modality are believed to be the same and take an *amodal* form, meaning that they are not tied to a given modality (i.e. modality-neutral).

Thus, visible speech is considered to hold the same properties (gestures), and is treated in the same way, as auditory speech. Data supporting this aspect of the gestural approach have come from numerous studies showing that speech information from the visual modality can have strong effects on perceptual and neurophysiological responses

to auditory speech (McGurk & MacDonald, 1976; and see Rosenblum, 2005, for a review).

With regard to speech alignment phenomena, the gestural approach provides an alternative explanation to that of the episodic encoding account. Unlike the episodic encoding theory which places memory as an integral part of speech, the gestural theories do not factor memory heavily into speech perception (Fowler, 2004; Fowler, Brown, Sabadini, & Weihing, 2003; Sancier & Fowler, 1997; Shockley, Sabadini, & Fowler, 2004).

The gestural theories propose that speech (and non-speech) alignment works through a perception-production link. Within this theory, the perception of behaviors leads to the production of similar behaviors; what you see influences what you do. This idea is consistent with the inadvertent imitation observed in non-speech alignment, where people subtly shift their body movements and facial expressions to match their conversational partner's (Chartrand & Bargh, 1999). This idea is also consistent with neurophysiological evidence for mirror neurons: specialized cells which have been found to be activated when one is engaged in a motor behavior and when that same behavior is perceived being performed by another person (e.g., Fadiga, Fogassi, Povesi, & Rizzolatti, 1995; Rizzolatti, Fadiga, Gallese & Fogassi, 1996).

With regard to speech alignment, the critical units for both perception and production are gestures (articulations), thus perception is thought to naturally facilitate—or prime—production. A perceiver's speech productions are therefore likely to be influenced by the talker-specific characteristics of the speech that was just perceived.

Given that the system works with gestural rather than acoustic dimensions, talker-specific gestural properties can be conveyed, and influential via visual, as well as auditory speech. This perspective is consistent with the recent findings showing visual influences on speech alignment mentioned above (Miller et al., 2010; Sanchez et al., 2010).

3.2.3 Episodic Gestures

Speech alignment can be adequately explained by both episodic encoding and gestural theories. In fact, the two explanations may be compatible; both theories assume that perceived talker information influences articulatory responses, although this talker information takes a gestural form from the perspective of the gestural approach, and typically, auditory form for the episodic approach. However, there is nothing endemic to the episodic approach that it cannot be modified to work with gestural primitives. In fact, Goldinger (1998) acknowledges that episodic encoding can work with gestures. He notes that one especially intriguing aspect of gestural theories is the idea of a resonance between the perceiver and the environment. The idea of resonance fits well with episodic encoding, where it is thought to occur between memories of the perceiver and the information signals in the environment. Along these lines, Sheffert and Fowler (1995) as well as Miller et al. (2010), have suggested that the episodes may be comprised of talker-specific gestural information of lexical items.

The results of this Dissertation will bear on episodic encoding, the gestural theories, and the notion of gestural episodes. These theories bear on all experiments within this Dissertation, as they explain alignment in general, the nature of some of the

relevant talker-specific characteristics influential in alignment, and how a model may be similarly influence multiple shadowers. The implications for the theories, with respect to the results obtained in the experiments, will be discussed (see section 7.1).

Chapter 4

The Current Study

Past investigations on alignment have aimed to address whether people would shift their speech in the direction of a model (Goldinger, 1998; Goldinger & Azuma, 2004; Shockley, Sabadini, & Fowler, 2004; Namy, Nygaard, & Sauerteig, 2002; Pardo, 2006; Nielsen, 2008; Miller, Sanchez, & Rosenblum, 2010; Sanchez, Miller, & Rosenblum, 2010). However, it is unknown whether multiple people align to a given model in similar ways or whether people uniquely alter their speech based on different characteristics of that model. It is also unknown whether people align in similar ways to a model when only hearing or only seeing the model's speech.

The core questions of the Dissertation are: *Do shadowers of the same model sound more similar to each other than they do to shadowers of a different model? Does the sensory modality of the shadowed speech affect the perceptual similarity between the shadowers of the same model?*

4.1 Relevance of the Dissertation Questions

Although the recent computational model by Fagyal et al. (2010) seems to suggest that people might be similarly influenced by a talker, this has not been directly tested in the laboratory.

In fact, it might be assumed that shadowers align to different dimensions of a model's speech given that speech produced by a single individual is so varied. For example, the words and phonemes produced by a single talker in two different instances are found to have different acoustical and articulatory patterns (e.g., Perkell, Zandipour, Matthies, & Lane, 2002). In addition, it has been found that a talker's vowel space, which consists of information from both the first and second formant, is quite varied for a specific vowel, from utterance to utterance (Tsao, Weismer, & Iqbal, 2006). On the perceptual side, it is well known that different listeners will attend to different acoustic cues of the same utterance when recognizing the same phoneme (Crystal & House, 1988; Newman, Clouse, & Burnham, 2000; Perkell, Matthies, Tiede, Lane, et al., 2004).

Thus, if perceivers weigh cues differently in perception, in addition to the varied nature of the utterances spoken by even a single person, it is reasonable to conclude that shadowed speech productions would also vary between shadowers of the same model. Notwithstanding, dialects represent a case where different people, despite the varied nature of perception and production, speak in a characteristically similar way. As stated, Fagyal et al.'s (2010) computational model of dialect change suggests that a highly-connected charismatic person (in the laboratory case, the shadowed model) can shift the speaking style of those who directly and indirectly perceive the model's speech. Thus,

although it would seem that the way people produce and perceive speech is quite varied, even for the same person's utterances and perceptions, this Dissertation challenges this notion and attempts to identify if perceivers of the same model's speech alter their produced speech in similar ways. This Dissertation also goes further, by identifying the nature of the influential speech information in alignment by investigating whether the *visual-only* speech of a model similarly shifts the speech of shadowers of the same model. The effect of perceiving auditory-only and visual-only speech of a model is also compared to identify whether the produced speech of shadowers of a model are altered in a similar way.

Chapter 5

Experiment Series 1

The speech produced (e.g., Perkell, Zandipour, Matthies, & Lane, 2002; Tsao, Weismer, & Iqbal, 2006) and perceived (Crystal & House, 1988; Newman et al., 2000; Perkell et al., 2004) by an individual is quite varied, yet the presence of dialects suggest that different people can speak in similar ways. The aim of Experiment Series 1 was to establish whether shadowers of auditory speech are more similar to shadowers of the same model than shadowers of a different model. The experiments within this series also serve to test some alternative explanations for obtaining the anticipated results. Acoustical analyses are also conducted to identify some similar dimensions upon which multiple shadowers of the same model might converge.

5.1 Experiment 1a

It is currently not known whether multiple shadowers shift their speech in the direction of a model in similar ways, or whether this shift is unique for each shadower for a given model. Experiment 1a tests this.

Perceptual raters were asked to judge the relative similarity of shadowers, two of whom shadowed the same model and one who shadowed a different model. It is hypothesized that if shadowers of the same model shift their speech in similar ways (or on similar dimensions), then perceptual raters should be sensitive to this similarity and judge them as being more similar than those who shadowed a different model.

If these results are obtained it would suggest that raters matched tokens based on some common auditory information across utterances produced by shadowers of the same model. This could mean that shadowers are aligning to some of the dimensions of the model's talker-specific characteristics.

5.1.1 Method

5.1.1.1 Participants

Three sets of participants engaged in this experiment, models, shadowers, and raters. All participants were native speakers of American English and reportedly had good hearing and good or corrected vision. All participants were recruited from the University of California, Riverside.

Models. The models in this experiment were two female graduate students who were compensated for their time at a rate of $20 per hour. Both models had similar linguistic backgrounds; both were Native Californians who were not fluent in a second language.

Shadowers. The shadowers in this experiment were eight females recruited from undergraduate psychology classes for course credit. Four shadowers were exposed to the

utterances of each model. Women shadowers were selected in order to match the gender of the models. The rationale for running female-only shadowers stems from the subtle nature of speech alignment. The current study, is considered a bit more difficult for perceptual raters than past studies. In the current study, three different voices are used in each AXB trial, whereas past studies have not used more than two different voices. Also, it was thought that if any similarities between shadowers of the same model could be observed, it would be through the speech of women, given evidence that suggests that women tend to align more than men in the context of a shadowing task (Namy et al., 2002).

