Perception, 2011, volume 40, pages 1457 – 1466

doi:10.1068/p7071

Visual influences on interactive speech alignment

James W Dias, Lawrence D Rosenblum¶
Department of Psychology, University of California, Riverside, Riverside, CA 92521, USA;
e-mail: lawrence.rosenblum@ucr.edu

Received 12 July 2011, in revised form 17 November 2011

Abstract. Speech alignment describes the unconscious tendency to produce speech that shares characteristics with perceived speech (eg Goldinger, 1998 *Psychological Review* **105** 251 – 279). In the present study we evaluated whether seeing a talker enhances alignment over just hearing a talker. Pairs of participants performed an interactive search task which required them to repeatedly utter a series of keywords. Half of the pairs performed the task while hearing each other, while the other half could see and hear each other. Alignment was assessed by naive judges rating the similarity of interlocutors' keywords recorded before, during, and after the interactive task. Results showed that interlocutors aligned more when able to see one another suggesting that visual information enhances speech alignment.

1 Introduction

The unconscious tendency for humans to spontaneously imitate the behaviors of those with whom they interact has been found for many social behaviors (eg Giles et al 1991). Interacting participants will subtly converge physical behaviors, such as posture (Condon and Ogston 1967), gesture (Mauer and Tindall 1983), head nodding, and facial expression (Hale and Burgoon 1984). Participants will also align to linguistic characteristics, such as utterance length (Matarazzo et al 1968), response latency (Cappella and Planalp 1981), pausing frequency (Jaffe and Feldstein 1970), information density (Aronsson et al 1987), and self-disclosure (Ehrlich and Graeven 1971). In adult conversational interactions, individuals' speech will also subtly align to the acoustical characteristics of the individual with whom they are engaged. These aligning acoustic characteristics include speech rate (Street 1984), vocal intensity (Natale 1975), as well as other less-measurable characteristics (Pardo 2006). Chartrand and Bargh (1999) have described these general unconscious tendencies as a "chameleon effect"; an appropriate description, as it exemplifies the apparent automatic and passive nature of such alignment to the behaviors of others.

In one example of speech alignment, Pardo (2006) had dyads work together to solve a mapping task with one participant giving directions to the other. The participants were separated by a wall, allowing for auditory communication only. In order to solve the task, participants needed to utter a number of keywords, multiple times during the interaction. Subjects also uttered these keywords before and after the interaction (as they read from a list) and these utterances, along with the interactive task, were audio recorded.

To evaluate the degree to which the interlocutors of each dyad aligned to one-another's speech, an AXB rating paradigm was employed (eg Goldinger 1998; Miller et al 2010; Shockley et al 2004). In such a paradigm, one interlocutor's (I_1) pre-interaction, mid-interaction, and/or post-interaction utterances are compared to the mid-interaction utterances of the other interlocutor (I_2) in the dyad. Naive raters then judge which of I_1's utterances sound more like those of I_2. If I_1's post-interaction tokens are rated as more similar to I_2's interaction tokens, or if I_1's interaction tokens are rated as more similar to I_2's interaction tokens, it is considered evidence of speech alignment.

¶ Author to whom all correspondence should be addressed.

[The AXB rating evaluation of alignment has been widely used over acoustical measures of alignment due to the complex task of measuring to which of the many acoustical characteristics participants might align (Goldinger 1998). The AXB rating method also serves to establish that alignment occurs in a perceptually-relevant manner (eg Goldinger 1998).] Pardo (2006) observed that interlocutors did align to each other's speech, and this effect was modulated by participant sex and social role within the interaction.

Alignment does not only occur in a social context. Participants asked to shadow the utterances of a prerecorded model will also align to the speech of that model (Goldinger 1998; Goldinger and Azuma 2003; Namy et al 2002; Shockley et al 2004). The shadowing paradigm consists of a participant quickly saying out loud the words presented to them by a prerecorded model. The design is such that participants are never asked to 'imitate' or even 'repeat' what they hear. This acts to avoid conscious imitation of the perceived signal. The participants' shadowed or post-shadowing utterances are then rated against baseline utterances taken prior to the shadowing task within an AXB rating paradigm. Using a shadowing paradigm, alignment has been found for shadowed heard speech in a number of contexts (eg Goldinger 1998; Goldinger and Azuma 2003; Namy et al 2002; Shockley et al 2004).

