UC Riverside

UC Riverside Previously Published Works

Title

Is speech alignment to talkers or tasks?

Permalink

https://escholarship.org/uc/item/8378b228

Journal

Attention, Perception, & Psychophysics, 75(8)

ISSN

1943-3921

Authors

Miller, Rachel M
Sanchez, Kauyumari
Rosenblum, Lawrence D

Publication Date

2013-11-01

DOI

10.3758/s13414-013-0517-y

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

NIH Public Access
Author Manuscript
Atten Percept Psychophys. Author manuscript; available in PMC 2014 November 01.

Published in final edited form as:
Atten Percept Psychophys. 2013 November ; 75(8): . doi:10.3758/s13414-013-0517-y.

Is speech alignment to talkers or tasks?

Rachel M. Miller, **Kauyumari Sanchez**, and **Lawrence D. Rosenblum**
University of California, Riverside

Abstract

Speech alignment, the tendency of individuals to subtly imitate each other's speaking style, is often assessed by comparing a subject's baseline and shadowed utterances to a model's utterances often through perceptual ratings. These types of comparisons provide information about the occurrence of a change in subject's speech, but do not indicate that this change is towards the *specific* shadowed model. Three studies investigated whether alignment is specific to a shadowed model. Experiment 1 involved the classic baseline to shadowed comparison to confirm that subjects did, in fact, sound more like their model when they shadowed, relative to any pre-existing similarities between a subject and model. Experiment 2 tested whether subjects' utterances sounded more similar to the model they had shadowed or to another unshadowed model. Experiment 3 examined whether subjects' utterances sounded more similar to the model they had shadowed or to another *subject* who shadowed a different model. Results of all experiments revealed that subjects sounded more similar to the model they had shadowed. This suggests that shadowing-based speech alignment is not just a change; it is a change in the direction of the shadowed model, specifically.

Speech alignment describes the tendency of talkers to subtly imitate the speaking style of the person to whom they are talking (Goldinger, 1998; Goldinger & Azuma, 2004; Miller, Sanchez, & Rosenblum, 2010; Namy, Nygaard, & Sauerteig, 2002; Pardo, 2006; Shockley, Sabadini, & Fowler, 2004, Sanchez, Miller, & Rosenblum, 2010, Sanchez, 2011; Dias & Rosenblum, 2011; Nielsen, 2011). The phenomenon has been demonstrated in different empirical contexts including interactive talker tasks, as well as in word shadowing tasks. Further, it has been demonstrated not only when spoken words are presented auditorily, but when they are presented visually – in a lipreading task (Gentilucci & Bernardis, 2007; Miller et al., 2010; Sanchez et al., 2010, Sanchez, 2011).

Generally, researchers have concluded that subjects will produce speech that has become more like that of the talker (model) with whom they interacted, or whom they shadowed. This conclusion is often based on comparisons between pre-task (or baseline) speech, often produced by subjects as they read words prior to the critical alignment task; and post-task speech that is produced during or after the alignment task (but see Gregory, Dagan, & Webster, 1997; Gregory, Green, Carrothers, Dagan, & Webster, 2001; Levitan & Hirschberg, 2011). Alignment is said to occur when the words uttered during the interaction or shadowing task are judged, or measured, as more similar to those of the model than are the baseline words spoken by the subject during the pre-alignment task.

However, in being based on comparisons between a subject's own utterances, the alignment findings which have used baseline comparisons cannot definitively show that subjects sound more like the *specific* model with whom they interacted, or whom they shadowed. The current experiments were designed to examine this possibility using an AXB rating task.

Correspondences should be addressed to Lawrence D. Rosenblum, Department of Psychology, 900 University Ave., University of California, Riverside, CA 92521.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Speech Alignment Paradigms

It has long been reported that talkers subtly change their speech patterns to be more like the person with whom they are talking. While social and situational factors can influence its prominence (e.g., Giles & Coupland, 1991; Gregory & Webster, 1996; Pardo, Jay, & Krauss, 2010), interlocutors have been shown to partially match each other's speech rate, accent, frequency/amplitude contours, and vocal intensity (e.g., Giles, Coupland, & Coupland, 1992; Gregory, 1990; Harrington, et al. 2000; Natale, 1975; Sancier & Fowler, 1997). Alignment has also been observed at the word and phoneme level in a variety of experimental settings (Goldinger, 1998; Goldinger & Azuma, 2004; Miller et al., 2010; Namy et al., 2002; Pardo, 2006; Shockley et al., 2004, Sanchez et al., 2010, Sanchez, 2011; Dias & Rosenblum, 2011; Nielsen, 2011).

For example, alignment has been found in socially-isolated tasks. In one of the first empirical demonstrations of the effect, Goldinger (1998) asked subjects who were isolated in a sound booth to shadow (produce as quickly as possible) a series of recorded words spoken by a model. Subjects were not told to imitate, or even repeat what they heard, but to simply say the words quickly and clearly. Goldinger then asked naïve raters to judge the similarity between the model's target words and subjects' shadowed words, relative to *baseline* words read by subjects before the shadowing task. For this purpose, Goldinger implemented an AXB matching task in which words were presented to raters in sets of three, where the middle token (X) was always the word spoken by the model, while the words in the first (A) or third (B) position consisted of the shadower's baseline (read) word or shadowed word. Raters were asked to judge whether the shadower's shadowed or baseline word was a better imitation of the model's word (X). Results revealed that raters judged the shadowed words as better imitations of the model's words at greater than chance levels.

