

Articulatory imaging implicates prediction during spoken language comprehension

Eleanor Drake¹ · Martin Corley¹

Published online: 12 June 2015
© Psychonomic Society, Inc. 2015

Abstract It has been suggested that the activation of speech–motor areas during speech comprehension may, in part, reflect the involvement of the speech production system in synthesizing upcoming material at an articulatorily specified level. In this study, we explored that suggestion through the use of articulatory imaging. We investigated whether, and how, predictions that emerge during speech comprehension influence articulatory realizations during picture naming. We elicited predictions by auditorily presenting high-cloze sentence stems to participants (e.g., *When we want water we just turn on the . . .*). Participants named a picture immediately following each sentence-stem presentation. The pictures either matched (e.g., TAP) or mismatched (e.g., CAP) the high-cloze sentence-stem target. Throughout each trial, participants’ speech–motor movements were recorded via dynamic ultrasound imaging. This allowed us to compare articulations in the match and mismatch conditions to each other and to a control condition (simple picture naming). Articulations in the mismatch condition differed more from the control condition than did those in the match condition. This difference was reflected in a second analysis that showed greater frame-by-frame change in articulator positions for the mismatch than for the match condition around 300–500 ms before the onset of the picture name. Our findings indicate that comprehension-elicited prediction influences speech–motor production, suggesting that the speech production system is implicated in the representation of such predictions.

Keywords Language comprehension · Speech production · Psycholinguistics

Have you ever felt that somebody else’s words are on the tip of your tongue? When we listen to another person speaking, our own motor system is activated (Fadiga, Craighero, Buccino, & Rizzolatti, 2002; Pulvermüller et al., 2006; Watkins & Paus, 2004; Wilson, Saygin, Sereno, & Iacoboni, 2004; for reviews see Gambi & Pickering, 2013; Scott, McGettigan, & Eisner, 2009). This motor activation appears to reflect two levels of representation: referential resonance and communicative resonance (Fischer & Zwaan, 2008; Willems & Hagoort, 2007). *Referential resonance* describes activation elicited by the linguistic content of the listened-to material, and involves the representation or simulation of the motor acts referred to by the speakers (e.g., hearing “kick” activates leg areas: Hauk, Johnsrude, & Pulvermüller, 2004; Tettamanti et al., 2005). *Communicative resonance* describes activation related to the phonetic content, and involves representation or simulation of the motor activity involved in speech production itself (e.g., hearing /k^hik/ activates areas involved in the articulation of that sound stream: Fadiga et al., 2002; Pulvermüller et al., 2006). This study was concerned with the speech–motor activation associated with communicative resonance. We employed articulatory imaging to investigate the suggestion that, as well as reflecting the bottom-up processing of auditory material as it is encountered, communicative resonance additionally indexes top-down prediction of the to-be-heard materials (e.g., Pickering & Garrod, 2007; Schiller, Horemans, Ganushchak, & Koester, 2009).

There is substantial evidence that language comprehension involves prediction at a variety of levels (Altmann & Kamide, 1999; Federmeier & Kutas, 1999; Federmeier, McLennan, De Ochoa, & Kutas, 2002; Kamide, Altmann, & Haywood, 2003;

✉ Martin Corley
martin.corley@ed.ac.uk

¹ Department of Psychology, School of Philosophy, Psychology and Language Sciences, University of Edinburgh, 7 George Square, Edinburgh EH8 9JZ, UK

Rommers, Meyer, Praamstra, & Huettig, 2013; Rothermich & Kotz, 2013; for reviews, see Dikker & Pylkkänen, 2013; Federmeier, 2007). If activity in the speech–motor system were shown to be related to prediction, the speech production system would be implicated, as was suggested by Pickering and Garrod (2007). Predictions would need to be made at least at the level of phonological–phonetic speech sounds for relevant activation of the speech–motor system to ensue. The prediction of phonologically specified representations has been demonstrated during written language comprehension: In an RSVP reading study, participants displayed N400-indexed surprisal upon encountering the indefinite article (*a/an*) in a phonological form that was inappropriate given the predicted upcoming word (e.g., encountering *an* when strongly constrained to anticipate a noun with a consonant onset, such as *kite*; DeLong, Urbach, & Kutas, 2005). The question remains open whether speech-sound predictions are generated during *spoken* language comprehension, and, if so, whether the speech–motor system, and the production system more generally, are implicated.