Raters. The raters in this experiment were 16 participants (10 males, 6 females) recruited from undergraduate psychology classes for course credit.

5.1.1.2 Materials

The 74 words used in this experiment were derived from Shockley et al.'s (2004) alignment investigation and were also used in the Miller et al. (2010) study. The words consisted bi-syllabic words with frequencies of less than 75 occurrences per million (Kucera & Francis, 1967).

This word list was selected for two reasons. First, in anticipation of Experiment Series 2, it has been used successfully for visual-only (lip-read) shadowed speech (see Miller et al., 2010). Second, it consists of relatively low frequency words. It has been suggested (via the episodic encoding theory) and found that low frequency words are more likely to induce alignment (Goldinger, 1998; Goldinger & Azuma, 2004).

5.1.1.3 Stimuli

The stimuli in this experiment consisted of shadowed recordings obtained from participants in the shadowers group (see section 5.1.2. Procedure, for more information).

5.1.2 Procedure

The procedure of this experiment was different for each participant group (models, shadowers, and raters) and will be discussed separately.

Model Task. The two models were individually audio-video recoded producing the 74 words (see section 5.1.1.2 Materials, for more information) out-loud from the list in a sound-attenuated booth. The words were presented to the models as text on a teleprompter using the program PsyScope (Cohen, MacWhinney, Flatt, & Provost, 1993). Each word was presented at random with an inter stimulus interval (ISI) of 2500 milliseconds (ms). The models were instructed to say each word out-loud, quickly but clearly. The video recording captured the models' head and shoulders. The recordings were digitized and edited on a computer using Final Cut Software. Seventy-four audio-only (used in Experiments 1a, 1b, and the Acoustical Analyses) and 74 video-only tokens (used in Experiment 2a) were created.

Shadower Task. The eight shadowers individually engaged in first a baseline and then a shadowing task in a sound attenuated booth.

In the baseline task, the 74 words from the list were presented as text on a 21-in. Panasonic video-monitor positioned three feet from the shadower. Shadowers were instructed to say each word out-loud, quickly but clearly into a microphone (Shure Beta

58A). These utterances were digitally recorded and edited into individual words using the software Amadeus II (HairerSoft, 2008) that were amplitude adjusted. These utterances shall be referred to as the shadowers' baseline utterances (used in Experiments 1b and the Acoustical Analyses).

In the shadowing task, shadowers heard the words produced by a model over headphones (SonyMDR-V600). Shadowers were instructed to say each word they heard out-loud, quickly but clearly into a microphone. Shadowers heard the model's 74 words a total of two times from two different blocks. Their utterances from the second block were digitally recorded and edited into individual words that were amplitude adjusted. The second repetition of the word was used because there is evidence that alignment is increased with repetition to a stimulus (Goldinger, 1998; Goldinger & Azuma, 2004). These utterances shall be referred to as the shadowers' *shadowed utterances* (used in Experiments 1a, 1b, Acoustical Analyses, and 2b).

Rater Task. The 16 raters were asked to judge the relative similarity between the utterances of shadowers who shadowed the same model verses those shadowers who shadowed a different model. Perceptual ratings were achieved through use of an AXB (see section 2.1.1.1) task (Goldinger, 1998; Namy et al., 2002; Pardo, 2006; Shockley et al., 2004; Miller et al., 2010).

The 16 perceptual raters listened to 5 unique shadowers per experiment. The experimental session was divided into two stimulus blocks. The shadower in the X position did not change between the blocks. The shadower pairs in the A and B positions were different between the blocks. The shadowers in the A and B positions consisted of

a shadower who shadowed the same model and a shadower who shadowed a different model as the person in the X position. Each block consisted of 148 trials (74 words X 2 A-B positions).

Each trial was presented at random to the raters over headphones. Raters were asked to identify whether the first (A) or third (B) word sounded more similar in pronunciation to the second (X). The raters were instructed to press the key labeled "1" on the keyboard if the first word sounded more similar to the second or to press the key labeled "3" on the keyboard if the third word sounded more similar to the second.

5.1.3 Results and Discussion

The aim of this experiment was to investigate whether shadowers of the same model shifted (or aligned) their speech in similar ways with respect to a model whose utterances were perceived. If shadowers are judged as sounding more similar to those who shadowed the same model, then shadowers of a different model, this would suggest shadowers who perceive the same model speech may be aligning to the model in some similar ways.

The mean proportion for perceived alignment was calculated for each rater. The data reveals that raters judged shadowers of the same model as more similar (M = .594) at a higher proportion of the time than a shadower of a different model. These ratings were found to be statistically different from chance (.50) using a one-sample t-test, t (15) = 2.437, p = .028, Cohen's d effect size = 1.258. This suggests that shadowers of the same model can be similarly influenced by the speech they perceive, and that this influence is

evident in their speech productions. In addition, this also suggests that there are perceptible commonalities with which different shadowers alter their speech when they align to the same model. Thus, to answer the first question of this Dissertation, shadowers of the same model *do* seem to sound more similar to each other than they do to shadowers of a different model, at least when shadowing auditory speech.

To ensure that the results were not driven by the speech of a particular model, an independent samples t-test was conducted on the factor of model shadowed (Model 1 vs. Model 2). The results of this analysis did not find a significant difference between the shadowers of the models, t (14) = -1.425, p = .176, Cohen's d effect size = .762.

Finally, an additional test was conducted on the role of the position, A verses B, of the judged utterances words in the AXB task. The results of this analysis did not find a significant difference between the position of the utterances and its likelihood of being selected as more similar to the item in the X position, t (15) = -.096, p = .925, Cohen's d effect size = .050.

5.2 Experiment 1b

Experiment 1b serves to ensure that the results from Experiment 1a were due to actual shifts in speech production towards the model perceived, and not due to some chance assignment of shadowers to the specific models.

While unlikely, it could be that, by chance, the results of Experiment 1a were based on subjects being 'randomly' assigned to shadow the model which naturally sounded like them, even before shadowing. Thus, it could be that shadowers didn't sound

more like the model upon shadowing and that their natural, pre-shadowed speech already sounded like their model, as well as the other shadowers of this model. This could account for the results of Experiment 1a. To preclude this possibility and ensure that shadowers were truly aligning to their models, raters in Experiment 1b were asked to judge the relative similarity of each shadower's baseline and shadowed utterances relative to their model's speech. This test is, in fact, the 'traditional AXB alignment' test implemented by Goldinger and others (e.g., Goldinger, 1998; Goldinger & Azuma, Namy et al., 2002; Pardo, 2006; Shockley et al., 2004; Miller et al., 2010). If shadowers shifted their speech in the direction of the model, then raters should judge the shadowed utterances as being more similar to the model than the baseline utterances.

To further examine the issue of the shadowers' pre-shadowed speech, raters also judged the relative similarity of the baseline utterances of shadowers. On these trials, two of the baseline utterances were from shadowers of the same model, while the other baseline utterance was from a shadower of a different model. If shadowers of the same model inherently sounded like each other even prior to perceiving the model, then the baseline utterances of those who shadowed the same model should be rated as more similar to each other than the baseline utterance of a shadower who shadowed a different model. However, if shadowers of a model did not inherently sound like each other, then this would mean the results from experiment 1a could be attributed to a common shift in the shadowers' speech due to the influence of their given model.

5.2.1 Method

5.2.1.1 Participants

Raters. The participants in this experiment consisted of 32 (12 males, 20 females) perceptual raters recruited from undergraduate psychology classes for course credit. All participants were native speakers of American English and reportedly had good hearing and good or corrected vision. All participants were recruited from the University of California, Riverside.

5.2.1.2 Materials

The materials for this experiment were the same as those used in Experiment 1a.

5.2.1.3 Stimuli

The stimuli in this experiment consisted of all recordings obtained from two sets of participants, Models and Shadowers, from Experiment 1a. (see section 5.1.2. Procedure, for more information).

5.2.2 Procedure

In this experiment, two different AXB tests were employed. Half of the raters in this experiment engaged in a traditional AXB test (see 2.1.1.1). This was used to establish whether the shadowers' shadowed utterances (A) were perceptually more similar to the models' tokens (X) than are the shadowers' baseline utterances (B). In other words, this task was used to identify whether shadowers actually aligned to the shadowed model.