Speech alignment has also been found in interactive tasks. For example, Pardo (2006) examined alignment of words uttered by interlocutors during an interactive instruction task. The task involved one subject instructing a second to navigate a pencil around a map containing drawn landmarks. To successfully complete the task, subjects needed to produce statements containing a series of landmark label phrases (e.g., pine forest). Subjects had also produced these phrases before the interactive task, in order to provide baseline utterances for later comparison. Based on raters' AXB matching judgments, cross-speaker phrases produced during the interaction were found to be more similar than phrases produced before the task (see also Pardo et al., 2010).

In addition, speech alignment has been reported in a task involving neither shadowing nor interaction. In this experiment, talkers were found to speak more like models one week after exposure to the models' speech (Goldinger & Azuma, 2004). Talkers were first asked to read a list of words (pre-task), to then listen to models producing these words (listening task), and to again read the list of words one week later (post-task). To measure alignment, naïve raters made AXB judgments in which they compared the model's word (X) to the talker's pre- and post-task words (A or B). Raters found that the post-task words read by subjects were more similar to the models' words, despite having heard those words one week earlier. Finally, speech alignment can also be induced and enhanced with lipread speech, even in subjects with no formal lipreading experience (e.g., Dias & Rosenblum, 2011; Miller et al., 2010; Sanchez et al., 2010, Sanchez, 2011).

Theoretical Significance of Speech Alignment

Speech alignment phenomena have influenced theories of speech perception and production, as well as lexical access and social interaction. For example, findings that speech production can be spontaneously influenced by the specifics of perceived speech (as in shadowing

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

tasks) have been interpreted as support for the close link between the speech perception and production functions (Fowler, 2004; Fowler, Brown, Sabadini, & Weihing, 2003; Sancier & Fowler, 1997; Shockley et al., 2004). Another explanation of alignment comes from Goldinger (1998), who suggests that speech perception involves storage of highly-detailed episodes of speech events in a mental lexicon, which contain information about the word and about the model (e.g., idiolect) who produced the word. Subsequent presentations of that word activate stored traces, which then influence a *production* of that word that is more similar to the original talker's production (i.e., alignment).

Additional accounts of alignment reveal its importance as a social behavior that increases identification with others by reducing social distance (Babel, 2010; Giles & Coupland, 1991; Gregory & Webster, 1996; Pardo et al., 2010).

Common to all explanations of alignment, is the assumption that perceivers change their speech to be more like the *specific talker* they hear. Therefore, determining that talkers align to the actual speaker they perceive has theoretical import.

Evaluating Speech Alignment

There are two classes of methods used to evaluate the presence of speech alignment. One method involves acoustic or articulatory analyses, which are used to compare changes in factors (duration, voice onset time, vowel space, articulatory rate, lip kinematics) corresponding to a model's speech signal that may indicate alignment (e.g., Babel, 2010; Gentilucci & Bernardis, 2007; Honorof, Weihing, & Fowler, 2010; Mitterer & Ernestus, 2008; Sanchez et al., 2010; Shockley et al., 2004). For example, Shockley and his colleagues (Shockley et al., 2004; and see also Nielsen, 2011; and Sanchez et al., 2010) measured voice onset time (VOT) values of words produced during shadowing of tokens with lengthened VOTs. These authors observed that words spoken when subjects shadowed the VOT-extended words had longer VOTs than did baseline words spoken before the shadowing task. While these acoustical evaluations of speech alignment suggest that talkers can align to specific phonetic details, it is still unclear whether subjects are aligning to their *particular* models, or if there is a general shift in these phonetic dimensions. This is also true of studies (e.g., Pardo, et al. 2010) in which no *specific* acoustic manipulation is applied to the model stimuli or is predicted to be imitated. Without a specific acoustic/articulatory prediction, finding greater overall acoustic similarity between model and shadowed (vs. baseline) utterances could simply be a consequence of alignment tasks inducing utterances that have acoustic/articulatory properties that are more like those of other talkers *in general*.

However, a handful of studies using acoustic measures *have* implemented comparisons to determine if aligners change their speech in the direction of a specific talker (e.g., Gregory, et al. 1997; Gregory, et al. 2001; Levitan & Hirschberg, 2011). For example, Gregory and his colleagues (1997 colleagues (2001) have evaluated F0 convergence in interlocutors by comparing differences in F0 between subjects who actually interacted with one-another (partners) to differences in F0 between two subjects who did not interact (pseudo-partners). In general, F0 values of the actual partners were closer, suggesting that subjects did actually align toward the specific interlocutor with whom they were interacting.