One attempt to investigate these questions used a paradigm modeled on picture–word interference (PWI) studies (e.g., Damian & Dumay, 2007; Lupker, 1982; Meyer & Schriefers, 1991). In PWI studies, participants typically name a sequence of pictures while they are instructed to ignore printed words presented at the same time; the relationships between the words and pictures are systematically varied to demonstrate effects on the production of, for example, phonological overlap between the picture and word. In an investigation of prediction in spoken language comprehension, rather than presenting printed (or auditory) distractor words, Drake and Corley (2014) induced participants to predict that they would hear a word (such as *tap*) by presenting them with highly constraining spoken sentence fragments (*when we want water we just turn on the . . .*). Pictures were presented for naming as each fragment ended. In contrast to findings from PWI studies, phonological overlap between the predicted words and picture names was not found to have an effect on response times (RTs): Participants were no quicker to name a picture when its name partially overlapped with the predicted word (e.g., TAN) than when it did not (e.g., COAT; Drake & Corley, 2014; see also Severens, Ratinckx, Ferreira, & Hartsuiker, 2008).

A reasonable interpretation of evidence such as this is that speech sounds are not routinely predicted in the production system during spoken language comprehension, at least not to the extent that they affect the timing of responses; this was the conclusion that Drake and Corley (2014) reached. However, the time taken to name pictures may be an inappropriate measure to base such a conclusion on. To complete the task, participants had to decide *when* to speak, and they might have been able to make use of prosodic and timing cues from the spoken sentence fragments in order to do so (e.g., M. Wilson

& Wilson, 2005). To the extent that participants' speech timing was governed by extrinsic as well as intrinsic factors, subtle differences in naming latencies may have been difficult to detect, in contrast to PWI studies, in which no extrinsic timing information is available.

Another reason for treating Drake and Corley's (2014) behavioral evidence with caution is that evidence does appear to implicate motor areas in prediction more generally. However, this evidence is derived primarily from studies of representational momentum in the perception of human movement (e.g., Huber & Krist, 2004; Miall & Wolpert, 1996; Miall & Reckess, 2002; Verfaillie & Daems, 2002; see Pickering & Garrod, 2007, 2013, for discussions with respect to language comprehension). In order to directly investigate the involvement of the speech–motor system in the prediction of upcoming sounds, a more appropriate source of evidence than speech timing might be the articulatory movements that are the product of activation in the speech–motor areas on an ongoing basis. If this activation reflects, in any part, the prediction of upcoming speech sounds, then we should be able to find evidence for the activation in perturbations of speech-sound movements made during the time in which such predictions are active.

Spatio-temporal variability in the realization of phonemes has often been treated in psycholinguistic studies as motor noise, in part because speech–motor (phonetic) realizations of phonological representations are inherently variable (e.g., Mitra, Nam, Espy-Wilson, Saltzman, & Goldstein, 2011; Neiberg, Ananthakrishnan, & Engwall, 2008). Articulation is a dynamic, highly flexible process, which maps to its acoustic consequences in a complex, many-to-one manner. Importantly, however, it adapts online to changes in environmental, physical, linguistic, and psychological circumstances (Fowler, 2014; Garnier & Henrich, 2014; McMillan & Corley, 2010; Pianesi, 2007). As an utterance unfolds, speech–motor behavior is influenced by both recent and upcoming demands on the speech execution system. This can be observed in the phenomena of perseveratory and anticipatory co-articulation: For any phonological representation to be realized during overt speech, motor effectors (i.e., the articulators: tongue, lips, etc.) must be positioned appropriately within a target region associated with the intended acoustic output. However, placement within that target region is influenced by the articulator configurations required for preceding and upcoming speech (for a review, see Hardcastle & Hewlett, 1999). Perseveratory co-articulation may arise due to mechanical and inertial forces associated with the preceding context (Recasens, Pallarès, & Fontdevila, 1997; Tilsen, 2007), but anticipatory co-articulation can occur only when the speaker is able to “look ahead” and perceive the articulatory requirements of upcoming speech. Such anticipatory co-articulation is characteristic of competent adult speakers, and it is assumed that the anticipatory processes that give rise to it are necessary

for the fluent production of speech (Dang et al., 2004; Goffman, Smith, Heisler, & Ho, 2008; Katz, 2000; Lubker, 1981; Whalen, 1990). In the present study, we employed ultrasound speech imaging to investigate the articulatory consequences of predicting that you will *hear*, rather than *produce*, an upcoming sound.

