Psychon Bull Rev (2013) 20:115–119
DOI 10.3758/s13423-012-0328-5

BRIEF REPORT

Reading your own lips: Common-coding theory and visual speech perception

Nancy Tye-Murray · Brent P. Spehar · Joel Myerson · Sandra Hale · Mitchell S. Sommers

Published online: 7 November 2012

Abstract Common-coding theory posits that (1) perceiving an action activates the same representations of motor plans that are activated by actually performing that action, and (2) because of individual differences in the ways that actions are performed, observing recordings of one's own previous behavior activates motor plans to an even greater degree than does observing someone else's behavior. We hypothesized that if observing oneself activates motor plans to a greater degree than does observing others, and if these activated plans contribute to perception, then people should be able to lipread silent video clips of their own previous utterances more accurately than they can lipread video clips of other talkers. As predicted, two groups of participants were able to lipread video clips of themselves, recorded more than two weeks earlier, significantly more accurately than video clips of others. These results suggest that visual input activates speech motor activity that links to word representations in the mental lexicon.

Keywords Visual word recognition · Models of visual word recognition and priming · Motor control · Motor planning/programming

The discovery of "mirror neurons" that respond both when a movement is made and when the same movement is observed has increased interest in the interaction between action and perception (Fadiga & Craighero, 2003; Rizzolatti & Craighero, 2004) and has focused attention on people's sometimes surprising ability to recognize a recording of an earlier action as being self- rather than other-generated (e.g., Knoblich & Prinz, 2001; Repp & Knoblich, 2004). This ability is of interest because it provides support for the theory that individuals' motor plans and their perceptions of observed actions are represented in a common code (Hommel, Müsseler, Aschersleben, & Prinz, 2001; Prinz, 1997), a theory with particular relevance for speech perception (Galantucci, Fowler, & Turvey, 2006). Common-coding theory posits that when perceiving an action, an individual activates the same representations of motor plans that would be activated if one were actually performing, or even just planning, the action. Similarly, the motor theory of speech perception assumes that when individuals perceive speech, they do so by accessing the motor codes for speech gestures (for a review, see Galantucci et al., 2006). Although a strong version of the motor theory, in which this process provides the sole or primary basis for speech perception, has been shown to be incorrect, the notion that speech perception shares a common code with speech production endures and represents a special case of a more general common-coding perspective (for reviews, see Galantucci et al., 2006; Hickok, 2010).

Individuals differ in the ways that they execute the same actions, and these differences can have important perceptual consequences. According to common-coding theory, this is because perception is affected by the correspondence between the sensory "event code" for a perceived action and the motor "action code" that is accessed when perceiving that action. This correspondence is assumed to be greater when the action was one's own, as compared to when the action was executed by another. Knoblich and Prinz (2001), for example, reported that one week after people drew alphanumeric characters from both familiar and unfamiliar alphabets on a graphics tablet, they could distinguish visual recordings of their own pen trajectories from those drawn by

N. Tye-Murray (✉) · B. P. Spehar
Department of Otolaryngology,
Washington University School of Medicine,
Campus Box 8115, 660 South Euclid Avenue,
St. Louis, MO 63124, USA
e-mail: murrayn@ent.wustl.edu

J. Myerson · S. Hale · M. S. Sommers
Department of Psychology, Washington University,
St. Louis, MO, USA

others, even if the original drawings were done without visual feedback. When information about dynamic changes in pen velocity was eliminated, however, participants could no longer make such discriminations, suggesting that self-recognition depends on individual differences in the kinematics of actions. Similarly, skilled pianists can distinguish their previously recorded piano performances from those of other skilled pianists playing the same piece, even when the performances were recorded using a digital piano without auditory feedback to the pianist (Repp & Knoblich, 2004).

One might object, however, that people are accustomed to attending to the perceptual consequences of their actions, presumably because this feedback helps them calibrate their movements and maintain accuracy, and thus familiarity with one's own movements might provide the basis for self–other judgments. Accordingly, the present experiment was designed with two goals in mind: first, to minimize the role of familiarity, in order to provide a more stringent test of the common-coding theory, and second, to assess the role that access to motor codes plays in the understanding of speech actions. Because individuals have relatively little familiarity with the visual consequences of their speech actions (i.e., most of us rarely watch ourselves talk in a mirror), we had participants perform a lipreading task. We compared participants' abilities to identify spoken words when viewing previously recorded silent video clips of themselves and of other speakers. We hypothesized that if, as the common-coding theory posits, there is greater correspondence between the sensory "event code" for a perceived action and the motor "action code" when the action is one's own, resulting in greater activation of motor plans, then people should be able to lipread themselves better than they can lipread others.