More commonly, speech alignment, and especially phonetic alignment, is evaluated with the aforementioned rater-matching (AXB) judgments for which naïve raters are asked to judge the perceptual similarity between subjects' speech and the speech of a model, as compared against subjects' baseline words (e.g., Goldinger, 1998; Goldinger & Azuma, 2004; Miller et al., 2010; Namy et al., 2002; Pardo, 2006; Pardo, et al. 2010; Pardo, et al. 2012, Sanchez, 2011). There are some good arguments for why alignment has more often been evaluated using rater matches than acoustic analyses. For example, although acoustical analyses can

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

reveal clear changes in phonetically-relevant dimensions of the speech signal, pinpointing exactly which dimensions of speech are changing during alignment is problematic. It is not clear, for example, if changes in a single phonetic dimension are indicative of alignment or if alignment occurs due to changes in a combination of factors. Additionally, to the degree that alignment serves some sociolinguistic function, it makes some sense to evaluate its prominence using *perceptual* judgments (see also, Goldinger, 1998; Pardo et al., 2010). Thus, for all of the studies using AXB rating measures, and many of the those using acoustic measures, alignment to a model or conversant is often empirically defined as occurring when a subject's utterances produced during or after an alignment task are more like those of a model than are the subject's utterances produced before the task. Of course, investigating the potential differences between a baseline and post-task speech sample is an appropriate test to determine if a shadower's speech has, in fact, changed. However, by using comparison tokens derived from a subject's own baseline utterances, it is difficult to know whether alignment tasks make subjects sound more like the *specific* model whose speech they perceive or if there is simply a *general* change in the way in which they are producing speech.

The current study examines the specificity of model alignment as evaluated with perceptual rating (AXB) measures. This perceptual rating-based design takes a cue from the studies using acoustic measures to compare alignment between partners vs. between pseudo-partners (e.g., Gregory et al., 1997; Gregory et al., 2001; Levitan & Hirschberg, 2011). For our rating measures, raters will be tasked with determining whether a shadowed token sounds more like the tokens of the model that was shadowed or a pseudo-model who was not shadowed. If shadowed tokens are rated as sounding more like the model than pseudo-model, then this demonstration would provide complementary evidence to the more typical baseline-to-shadowing comparison showing that shadowers do change their speech to align with that of the specific talker their shadow. In this sense, the evidence would have implications for theories that incorporate alignment into explanations of speech perception, lexical access, and social interactions between interlocutors as discussed above.

Three experiments used AXB subject matching judgments to examine whether the change in a subject's speech is due to alignment in the direction of a specific shadowed model or if it reflects a more general change in a subject's speaking style. Experiment 1 used the classic baseline-shadowed speech comparison to establish that subjects' speech did change between baseline and shadowed utterances to be more like the model's utterance. Experiment 2 then tested whether a subject's shadowed utterances sounded more similar to the model they had shadowed relative to another, unshadowed model. Finally, Experiment 3 examined whether a subject's shadowed utterances sounded more similar to the model they had shadowed or to another subject who had shadowed a different model. Throughout these experiments, alignment was measured using perceptual measures.

Experiment 1

Experiment 1 was performed to establish that subjects would sound more like a model while shadowing, than before shadowing. For this purpose, the typical AXB-matching task was implemented in Experiment 1. Raters were asked to judge whether an utterance made prior to the shadowing task (baseline) or a shadowed utterance was more similar to the model (e.g., Goldinger, 1998). If it is confirmed that our shadowers do change their speech to be more like that of a model, we would then be poised to test (in Experiments 2 and 3) whether this change is in the direction of the *specific* model they shadowed.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Method

Participants

Two female and two male graduate students served as models in the experiment and produced the original word list to be shadowed (e.g., Shockley et al., 2004). Eight female and eight male undergraduates served as shadowing subjects in the experiment. The female shadowers shadowed the female models (with four shadowers assigned to each model), and the male shadowers shadowed the male models. Sixteen undergraduates (7 male) served as raters in the AXB matching task. All models, subjects, and raters were native speakers of American English with normal hearing and normal or corrected vision and ranged in age from 18 to 24 years. All four models were in their mid-twenties and were native English speakers. While the models were living in Southern California, only one (female) was from the area. One of the other models (female) grew up in Northern California, and the remaining two grew up in North Atlantic states. None had conspicuous accents, based on the experimenters' (untrained) impressions. The models were paid for their participation, while subjects and raters participated in order to partially fulfill a course requirement.

Materials and apparatus

A list composed of 74 bisyllable, low frequency English words were used as stimuli. The word list was derived from Shockley et al. (2004) and used by Miller et al. (2010). These words had frequencies of less than 75 occurrences per million (Kučera & Francis, 1967) and all began with the voiceless stop consonants /p/, /t/, or /k/.

All stimuli were presented to participants using PsyScope software. A SONY DSR-11 camcorder was used to videotape the models. Text (baseline) words were presented on a 20-in. video monitor positioned 3 ft in front of the participants. Auditory stimuli were presented through SONY MDR-V6 headphones. The models and subjects responded verbally into a Shure; Beta 58a microphone and were audio recorded at a 44000 Hz, 16 bit rate using software.

Procedure

The experiment took place in three phases. For all three phases, participants sat in a sound-attenuating chamber.

Phase 1—In Phase 1, two female models and two male models were videotaped each producing the list of 74 bisyllable words. The word list was presented to the models as text on a video monitor. The words were randomly presented at an interval of 1 word per second. Models were asked to speak the words "quickly, but clearly" into the microphone. These utterances were filmed using a high-quality camcorder and the audio component of these recordings were then edited on a computer to produce individual word presentations.