Ultrasound imaging allows dynamic recording of the movements of the tongue during speech and has been valuable in providing information about many aspects of articulation, including co-articulation (see Stone, 2005, for a comprehensive introduction to the technique; see Davidson, 2005, for an example of a study in which the technique was used to measure co-articulation in order to address a phonological question). Articulatory imaging is achieved by placing a Doppler transducer probe (similar to that used in fetal imaging) against the undersurface of the participant's chin. The transducer emits and receives very high-frequency sound waves (inaudible to humans). The sound waves sweep the midsagittal plane and are reflected at points where substance impedance changes (primarily at the tongue surface). The transducer then receives the reflected echoes. Because the speed of sound is constant, it is possible to determine the location coordinates of the surface boundary at which a reflection took place. The location coordinates are then converted into a visual image of the oral cavity in midsagittal section.

In the present study, we employed grayscale images. The intensity of reflections from any given location was represented on a scale from black (no reflection) to white (total reflection). The tongue surface appeared as a bright contour on screen, with the tongue root typically being pictured on the left of the screen and the tongue tip on the right. Changes in tongue position—for example, those associated with changes in the sound being articulated—are visible as movements of this contour. Sampling rates greater than 300 frames per second (fps) can be achieved. For our study, the data were acquired at a rate of 100 fps but were processed at a video rate of ~30 fps for reasons of tractability. This sampling rate allowed us to examine tongue position at key times determined from the auditory data (e.g., the onset of acoustically available speech), and also to investigate frame-to-frame changes in tongue position during the response latency period.

The ultrasound technique is well suited to psycholinguistically motivated studies, in that it provides a noninvasive and relatively low-cost way to capture tongue movements during speech. However, ultrasound data are notoriously noisy and are both time-consuming and complex to process. Typically, the processing of speech ultrasound data requires considerable manual labor to determine the location of the tongue surface (as opposed to other reflective surfaces) at any given point during an utterance. Although tongue surface contour tracking can be semi-automated, the algorithms

that permit this generally require guide information obtained through visual inspection and manual annotation of the image by the researcher (for further description and an example, see Pouplier, 2008). This increases the potential for researcher subjectivity and error to impact the findings, and, perhaps more significantly, limits the quantity of data that can be reasonably processed. This constraint means that, although data are captured dynamically, the analysis tends to be conducted only on a single frame per token. In the present study, we were not concerned with the absolute position of the tongue, but with whether articulation varied systematically as a function of the relationship between the predicted word and the articulated word. This meant that we were able to use and extend a fully automated analysis approach that does not rely on tongue contour tracing (McMillan & Corley, 2010). This approach has previously been used to investigate motor variability during the production of tongue-twisters. It represents each token with multiple frames, allowing the dynamics of articulation to be examined and compared across conditions.

We recorded the responses of eight new participants in an experiment that was closely related to that of Drake and Corley (2014). Predictions were elicited using high-cloze sentence stems, each of which strongly predicted a specific word (cf. DeLong et al., 2005). Presentation of the sentence stems was auditory (cf. Drake & Corley, 2014; Loerts, Stowe, & Schmidt, 2013); following each stem (e.g., *when we want water we just turn on the . . .*), participants named a picture that either matched the predicted word (TAP) or differed in onset (CAP), in a fully counterbalanced design. We used pictures because it has been suggested that written words have privileged access to articulation (e.g., Costa, Alario, & Caramazza, 2005). We anticipated that, in cases in which participants were anticipating *tap* but naming a CAP, activation of the speech-motor system related to prediction would affect the articulation of *cap*, such that its onset would be “less /k/-like” than when *cap* was predicted (*On his head he wore the school . . .*). To investigate this question, we measured the articulations of the same picture names when they were accompanied by no sentence stem, and therefore there was no potential interference from a predicted word. By calculating the differences between the articulations of picture names in the experimental and control conditions, we were able to establish whether articulation varied more from the control when participants anticipated that they would hear a mismatching word than when a matching word was predicted. By calculating the degree of movement over time in the matching and mismatching conditions, we were able to investigate whether one experimental condition would generate more movement than the other during specific periods of articulation.

Method

Participants

Eight participants (seven female, one male) between 21 and 40 years of age took part in the study. All participants were monolingual speakers of English, had normal or corrected-to-normal vision, and reported no positive history of hearing or speech-language difficulties. Participants were recruited from research pools at Queen Margaret University and the University of Edinburgh, were paid for their participation, and gave written informed consent in line with British Psychological Society guidelines. The study was granted ethical approval by the Psychology Research Ethics Committee of the University of Edinburgh (Approval No. 14-1213/1).