Method

Participants

A group of 20 adults (mean age = 25.7 years, *SD* = 7.1), all of whom were native English speakers and none of whom was a professional actor, participated in the study. All of the participants had normal hearing (pure-tone thresholds of 25 dB HL or better at 0.5, 1, and 2 KHz, as assessed using TDH-39 headphones and a calibrated audiometer), normal visual acuity (20/30 or better, as assessed using the Snellen Eye Chart), and normal visual contrast sensitivity (1.8 or better, as assessed using the Pelli–Robson Contrast Sensitivity Chart). The participants received $10/h for their participation.

Stimuli

Video of each participant, as seen from the shoulders up and illuminated with studio-quality halogen umbrella lights using a two-point lighting system, was recorded while the participant spoke multiple lists of sentences from the Build-A-Sentence (BAS) test (Tye-Murray, Sommers, & Spehar, 2007). The BAS software generates lists of 12 sentences each, with each list consisting of the same 36 high-frequency words (mean log frequency = 9.8, *SD* = 1.4; Balota et al., 2007) randomly organized into four types of sentences (see the Appendix). For the present study, 125 BAS lists of unique sentences—1,500 sentences in all—were generated as potential stimuli.

During recording, each sentence appeared on a teleprompter positioned directly above a Canon Elura camcorder. Participants were asked to read the sentence silently and then, when prompted by a green light, to speak the sentence as naturally as possible while looking directly at the camera. Video images were recorded directly to a PC hard drive for future editing, which removed the portions where participants read silently from the teleprompter. The first three lists were the same for each participant. After a participant was comfortable with the task, a unique set of sentence lists assigned to that participant was recorded. Two of the lists recorded by each participant were later selected to be used as test stimuli.

Procedure

Participation consisted of a 2-h recording session and a 2-h testing session. The participants were told that they would see video clips of themselves during the testing phase, but this was not emphasized as an important aspect of the study. To reduce the possibility that participants would remember any specific sentence that they might have spoken during the recording phase, we had participants record as many lists as time allowed during the initial session (*M* = 30.8 lists, or approximately 360 sentences; *SD* = 6.1 lists,), although only two unique lists from each person (24 sentences) were used as sources of stimuli for the testing session. Moreover, the average time between recording and testing was 31 days (minimum = 17 days).

To ensure that findings were not specific to one group of talkers, two groups of 10 participants each (five males and five females) were studied. The participants in each group were asked to lipread the set of BAS sentences spoken by the members of their group (including themselves and the nine others). Participants were tested individually in a sound-treated room while sitting approximately 0.5 m from a 17-in. video monitor. After each sentence was presented, the four possible sentence types and 36 possible words were displayed on the monitor to remind participants what the possible responses were, and the participants said aloud what they believed they had just lipread while an experimenter located outside the test booth recorded their responses.

To familiarize the participant with the test format, 20 practice sentences were presented. The first ten practice

sentences, one spoken by each participant in the group, were presented along with the audio of the talker's voice in order to refamiliarize the participant with the 36 BAS words. The second ten practice sentences included only the video of each of the ten talkers in order to familiarize the participant with the lipreading task. Participants were then tested on silent video clips of 240 sentences, 24 from each of the ten talkers in their group. These sentences were presented in the same random order (with the constraint that no talker was shown twice in a row) to all of the participants in a group.

Results

As may be seen in Fig. 1, the participants in each group were able to lipread themselves more accurately than they could lipread the others in their group [both $ts(9) > 2.75$, $ps < .05$]. Indeed, examination of the individual data revealed that 15 of the 20 participants could lipread themselves better than they could lipread their group. This may be seen in Fig. 2, which plots the accuracy with which individuals lipread themselves as a function of their general lipreading ability (data points above the dashed equality line indicate individuals who lipread themselves better than they lipread others), where general lipreading ability is indexed by the mean accuracy with which an individual lipread the others in the group. When the accuracy of lipreading oneself was regressed on general lipreading ability, examination of the parameters of the regression equation revealed that participants were approximately ten percentage points more accurate lipreading themselves than lipreading others, regardless of their lipreading ability: $Y = 10.1 + 1.02X$, $R^2 = .614$.