For this condition, each rater heard the tokens of one model and her four respective shadowers. Each rater heard two entire lists. Although each list originally contained 74 words, only 64 were used due to presentation malfunctions for some of the shadowers' baseline trials. In order to make the lists comparable, only the 64 tokens that each shadower had were used. Each rater heard 32 unique words per shadower. The presentation lists were created in a fashion to allow two raters per shadower to hear one set of 32 words, while the other two raters per shadower heard the other 32 words. There were 256 trials in total (32 words X 2 A-B positions X 4 shadowers = 256).

The other half of the raters rated baseline utterances in an AXB task. This condition goes further to ensure that the shadowers' speech of a given model is not inherently similar. In this AXB task comparisons were made on the baseline utterances of the shadowers of the same model (A and X) and shadowers of a different model (B) to see if the shadowers we had randomly assigned happened to be perceptually more similar to each other. If the shadowers' speech of a given model was inherently similar to each other and the model they shadowed, then raters should judge the baseline utterances of the shadowers who shadowed the same model as more similar than the baseline utterances of shadowers who shadowed a different model.

The set-up for this condition was identical to that of Experiment 1a (see 5.1.2 Procedure for details), except it used the shadowers' baseline utterances instead of the shadowers' shadowed utterances. Here, each block consisted of 128 trials (64 words X 2 A-B positions)

5.2.3 Results

The aim of this experiment was to ensure that the results from Experiment 1a were due to actual shifts in speech production toward those of the model perceived. This experiment tested two things: One was whether shadowers of the same model, actually aligned to the model as a function of shadowing, relative to their natural, pre-shadowed speech. The second was whether shadowers of the same model, by chance, naturally sounded like *each other*, even before experiencing the model's speech.

The mean proportion for perceived alignment obtained from both rating groups were calculated and compared against each other using an independent samples t-test. The analysis revealed a difference between the ratings groups, t (30) = -3.834, p = .001, Cohen's d effect size = 1.400. Given that the rating groups were different, each group was then separately compared against chance.

5.2.3.1 Group 1: Traditional AXB

The data reveals that raters judged a shadower's shadowed utterance as more similar (M = .609) to the model's speech at a higher proportion of the time than the shadower's baseline utterance. These ratings were found to be statistically different from chance (.50) using a one-sample t-test, t (15) = 3.838, p = .002, Cohen's d effect size = 1.982.

To ensure that the results were not a product of the align-ability of a particular model, an independent t-test was conducted on the factor of model shadowed (Model 1

vs. Model 2). The results of this analysis did not find a significant difference between the models, t (14) = -.538, p = .599, Cohen's d effect size = 0.320.

Finally, an additional test was conducted on the role of the position, A verses B, of the judged utterances words in the AXB task. The results of this analysis did not find a significant difference between the position of the utterances and its likelihood of being selected as more similar to the item in the X position, t (15) = -1.739, p = .126, Cohen's d effect size = 0.898.

These results replicate past experimental findings in speech alignment (e.g., Goldinger, 1998; Namy et al., 2002; Pardo, 2006; Shockley et al., 2004; Miller et al., 2010). Shadowers were influenced by the model's speech they perceived. This perception influenced their subsequent speech productions. These results seem to indicate that the shadowers' shadowed speech shifted in the direction of the model. These findings suggest that the shadowers in Experiment 1a may have indeed aligned to similar aspects of the model's speech.

5.2.3.2 Group 2: Baseline Comparisons

These data revealed that raters did not judge baseline utterances of shadowers of the same model as more similar (M = .46) at a higher proportion of the time than those of shadowers of the other model. These ratings were not found to be statistically different from chance (.50) using a one-sample t-test, t (15) = -1.399, p = .182, Cohen's d effect size = 0.722. Thus, the data suggests that the baseline utterances of shadowers of a given model were not inherently similar. This also suggests that the results from Experiment 1a

were due to a shift in the shadowers' utterances to the model (see 5.2.3.1) and that similar speech properties were shifted for shadowers of a given model.

5.3 Acoustical Analyses

Acoustical analyses attempted to identify some of the acoustic and, by implication, some of the articulatory properties that converged for the speech of shadowers who shadowed the same model. Identifying the articulatory dimensions that are found to be similar for shadowers of the same model may provide some insight as to the relevant speech information influencing different shadowers to align. Though only a few dimensions were measured, these analyses may help initiate further investigations of acoustical analyses in the speech alignment field. Investigating the acoustical similarities between shadowers of the same model may link the currently disconnected spheres of what dimensions are *perceived* to be similar based on rating judgments and the dimensions that *actually* change.

These analyses also provide information about whether alignment to these acoustic dimensions might be based on articulatory properties that are also *visible* in lip-read speech. In other words, these analyses sought to identify some of the articulatory dimensions that *shaped* the acoustics and whether these dimensions can also be seen. Thus, these acoustical signals were measured to identify some of the relevant articulatory dimensions for which shadowers of the same model were similar and whether the relevant dimensions were able to be perceived *visually*. If the relevant articulatory dimensions are able to be perceived visually, then this would suggest that shadowers of

the *visual* speech of a model should also be perceived as more similar to shadowers of the same model than those who shadowed a different visually perceived model (see Experiment Series 2). This would be predicted from a gestural view, given that speech information is not tied to any one modality. Within the gestural theory, speech is both auditory and visual and the information from either modality is the same: they have a common currency. Still, this hypothesis would not be inconsistent with an episodic account, if the speech information retained in the episode can be composed of visual talker-specific gestural information. In this sense, the results of the experiment could help motivate Experiment Series 2.

.

5.3.1 Procedure

Each shadowed utterance provided by the shadowers from Experiment 1a were measured on six acoustical dimensions, five of which are produced by articulatory dimensions potentially observable in visual speech. These dimensions were duration, fundamental frequency (F0), the first formant of the first vowel (V1F1), the second formant of the first vowel (V1F2), the first formant of the second vowel (V2F1), and the second formant of the second vowel (V2F2). These dimensions were selected based on investigations that have suggested that they might play a role in alignment (Goldinger, 1998; Babel, 2010). All measurements were conducted using the software program Praat (Boersma & Weenink, 2008).

5.3.1.1 Duration

Duration is the length of an utterance. The length of an utterance, relates to the total time for its sound and visible articulation to unfold. In this sense, duration can both be heard and seen in a visible utterance.

The durations were measured for shadowers' shadowed utterances. Research assistants were instructed to highlight the length of the word, from acoustic onset to offset, to obtain a duration value, using the Praat software. This dimension was selected because it can be seen and because Goldinger (1998) had suggested that it may be influential to speech alignment.

5.3.1.2 Fundamental Frequency (f0)

Fundamental Frequency is the lowest, or first, harmonic frequency (Borden & Harris, 1984). Although this dimension cannot readily be perceived visually, it is of interest given that Goldinger (1998) suggested that it (along with duration) is a promising dimension that may influence speech alignment. Research assistants were instructed to record the f0 of each utterance after recording the value for duration. The token was to maintain its highlighted state from the duration recording to achieve the value for f0.

5.3.1.3 First Formant (F1)

Formants refer to the resonance of the human vocal tract. The first formant may be used to identify particular vowels and is affected by changes in the opening of the mouth and is thus considered to be not only auditory, but could be visible as well (Borden & Harris, 1984). The F1 of the first and second vowel of the bi-syllabic stimuli words were measured. Research assistants were instructed to locate and highlight the first

vowel by both listening to the token and by observing the spectrogram from the Praat program. They were then instructed to click on the mid-point (or center) or the vowel and record the first formant. This process was repeated for the second vowel of each (bi-syllable) word.

5.3.1.3 Second Formant (F2)

The second formant is also relevant in identifying differences between vowels. Although the second formant is affected by changes *within* the mouth (Borden & Harris, 1984), there is evidence that F2 is also visible (Remez, Fellows, Pisoni, Goh, & Rubin, 1998). Measurements for F2 were conducted identically to the measurements of F1.

5.3.2 Results and Discussion

Analyses were conducted to identify the articulatory similarities between shadowers of the same model. Differences scores were calculated to test for the similarity of shadowers who shadowed the same model verses those who shadowed a different model. For each shadower, two sets of values, "in-group" (depicting those who shadowed the same model) and "out-group" (depicting those who shadowed a different model) scores, were calculated in the following way: In-group scores were obtained by taking the difference between a given shadower of a model and each of the three other shadowers of that *same* model. These values were then averaged. Out-group scores were obtained by taking the difference between a given shadower of a model and each of the four shadowers who shadowed a *different* model. These values were also averaged. All

analyses in this section were conducted with these values. These scores were used in a set of paired-samples t-tests on each acoustical factor and are depicted in Table 1.