Materials

Twelve pictures were used as the experimental items, and a further two pictures were used as practice items. We selected experimental pictures so that the picture names were single-syllable and represented the six rimes /-æn, -æp, -eɪp, -eɪk, -əʊn, -əʊst/, each paired once with the onset /t-/ and once with the onset /k-/ (i.e., *can, tan, cap, tap, cape, tape, cake, take, cone, tone, coast, and toast*). For each picture, we generated three sentence stems that each predicted that picture name as their high-cloze final item (all cloze likelihoods $\geq .8$ on pretest). The sentence cloze probability was determined via an online pretest involving ten participants who did not take part in the main experiment. The sentence stems were presented auditorily, and participants were instructed to type in the word that they felt best completed the sentence. Typed responses were coded as either “target” (the intended high-cloze word) or “other,” and only sentence stems that elicited the target response from at least eight of the ten participants were included in the main experiment. The auditory sentence-stem stimuli were recorded as they were spoken by a female speaker of British English, who was a trained phonetician. The sentences were recorded complete with the predictable final words in order to achieve typical prosodies. The final words were subsequently excised from the recordings, to produce 36 sentence stems (mean speaking rate = 3.92 syllables/s; mean sentence-stem duration = 3.10 s, range = 1.90–5.29 s; see [Appendix](#) for the full list of experimental sentence stems). The final, high-cloze item was omitted from all sentence-stem recordings.

Procedure

Participants wore an ultrasound probe throughout the experiment. The probe was secured directly against the undersurface of the chin using a proprietary helmet (Articulate Instruments 2008: www.articulateinstruments.com/ultrasound-products/). This allowed us to record the movement of the tongue

within the oral cavity during each trial (the tongue is the key active supralaryngeal articulator). Ultrasound images were recorded at a rate of ~30 fps, with acoustic data being simultaneously captured via Articulate Assistant Advanced (Articulate Instruments 2012; for details, see Wrench & Scobbie, 2008).

The experiment was presented on a laptop using the DMDX software (Forster & Forster, 2003). Participants were trained on the correct name for each picture prior to the experiment, to ensure that any articulatory differences could be ascribed to competition between the predicted word and the picture name, rather than to uncertainty concerning the name of the picture. All participants named the pictures with 100 % accuracy by the end of the training phase (which consisted of three exposures to each picture).

In all blocks, trial presentation was randomized via the presentation software. In the first experimental block, participants named each picture aloud once. They viewed a fixation point in the center of the screen for 2.9 s immediately prior to the presentation of each picture to be named, and were instructed to name the pictures as soon as they could, but to make sure that their naming was accurate.

In Blocks 2 and 3, the participants again viewed a fixation point immediately prior to the presentation of each picture for naming, but this time while listening to an auditory sentence stem. In all trials the picture was presented immediately after the end of the auditory sentence stem. The sentence stems and pictures were paired together within trials, so that on half of the trials the sentence stem predicted the picture name (i.e., *match condition*: e.g., *on his head the boy wore the school . . . CAP*). On the other half of the trials, the sentence stem predicted a name that rhymed with the name of the picture presented for naming (i.e., *mismatch condition*: e.g., *Jimmy used a washer to fix the drip from the old leaky . . . CAP*). All sentence stems were presented once in each experimental condition. The condition in which a sentence stem was first encountered was counterbalanced across participants, who encountered equal numbers of match and mismatch trials in each of Blocks 2 and 3.

Block 4 was identical to Block 1 (i.e., simple picture naming following a fixation point). The trials from Blocks 1 and 4 formed the control condition. Each participant named each picture eight times in total (twice in the control condition, three times in the match condition, and three times in the mismatch condition). In all blocks, participants followed the same instructions: to name the picture as quickly and accurately as they could. Including setup, the experiment lasted approximately 30 min.

Data treatment and analysis approach

The ultrasound data for each token recorded consisted of a sequence of black-and-white video frames. Each frame

contained 141,824 pixels, which ranged in luminance from 0 (black) to 255 (white). To make the analysis tractable, we first calculated the average luminance of each 8×8 grid of pixels, resulting in a 2,240-pixel pixelized image.