Interestingly, participants' general lipreading ability was unrelated to how accurately they were lipread by others ($r = -.02$). When these two independent factors, participants' lipreading ability and how accurately they could be lipread (denoted by X_1 and X_2, respectively), were incorporated into a simple, two-parameter linear model ($Y = b_0 + X_1 + b_1{*}X_2$), this model accounted for 78.7 % of the variance in the accuracy with which participants lipread themselves. Thus, rather than being due to measurement error or individual variation in the benefit from lipreading oneself, much of the variation in the differences in accuracy between participants lipreading themselves versus others could be accounted for simply by how accurately they could be lipread.

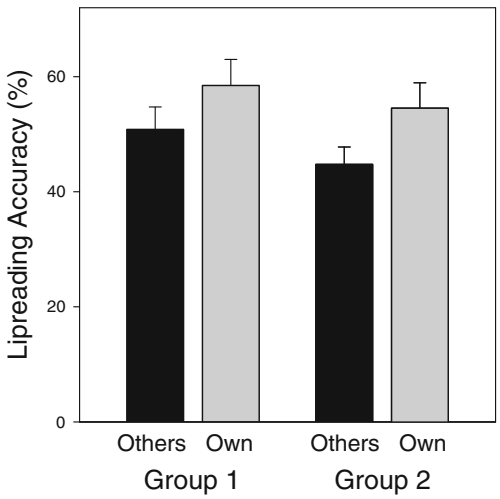

Fig. 1 Accuracy (i.e., percentages of words correct) with which participants read the lips of the others in their group, as compared with the accuracy with which they read their own lips. Error bars indicate standard errors of the means

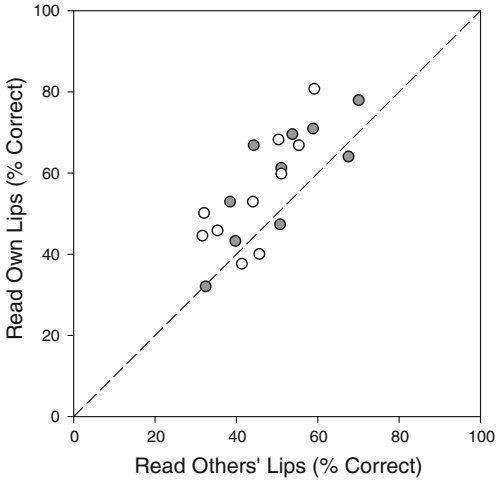

Fig. 2 Accuracy (i.e., percentages of words correct) with which participants read their own lips as a function of the accuracy with which they read the lips of the others in their group. Dark circles represent the data from participants in Group 1, and white circles represent the data from participants in Group 2

This finding is important, because it suggests that at the individual level, the most precise way to estimate the effect of lipreading oneself versus others can be obtained by comparing one's accuracy when lipreading oneself to one's accuracy when lipreading another person who is equally easy to lipread. Accordingly, we ranked participants on the basis of how accurately they were lipread by others in their group, and then, where possible (i.e., for all participants except those ranked at the very top and bottom), we compared participants' ability to lipread themselves with their ability to lipread, on average, the individuals directly above and below them in the ranking. In all cases, the participants lipread themselves more accurately, and the mean difference in accuracy was 14.0 percentage points.

Discussion

According to the common-coding theory, perceiving an action activates the same representations of motor plans that would be activated if one were actually producing that action (Hommel et al., 2001; Prinz, 1997). Furthermore,

observing one's own previous behavior activates motor plans to an even greater degree than does observing someone else's behavior, and this, according to the common-coding theory, makes it possible to predict the effects of one's own actions more accurately than the effects of others' actions (Knoblich & Flach, 2001; Knoblich, Seigerschmidt, Flach, & Prinz, 2002). We hypothesized that if observing one's own previous actions activates motor plans to a greater degree than does observing others' actions, and if these activated plans contribute to perception, then people should be able to lipread silent video clips of their own previous utterances more accurately than they can lipread video clips of other talkers.