Phase 2—Phase 2 of the experiment consisted of having the 16 subjects (8 female) participate in three tasks in the following order: (1) baseline word production (text reading), (2) audio shadowing of 74 words, and (3) a second block of audio shadowing the same 74 words.

For the baseline word task, the subjects were audio recorded producing the original word list, which they read from a video monitor. The words were presented individually at one second intervals. Subjects were asked to say the words they saw "quickly, but clearly" into the microphone.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

For shadowing task blocks, the subjects were audio-recorded shadowing a model's 74 audio words, which they heard over headphones. Subjects were gender-matched to models (four subjects per model) and were required to say each word they heard "quickly, but clearly" into the microphone (e.g., Shockley et al., 2004; Miller et al., 2010). Subjects were never asked to imitate the model or to "repeat" the words. All shadowed utterances were recorded directly onto the computer. Utterances from the second shadowing block were later edited to create 74 audio-shadowed tokens for comparison purposes in Phase 3. Tokens from only the second shadowing block were used in the present studies. Previous research suggests that the number of instances a word is shadowed increases the strength of alignment, possibly because of the accumulated influence of a talker's specific information in a word's lexical trace (e.g., Goldinger, 1998). Because alignment findings are often subtle, it was felt that the use of second block shadows would allow for the most sensitive test of this phenomenon (e.g., Goldinger, 1998).

Phase 3—For each AXB triad, the sixteen raters judged whether a model's utterance was more similar to a subject's shadowed utterance or to a subject's baseline (pre-task) utterance. Each triad contained presentations of the same word (e.g., cabbage) produced once by the subject reading text words from a monitor (i.e., baseline); once by the model who was shadowed by the main subject, and once by that subject shadowing the model. Throughout the experiment, the model's utterances always appeared as the middle, X token. The subject's baseline utterances appeared either in the A (first) or B (third) position and the subject's shadowed utterances appeared in the remaining A or B position. The A and B positions were counterbalanced.

During the task, a rater heard a total of six different voices (two models, two subjects who shadowed one model, two subjects who shadowed the other model). The 74-word list was split into two sets of 37 words (e.g., Set 1 contained cable, Set 2 contained camel) again to reduce the chances of rater fatigue. Each script contained a total of 74 words shadowed after Model A and 74 words shadowed after Model B, but each shadower was only heard producing either Set 1 or Set 2 (e.g., Shadower 1 who shadowed Model A produced Set 1, Shadower 2 who shadowed Model A produced Set 2, Shadower 3 who shadowed Model B produced Set 1, Shadower 2 who shadowed Model B produced Set 2). Each rater judged a total of 296 separate triads composed of two sets of 74 words (one set per model-shadower pairing) with two different orderings of the triads (once with the baseline word in A position, once in B position). Raters only made judgments either for female or male model/subject combinations. Two raters made judgments for each script, meaning that any given subject's speech was rated by a total of four raters. Although previous research has had raters make within-subject AXB judgments (i.e., they only heard one subject and the model that was shadowed; Miller et al., 2010), the presentation procedure was used to stay consistent with the subsequent two experiments in this study (see below).

These triads were randomly presented to raters auditorily over SONY MDR-V6 headphones. Raters were asked to choose which of the words, the first or third, sounded more similar in pronunciation to the second. Pronunciation instructions were employed to reduce judgments based on extraneous information (e.g., background noise; Pardo, 2007). 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.

Results and Discussion

Means were calculated for subjects as determined by the number of shadowed utterances chosen by raters as sounding more like those of the model. The mean percentage of

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

shadowed tokens considered to be pronounced more like the models' tokens than the baseline tokens, was 58.9%. A one-sample t-test was used to compare the mean AXB rating against chance (50%)[1]. This test revealed that the shadowed tokens were judged to be pronounced more like the models' tokens than were the baseline tokens $t(15) = 3.89$, $p = .001$. An item analysis was also conducted to determine whether these effects were based on the influence of a few tokens. This analysis revealed that again, shadowed tokens were chosen more often as matches than baseline tokens, $t(71) = 11.076$, $p < .001$, suggesting that these alignment results were not simply due to a few of the word tokens.

An ANOVA was conducted to determine whether alignment occurred more to a given gender or model. Neither of these comparisons revealed a significant effect at the $p < .05$ level.

To ensure our results were due to a shift in speech in the direction of the model shadowed and not simply an artifact due to specific attributes associated with the raters or shadowers, or certain words we used, we implemented a linear mixed effects model using R (R Development CoreTeam, 2009) and the R packages lme4 (Bates & Maechler, 2009) and languageR (Baayen, 2009; cf. Baayen, 2008). We used Rater, Shadower, and Word as random effects (see Baayen, Davidson, & Bates, 2008). We used Model as our fixed effect. Our predictor variable was our alignment score, which was whether a rater judged a shadower as similar to the model or a competitor. All levels of Model were found to predict the alignment score in a positive direction, though only three out of four were able to do so significantly ($p < .05$). Thus, for all but one of our models, shadowed utterances were rated as significantly more similar to the models' speech than were baseline utterances, with the non-significant Model trending in the expected direction.

The results of this standard AXB-matching task suggest that the subjects' shadowed speech did sound more like the shadowed model than they did before hearing that model (baseline speech).