In order to analyze the pixelized ultrasound images, we first inspected the relevant audio file, independent of the visual data and blind to the experimental condition, using Audacity (2014; <http://audacity.sourceforge.net>). We identified two key moments during the participants' productions of each word: the acoustic release of the onset consonant, and the end of the vowel. The acoustic response latency was taken to be the time between stimulus (picture) presentation and the acoustic burst of the onset consonant of the picture name (i.e., the onset of a target-specific acoustic signal visible in the waveform). It was possible to determine this period for seven of the eight participants.¹ Time points based on the audio recordings were also used to select portions of each video for further analysis. Different portions of the video were used for different purposes, as will be described in the relevant sections below. Once a portion of video had been extracted, it was expanded or contracted to a standardized number of video frames, using an averaging algorithm. This allowed us to control for slight differences in video frame rates and articulation timings.

Differences between the standardized video sequences were calculated using the Delta technique (McMillan & Corley, 2010). The pixels in each frame were represented as a 2,240-dimensional vector, with each dimension taking values between 0 (black) and 255 (white). Differences between pairs of frames were calculated as the Euclidean distances between the vector representations; and differences between sequences of frames were calculated as the average by-pair Euclidean distance. This quantity, in arbitrary units, will be referred to as the *Delta distance*.

When analyzing the ultrasound video, we initially generated a data quality metric by calculating “discrimination scores” for the data recorded in each session (see the following section). These discrimination scores were then used as weighting factors in a series of regressions examining the effects of the experimental manipulations (with the consequence that observations with higher discrimination scores had more influence on the reported outcome). In our weighted statistical analyses, we first examined the degree to which participants' productions were affected by auditory sentential context, by comparing the degrees to which their experimental articulations varied from control articulations in the matching and mismatching conditions (see the Differences Between

Conditions section). Second, we examined the degree of movement over the time course of each articulation, allowing us to examine the time course of articulatory differences due to the context (see the Time Course of Differences section).

Recording quality

A problem with ultrasound recordings of articulatory movements is that they can vary greatly in quality, due to individual differences in the tongue and oral cavity, noise in the recordings, and ultrasound probe slippage, among other factors. However, such differences are difficult to detect at recording time.

In the present study, we reduced the impact of this issue by generating discrimination scores. We conceptualized recording quality as the ability to discriminate between the six different consonant–vowel (CV) onsets used throughout the present paradigm (/kæ/, /keɪ/, /kəʊ/, tæ/, /teɪ/, and /təʊ/). We used the ultrasound video beginning 0.1 s before consonant onset (acoustic burst) and ending at the offset of the steady-state vowel. The relevant section of each video was digitized and quantized into eight frames, each of which represented approximately 33 ms of recorded time. For each participant, we then created a table of the Delta distances between each possible pair of articulations. Initially, we used multidimensional scaling (Gower, 1966; Mardia, 1978) over two dimensions to visualize the relationships between a participant's recordings. For illustration, Fig. 1 shows data from the participants we judged, by visual inspection, to have produced the “best” and “worst” recordings (i.e., the least and the most noisy ones). Whereas the left-hand plot clearly shows that ultrasound analysis using the Delta approach is capable of distinguishing articulations, the right-hand plot shows that this capability is at the mercy of the noise that is inherent in ultrasound recordings.

In order to deal with this problem, we generated a discrimination score for each participant and each CV onset, designed to calculate how well a given CV, such as /keɪ/, could be discriminated from the other CVs in the experiment (here, /kæ/, /kəʊ/, tæ/, /teɪ/, and /təʊ/). These calculations were based on articulations from the control conditions only, since we predicted additional variability in articulation in the experimental conditions. Using the tables of Delta distances calculated above, we divided the mean distance between the control articulations of words that *did not* share a given CV onset by the mean distance between words that *did* share that onset. The more discriminable the words sharing an onset were from the other words, the higher was the discrimination score. The discrimination scores ranged from 1.25 to 2.43 ($M = 1.62$, $SD = 0.29$). Table 1 shows the discrimination scores calculated for the participants shown in Fig. 1, which include the highest and lowest scores obtained.

¹ The onset of picture presentation was determined as being the point at which the acoustic presentation of the sentence stem stopped. In the case of the one participant excluded from the response time data, the recording of the sentence-stem presentation was not loud enough to permit reliable annotation.