Additional tests were conducted on identifying differences between the various vowels, which may have affected the formant values. This is an appropriate test because it may be the case that some vowels shift a shadower's speech in one direction, while a different vowel might lead to an opposite shift. Thus the analysis on the *average* formant value may not be able to reflect the true differences that might occur. The results of the vowel analyses are displayed in Table 2, depicting tests of the first vowel, and Table 3, depicting tests of the second vowel.

Table 1

T-tests of Acoustical Factors Assessing Similarity of Shadowers of the Same Model

	Model Shadowed			
	Same (In-group)	Different (Out-group)	*t*	*df*
Duration (in seconds)	0.033 (0.006)	0.079 (0.024)	-4.928 *	7
F0 (in hertz, hz)	22.923 (10.0175)	29.192 (13.813)	-1.052	7
V1F1 (hz)	62.766 (36.977)	83.724 (45.843)	-1.008	7
V1F2 (hz)	127.193 (40.212)	115.882 (28.421)	0.753	7
V2F1 (hz)	38.466 (17.817)	39.639 (15.471)	-0.161	7
V2F2 (hz)	148.0157 (43.680)	121.456 (31.755)	2.088	7

Note. * = $p \leq .05$. Standard Deviations appear in parentheses below means.

Table 2

T-tests of Acoustical Factors: Similarity of Shadowers of the Same Model: Initial Vowel

		Model Shadowed			
Formant	Vowel (IPA)	Same (In-group)	Different (Out-group)	*t*	*df*
F1	/eɪ/	26.141 (12.457)	30.614 (7.319)	-0.763	7
F2	/eɪ/	204.344 (119.898)	180.611 (86.210)	0.744	7
F1	/æ/	145.296 (94.832)	157.287 (67.529)	-0.313	7
F2	/æ/	128.188 (34.492)	116.379 (44.075)	0.850	7
F1	/ɑ/	88.531 (30.920)	75.609 (35.347)	1.360	7
F2	/ɑ/	192.078 (46.483)	150.876 (33.357)	8.147 ***	7
F1	/ɒ/	81.648 (19.104)	182.976 (61.315)	-3.802 *	7
F2	/ɒ/	135.151 (49.086)	119.244 (27.015)	0.968	7
F1	/ə/	129.585 (83.617)	271.752 (108.377)	-3.336 *	7
F2	/ə/	129.318 (92.590)	117.536 (62.732)	0.397	7
F1	/i/	53.792 (26.278)	44.5855 (25.531)	1.611	7

F2	/i/	279.695 (194.421)	282.659 (106.002)	-0.047	7
F1	/ɛ/	83.461 (23.967)	129.092 (54.684)	-2.029	7
F2	/ɛ/	163.445 (30.584)	136.284 (33.636)	2.201	7
F1	/ɪ/	51.420 (31.273)	54.626 (9.680)	-0.296	7
F2	/ɪ/	190.887 (84.127)	178.775 (52.7187)	0.370	7
F1	/aɪ/	137.688 (48.829)	132.501 (33.811)	0.248	7
F2	/aɪ/	94.466 (32.389)	81.701 (14.898)	1.103	7
F1	/oʊ/	39.470 (14.441)	33.818 (18.160)	3.673 *	7
F2	/oʊ/	98.070 (32.002)	225.160 (73.108)	-5.050 *	7
F1	/ɜ/	59.092 (22.658)	55.801 (6.345)	0.413	7
F2	/ɜ/	132.271 (51.898)	116.976 (38.602)	1.436	7
F1	/aʊ/	92.0813 (43.144)	131.519 (49.342)	-1.393	7
F2	/aʊ/	163.287	243.155	-1.451	7

		(105.286)	(87.124)		
F1	/u/	43.767 (10.559)	34.446 (6.939)	4.638 *	7
F2	/u/	218.381 (61.511)	280.066 (135.979)	-1.162	7
F1	/ʌ/	91.355 (66.375)	154.480 (75.666)	-1.819	7
F2	/ʌ/	139.560 (61.882)	126.363 (30.806)	0.685	7
F1	/ʊ/	48.404 (18.172)	40.607 (13.367)	1.875	7
F2	/ʊ/	285.925 (121.529)	263.876 (53.240)	0.511	7

Note. * = $p \leq .05$, *** = $p < .001$. Standard Deviations appear in parentheses below means.

Table 3

T-tests of Acoustical Factors: Similarity of Shadowers of the Same Model: Second Vowel

		Model Shadowed			
Formant	Vowel (IPA)	Same (In-group)	Different (Out-group)	*t*	*df*
F1	/æ/	104.134 (66.974)	130.715 (39.266)	-0.909	7
F2	/æ/	60.207 (19.015)	52.154 (20.902)	1.586	7
F1	/ə/	36.067 (19.041)	42.155 (10.300)	-0.756	7
F2	/ə/	149.571 (41.889)	120.015 (30.0857)	2.508 *	7
F1	/i/	33.618 (21.442)	30.810 (22.300)	0.434	7
F2	/i/	240.062 (133.219)	206.765 (106.359)	0.929	7
F1	/ˈə/	58.0621 (53.211)	54.064 (51.689)	0.254	7
F2	/ˈə/	190.617 (62.123)	150.433 (48.942)	4.282 *	7
F1	/ɪ/	44.012 (9.840)	42.490 (20.144)	0.275	7
F2	/ɪ/	168.218 (78.674)	173.421 (88.248)	-0.141	7

F1	/ɜ/	38.388 (20.743)	43.260 (17.271)	-0.528	7
F2	/ɜ/	201.840 (72.876)	170.041 (51.759)	1.221	7
F1	/u/	44.422 (14.746)	44.187 (12.018)	0.061	7
F2	/u/	215.577 (97.421)	210.134 (97.632)	0.137	7
F1	/ʌ/	119.128 (100.703)	188.627 (82.287)	-1.388	7
F2	/ʌ/	144.886 (29.744)	120.190 (10.679)	2.929 *	7
F1	/yə/	101.617 (36.624)	100.039 (34.858)	0.079	7
F2	/yə/	195.103 (85.540)	253.142 (151.677)	-1.764	7

Note. * = $p \leq .05$, *** = $p < .001$. Standard Deviations appear in parentheses below means.

The acoustical analyses were performed to identify some of the acoustical dimensions that may be relevant in alignment. These analyses were conducted on the stimuli from Experiment 1a, which were selected based on their past success in visual alignment experiments (Miller et al., 2010). They were not specifically selected for a rigorous investigation of the acoustic dimensions. As a consequence, the design is not balanced for vowel phoneme or position, nor is it representative of all possible vowel phonemes. Given evidence that people do not align equally to vowels (Babel, 2010), each vowel was examined separately. Additionally, due to the low sample size, the planned nature of the analysis, and the descriptive nature of the tests, corrections to the tests were not performed. The results of the acoustical analyses should be taken as *preliminary* evidence for the bridging between what is perceived to change and what actually changes in the speech signal.

Using these acoustic dimensions, it was found that overall, the shadowers of a given model were more similar to the shadowers who shadowed the same model than those who shadowed a different model on some dimensions. These dimensions included duration, and for *certain* vowels, V1F1, V1F2, V2F1 and V2F2. These findings are in line with the few alignment studies that have investigated the acoustical dimensions of alignment (Goldinger, 1998; Babel, 2010).

Of the dimensions that were found to be significant, all of them are considered to be observable visually. Although the dimensions measured were not an exhaustive list of factors by any means, this does provide some promising preliminary evidence that some

articulatory dimensions are indeed more similar after being exposed to a model and that these factors may be perceived visually.

Thus it could be that the pattern of results from Experiment 1a may also be found when shadowers shadow the *visual* speech of models. If the relevant articulatory dimensions are able to be perceived visually, then this would suggest that shadowers of *visual* speech of a model should also be perceived as more similar to shadowers of the same model than those who shadowed a different visually perceived model. In addition, this would also suggest that shadowers of the same model's speech should be similar when perceived auditorily-only *and* visually-only than shadowers of a different model. This would be predicted from a gestural position, given that speech is not tied to any one modality. Speech is both auditory and visual and the information from either modality is essentially the same; they have a common currency. This hypothesis is also consistent with an episodic account, if the speech information retained in the episode can be composed of visual talker-specific information. This idea was further investigated in this Dissertation via Experiment Series 2.