Experiment 2

Having established that shadowers' shadowed tokens sound more like the model than do their baseline tokens, Experiment 2 tested if shadowers do truly align to the *specific* model they shadowed. This was accomplished by testing whether shadowers sound more like the model they shadowed or a different model whom they did not shadow. In this experiment, raters were asked to judge the relative similarity of a shadower's utterances to the utterances of two models. If shadowers truly align to the model's speech they shadowed, then raters should judge the shadower as sounding more like the shadowed model than the non-shadowed model.

Method

Participants

The graduate student models and undergraduate shadowers were the same as in Experiment 1. Sixteen new undergraduates (11 male) served as raters for the AXB matching task. All raters were native speakers of American English with normal hearing and normal or corrected vision. None had participated in Experiment 1. All raters participated in order to partially fulfill a course requirement.

[1]The use of t-tests on percentage correct scores is possible for our data because it falls between the 30-70% range as suggested by Jaeger (2008) and Dixon (2008).

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Materials and apparatus

All materials and apparati were the same as in Experiment 1.

Procedure

The model recording and shadowing phases of the experiment were the same as in Experiment 1 (above).

For the rating phase, sixteen raters judged whether a subject's shadowed utterance was more similar to the utterance of the model they had shadowed (shadowed model) than it was to an utterance of the other model (comparison model) of the same gender. Stimuli were presented to raters as triads, so that a subject's shadowed utterances appeared as the middle, X token. The shadowed model's utterances appeared either in the A (first) or B (third) position and the comparison model's utterances appeared in the remaining A or B position. Position was counterbalanced across an experimental session.

In order to reduce the chances of rater judgments being based on mere similarity between a given subject and a given model's natural voices, each rater heard a total of six different voices during the rating phase: two models (e.g., Model A, Model B), two subjects who shadowed Model A, and two subjects who shadowed Model B.

The 74-word list was split into two sets of 37 words (e.g., Set 1 contained the word 'cabbage', Set 2 contained 'cable') to keep the number of trials reasonable (i.e., in order to keep the raters from becoming fatigued by the task). Each script contained a total of 74 words shadowed after Model A and 74 words shadowed after Model B, but each shadower was only heard producing either Set 1 or Set 2 (e.g., Shadower 1 who shadowed Model A produced Set 1, Shadower 2 who shadowed Model A produced Set 2, Shadower 3 who shadowed Model B produced Set 1, Shadower 4 who shadowed Model B produced Set 2). Because each script represented only half the words produced by each shadower, a total of four raters judged a shadower's words (two for Set 1 words, two for Set 2 words. Each rater judged a total of 296 separate triads composed of two sets of 74 words (one set per model-shadower pairing) with two different orderings of the triads (once with the model word in A position, once in B position). Raters only made judgments either for female or male model/subject combinations.

As in Experiment 1, raters were asked to choose which of the words, the first or third, sounded more similar in pronunciation to the second. 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.

Results and Discussion

Scoring of the AXB rating task revealed that the mean percentage of subjects' shadowed tokens considered to be pronounced more like the shadowed model's tokens (than like the comparison model's tokens) was 64%. A one-sample t-test comparing the mean AXB rating against chance (50%) revealed that raters judged the subjects' shadowed utterances to be pronounced more like the model they shadowed, $t(15) = 6.04$, $p < .0001$. An item analysis also revealed that the shadowed model's tokens were chosen more often as matches than the non-shadowed model, $t(73) = 17.233$, $p<.001$, suggesting that these alignment results were not simply due to a few of the word tokens.

An ANOVA was also conducted to determine whether alignment differed depending on the gender of the shadowers (and models) or on the specific model shadowed. Neither of these comparisons revealed a significant difference at the $p < .05$ level.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

As for Experiment 1, a linear mixed effects model was conducted using Rater, Shadower, and Word as random effects (see Baayen et al., 2008). We used Model as our fixed effect. Our predictor variable was our alignment score, which was whether a rater judged a shadower as similar to the model or a competitor. All levels of Model were found to be statistically significant in the positive direction. Thus, for all of the models, shadowers were rated as more similar to the model than the competitor.

Subjects were judged as sounding more similar to the model whom they shadowed than another model they did not shadow. These results suggest that alignment is in the direction of the perceived model and not simply a general change in the way a subject produces speech during a shadowing task. The fact that subjects in Experiment 1 were shown to change their speech between baseline and shadowing suggests that the results of Experiment 2 were not a consequence of subjects somehow being randomly assigned to models. Thus taken together, the findings of Experiments 1 and 2 suggest that, as evaluated by perceptual ratings, shadowers do change their speech toward the specific model they are shadowing.

One additional test was conducted using shadower model and non-model comparisons. In this final test, the non-model utterances were comprised of shadowed tokens spoken by subjects who had shadowed another model. Thus, raters were tasked with judging whether a shadowed token sounded more like the model's token from which it was shadowed, or the same word spoken by another shadower.