Though auditory information alone is sufficient to induce alignment (Giles et al 1991; Natale 1975), visual speech (lipreading) information has also been found effective in this regard. Miller et al (2010; and see also Gentilucci and Bernardis 2007) found that participants shadowing words spoken by a silent video-recorded model aligned to that model (as evaluated by AXB judgments of the produced auditory speech). Subsequently, participants' shadowed acoustic tokens were rated as more similar to the model's silent video-recorded tokens than the participants' baseline acoustic tokens in a crossmodal matching task (Miller et al 2010). Sanchez et al (2010) examined the modulating effect that visible articulatory rates have on alignment to voice onset time (VOT) and discovered that alignment was systematically influenced by the visual information, with faster rates resulting in greater rated alignment. Evidence that visual speech can modulate alignment has relevance to theories of both the nature of episodic lexical memory (eg Goldinger 1998), as well as the form of information thought to be influential in priming speech-production responses (eg Fowler 2004).

Findings that visual speech can induce alignment are also consistent with research showing that visible articulatory information generally plays an important role in speech perception (eg Rosenblum 2008). Visual information modulates audio speech perception (McGurk and MacDonald 1976) and provides an advantage of audiovisual information over audio alone (Sumby and Pollack 1954). Naive raters associate acoustical speech information for a talker with the talker-specific visible articulations of that talker in a crossmodal talker-matching task (Lachs and Pisoni 2004a, 2004b). Participants are better able to identify speech in noise from a talker they were previously trained to lip-read than from a novel talker (Rosenblum et al 2007). These last two findings suggest that not only can visual speech provide information about what is being said, it can also provide information for a talker's particular style of speaking—information likely important for speech alignment to occur (eg Miller et al 2010).

To summarize, visual information has been found to induce speech alignment in a shadowing context (Miller et al 2010; Sanchez et al 2010). In addition, participants interacting in a live context with visual access to one another have been found to align to one another's physical behaviors, such as posture (Condon and Ogston 1967), gesture (Mauer and Tindall 1983), head nodding, and facial affect (Hale and Burgoon 1984). However, the modulating effects of visual information in a live interactive context on speech alignment have not yet been explored. Taking into account the importance of visual information to the speech process (eg Rosenblum 2008) as well as the salience of general visual information in inducing non-speech related behavioral alignment (eg Giles et al 1991), it could very well be that visual information for an interlocutor can enhance speech alignment. In other words, the speech of interlocutors interacting audiovisually may align more than the speech of those interacting, while only being able to hear one another (as in the study of Pardo 2006).

This hypothesis is tested in the current investigation by comparing two groups of dyads, each completing an interactive task. One group completed the task with access to only auditory information from the interlocutor, while the other group completed the task with access to both auditory and visual information from the interlocutor. Speech tokens were isolated from interlocutors' repeated carrier phrases recorded prior to and following the interactive task. The final, end-clause speech tokens of each interlocutor were also isolated out of recordings taken during the interaction task itself. Alignment was then assessed by naive raters in an AXB paradigm (for the reasons described above). Raters were tasked with reporting whether the posttask or interaction utterances of interlocutor 1 (I_1) were more similar to the posttest or interaction utterances of interlocutor 2 (I_2) above and beyond the pretest utterances of I_1. The general structure of the experiment followed the methodologies used by Pardo (2006). If visual information can enhance speech alignment during an interactive task, then interlocutors with audiovisual access to one another should align more than those with only auditory access to one another.

2 Methods

2.1 *Phase I: Interaction elicitation*

2.1.1 *Participants.* Twenty female undergraduate students from the University of California, Riverside Human Subjects Pool participated in the interaction procedure. Females were used as interlocutors due to their somewhat greater likelihood of alignment reported in the literature (Namy et al 2002). These twenty participants were organized into ten pairs. Pairs were selected to ensure that no two interlocutors had any prior social interactions with one another. Five of the pairs participated in the audiovisual condition while the other five participated in the audio-only condition.

2.1.2 *Interaction task.* The interaction methodology was designed to accommodate three primary requirements. First, the apparatus provided the interlocutors with either a full, unobstructed view of their partners (audiovisual group), or a completely occluded view while not obstructing their audible speech signals (audio-only group). Second, the apparatus ensured that interlocutors attended to their partners during the full extent of the interaction task. Third, the task itself was designed to maximize the number of utterances of target speech tokens spoken between interlocutors.

Materials. For the task, participants were seated between two 54 inch high × 72 inch long black curtains, making a 24 inch wide corridor (figure 1). The purpose of the curtain system was to ensure that the audiovisual group interlocutors visually attended to their partner who was seated, facing them, 55 inches away on the other side of the corridor. Those participants interacting audiovisually had full view of their partner across from them, including her head, torso, and legs. Informal observations revealed that participants in the audiovisual condition did continually look directly at each other during a majority of their interactions.