Chapter 6

Experiment Series 2

Most tests of speech alignment have focused primarily on auditory speech. To date, there are only a few studies (Sanchez et al., 2010; Miller et al., 2010; Gentilucci & Bernardis, 2007) in the literature that have investigated the role of visual speech in alignment. In addition, of the speech theories that account for speech alignment, only the gestural theory has openly made claims as to the role of visual speech. If visual speech affects speech production (alignment) in a similar way as auditory speech, then shadowers who perceive the visual speech (via lip-reading) of the same model, should sound more similar to each other than the shadowers of a different lip-read model.

6.1 Experiment 2a

Experiment 2a implements a test of shadowed speech that is perceived visually, via lip-reading. This test was identical to Experiment 1a, but tested the similarity of shadowers of visually perceived speech who had the same model as compared to those who lip-read a different model.

Perceptual raters will be asked to judge the relative similarity of shadowers, two of whom shadowed the same model via lip-reading, while another shadowed a different model via lip-reading. It is hypothesized that if lip-reading shadowers of the same model shift their speech in similar ways (or on similar dimensions), then perceptual raters should be sensitive to this similarity and judge them as being more similar than those who shadowed a different model.

If these results are obtained it would suggest that raters matched tokens based on some common information across utterances produced by shadowers of the same visually perceived model. This could mean that shadowers are aligning to some of the same dimensions of the model's talker-specific characteristics.

6.1.1 Method

6.1.1.1 Participants

As in Experiment 1, three sets of participants engaged in this experiment: models, shadowers, and raters. All participants were native speakers of American English and reportedly had good hearing and good or corrected vision. All participants were recruited from the University of California, Riverside.

Models. The models in this experiment were the same from Experiment Series 1.

Shadowers. The shadowers in this experiment were eight females recruited from undergraduate psychology classes for course credit. Four shadowed one model, and four shadowed the other model.

Raters. The raters in this experiment were 16 participants (3 males, 13 females) recruited from undergraduate psychology classes for course credit.

6.1.1.2 Materials

The materials for this experiment were the same as those used in Experiment 1a.

6.1.1.3 Stimuli

The stimuli in this experiment consisted of "shadowed" recordings obtained from participants in the shadowers group (see section 6.1.2. Procedure, for more information).

6.1.2 Procedure

Shadowers. The eight shadowers individually engaged in a baseline and a shadowing task in a sound attenuated booth. Each shadower first engaged in the baseline task, followed by the shadowing task.

The baseline task for these participants was identical to the baseline task in Experiment 1a (see 5.1.2 Procedure for details).

In the shadowing task, shadowers lip-read the words produced by a model on a video monitor. Lip-reading was achieved by employing a two-alternative forced choice task (2AFC) (Miller et al., 2010). On each trial shadowers first saw two words on a computer screen ('tennis' 'table'). Immediately after, participants were presented with a model's articulating face silently producing one of the two words. Shadowers were instructed to say the lip-read word out-loud, quickly but clearly. The word pairings were matched according to initial sound of the word.

Each model was shadowed by four shadowers. Shadowers were instructed to say each word they lip-read out-loud, quickly but clearly into a microphone. Shadowers lip-read the model's 74 words a total of two times from two different blocks. Their utterances from the second block were digitally recorded and edited into individual words that were amplitude adjusted. These utterances shall be referred to as the shadowers' shadowed utterances (used in Experiments 2a and 2b).

Raters. The set-up for this experiment was identical to Experiment 1a (see 5.1.2 Procedure for details), except it used the shadowed utterances to visual-only speech. As in Experiment 1a, each block consisted of 148 trials (74 words X 2 A-B positions).

6.1.3 Results and Discussion

The aim of this experiment was to investigate whether shadowers of the same model shifted (or aligned) their speech in similar ways with respect to a model whose utterances were perceived visually. If shadowers are influenced in the same way by a model's speech, then raters should judge the shadowers who perceived the same model as more similar than those who shadowed a different model.

The mean proportion for perceived alignment was calculated for each rater. The data reveals that raters judged shadowers of the same model as more similar (M = .554) at a higher proportion of the time than a shadower of a different model. These ratings were found to be statistically different from chance (.50) using a one-sample t-test, t (15) = 2.672, p = .017, Cohen's d effect size = 1.380. Thus, the data suggests that shadowers of the same model are influenced by the visual speech they perceive, and that

this influence is evident in their speech productions. In addition, this also suggests that shadowers who lip-read the same model's speech are influenced in a similar manner. Thus, a given model's speech is influential on a given perceiver's eyes (or ears) in a similar way.

To ensure that the results were not driven by the results of a particular model, an independent samples t-test was conducted on the factor of model shadowed (Model 1 vs. Model 2). The results of this analysis did find a significant difference between the models, t (14) = -2.248, p = .041, Cohen's d effect size = 1.202. Visual shadowers of Model 2 (M = .595) were rated as sounding more like their fellow shadowers who shadowed the same model than visual shadowers of Model 1 (M = .514). In light of the differences between the models, the shadowers of the models were compared against chance separately. The shadowers of Model 1 were not found to be similar. Their scores were not significantly different from chance, t (7) = .618, p = .556. However, those who shadowed the visual face of Model 2 were found to be significantly different from chance, t (7) = 3.334, p = .012, Cohen's d effect size = 2.521, indicating that shadowers of that model were found to be similar. Thus, the visual shadowers of Model 2 shifted their speech in a similar way compared to shadowers of the other model.

It is unclear why the shadowers of a particular lip-read model were found to be more similar to each other than the shadowers who shadowed the other model. This difference is somewhat surprising given that the groups of shadowers in Experiment 1a who *heard* the speech of the models were not considered different with respect to their similarity.

Perhaps the differences stem from the ease or difficulty of lip-reading. It is possible that one of the models was easier to lip-read than the other, thus the shadowers of that model may have had an easier time in producing the speech perceived. Although there were no differences in the shadowing accuracy between shadowers of different models, there may have been some subtle differences between the shadowers' ease of lip-reading the models. For instance, if the model was more difficult to lip-read, this may then leave the shadower with some ambiguity as to the nature of the articulation, leading to poor alignment. What is more, it is also possible that social factors were involved in the differences of the shadowers of a given model. It is possible that one of the models was perceived to be more attractive, while the other model was not given that assessment. If shadowers perceived a model as being attractive, it may have motivated them to align more to that model. These shadowers might then be judged as being more similar because they aligned with a greater fidelity than the other shadowers.

Finally, an additional test was conducted on the role of the position, A verses B, of the judged utterances words in the AXB task. The results of this analysis did not find a significant difference between the position of the utterances and its likelihood of being selected as more similar to the item in the X position, t (15) = 1.170, p = .260, Cohen's d effect size = 0.604.

Nevertheless, the results of this experiment suggest that shadowers of the same lip-read model can align in a similar way, at least for one of the models. These findings compliment the results for auditory shadowed speech. These results are in line with the gestural approach and an episodic approach that includes visual speech episodes. The

current experiment does not, however, suggest that the speech information garnered auditorily and visually are similar, as the gestural theory suggests. Experiment 2b will address this proposition.

6.2 Experiment 2b

Experiment 2b aims to identify whether the talker-specific information that is influential to shadowers of the same model's speech is similar when perceived auditorily-only *and* visually-only.

Perceptual raters were asked to judge the relative similarity of shadowers, both of whom shadowed the same model, but with each perceiving the model's speech using a different modality (auditory or visual). It is hypothesized that if the perceived talker information is similar across the modalities, then perceptual raters should be sensitive to this common information of shadowers of the same speech. Raters should thus judge shadowers of the same model as being more similar to each other regardless of how the information was shadowed (auditorily or visually) than a shadower of a different model.

6.2.1 Method

6.2.1.1 Participants

Raters. The participants in this experiment consisted of 32 (thirteen males, nineteen females) perceptual raters recruited from undergraduate psychology classes for course credit. All participants were native speakers of American English and reportedly

had good hearing and good or corrected vision. All participants were recruited from the University of California, Riverside.

6.2.1.2 Materials

The materials for this experiment were the same as those used in Experiment 1a.