The present results were consistent with this hypothesis: Individuals lipread silent video clips of themselves, recorded on average a month earlier, significantly better than they lipread video clips of others. These results provide strong support for the common-coding theory, adding to previous evidence from self–other judgments regarding recorded actions (e.g., Knoblich & Prinz, 2001; Repp & Knoblich, 2004), and they are consistent with the hypothesized role of activated motor plans in action understanding. Notably, the present study focused on actions that people (other than, perhaps, professional actors) rarely observe themselves perform in order to minimize the potential role of familiarity with the characteristics of one's own movements in the perception of self- and other-generated actions. Nevertheless, participants were consistently more accurate when they were lipreading themselves, and this difference in accuracy between lipreading oneself and lipreading others was independent of general lipreading ability.

These findings suggest that the common-coding principle extends to visual perception of speech gestures (lipreading) and implies that watching someone talk activates an observer's speech motor plans. Consistent with this view, a region of the left inferior frontal gyrus (Broca's area, which is classically associated with speech production) has been shown to be activated not only by both speech production and the perception of auditory speech stimuli (Pulvermüller et al., 2006; Wilson, Saygin, Sereno, & Iacoboni, 2004), but also by silent videos of oral speech movements. Importantly, this cortical region is more activated by videos of speech movements than by videos of nonspeech oral movements (Fridriksson et al., 2008; Hall, Fussell, & Summerfield, 2005; Turner et al., 2009). This suggests that the region is not part of a generalized mirror neuron system. Rather, it appears to be specialized for the production and perception, both auditory and visual, of lexical stimuli (Kotz et al., 2010), properties that are consistent with the common coding of speech production and perception.

We assume that individuals have unique motor signatures for speech gestures, just as they do for other kinds of actions, and that because of this, there is a greater correspondence between the representations of visually perceived speech gestures and the representations of speech motor plans when the talker and the observer are the same person. This correspondence results in greater activation of imitative motor plans, thereby generating more accurate expectations that contribute to better visual speech perception.

Previous evidence that greater correspondence between the representations of perceived actions and of motor plans leads to more accurate expectations regarding unfolding action has come from studies of handwriting and dart throwing (Knoblich & Flach, 2001; Knoblich et al., 2002). In the study by Knoblich and Flach, for example, participants watched video clips, recorded one week earlier, of themselves as well as of other participants throwing darts. They were able to "predict" where a dart had landed more accurately when they had been the thrower than when someone else had been the thrower, just as participants in the present study were able to identify what had been said more accurately when they had been the speaker.

Participants in the Knoblich and Flach (2001) study were told whether they had been the thrower in each video clip, and participants in the present study undoubtedly could recognize themselves in the video clips that they viewed. Although this might somehow have contributed to the results in both studies, participants' predictions in the Knoblich and Flach study were more accurate when they had been the thrower, regardless of whether or not their heads were visible in the video clips, suggesting that the awareness of watching oneself plays a relatively small role, as compared to the correspondence between the representations of motor plans and observed actions. Similarly, although a role for memory in the present study cannot be completely ruled out, the sheer number of sentences that each participant would have had to remember suggests that any contribution of memory for past utterances would have been very small.

What is important about these studies comparing individuals' ability to make judgments about their own and others' recorded actions is obviously not the phenomenon itself: Most people rarely need to make such judgments. Rather, what is important is the implications of this phenomenon for how actions are represented and—particularly in the case of speech actions—how they are understood. We suggest that just as word representations in the mental lexicon can be directly activated by both auditory and visual input (Summerfield, 1992; Tye-Murray et al., 2007), they can also be activated by speech motor activity, which itself can be evoked by auditory input or, as in the present study, by visual input. In short, visual speech perception involves both visual and motor activation of word representations in the mental lexicon. Although visual and motor activation may not play a major role in speech perception under good listening conditions, it can be extremely important when auditory input is absent or degraded (Hickok, Houde, &

Rong, 2011), whether this is because of the situation or because of age or other individual differences in hearing ability (Tye-Murray et al., 2008).

Author note This research was supported by National Institutes of Health Grant No. AG018029.

Appendix

The four sentence structures used in the BAS test (Tye-Murray et al., 2007)

1) The ___ watched the ___ .
2) The ___ watched the ___ and the ___.
3) The ___ and the ___ watched the ___ .
4) The ___ and the ___ watched the ___ and the ___.