Those participants interacting in the audio-alone condition were separated by a 54 inch high × 24 inch wide section of Guilford of Maine Fabric Anchorage Series onyx speaker grill cloth (Class 1 or A, per ASTM E84; Acoustically Transparent) positioned halfway between the two interlocutors (figure 1). This speaker grill cloth allowed sound to pass through (virtually) unobstructed, but completely occluded the view between interlocutors.

The materials for the interaction task itself included two 12 inch × 14 inch × 2 inch grid boxes, each divided into a total of nine 4 inch × 4⅔ inch open cells. Within each cell

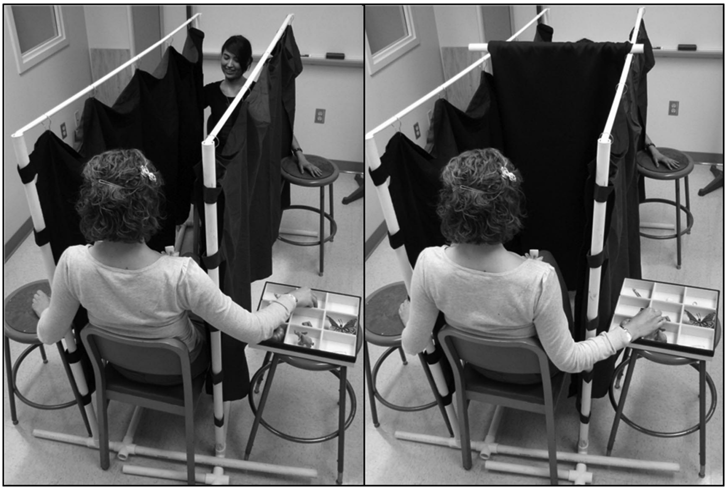

Figure 1. Interlocutors during the interaction task under two conditions: (a) audiovisual interaction and (b) audio-only interaction.

was one of nine small items that could be easily identifiable by touch (figure 1). These items were chosen for being nameable in a distinct manner with a two-syllable minimum utterance. The items used were a paperclip, small plastic butterfly, rubber band, eraser, crayon, battery, small plastic elephant, cotton ball, and small plastic pumpkin.

Procedure. Each participant received one grid box, which was positioned on a pedestal outside of the right side curtain, where it could not be seen. Participants were instructed to interact with their respective grids by feeding their hand through the curtain, where they would manipulate the grid by touch alone. Their other hand rested on a second pedestal positioned outside of the curtain on their left side. Having both hands outside of the curtains acted to eliminate visible hand gesturing.

For the interaction task, an experimenter placed the pieces in each grid box in pseudo-random positions so that their arrangement was different for the two participants of each pair. Participants were instructed to work together to manipulate the pieces in their grids so that the pieces' positions matched those of their partner's. Participants were also told that only one item could be moved at a time, and no more than one item could occupy any given cell at one time. The participants alternated who moved their pieces, such that for each run of the task, only one participant moved her pieces to match her partner's grid. Participants were encouraged to speak freely in working out their solutions for the puzzle. Toward the end of each run of the task, the participants were asked to report the locations of the pieces in their grid to ensure that both arrangements were the same. Upon completion, the participants removed their hands from the grids, the pieces were again placed in pseudo-random locations in the grid, and the task began again. Each pair of interlocutors completed the task four times, taking an average of 5 min per round. The entire interaction was recorded with two lapel microphones and a Roland Edirol R-9 digital audio recorder. The target words were later downloaded to a computer with Amadeus II software at 44 kHz, 16 bit (Hairer 2007).

2.2 *Pretest – posttest*

Materials. Pretest/posttest items consisted of nine different carrier phrases which included the nine token items (eg butterfly, paperclip) used in the search-task procedure. Photographs of the tokens were randomly presented to each participant twice in a sound-attenuated booth on a computer screen via PsyScope software (Cohen et al 1993). As each photograph was presented, participants would say the carrier phase, including the token and the number it was presented. Carrier phrases were organized such that tokens were positioned mid-sentence: “The [TOKEN] is number [#].” (Pardo 2006).

Procedure. Participants were tasked with reading each carrier phrase out loud as they were presented on screen. Participants were instructed to read the sentences in a clear and normal manner. Verbal responses were digitally recorded with a Shure SM57 microphone and Amadeus II software at 44 kHz, 16 bit (Hairer 2007). Participants completed this task as a pretest prior to the interaction procedure and again as a posttest after.

The entire experiment, including pretest, interaction task, and posttest, took about 1 h for each participant to complete.