6.2.1.3 Stimuli

The stimuli in this experiment consisted of "shadowed" recordings obtained from participants in the shadowers groups from Experiment 1a and Experiment 2a (see sections 5.1.2 and 6.1.2. Procedure, for more information).

6.2.2 Procedure

Raters. The 32 perceptual raters listened to 6 unique shadowers per experiment. This experiment was divided into two rating blocks. The shadower in the X position was different between the blocks, but both had shadowed the same model, though through a different modality. Thus, for a given block, the shadower in the X position had auditorily-only or visually-only shadowed the model's speech. The shadower pairs in the A and B positions were also different between the blocks and had shadowed a different modality from the shadower in the X position. The shadowers in the A and B positions consisted of a shadower who shadowed the same model and a shadower who shadowed a different model as the person in the X position. Each block consisted of 148 trials (74 words X 2 A-B positions).

6.2.3 Results and Discussion

The aim of this experiment was to investigate whether shadowers of the same model shifted their speech in perceptually similar ways, regardless of how the model was perceived (auditorily or visually). If shadowers of the same model's speech are similarly influenced despite differences in the modality used, then this would suggest that shadowers are influenced by and aligning to some amodal gestural talker-specific properties. Thus, raters should judge the shadowers of the same model as more similar than shadowers of a different model.

The mean proportion for perceived alignment was calculated for each rater. The data reveals that raters judged shadowers of the same model as more similar (M = .535) at a higher proportion of the than a shadower of a different model. These ratings were found to be statistically different from chance (.50) using a one-sample t-test, t (31) = 4.366, p = .001, Cohen's d effect size = 1.568. These results suggest that shadowers of the same model are influenced in similar ways by the speech they perceive, regardless of *how* the speech was perceived (e.g. auditory-only or visual-only). This influence is evident in the shadowers' speech productions.

To ensure that the results were not driven by a particular model, an independent samples t-test was conducted on the factor of model shadowed (Model 1 vs. Model 2). The results of this analysis did not find a significant difference between the models, t (30) = -.131, p = .897, Cohen's d effect size = 0.048.

Finally, an additional test was conducted on the role of the position, A verses B, of the judged utterances words in the AXB task. The results of this analysis did not find a

significant difference between the position of the utterances and its likelihood of being selected as more similar to the item in the X position, t (31) = 1.885, p = .069, Cohen's d effect size = 0.677.

These results suggest that not only are there similarities to how shadowers align to a given model's speech, but that they align in a similar way regardless of whether the model was perceived auditorily or visually. These findings can be addressed by the gestural theories of speech, given that speech information is thought to be amodal, meaning the information is not tied to a given modality. These findings have ramifications on how the episodic theory is considered. This will be notion will be discussed further in the general discussion (7.1).

Chapter 7

General Discussion

The aim of this Dissertation was to assess how talker-specific characteristics influence speech alignment by addressing the following questions: *Do shadowers of the same model sound more similar to each other than they do to shadowers of a different model? Does the sensory modality of the shadowed speech affect the perceptual similarity between the shadowers of the same model?* These questions were addressed by two series of experiments, using perceptual judgments and acoustical analyses, where shadowers perceived the speech of a model either auditorily or visually.

In Experiment 1a, it was found that perceptual raters judged the shadowed utterances of those who shadowed the auditory utterances of the same model as sounding more similar than the shadowed utterances of those who shadowed a different model. The results of Experiment 1a suggest that shadowers of the same model shifted their speech in some similar ways. Thus, shadowers of the same model are influenced in a similar manner and consequently sound alike.

In Experiment 1b it was found that raters were more likely to judge the shadowed utterance of a shadower as more similar to the model shadowed than the shadower's

baseline utterance. The results of Experiment 1b suggests that shadowers are influenced by the speech they perceive. The perception of a model's speech can lead to changes in how one produces speech. This change in speech production seems to be a shift from how one speaks in the direction of the model perceived.

In Experiment 1b it was also found that the *baseline* utterances of shadowers who shadowed the same or different model were not considered to be consistently similar. The results of Experiment 1b suggests that the shadowers of a given model did not originally sound like each other (before shadowing). This finding gives additional support to the results of Experiment 1a. It seems that shadowers similarly shifted their speech from the influence of a perceived model. These shadowers thus sounded more alike because of this shift.

In the acoustical analyses, it was found that shadowers of the same model are more similar to each other after perceiving the same model. Moreover, it was found that shadowers of the same model were more similar to each other along articulatory dimensions that are able to be *seen*. This suggests that shadowers similarly shift their speech to articulatory dimensions that contain information which can be characterized as amodal.

In Experiment 2a it was found that perceptual raters judged the shadowed utterances of those who shadowed the lip-read speech from the same model as more similar than those who shadowed the lip-read speech of a different model, at least for one of the models. These results suggest that, like auditory speech, visual speech can sometimes influence speech productions of those who perceived the same model, in a

similar way. Thus, shadowers of the same lip-read model can be influenced in a similar manner and consequently sound alike.

In Experiment 2b it was found that perceptual raters judged the shadowed speech of those who shadowed the same model as being more similar than those who shadowed a different model, regardless of whether the model was perceived auditorily or visually. This suggests that the information shadowed auditorily contains some of the same talker-specific information as information perceived visually. In other words, shadowers of a model can be influenced in similar ways, regardless of how the information was perceived (auditorily or visually).

7.1 Theoretical Implications

The experiments conducted for this dissertation have theoretical implications. Obtaining evidence that visually perceived speech shifts speech productions in a similar way to auditory perceived speech (Experiment 2a) suggests that talker-specific information can be amodal and can be transmitted through both modalities. This is in line with the findings obtained by Miller, Sanchez, and Rosenblum (2010) who found that shadowing the voice or visual face of a model leads to a shift in the shadower's speech in the direction of the model, compared to the shadower's baseline utterance.

Furthermore, within this Dissertation, it was found that shadowed speech from a visually perceived model and an auditorily perceived model similarly affected the speech of different shadowers (Experiment 2b). This suggests that some of this information from the different modalities is inherently similar. This suggests that some talker-

specific information consists of amodal gestures, meaning that it is not tied to a specific modality. In this sense, these findings are consistent with the tenants of the gestural theories (Liberman & Mattingly, 1985, 1989; Fowler, 1986).

These results also have implications for how the episodic theory is considered. Given that the episodic theory has largely been considered an auditory theory, the observed results could further suggest that talker-specific characteristics may be stored in visual, in addition to auditory, episodes of talker-specific information. Moreover, these findings further alter the current notion of the nature of a speech episode by suggesting that talker-specific characteristics may be stored in *amodal gestural episodes*. The current findings have also furthered the current theoretical landscape by finding some initial evidence of some articulatory dimensions that are influential in alignment, where the relevant dimensions were able to both be heard and seen.

Thus, if the theory of episodic encoding incorporates talker-specific information that takes an amodal gestural form, then it would be able to account for the visual speech alignment results as well as other findings from the gestural approach. However, reconsidering the stored episodes as gestures would also make a number of predictions about the longer-term influences of talker-specific information on speech perception, memory, and production. For example, if retained episodes take a gestural form, one would expect to see visual and sub-lexical talker-specific influences on perceiving, remembering, and producing speech.

Moreover, finding that shadowers of the same model are influenced in a similar way both when hearing and seeing a model has implications for understanding how

dialects are formed and spread. The results of this Dissertation may serve as an early step in understanding the mechanisms of dialect change, where a given talker's speech influences those who directly perceive it. The results in this Dissertation are consistent with the computational model of the spread of dialects proposed by Fagyal et al. (2010). In the model, a single person's speech has the opportunity of altering the speech of society, given certain conditions (e.g. the person is charismatic and well connected). This Dissertation provides initial evidence that indeed a single person (model) can influence the speech of multiple people (shadowers) in a similar way.

As mentioned, the observed results suggest that talker-specific characteristics may be stored in *amodal gestural episodes*, and that these stored episodes influence produced speech. Given that multiple shadowers of the same model were similarly influenced by the speech of a model, then the information within the amodal gestural episodes could be similar for the multiple shadowers of a model. Although this represents only a microcosm of speech, it does have relevance with respect to dialects. In essence, it suggests that dialects are based on the direct and indirect transference of amodal talker-specific gestural characteristics. Thus, returning to Larry, the well connected charismatic figure, his amodal gestural characteristics influence not only Chris and Steve's speech, but also the speech of others with whom Chris and Steve interact.