The 36 words used in the BAS test

1) bear	10) dog	19) goat	28) son
2) bird	11) dove	20) guest	29) team
3) boys	12) duck	21) men	30) toad
4) bug	13) fawn	22) mice	31) tribe
5) cat	14) fish	23) mole	32) troop
6) cook	15) fox	24) moose	33) whale
7) cop	16) frog	25) saint	34) wife
8) cow	17) geese	26) seal	35) wolf
9) deer	18) girls	27) snail	36) worm

References

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., . . . Treiman, R. (2007). The English Lexicon Project. *Behavioral Research Methods*, *39*, 445–459. doi:10.3758/BF03193014

Fadiga, L., & Craighero, L. (2003). New insights on sensorimotor integration: From hand action to speech perception. *Brain and Cognition, 53,* 514–524.

Fridriksson, J., Moss, J., Davis, B., Baylis, G. C., Bonilha, L., & Rorden, C. (2008). Motor speech perception modulates the cortical language areas. *NeuroImage, 41,* 605–613. doi:10.1016/j.neuroimage.2008.02.046

Galantucci, B., Fowler, C. A., & Turvey, M. T. (2006). The motor theory of speech perception reviewed. *Psychonomic Bulletin & Review, 13,* 361–377. doi:10.3758/BF03193857

Hall, D. A., Fussell, C., & Summerfield, A. Q. (2005). Reading fluent speech from talking faces: Typical brain networks and individual differences. *Journal of Cognitive Neuroscience, 17,* 939–953.

Hickok, G. (2010). The role of mirror neurons in speech and language processing. *Brain and Language, 112,* 1–2. doi:10.1016/j.bandl.2009.10.006

Hickok, G., Houde, J., & Rong, F. (2011). Sensorimotor integration in speech processing: Computational basis and neural organization. *Neuron, 69,* 407–422. doi:10.1016/j.neuron.2011.01.019

Hommel, B., Müsseler, 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. doi:10.1017/S0140525X01000103

Knoblich, G., & Flach, R. (2001). Predicting the effects of actions: Interactions of perception and action. *Psychological Science, 12,* 467–472.

Knoblich, G., & Prinz, W. (2001). Recognition of self-generated actions from kinematic displays of drawing. *Journal of Experimental Psychology: Human Perception and Performance, 27,* 456–465.

Knoblich, G., Seigerschmidt, E., Flach, R., & Prinz, W. (2002). Authorship effects in the prediction of handwriting strokes: Evidence for action simulation during action perception. *Quarterly Journal of Experimental Psychology, 55A,* 1027–1046. doi:10.1080/02724980143000631

Kotz, S. A., D'Ausilio, A., Raettig, T., Begliomini, C., Craighero, L., Fabbri-Destro, M., . . . Fadiga, L. (2010). Lexicality drives audio–motor transformations in Broca's area. *Brain & Language*, *112*, 3–11. doi:10.1016/j.bandl.2009.07.008

Prinz, W. (1997). Perception and action planning. *European Journal of Cognitive Psychology, 9,* 129–154. doi:10.1080/713752551

Pulvermüller, F., Huss, M., Kheri, F., del Prado, M., Martin, F., Hauk, O., & Shtyrov, Y. (2006). Motor cortex maps articulatory features of speech sounds. *Proceedings of the National Academy of Sciences, 103,* 7865–7870.

Repp, B. H., & Knoblich, G. (2004). Perceiving action identity: How pianists recognize their own performances. *Psychological Science, 15,* 604–609. doi:10.1111/j.0956-7976.2004.00727.x

Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience, 27,* 169–192. doi:10.1146/annurev.neuro.27.070203.144230

Summerfield, Q. (1992). Lipreading and audio-visual speech perception. *Philosophical Transactions of the Royal Society B, 335,* 71–78. doi:10.1098/rstb.1992.0009

Turner, T. H., Fridriksson, J., Baker, J., Eoute, D., Jr., Bonilha, L., & Rorden, C. (2009). Obligatory Broca's area modulation associated with passive speech perception. *NeuroReport, 20,* 492–496. doi:10.1097/WNR.0b013e32832940a0

Tye-Murray, N., Sommers, M., & Spehar, B. (2007). Auditory and visual lexical neighborhoods in audiovisual speech perception. *Trends in Amplification, 11,* 233–241. doi:10.1177/1084713807307409

Tye-Murray, N., Sommers, M., Spehar, B., Myerson, J., Hale, S., & Rose, N. S. (2008). Auditory–visual discourse comprehension by older and young adults in favorable and unfavorable conditions. *International Journal of Audiology, 47*(Suppl. 2), S31–S37. doi:10.1080/14992020802301662

Wilson, S. M., Saygin, A. P., Sereno, M. I., & Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nature Neuroscience, 7,* 701–702.