2.3 *Phase II: Alignment assessment*

Participants. One hundred undergraduates from the University of California, Riverside Human Subjects Pool (seventy-one female, age range 17 – 34 years) acted as naive raters.

Materials. Utterances of the nine token items used in the interactive task were isolated out of each interlocutor’s digital recordings using Final Cut Pro 5. Pretest and posttest tokens were taken from the repeated carrier phrase utterances. Tokens taken during the interaction task were isolated out of the end-clause of each token item uttered last during the 20 min interaction (Pardo 2006).

Procedure. Using an AXB paradigm, raters were tasked with making one of two comparisons. One group of raters, $n = 50$, compared an interlocutor’s last-interaction-utterance token with their partner’s pretest and last-interaction-utterance token. For example: A = (I_1’s pretest token ‘paperclip’), X = (I_2’s last-interaction-utterance token ‘paperclip’), B = (I_1’s last-interaction-utterance token ‘paperclip’). This pretest-to-interaction-utterance AXB comparison was borrowed from Pardo (2006), and tests whether interlocutors’ utterances sound more like each other as they interact, relative to utterances they produce on their own, before interacting.

The other group, $n = 50$, compared an interlocutor’s posttest tokens with their partner’s pretest and posttest tokens. For example, A = (I_1’s pretest token ‘paperclip’), X = (I_2’s posttest token ‘paperclip’), B = (I_1’s posttest token ‘paperclip’). This pretest-to-posttest-utterance AXB comparison was not conducted by Pardo (2006), but was designed to act as a control to ensure that ratings of similarity for the pretest-to-interaction comparison were not simply the result of putative differences between prompted and spontaneous speech. Given that pretest and posttest utterances are both prompted stimuli, then finding greater similarity across interlocutors’ posttest tokens could not be attributable to ‘instructed’ differences behind the utterances (spontaneous versus prompted). In addition, finding greater rated similarity for posttest tokens would suggest that interlocutors’ alignment actually lasted after the interaction itself was completed, for at least a brief period.

Regardless, it is predicted that if alignment is occurring, then interlocutors will sound more like each other during or just after the interaction, as assessed by the pretest-to-interaction and pretest-to-posttest-utterance AXB tests, respectively. Further, if having access to visual information enhances alignment, then both AXB tests should reveal greater alignment for the AV over AO group.

All raters were tasked with indicating which token, A or B, sounded more like the target, X. Each rater rated utterances from a single pair of interlocutors across 72 trials, counterbalancing for order effects and target X, with one repetition of each combination (9 token items, 2 target speakers, 2 possible orders of A and B, 2 instances of each combination). If a posttest or last-interaction-utterance token was chosen as sounding more similar to the target X, it was considered indicative of speech alignment. Responses were recorded by having subjects indicate their A or B choices on an ioLab Systems USB Response Box. Stimulus presentation and response recordings were executed with PsyScope software (Cohen et al 1993).

3 Results

The observed proportion (c) of alignment rated responses for each rater was calculated by dividing the number of ratings of alignment [when I_1's posttest tokens are rated as more similar to I_2's posttest tokens or when I_1's (last) interaction utterances are rated as more similar to I_2's (last) interaction utterances] by the total number of ratings. An additional measure of the data, the detection parameter (d'), was calculated using a two-alternative forced-choice model ($d' = Z[c_{(\mathrm{SN})}] + Z[c_{(\mathrm{NS})}]$). The AXB position (A versus B) bias parameter ($\log\beta$) was also calculated $\{\log\beta = 0.5(Z^2[c_{(\mathrm{SN})}]$ $-Z^2[c_{(\mathrm{NS})}])$ where the parameter $c_{(\mathrm{SN})}$ was the proportion of alignment rated responses when, within the AXB paradigm, A is the posttest or last-interaction utterance, and the parameter $c_{(\mathrm{NS})}$ was the proportion of alignment rated responses when B is the posttest or last-interaction utterance$\}$.

Table 1. One sample t-tests (1-tailed) of dependent variables across groups.