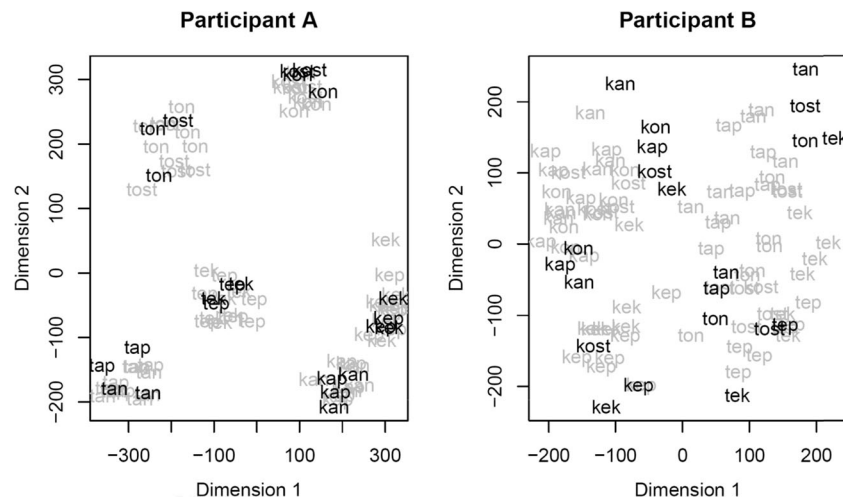


Fig. 1 Multidimensional scaling of Delta differences between all articulations produced by each of two participants, measured from 0.1 s before consonant onset to vowel offset. The plots show (left) that the Delta technique is highly capable of distinguishing articulations, but

that (right) ultrasound recordings can be subject to noise. Words in black represent recordings from the control condition, and words in gray correspond to the experimental conditions

Analyses of the treated data were all conducted using linear mixed-effects models with maximally specified random-effect structures (following Barr, Levy, Scheepers, & Tily, 2013).

the calculation of discrimination scores described in the Method section above).

Results

Individual audio and ultrasound recordings were obtained of each participant’s articulatory movements during each trial and were digitized to video. Each participant produced 96 picture names (24 in the control condition and 72 in the experimental conditions). Of the resulting 768 recordings, 27 (3.5 %) were discarded because of failures either to record audio or to properly register the ultrasound. We found no difference between the conditions in the proportions of recordings removed [$\chi^2(2) = 2.62, p = .27$]. The remaining 741 recordings were used in all subsequent analyses (including

Response latencies

When participants named pictures in the match condition (mean RT = 515 ms, SE = 11 ms), the acoustic burst occurred sooner than in the mismatch condition (mean RT = 632 ms, SE = 12 ms) or the control condition (mean RT = 606 ms, SE = 15 ms). We conducted a mixed-effect regression analysis of the effect of sentential context (match, mismatch, or control) on RTs. The model included both intercepts and slopes, which could vary by participant and by picture name. The random effects for intercepts and slopes were allowed to correlate. This constitutes the maximal justified random-effect structure, in line with recent recommendations for confirmatory hypothesis testing (Barr et al., 2013). Using orthogonal contrasts, the model confirmed that the response latencies in the match condition were significantly shorter than in the mismatch and control conditions ($\beta = 110, SE = 28, t = 3.96$), and that RTs did not differ between the mismatch and control conditions ($\beta = 20, SE = 31, t = 0.63$). This pattern replicated the patterns for the relevant conditions reported in Drake and Corley (2014).

Table 1 Discrimination scores by consonant–vowel (CV) onset for Participants A and B (see also Fig. 1). Scores are calculated from the control articulations only and represent the degree to which articulation of a given CV can be distinguished from other CVs in the ultrasound recordings

| CV | CV(IPA) | Participant A | Participant B |
|----|---------|---------------|---------------|
| ke | (/keɪ/) | 2.39 | 1.3 |
| ko | (/kəʊ/) | 2.43 | 1.42 |
| ka | (/kæ/) | 2.32 | 1.32 |
| te | (/teɪ/) | 2.00 | 1.36 |
| to | (/təʊ/) | 1.97 | 1.28 |
| ta | (/tæ/) | 2.14 | 1.25 |

Ultrasound analysis

All regression models reported here were weighted (Carroll & Ruppert, 1988) using the CV-specific discrimination scores described in the Method section. To avoid misrepresenting the effective power of the experiment, discrimination scores were scaled to a geometric mean of 1. This allowed recordings that were better able to capture relevant differences between

the control articulations to have greater influence on the outcomes of the analyses, without arbitrarily excluding recordings that may have been of poorer quality. In this context, it should be noted that the discrimination measure is independent of within-cell variation about the mean (correlation: $r = -.01$).