Besides providing a conceptual replication of Experiment 2, this last experiment was designed to test whether alignment could overcome whatever commonalities occur when words are produced with the same instructions. In the traditional rater matching procedure (e.g., Experiment 1), both the model and subject baseline utterances are produced by reading text words from a screen. Subjects' shadowed utterances, on the other hand, are produced by having subjects listen to a word and then say the word out loud quickly and clearly (i.e., shadow). Despite the commonality between baseline and model word production, alignment is strong enough so that the shadowed word sounds more like the model's (read) word. The question arises of whether alignment to a model can overcome the inherent similarity between *two* shadowed utterances. In other words, will a shadowed utterance sound more like the model's read utterance on which it was based than it would another subject's *shadowed* utterance? It could very well be that two shadowed utterances will naturally sound more like each other than a shadowed and read utterance, simply because of the task commonality. However, alignment to a model could be strong enough to overcome the task commonality. This question was examined in Experiment 3.

Experiment 3

Experiment 3 tested the possibility that two utterances produced by two different shadowers might be judged as more similar than either would to the models' utterances that were shadowed. Raters in Experiment 3 were asked to judge the relative similarity of a subject's shadowed utterances to the model they shadowed versus the shadowed utterances of another subject who shadowed a different model. If the act of shadowing speech produces utterances that sound overwhelmingly similar, then raters should judge the two shadowed utterances as more similar to each other, than to the model's read utterance. If, on the other hand, the alignment produced during shadowing is strong enough to offset any inherent similarity between utterances both produced during shadowing, then raters should judge the shadowed utterance as sounding more like the model's utterance on which it was based, than another shadowed utterance.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Method

Participants

The graduate student models and undergraduate shadowers were the same as those used in Experiments 1 and 2. Thirty-two new undergraduates (25 female) served as raters in an AXB matching task. All raters were native speakers of American English with normal hearing and normal or corrected vision. None had participated in Experiment 1 or 2 and all participated in order to partially fulfill a course requirement.

Materials and apparatus

All materials and apparati were the same as in Experiment 1.

Procedure

The experiment used the shadow and model word stimuli borrowed from Phases 1 and 2 of Experiments 1 and 2 (see above). The thirty-two naïve raters judged whether a subject's shadowed utterance was more similar to the shadowed model's utterance or to an utterance produced by a subject who shadowed the other model (of the same gender).

Thus, each triad contained presentations of the same word (e.g., cable) produced once by the subject shadowing the model (main subject), once by the model who was shadowed by the main subject, and once by a subject who shadowed the other model (comparison subject). Throughout the experiment, the main subject's shadowed utterances appeared as the middle, X token. The model's utterances appeared either in the A (first) or B (third) position and the comparison subject's shadowed utterances appeared in the remaining A or B position. The A and B positions were counterbalanced.

During the task a rater heard a total of six different voices (two models, two main subjects, and two comparison subjects). This procedure was chosen over simply presenting one main subject in order to reduce the possibility of judgments based on general similarities between the model's voice and the main subject's voice. Additionally, each main subject and comparison subject pairing was reversed in a separate script presented to different raters. Two raters made judgments for each script, meaning that any given subject's speech was rated by a total of four raters. Raters again only made judgments either for female or male model/subject combinations.

Each rater judged a total of 296 separate triads composed of two sets of 74 words (one set per main subject) with two different orderings of the triads (once with the model word in A position, once in B position). These triads were randomly presented to raters auditorily over headphones. As in Experiment 1, raters were asked to choose which of the words, the first or third, sounded more similar in pronunciation to the second. 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.

Results and Discussion

Means were calculated for subjects as determined by the number of model utterances chosen as sounding more like those of the main subject. The mean percentage of main subjects' shadowed tokens considered to be pronounced more like the models' tokens than the comparison subjects' shadowed tokens, was 63.1%. Again, a one-sample t-test was used to compare the mean AXB rating against chance (50%). This test revealed that the main subjects' shadowed tokens were judged to be pronounced more like the models' tokens than were the comparison subjects' tokens $t(15) = 5.27$, $p < .0001$. An item analysis also revealed

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

that the models' tokens were chosen more often than the comparison subjects' shadowed tokens $t(73) = 21.6096$, $p<.001$, suggesting that these alignment results were not simply due to a few of the word tokens. Additional tests were conducted to determine whether alignment differed depending on the gender of the shadowers (and models) or on the specific model shadowed. Neither of these comparisons revealed a significant difference at the $p < .05$ level.

Again, we conducted a linear mixed effect analysis using Rater, Shadower, and Word as random effects (see Baayen et al., 2008). We used Model as our fixed effect. Our predictor variable was our alignment score, which was whether a rater judged a shadower as similar to the model or a competitor. All levels of Model were found to be statistically significant in the positive direction. Thus, for all of the models, shadowers were rated as more similar to the model than the competitor.

Thus, raters judged utterances produced during shadowing as sounding more similar to the model's read utterance on which it was based, than on another utterance produced during shadowing. These results suggest that shadow-based alignment, as evaluated perceptually, is strong enough to offset the commonalities inherent in utterances produced during the same shadowing task.