AXB comparison	Interaction condition	DV	N	M	SEM	t_{24}	p	r
Pretest-to-AO	c	25	0.646	0.027	5.399***	0.000	0.741	
interaction		d'	25	0.862	0.181	4.753***	0.000	0.696
utterances		$\log\beta$	25	0.066	0.064	1.040	0.155	0.208
	AV	c	25	0.729	0.031	7.322***	0.000	0.831
		d'	25	1.456	0.224	6.493***	0.000	0.798
		$\log\beta$	25	0.337	0.071	4.736***	0.000	0.695
Pretest-to-	AO	c	25	0.524	0.011	2.240*	0.018	0.416
posttest		d'	25	0.129	0.059	2.205*	0.018	0.410
utterances		$\log\beta$	25	−0.023	0.021	−1.113	0.139	0.222
	AV	c	25	0.558	0.020	2.965**	0.004	0.518
		d'	25	0.307	0.105	2.919**	0.004	0.512
		$\log\beta$	25	0.002	0.030	0.064	0.475	0.013

Note: c (mean probability correct) was tested against a criterion of chance probability (0.50). The DV's d' and $\log\beta$ were tested against a criterion of no detection and no bias respectively (0.00). $^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Raters of pretest-to-interaction tokens detected levels of alignment between audio-only (AO) interlocutors significantly above chance ($c = 0.646$, SE $= 0.027$, $t_{24} = 5.399$, $p < 0.001$, $r = 0.741$; $d' = 0.862$, SEM $= 0.181$, $t_{24} = 4.753$, $p < 0.001$, $r = 0.696$), with no significant levels of response bias ($\log\beta = 0.066$, SEM $= 0.064$, $t_{24} = 1.040$, $p = 0.155$, $r = 0.208$) (see table 1). These findings of AO interlocutors replicate Pardo's (2006) pretask to task-utterance findings. Ratings of audiovisual (AV) interlocutors also showed detectable levels of alignment significantly above chance ($c = 0.729$, SEM $= 0.031$, $t_{24} = 7.322$, $p < 0.001$, $r = 0.831$; $d' = 1.456$, SEM $= 0.224$, $t_{24} = 6.493$, $p < 0.001$, $r = 0.798$), table 1. There was a significant level of response bias ($\log\beta$ $= 0.337$, SEM $= 0.071$, $t_{24} = 4.736$, $p < 0.001$, $r = 0.695$), indicating that participants were biased to A responses in the AXB task (table 1). A t-test revealed that ratings for AV tokens show significantly greater alignment than those for AO alignment ($\Delta c = -0.083$, SED $= 0.041$, $t_{48} = -2.013$, $p < 0.05$, $r = 0.276$; $\Delta d' = -0.595$, SED $= 0.288$, $t_{48} = -2.063$, $p < 0.05$, $r = 0.283$), and that response bias for AV raters was greater than for AO raters ($\log\beta = -0.271$, SED $= 0.096$, $t_{48} - 2.833$, $p < 0.01$, $r = 0.375$), table 2.

Turning to the pretest-to-posttest token comparison, this measure also revealed levels of alignment for AO interlocutors above chance ($c = 0.524$, SEM $= 0.011$, $t_{24} = 2.240$, $p < 0.05$, $r = 0.416$; $d' = 0.129$, SEM $= 0.059$, $t_{24} = 2.205$, $p < 0.05$, $r = 0.410$), with no significant level of response bias ($\log\beta = -0.023$, SEM $= 0.021$,

Table 2. Independent samples *t*-tests (1-tailed) comparing AO versus AV alignment.

AXB comparison	DV	N	$\Delta(M_{\mathrm{AO}} - M_{\mathrm{AV}})$	SED	t_{48}	p	r
Pretest-to-interaction utterances	c	50	−0.083	0.041	−2.013*	0.025	0.276
	d'	50	−0.595	0.288	−2.063*	0.023	0.283
	$\log\beta$	50	−0.271	0.096	−2.833**	0.004	0.375
Pretest-to-posttest utterances	c	50	−0.034	0.022	−1.506	0.070	0.210
	d'	50	−0.177	0.120	−1.475	0.074	0.206
	$\log\beta$	50	−0.025	0.036	−0.688	0.247	0.098

$^*p < 0.05$, $^{**}p < 0.01$.

$t_{24} = -1.113$, $p = 0.139$, $r = 0.222$); see table 1. Levels of alignment for AV interlocutors was significantly above chance levels ($c = 0.558$, SEM $= 0.020$, $t_{24} = 2.965$, $p < 0.01$, $r = 0.518$; $d' = 0.307$, SEM $= 0.105$, $t_{24} = 2.919$, $p < 0.01$, $r = 0.512$), with no significant level of response bias ($\log\beta = 0.002$, SEM $= 0.030$, $t_{24} = 0.064$, $p = 0.475$, $r = 0.013$), table 1. These results suggest that the observed alignment for the pretest-to-interaction measure was not simply the result of differences between prompted and spontaneous speech. A *t*-test revealed that AV levels of alignment were marginally greater than AO levels ($\Delta c = -0.034$, SED $= 0.022$, $t_{48} = -1.506$, $p = 0.070$, $r = 0.210$; $\Delta d' = -0.177$, SED $= 0.120$, $t_{48} = -1.475$, $p = 0.074$, $r = 0.206$), with no significant difference in response bias ($\log\beta = -0.025$, SED $= 0.036$, $t_{48} = -0.688$, $p = 0.247$, $r = 0.098$), table 2.