Moreover, the changes in a dialect that occur due to a person's amodal gestural characteristics would presumably be evident in words that are relatively rare. Finding dialectical changes for rarely occurring words would be in line with an episodic gestural theory and would also be supported by the language change literature. Within the area of

language change it has been found that frequency influences language production, use and acquisition (Diessel, 2007; Bod, Hay, & Jannedy, 2003; Bybee & Hopper, 2001), where common items are found to be resistant to change, but rare items tend to be more susceptible to change. Further investigations may identify the presence of indirect amodal talker-specific characteristics that influences a given talker's speech on a community, which may lead to a change in dialect.

7.2 Directions for Future Work

If speech episodes contain gestural, rather than simply auditory information, some interesting predictions arise. Evidence suggesting that speech information takes an amodal gestural form should be observable in the other phenomena explained by episodic theories. These phenomena would include the talker effects on recognition word memory. If, for example, episodes contain gestural rather than auditory word information, they could be established by visual speech information. Moreover, regardless through which modality the gestural episodes are established, they should be able to influence speech perception in either modality. In fact, there is evidence for talker effects across modalities. Rosenblum et al., (2007) found that being familiarized with a talker's visible articulating face in a lip-reading task, later provides facilitation when hearing that same talker's speech in white noise. Thus, talker-specific information learned from one modality can be utilized in the service of a different modality. This could mean that talker-specific characteristics are carried through, and stored as gestural information.

There is another implication of the stored episodes taking a gestural, rather than auditory word form. In that gestures encode speech movements rather than lexical items per se, the evidence for episodic encoding should work with sub-lexical, as well as lexical units.

Recent investigations (Shockley et al., 2004; Nielsen, 2008; Sanchez et al., 2010) in speech alignment have found evidence for alignment to sub-lexical stimuli on a phonetically relevant dimension, voice onset time (VOT). The Nielsen (2008) study also found evidence that alignment to VOTs generalized on gestural similarity. In her examination, participants were asked to listen to words that varied on lexical frequency. These heard words all had initial /p/s whose VOTs were extended. After the listening task, participants read text words out-loud (as they did during a baseline phase). Alignment to words that were heard and the presence of generalization to words that were not heard were investigated. It was found that alignment (as measured by VOT) was greatest for previously heard low frequency items, which the episodic theory would predict. However, it was also found that /k/ text items that were read out-loud also had extended VOTs as compared with the baseline recoding. The generalization of VOT lengthening from /p/ items to /k/ items was thought to be due to a generalization based on sub-lexical episodes (note: both /p/ and /k/ share a common articulatory gesture). This suggests that not only are lexical units stored in episodes, but that perhaps there may be episodes for sub-lexical gestures. However, it is unknown whether the same effects (generalization) would be found if visual information was used, instead of auditory only presentations, as Sanchez et al. (2010) found.

The current findings may also be relevant in investigations on dialect formation. An examination of how second-hand or indirect perceivers of a talker's speech may identify the key components of speech that lead to lasting change in a community. Given the current findings in light of Fagyal et al.'s (2010) computational model of dialect change, it is reasonable to predict that shadowers of shadowers of a model may might not only sound like the original shadower of a model, but also the model as well, especially for rarely occurring words. In acknowledgment that speech alignment results are rather sensitive here, perceptual raters and acoustical analyses may benefit each other in understanding and identifying similarities.

7.3 Practical Implications

This research also has practical implications. The notion of talker-specific information that is not necessarily tied to a particular modality may impact the technology involved in systems used in teleconferencing and speech/talker identification systems. Advancement in these areas may help in the identification and isolation of critical talker-specific gestures relevant to a given talker. In addition, the speech alignment methodology may be used to benefit education (both traditional and web-based) and especially second language learning. Finally, this research also has implications for the deaf and visually impaired communities. For example, the creation of comprehensive programs and curriculum that converge on the basic notion that speech takes an amodal form may facilitate communication as well as language acquisition for these groups.

7.4 Conclusion

Like a chameleon, humans are ever resilient, ever adaptable, and ever changing. This is evident in many ways, even in produced speech. The *color* of a shadower's utterances can be influenced in similar ways by the speech that is heard and seen. In fact, the voice and face of a talker (or model) can similarly influence the speech of multiple shadowers. The ramifications of the spreading of a given talker's speech may impact not only the direct perceivers of that speech, but may also influence others in a talker's community, possibly leading to a dialectical change. Regardless as to whether the purpose of the alignment process is to promote understanding in communication or social relations, it is clear that the influential speech information responsible is not tied to a particular modality. What you hear is what you see and subsequently what you say.

References

Abercrombie, D. (1967). *Elements of general phonetics*. Chicago:IL: Aldine Publishing Company.

Algeo, J. (Ed.). (2001). The Cambridge history of the English language: Volume VI: English in North America. Cambridge, UK: University Press.

Arnold, P. & Hill, F. (2001). Bisensory augmentation: A speechreading advantage when speech is clearly audible and intact. *British Journal of Psychology*, *92*, 339-55.

Azuma, S. (1997). Speech accommodation and Japanese Emperor Hirohito*, Discourse & Society, 8*. 198-202.

Babel, M. (2010). Dialect divergence and convergence in New Zealand English. *Language in Society*, *39*, 437-456.

Bod, R., Hay, J., & Jannedy, S. (Eds.). (2003). *Probabilistic linguistics*. Cambridge, MA: MIT Press.

Boersma, P., & Weenink, D. (2008). Praat: Doing phonetics by computer (Version 5.0.32) [Computer program]. Retrieved January 2008 from www.praat.org/.

Borden, G. J., & Harris, K. S. (1984). Speech science primer: Physiology, acoustics, and perception of speech (2nd.Ed). Baltimore, MD: Waverly Press, Inc.

Bourhis, R., & Giles, H. (1977). The language of intergroup distinctiveness. In Howard Giles (ed.), Language, ethnicity, and intergroup relations, 119-136. London: Academic Press.

Bybee, J., & Hopper, P. (Eds.). (2001). Frequency and the emergence of linguistic structure. Amsterdam: John Benjamins.

Chambers, J. K. (1992). Dialect Acquisition, *Language, 68*, 673-705.

Chartrand, T. L., & Bargh, J. A. (1999). The chameleon effect: The perception-behavior link and social interaction. *Journal of Personality and Social Psychology*, *76*, 893-910.

Cohen, J., MacWhinney, B., Flatt, M., & Provost, J. (1993). PsyScope: An interactive graphic system for designing and controlling experiments in the psychology laboratory using Macintosh computers. *Behavior Research Methods, Instruments, & Computers*, *25*, 257-271.

Coupland, N. (1984). Accommodation at work: Some phonological data and their implications. *International Journal of the Sociology of Language, 46*, 49-70.

Craik, I. M., & Kirsner, K. K. (1974). The effect of speaker's voice on word recognition. *Quarterly Journal of Experimental Psychology*, *26*, 274-284.

Crystal, T. H., & House, A. S. (1988). The duration of American-English stop consonants: An overview. *Journal of Phonetics*, *16*, 285–294.

Diessel, H. (2007). Frequency effects in language acquisition, language use, and diachronic change, *New Ideas in Psychology, 25*, 108-127.

Drager, K. (2006). Social categories, grammatical categories, and the likelihood of "like" monophtongization. Proceedings Australian International Conference on Speech Science & Technology, 11th, Auckland, pp. 384–87. Auckland: University of Auckland Press.

Eckert, P. (1996). Vowels and nail polish: The emergence of linguistic style in the preadolescent heterosexual marketplace. Gender and Belief Systems: Proceedings BerkeleyWomen and Language Conference, 4th, Berkeley, pp. 183–90, ed. Warner, Ahlers, Bilmes, Oliver, Wertheim, & Chen. Berkeley: Berkeley Women and Language Group.

Fadiga, L., Craighero, L., Buccino, G., & Rizzolatti, G. (2002). Speech listening specifically modulates the excitability of tongue muscles: A TMS study. *European Journal of Neuroscience*, *15*, 399–402.

Fadiga, L., Fogassi, L., Povesi, G., & Rizzolatti, G. (1995). Motor facilitation during action observation: A magnetic stimulation study. *Journal of Neurophysiology*, *73*, 2608–2611.