General Discussion

The present study examined if in perceptually-evaluated alignment, shadowers' speech changes in the direction of a *specific* talker. Alignment has been referred to as the subtle tendency of interlocutors to sound more similar to each other and is thought to involve a change in speech in the direction of a specific talker. As stated, individuals have been thought to align in both interactive (Pardo, 2006) and shadowing (e.g., Goldinger, 1998; Miller et al., 2010; Sanchez et al., 2010) contexts, as well as after simply listening to talkers (Goldinger & Azuma, 2004, Nielsen, 2011). Shadowers have even been shown to align to visual speech information (e.g., Gentilucci & Bernardis, 2007; Miller et al., 2010; Sanchez et al., 2010, Sanchez, 2011). While some of the acoustically-evaluated alignment studies have shown that talkers align to the specific interlocutor to whom they're talking (Gregory, et al. 1997; Gregory, et al. 2001; Levitan & Hirschberg, 2011), the perceptually-evaluated studies had not established if the shadower's speech aligned towards a specific talker. All of the perceptually-evaluated speech alignment demonstrations have used baseline utterances as comparison stimuli, which, while an appropriate method to determine that a shadower's speech has changed, could not establish model-specificity. The current results provide evidence that in perceptually-evaluated shadowing tasks, alignment does make a shadower sound specifically like the shadowed model, as opposed to another unshadowed model (Experiment 2 and 3) and relative to a pre-shadowing, baseline utterance (Experiment 1). Further, this alignment is strong enough to override any shadowing-task specific similarities in produced speech (Experiment 3). In this sense, the current results are supportive that, at least for shadowing, alignment is to specific talkers rather than simply to task.

These results should be reassuring to researchers who have incorporated perceptually-evaluated speech alignment results into their theories. As mentioned, speech alignment phenomena are supportive of a behavioral and neurophysiological coupling of perception and action (e.g., Fadiga, Fogassi, Povesi, & Rizzolatti, 1995; Fadiga, Craighero, Buccino, & Rizzolatti, 2002; Fowler, 2004; Fowler et al., 2003; Hecht, Vogt, & Prinz; 2001; Sancier & Fowler, 1997; Shockley et al. 2004; but see Lotto, Hickok, & Holt, 2009; Scott, McGettigan, & Eisner, 2009). Speech alignment phenomena are also consistent with episodic models of speech perception which proffer the encoding of highly-detailed traces of speech events that later influence productions (e.g., Goldinger, 1998). Alignment has also been explained by

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

referring to its importance in facilitating social interaction between interlocutors (Babel, 2010; Giles & Coupland, 1991; Gregory & Webster, 1996; Pardo, et al., 2010, Dias & Rosenblum, 2011). Inherent in all of these theories is the assumption that speech alignment is occurring to characteristics of a *specific* talker's speech (e.g., idiolect, accent) and is not simply representing a general change occurring when an individual shadows speech. The current results are consistent with this assumption, and therefore supportive of these theories.

The present study provides evidence that in perceptually-evaluated shadowing experiments, talker-specific alignment is occurring. Still, additional questions remain about other alignment paradigms that use perceptual measures. For example, it is unclear whether talker-specific alignment is occurring in perceptually-rated interactive alignment experiments (e.g., Pardo, 2006; 2010). It is also unclear whether alignment occurs to specific talkers when subjects do not utter words until *days after* they hear talkers say those words, as in the Goldinger and Azuma (2004) study. In both types of studies, perceptually-judged comparison stimuli are composed of subjects' own baseline utterances. Future research examining interactive and delayed reading tasks could easily examine if participants align to a *specific* talker by adding an AXB test involving comparison tokens from another model's speech (as in Experiment 2). A similar approach could be used to determine if alignment based on visible speech (Dias & Rosenblum, 2011; Miller et al., 2010; Sanchez et al., 2010, Sanchez, 2011) is to the specific model perceived.

Regardless, in showing that shadowers do truly change their speech to sound more like the shadowed versus an unshadowed model, the current results are suggestive that perceivers do align to the specific talkers they perceive.

Acknowledgments

This research was supported by NIDCD Grant 1R01DC008957-01.

References

Babel M. Dialect divergence and convergence in New Zealand English. Language in Society. 2010; 39:437–456.

Baayen, RH. Analyzing Linguistic Data: A Practical Introduction to Statistics Using R. Cambridge: Cambridge University Press; 2008.

Baayen RH. Language R: Data sets and functions with "Analyzing Linguistic Data: A practical introduction to statistics". R package version 0.955. 2009

Baayen RH, Davidson DJ, Bates DM. Mixed-effects modelling with crossed random effects for subjects and items. Journal of Memory and Language. 2008; 59:390–412.

Dias JW, Rosenblum LD. Visual influences on interactive speech alignment. Perception. 2011; 40:1457–1466. [PubMed: 22474764]

Dixon P. Models of accuracy in repeated-measures design. Journal of Memory and Language. 2008; 59:447–456.

Fadiga L, Craighero L, Buccino G, Rizzolatti G. Speech listening specifically modulates the excitability of tongue muscles: A TMS study. European Journal of Neuroscience. 2002; 15:399–402. [PubMed: 11849307]

Fadiga L, Fogassi L, Povesi G, Rizzolatti G. Motor facilitation during action observation: A magnetic stimulation study. Journal of Neurophysiology. 1995; 73:2608–2611. [PubMed: 7666169]

Fowler, CA. Speech as a supermodal or amodal phenomenon. In: Calvert, GA.; Spence, C.; Stein, BE., editors. The Handbook of Multisensory Processing. Cambridge, MA: MIT Press; 2004. p. 189-201.