A final test examined whether alignment ratings differed between the pretest-to-interaction and pretest-to-posttest AXB comparison measures. For this purpose, a two (pretest-to-interaction and pretest-to-posttest AXB comparisons) by two (AO and AV interaction conditions) factorial ANOVA was conducted using observed proportion (c) of alignment rated responses as the dependent variable. A main effect of AXB comparison was found, between pretest-to-interaction ($M = 0.688$, SEM $= 0.013$) and pretest-to-posttest ($M = 0.541$, SEM $= 0.011$), $F_{1,96} = 38.601$, $p < 0.001$, $\eta_{\mathrm{p}}^2 = 0.287$. A main effect was also found for interaction condition, between AO ($M = 0.585$, SEM $= 0.017$) and AV ($M = 0.644$, SEM $= 0.022$), $F_{1,96} = 6.188$, $p < 0.05$, $\eta_{\mathrm{p}}^2 = 0.061$. However, there was no significant interaction of AXB comparison by interaction condition, $F_{1,96} = 1.101$, $p = 0.297$, $\eta_{\mathrm{p}}^2 = 0.011$.

The same two-by-two factorial ANOVA was conducted using d' as the dependent variable. Again, a main effect of AXB comparison was found, between pretest-to-interaction ($M = 1.159$, SEM $= 0.149$) and pretest-to-posttest ($M = 0.218$, SEM $= 0.061$), $F_{1,96} = 36.267$, $p < 0.001$, $\eta_{\mathrm{p}}^2 = 0.274$. A main effect of interaction condition was also found, between AO ($M = 0.495$, SEM $= 0.108$) and AV ($M = 0.882$, SEM $= 0.148$), $F_{1,96} = 6.108$, $p < 0.05$, $\eta_{\mathrm{p}}^2 = 0.060$. There was no significant interaction of AXB comparison by interaction condition, $F_{1,96} = 1.783$, $p = 0.185$, $\eta_{\mathrm{p}}^2 = 0.018$.

4 Discussion

Ratings across both AXB tests suggest that auditory information alone is sufficient for speech alignment to occur in an interactive context. These audio-only findings replicate those of Pardo (2006), who observed similar levels of alignment, albeit in the context of a different interactive task.

Speech alignment was also significantly detected for audiovisual interlocutors across both AXB test types. Comparing audio-only ratings to audiovisual ratings, the audiovisual interlocutors, in general, seemed to align more to one another's speech. When comparing pretest to (last) interaction-utterances, audiovisual ratings were significantly greater than audio-only ratings. When comparing pretest to posttest utterances, audiovisual ratings were also (marginally) greater than audio-only ratings, indicating this difference was not simply the result of the putative differences between prompted and spontaneous speech.

Regarding the two AXB measures of alignment, rated alignment was observed to be greater for the pretest-to-interaction comparison than for the pretest-to-posttest comparison. There are at least two possible reasons for this difference. The difference could indicate that for the pretest-to-interaction comparison, the greater perceived similarity across interlocutors' interaction tokens was, in fact, based partly on inherent differences between spontaneous and prompted speech. However, the greater alignment ratings for pretest-to-interaction tests may also indicate that alignment was strongest during the live interaction and then partially faded—without completely disappearing—over the short period between when the interaction ended and the posttest utterances were recorded. Future research can be designed to test these explanations.

Regardless, the overall findings show that visual information for a talker can enhance speech alignment in an interactive context. This is consistent with other research showing that visual speech, both on its own and in an audiovisual context, can induce shadowing-based alignment (Miller et al 2010; Sanchez et al 2010) and that visual information for an interlocutor can induce behavioral alignment outside of speech perception (eg Giles et al 1991).

The design of this experiment cannot thoroughly address the specific visual information that acts to enhance speech alignment. Interlocutors in the audiovisual group had a full, unobstructed view of each other's heads, torsos, and legs. This means that visible speech articulations, eye movements, upper body motions, leg positions, postures, or any combination of these visible dimensions could have been influential in alignment. For example, being able to see an interlocutor's eyes could create a greater degree of engagement, which could potentially induce greater speech alignment.