Experimental findings: differences between conditions

The effect of context on articulation was investigated by comparing articulations in the experimental conditions to the reference articulations from the control condition. Here, we were primarily interested in the production of the onset consonant /k/ or /t/, since the vowels in the picture names never differed from the vowels predicted by context. Accordingly, we extracted ultrasound video starting half a second before the consonant onset and ending at the consonant release (approximately 17 frames of video at 30 fps). All recordings were averaged into 17 frames; for each participant, we then proceeded as follows. First, we created participant-specific reference articulations of the onset consonants /k/ and /t/, by averaging the luminance of each of the 2,240 pixels, frame by frame, for all of the 17-frame sequences representing control articulations of words beginning with /k/ or /t/. We then calculated a Delta score for each individual articulation produced in the experimental conditions, representing the (mean frame-by-frame Euclidean) difference between a particular onset articulation and that participant's mean control articulation of the same onset (see McMillan & Corley 2010).²

The Delta scores thus obtained were subjected to a mixed-effect regression analysis, examining the effects of onset (/k/ or /t/) and of context (match or mismatch) on deviance from the mean control articulation. Together with these fixed effects and their interaction, our model included intercepts that could vary randomly by participant and by picture name. The slopes associated with each fixed effect and the interaction could vary by participant, and the slope associated with context could vary by picture name. Random effects for the intercepts and slopes were allowed to correlate. This model therefore included the maximal justified random-effect structure (Barr et al., 2013). Predictors were centered about their means prior to analysis. We considered that the coefficients differed reliably from zero where $|t| > 2$. Because our conclusions were based on model coefficients, we fit models using restricted maximum likelihood, to reduce the probability of Type I errors.

² Due to the nature of ultrasound recordings, a number of pixels in each frame are more-or-less randomly gray. However, pixels at clear physiological junctures tend to be more deterministically colored, and there are likely to be similarities in the luminance patterns across frames for similar tongue positions within a given speaker. Similarities between pixels will tend to reduce Delta values, allowing us to distinguish signal from noise.

The discrimination-score-weighted regression showed a numerical tendency for participants to produce /t/ onsets that differed more from the participant-specific /t/ controls than their /k/ productions differed from the /k/ controls, but this trend failed to reach significance ($\beta = 9.98$, $t = 1.73$). Participants were, however, affected by sentential contexts, such that onsets produced in the mismatching condition differed more from their controls than did those produced in the matching condition ($\beta = 10.89$, $t = 2.15$). The effects of context did not differ by onset consonants ($t = 0.58$).³ Table 2 gives full details of the regression model.

Experimental findings: time course of differences

In order to investigate the time course of articulation, we extracted standardized ultrasound videos corresponding to the period from 1 s before consonant onset to consonant release (approximately 32 frames of video at 30 fps). Using the same vectorizations as for the Delta calculations, we then calculated Euclidean distances between successive frames of the standardized ultrasound video, producing a sequence of 31 interframe values that represented the moment-by-moment degree of movement for a particular articulation. These values were related to “speed” of articulatory movement rather than “velocity,” since they did not include information on the direction of movement. However, it was possible to determine at which points in time participants' tongues tended to be moving quickly between frames, and at which points they were more stationary; these values were then used to provide plots of the speed of tongue movements over time in different experimental conditions.

The Euclidean distances between successive frames of ultrasound were compared by experimental condition, using a series of mixed-effect regression analyses investigating the effects of onset, context, and their interaction at each time point. The models were fit using restricted maximum likelihood and included maximally justified random effects, as we described in the previous section, and were weighted by the CV-specific difference scores. Effects were considered reliable when $|t| > 2$.

No interactions were apparent between onset and context at any time point. Effects of onset were found from approximately -217 to -117 ms and from -83 to -50 ms, reflecting more frame-to-frame movement for /k/ at the earlier epoch and more movement for /t/ just prior to consonant release. Effects of context were found from approximately -483 to -283 ms and from -50 to -17 ms, in each case reflecting more frame-to-frame movement in the mismatch condition.

³ Regression without weights showed the same general pattern of results, although the difference between onsets did reach significance: /t/s differed more from their controls than did /k/s ($\beta = 10.99$, $t = 2.08$), mismatching onsets differed more than matching onsets ($\beta = 10.94$, $t = 2.08$), and there was no interaction ($t = 0.64$).