Fowler CA, Brown JM, Sabadini L, Weihing J. Rapid access to speech gestures in perception: Evidence from choice and simple response time tasks. Journal of Memory & Language. 2003; 49(3):396–413. [PubMed: 20622982]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Gentilucci M, Bernardis P. Imitation during phoneme production. Neuropsychologia. 2007; 45:608–615. [PubMed: 16698051]

Giles, H.; Coupland, J.; Coupland, N. Accommodation theory: Communication, context, and consequences. In: Giles, H.; Coupland, J.; Coupland, N., editors. Contexts of accommodation: Developments in applied sociolinguistics. Cambridge: Cambridge University Press; 1991. p. 1-68.

Goldinger SD. Echoes of echoes? An episodic theory of lexical access. Psychological Review. 1998; 105:251–279. [PubMed: 9577239]

Goldinger SD, Azuma T. Episodic memory reflected in printed word naming. Psychonomic Bulletin. 2004; 11(4):716–722.

Gregory SW. Analysis of fundamental frequency reveals covariation in interview partners' speech. Journal of Nonverbal Behavior. 1990; 14:237–251.

Gregory SW, Dagan K, Webster SW. Evaluating the relation of vocal accommodation in conversation partners' fundamental frequencies to perceptions of communication quality. Journal of Nonverbal Behavior. 1997; 27(1):23–43.

Gregory SW, Green BE, Carrothers RM, Dagan KA, Webster SW. Verifying the primacy of voice fundamental frequency in social status accommodation. Language & Communication. 2001; 21:37–60.

Gregory SW, Webster S. A nonverbal signal in voices of interview partners effectively predicts communication accommodation and social status perceptions. Journal of Personality and Social Psychology. 1996; 70:1231–1240. [PubMed: 8667163]

Harrington J, Palethorpe S, Watson C. Does the Queen speak the Queen's English? Nature. 2000; 408:927–928. [PubMed: 11140668]

Hecht H, Vogt S, Prinz W. Motor learning enhances perceptual judgment: A case for perception-action transfer. Psychological Research. 2001; 63:3–14. [PubMed: 11505611]

Honorof DN, Weihing J, Fowler CA. Articulatory events are imitated under rapid shadowing. Journal of Phonetics. 2011; 39(1):18–38. [PubMed: 23418398]

Howell P, Kadi-Hanifi K. Comparison of prosodic properties between read and spontaneous speech material. Speech Communication. 1991; 10:163–169.

Jaeger TF. Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. Journal of Memory and Language. 2008; 59:434–446. [PubMed: 19884961]

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

Laan GPM. The contribution of intonation, segmental durations, and spectral features to the perception of a spontaneous and a read speaking style. Speech Communication. 1997; 22:43–65.

Levitan R, Hirschberg J. Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. Proceedings of Interspeech 2011. 2011 Florence, Italy, August 2011.

Lotto AJ, Hickok GS, Holt LL. Reflections on mirror neurons and speech perception. Trends in Cognitive Sciences. 2009; 13(3):110–114. [PubMed: 19223222]

Mitterer H, Ernestus M. The link between speech perception and production is phonological and abstract: Evidence from the shadowing task. Cognition. 2008; 109(1):168–173. [PubMed: 18805522]

Miller RM, Sanchez K, Rosenblum LD. Alignment to visual speech. Attention, Perception & Psychophysics. 2010; 72(6):1614–1625.

Namy LL, Nygaard LC, Sauerteig D. Gender differences in vocal accommodation: The role of perception. Journal of Language and Social Psychology. 2002; 21(4):422–432.

Natale M. Convergence of mean vocal intensity in dyadic communication as a function of social desirability. Journal of Personality and Social Psychology. 1975; 32:790–804.

Nielsen K. Specificity and abstractness of VOT imitation. Journal of Phonetics. 2011; 39:132–142.

Pardo JS. On phonetic convergence during conversational interaction. Journal of the Acoustical Society of America. 2006; 119:2382–2393. [PubMed: 16642851]

Pardo JS, Gibbons R, Suppes A, Krauss RM. Phonetic convergence in college roommates. Journal of Phonetics. 2010; 40(1):190–197.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Pardo JS, Jay IS, Krauss RM. Conversational role influences speech imitation. Attention, Perception, & Psychophysics. 2010; 72(8):2254–2264.

R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2009. http://www.R-project.org

Sanchez, K. Doctoral Dissertation. University of California; Riverside: 2011. Do you hear what I see? The voice and face of a talker Alignment Talkers or Task 25 similarly influence the speech of multiple listeners.

Sanchez K, Miller RM, Rosenblum LD. Visual influences on alignment to voice onset time. Journal of Speech, Language, and Hearing. 2010; 53:262–272.

Sancier ML, Fowler CA. Gestural drift in a bilingual speaker of Brazilian Portuguese and English. Journal of Phonetics. 1997; 25:421–436.

Scott SK, McGettigan C, Eisner F. A little more conversation, a little less action: candidate roles for the motor cortex in speech perception. Nature Reviews Neuroscience. 2009; 10:295–302.

Shockley K, Sabadini L, Fowler CA. Imitation in shadowing words. Perception & Psychophysics. 2004; 66(3):422–429. [PubMed: 15283067]