However, it could very well be that having access to visible speech information—seeing the speech articulations of the partner—helped enhance alignment. As discussed, there is already evidence that talker-specific information is available in visible speech (Lachs and Pisoni 2004a, 2004b; Rosenblum et al 2002, 2007), and that this information can induce alignment in a shadowing context (Miller et al 2010; Sanchez et al 2010). Thus, interlocutors in the audiovisual group may have aligned to each other's speech to a greater degree because they were able to see each other's articulatory movements, which provided additional talker-specific information.

Follow-up experiments could be designed to isolate visible speech articulator information in live context, to examine this possibility. This could be accomplished using a combination of video cameras and real-time image manipulation. Alternatively, isolating live interactive speech might be accomplished using a facial point-light technique (Rosenblum et al 1996). If it is found that isolated visible articulatory information can enhance speech alignment, such findings will add to the growing body of evidence (Miller et al 2010; Sanchez et al 2010) indicating that visual speech can induce alignment. As intimated, this evidence would have theoretical implications for our understanding of the nature of lexical memory (eg Goldinger 1998), and how speech perception can help prime production (eg Fowler 2004).

Regardless of whether visible speech information is most relevant in this experiment, the current results do show that alignment can be enhanced with the addition of crossmodal information. In this more general sense, the finding is consistent with decades of research showing crossmodal enhancement of perception and behavior (see Calvert et al 2004, for a review). Besides the aforementioned example of visible speech enhancing comprehension of degraded auditory speech (eg Sumby and Pollock 1954), adding cross-sensory information for determining an object's presence (eg Diederich and Colonius 2004), identity (eg Giard and Peronnet 1999), and location (eg Vroomen and de Gelder 2000) can enhance speed and accuracy of responses. Cross-sensory information is also known to enhance detection of odors (eg Gottfried and Dolan 2003), and the flavor of food (eg Dalton et al 2000). The current results show that adding cross-modal (visual) information to an auditory speech signal can enhance the degree to which observers inadvertently imitate the speech they perceive.

The current results also provide a new example of visual information modulating behavioral alignment between individuals. As stated, visually based behavioral alignment of posture (Condon and Ogston 1967), gesture (Mauer and Tindall 1983), head nodding and facial expression (Hale and Burgoon 1984) have all been observed between interacting subjects (eg Giles et al 1991). The current research shows that speech imitation is yet another type of interactive alignment that can be visually modulated.

More generally, these findings add to the burgeoning evidence for the ubiquity of imitation in human interactions. From birth, infants are known to imitate the facial expressions and articulator positions of heard and seen adults (Chen et al 2004; Meltzoff and Moore 1997) and, as language development continues, infants will audibly imitate the vocalizations of heard speech (Kuhl and Meltzoff 1993). In adulthood, this alignment continues in language and other behaviours, like facilitating smoother and more successful interpersonal interactions (eg Chartrand and Bargh 1999). The ubiquity and importance of alignment has led many researchers to argue that one primary purpose of mirror neuron systems is to facilitate inadvertent imitation (eg Iacoboni 2009). If true, the current research could suggest that these systems can make use of multisensory input to help enhance interpersonal alignment.

References

Aronsson K, Jonsson L, Linell P, 1987 "The courtroom hearing as a middle ground: Speech accommodation by lawyers and defendants" *Journal of Language and Social Psychology* **6** 99 – 115

Calvert G, Spence C, Stein B E, 2004 *The Handbook of Multisensory Processes* (Cambridge, MA: MIT Press)

Cappella J N, Planalp S, 1981 "Talker and silence sequences in informal conversations III: Interspeaker influence" *Human Communication Research* **7** 117 – 132

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

Chen X, Striano T, Rakoczy H, 2004 "Auditory – oral matching behavior in newborns" *Developmental Science* **7** 42 – 47

Cohen J D, MacWhinney B, Flatt M, Provost J, 1993 "PsyScope: An interactive graphic system for designing and controlling experiments in the psychology laboratory using Macintosh computers" *Behavioral Research Methods, Instruments, and Computers* **25** 257 – 271

Condon W S, Ogston W D, 1967 "A segmentation of behaviour" *Journal of Psychiatric Research* **5** 221 – 235

Dalton P, Doolittle N, Nagata H, Breslin P A S, 2000 "The merging of the senses: Integration of subthreshold taste and smell" *Nature Neuroscience* **3** 431 – 432

Diederich A, Colonius H, 2004 "Bimodal and trimodal multisensory enhancement: Effects of stimulus onset and intensity on reaction time" *Perception & Psychophysics* **66** 1388 – 1404

Ehrlich H J, Graeven D B, 1971 "Reciprocal self-disclosure in a dyad" *Journal of Experimental Social Psychology* **7** 389 – 400