Fagyal, Z., Swarup, S., Escobar, A., Gasser, L., & Lakkaraju, K. (2010). Centers and peripheries: Network roles in language change, *Lingua, 120,* 2061-2079.

Fowler, C. A. (2004). Speech as a supermodal or amodal phenomenon. In Calvert, G. A., Spence, C., & Stein, B. E. (eds.). *The Handbook of Multisensory Processing*, 189-201,Cambridge, MA: MIT Press.

Fowler, C. A., Brown, J. M., Sabadini, L., & Weihing, J. (2003). Rapid access to speech gestures in perception: Evidence from choice and simple response time tasks. *Journal of Memory & Language*, *49*, 396-413.

Garrod, S., & Pickering, M. J. (2004). Why is conversation so easy? *Trends in Cognitive Sciences*. *8*, 8-11.

Gentilucci, M. & Bernardis, P. (2007) Imitation during phoneme production. *Neropsychologia*, *45*, 608-615.

Gerstman, H. (1968). Classification of selfnormalized vowels. *IEEE Transactions on Audio and Electoacoustics*, *16*, 630-640.

Gibson, J. (1966). The senses considered as perceptual systems. Boston: Houghton-Mifflin.

Giles, H., Coupland, J., & Coupland, N. (1991). Accommodation theory: Communication, context and consequences. In H. Giles, J. Coupland, & N. Coupland (Eds.), *Contexts of accommodation: Developments in applied sociolinguistics* (pp. 1-68). Cambridge: Cambridge University Press.

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

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105*, 251-79.

Goldinger, S. D., Kleider, H. M., & Shelley, E. (1999). The marriage of perception and memory: Creating two-way illusions with words and voices. *Memory & Cognition*, *27*, 328-338.

Goldinger, S. D., Pisoni, D. B., & Logan, J. S. (1991). On nature of talker variability effects on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*, 152-162.

Goldinger, S. D., & Azuma, T. (2004). Episodic memory reflected in printed word naming. *Psychonomic Bulletin & Review*, *11*, 716-722.

Grant, K. W., & Seitz, P. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *Journal of the Acoustical Society of America*, *108*, 1197-1208.

Gregory, S. W. (1990). Analysis of fundamental frequency reveals covariation in interview partners' speech. *Journal of Nonverbal Behavior*, *14*, 237-251.

Gregory, S. W., & Webster, S. (1996). A nonverbal signal in voices of interview partners effectively predicts communication accommodation and social status predictions. *Journal of Personality and Social Psychology, 70*, 1231–1240.

HairerSoft. (2008). Amadeus II [Sound editing software]. Retrieved January 2008 from www.hairersoft.com.

Hommel, B., Musseler, J., Aschersleben, G., & Prinz, W. (2001). The theory of event coding (TEC): A framework for perception and action planning. *Behavioral and Brain Sciences*, *24*, 849–878.

Joos, M. A. (1948). Acoustic Phonetics. *Language*, *24* (Suppl. 2), 1-136.

Kim, J., & Davis, C. (2004). Integrating the audio-visual speech detection advantage. *Speech Communication*, *44*, 19-30

Kučera, H., & Francis, W. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

Lachs, L., McMichael, K., & Pisoni, D. B. (2000). Speech perception and implicit memory: Evidence for detailed episodic encoding of phonetic events. *Research on Spoken Language Processing*, *Progress Report No. 24*. Bloomington IN: Indiana University.

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

Meltzoff, A. N., & Moore, M. K. (1997). Explaining facial imitation: A theoretical model. *Early Development & Parenting*, *6*, 179-192.

Miller, R. M., Sanchez, K., & Rosenblum, L. D. (2010). Alignment to visual speech information. *Attention, Perception, & Psychophysics*, *72*, 1614-1625.

Mills, A.E. (1987). The development of phonology in the blind child (pp. 145-162). In B. Dodd and R. Campbell Eds, *Hearing by eye: The psychology of lip reading*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Namy, L. L., Nygaard, L. C., & Sauerteig, D. (2002). Gender difference in vocal accommodation: The role of perception. *Journal of Language and Social Psychology*, *21*, 422-432.

Natale, M. (1975). Convergence of mean vocal intensity in dyadic communication as a function of social desirability. *Journal of Personality & Social Psychology*, *32*, 790-804

Natale, M., (1975). Social desirability as related to convergence of temporal speech patterns. *Perceptual & Motor Skills*, *40*(3), 827-830

Newman, R. S., Clouse, S. A., Burnham, L. J. (2001). The perceptual consequences of within-talker variability in fricative production. *Journal of the Acoustical Society of America*, *109*, 1181-1196.

Nielsen, K. Y. (2008). Word-level and feature-level effects in phonetic imitation. *Unpublished doctoral Dissertation*, University of California, Los Angeles.

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Percept Psychophysics*, *60*(3),355-376.

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

Palmeri, T. J., Goldinger, S. D., & Pisoni D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*, 309-328.

Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *Journal of the Acoustical Society of America, 119*, 2382–2393.

Perkell, J. S., Zandipour, M., Matthies, M. L., & Lane, H. (2002). Economy of effort in different speaking conditions. I. A preliminary study of intersubject differences and modeling issues. *Journal of the Acoustical Society of America*, *112*, 1627-1641.

Perkell, J. S., Matthies, M. L., Tiede, M., Lane, H., Zandipour, M., Marrone, N., & Stockmann, E., Guenther, F. H. (2004). The distinctness of speakers' /s/-/s/ contrast is related to their auditory discrimination and use of articulatory saturation effect. *Journal of Speech, Language, and Hearing Research*, *47*, 1259-1269.

Pickering, M. J., Garrod, S., (2004). Toward a mechanistic psychology of dialogue. *Behavioral & Brain Sciences*, 27(2), 169-226.

Reisberg, D., McLean, J., & Goldfield, A. (1987). Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli. In B. Dodd & R. Campbell (Eds), *Hearing by eye: The psychology of lip reading*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Remez, R. E., Fellows, J. M., Pisoni, D. B., Goh, W. D., & Rubin, P. E. (1998). Mutlimodal perceptual organization of speech: Evidence from tone analogs of spoken utterances, *Speech Communication*, *26*, 65-73.

Rizzolatti, G., Fadiga, L., Gallese, V., & Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Cognitive Brain Research*, *3*, 131–141.

Rosenblum, L.D. (2005). The primacy of multimodal speech perception. In D. Pisoni & R. Remez (Eds.). *Handbook of Speech Perception*. Blackwell: Malden, MA. pp. 51-78

Rosenblum, L. D, Miller, R., & Sanchez, K. (2007). Lipread me now, hear me better later: Crossmodal transfer of talker familiarity effects. *Psychological Science*, *18*, 392-396.

Rosenblum, L. D., Niehus, R. P., & Smith, N. M. (2007). Look who's talking: Recognizing friends from visible articulation, *Perception, 36*, 157-159.

Rosenblum, L. D., Yakel, D. A., Baseer, N., Panchal, A., Nodarse, B. C., & Niehus, R. P. (2002). Visual speech information for face recognition, *Perception & Psychophysics, 64*, 220-229.

Sanchez, K., Miller, R. M., & Rosenblum, L. D. (2010). Visual influences on alignment to voice onset time. *Journal of Speech, Language, and Hearing Research*, *53*, 262-272.

Sancier, M. L., & Fowler, C. A. (1997). Gestural drift in a bilingual speaker of Brazilian Portuguese and English. *Journal of Phonetics*, *25*, 421-436.

Sheffert, S. M., & Fowler, C. A. (1995). The effects of voice and visible speaker changes on memory for spoken words. *Memory & Language*, *34*, 665-685.

Shockley, K., Sabadini, L., & Fowler, C. A. (2004). Imitation in shadowing words. *Perception & Psychophysics, 66*, 422-429.

Summerfield, Q., & Haggard, M. P. (1973). Vocal tract normalization as demonstrated by reaction times. *Report on Research in Progress in Speech Perception*, *2*, 1-12. Belfast, Ireland: The Queen's University of Belfast, Department of Psychology.

Trudgill, P. (2008). Colonial dialect contact in the history of European languages: On the irrelevance of identity to new-dialect formation, *Language in Society, 37*, 241-280.

Tsao, Y., Weismer, G., & Iqbal, K. (2006). The effect of intertalker speech rate variation on acoustic vowel space. *Journal of the Acoustical Society of America*, *119*, 1074-1082.