Fowler C A, 2004 "Speech as a supramodal or amodal phenomenon", in *The Handbook of Multisensory Processes* Eds G A Calvert, C Spence, B E Stein (Cambridge, MA: MIT Press) pp 189 – 202

Gentilucci M, Bernardis P, 2007 "Imitation during phoneme production" *Neuropsychologia* **45** 608 – 615

Giard M H, Peronnet F, 1999 "Auditory – visual integration during multimodal object recognition in humans: A behavioral and electrophysiological study" *Journal of Cognitive Neuroscience* **11** 473 – 490

Giles H, Coupland N, Coupland J, 1991 "Accommodation theory: Communication, context, and consequence", in *Contexts of Accommodation* Eds H Giles, J Coupland, N Coupland (New York: Press Syndicate of the University of Cambridge) pp 1 – 162

Goldinger S D, 1998 "Echoes of echoes? An episodic theory of lexical access" *Psychological Review* **105** 251 – 279

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

Gottfried J A, Dolan R J, 2003 "The nose smells what the eye sees: Crossmodal visual facilitation of human olfactory perception" *Neuron* **39** 375 – 386

Hairer M, 2007 *Amadeus II (version 3.8.7)* [Kenilworth: HairerSoft; retrieved from http://www.hairersoft.com/Amadeus.html]

Hale J L, Burgoon J K, 1984 "Models of reactions to changes in nonverbal immediacy" *Journal of Nonverbal Behavior* **8** 287 – 314

Iacoboni M, 2009 "Imitation, empathy, and mirror neurons" *Annual Review of Psychology* **60** 653 – 670

Jaffe J, Feldstein S, 1970 *Rhythms of Dialogue* (New York: Academic Press)

Kuhl P K, Meltzoff A N, 1993 "Infant vocalizations in response to speech: Vocal imitation and developmental change" *Journal of the Acoustical Society of America* **100** 2425 – 2438

Lachs L, Pisoni D B, 2004a "Cross-modal source information and spoken word recognition" *Journal of Experimental Psychology: Human Perception and Performance* **30** 378 – 396

Lachs L, Pisoni D B, 2004b "Crossmodal source identification in speech perception" *Ecological Psychology* **16** 159 – 187

McGurk H, MacDonald J, 1976 "Hearing lips and seeing voices" *Nature* **264** 746 – 748

Matarazzo J D, Weins A N, Matarazzo R G, Saslow G, 1968 "Speech and silence behaviour in clinical psychotherapy and its laboratory correlates", in *Research in Psychotherapy* volume 3, Eds J Schlier, H Hunt, J D Matarazzo, C Savage (Washington, DC: American Psychological Association) pp 347 – 394

Mauer R E, Tindall J H, 1983 "Effects of postural congruence on client's perceptions of counselor empathy" *Journal of Counseling Psychology* **30** 158 – 163

Meltzoff A N, Moore M K, 1997 "Explaining facial imitation: A theoretical model" *Early Development and Parenting* **6** 179 – 192

Miller R M, Sanchez K, Rosenblum D, 2010 "Alignment to visual speech information" *Attention, Perception, & Psychophysics* **72** 1614 – 1625

Namy L L, Nygaard L C, Sauerteig D, 2002 "Gender differences 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 and Social Psychology* **32** 790 – 804

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

Rosenblum L D, 2008 "Speech perception as a multimodal phenomenon" *Journal of the Association of Psychological Science* **17** 405 – 409

Rosenblum L D, Johnson J A, Saldana H M, 1996 "Point-light facial displays enhance comprehension of speech in noise" *Journal of Speech and Hearing Research* **39** 1159 – 1170

Rosenblum L D, Miller R M, Sanchez K, 2007 'Lip-read me now, hear me better later" *Journal of the Association of Psychological Science* **18** 392 – 396

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

Shockley K, Sabadini L, Fowler C A, 2004 "Imitation in shadowing words" *Perception & Psychophysics* **66** 422 – 429

Street R L J, 1984 "Speech convergence and speech evaluation in fact-finding interviews" *Human Communication Research* **11** 139 – 169

Sumby W H, Pollack I, 1954 "Visual contribution of speech intelligibility in noise" *Journal of the Acoustical Society of America* **26** 212 – 215

Vroomen J, de Gelder B, 2000 "Sound enhances visual perception: Cross-modal effects of auditory organization on vision" *Journal of Experimental Psychology: Human Perception and Performance* **26** 1583 – 1590

ISSN 0301-0066 (print) ISSN 1468-4233 (electronic)

PERCEPTION

VOLUME 40 2011

www.perceptionweb.com