Table 2 Differences between conditions: Details of the Context × Onset model

| | Effect in Delta (<i>SE</i>) |
|------------------------------|-------------------------------|
| (Intercept) | 286.87 (8.99) |
| Context (match vs. mismatch) | 10.89 (5.05) |
| Onset (/k/ vs. /t/) | 9.98 (5.76) |
| Context × Onset | 5.62 (9.63) |
| | Value |
| AIC | 5,930.20 |
| BIC | 6,008.07 |
| Log likelihood | −2,947.20 |
| Deviance | 5,894.20 |
| Num. observed | 559 |
| Num. groups: Word | 12 |
| Num. groups: Subject | 8 |
| Variance: Intercept/word | 39.16 |
| Variance: Context/word | 63.73 |
| Variance: Intercept/subject | 589.61 |
| Variance: Context/subject | 39.37 |
| Variance: Onset/subject | 38.03 |
| Variance: C×O/subject | 82.64 |
| Variance: Residual | 2,159.63 |

Because of the cumulative risk of a Type I error associated with multiple independent tests of this nature, we did not consider the isolated significant differences further, but instead focused on early time points when there were clusters of differences. Figure 2 illustrates the differences between conditions over the time course of articulation.

Discussion

We recorded ultrasound images of tongue movements while participants named pictures. In one experimental condition, the name of the picture matched the most likely continuation of a spoken sentence stem that the participant had just heard; in another, the picture name was a mismatching name that began with a different consonant. We used two approaches to compare articulation across the matching and mismatching conditions, and found in both cases that articulation prior to the consonant release differed between conditions. The by-condition difference confirms that predictions made as a listener can affect production. More specifically, the finding demonstrates that prediction from another’s speech affects the motor execution of one’s own speech, suggesting that top-down prediction can involve simulation of the motor activity involved in speech production.

In the first analysis, we compared summarized articulatory movements directly, and found that participants’ articulations of word onsets differed more from their average articulations in a control condition when a mismatching word was predicted than when the prediction was of a matching word. One potential account of this process might be that, as compared to the faster matching condition, articulation was simply slowed in the mismatching conditions. Under this hypothesis, the apparent differences in variability would, in fact, be due to differences in the timings of articulatory gestures. However, two aspects of the data militate against this view. The first is that the differences in naming latencies between the match, mismatch, and control conditions did not align with the differences in articulations. Both the mismatch and control conditions resulted in slower naming latencies than the match

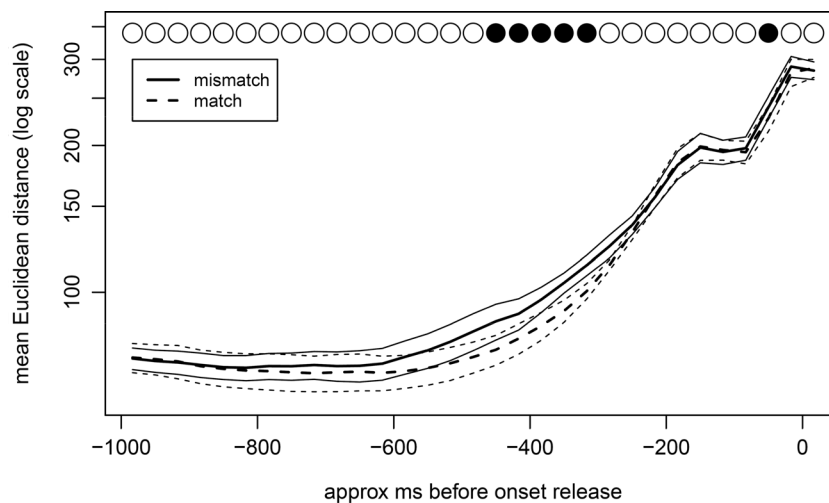


Fig. 2 Articulatory movement over time when producing the onsets of picture names that match or do not match predictions from the sentence stem. Time 0 represents the release of the /t/ or /k/ onset. Lines show the Euclidean distances between vectors of pixel intensity for successive (normalized) frames of the ultrasound video, together with by-

participants standard errors. The y-axis is log-scaled to help with viewing the differences. Filled circles correspond to transitions at which a significant difference is visible (at $|t| > 2$ for mixed models weighted by discrimination scores, with participants and words as random effects) between the mismatched and matched onset productions

condition. If articulation speed explained the differences in the first analysis, then the mismatch articulation should have been more similar to the control than to the match articulation: In fact, the opposite was the case.

The second argument against an account based on speed of articulation comes from the second analysis, in which we inspected the frame-by-frame degree of movement involved in each articulation. We achieved this by measuring the differences between consecutive frames of each ultrasound video in the experimental conditions. The resultant measurements encompassed the second or so leading up to the acoustic release of each onset consonant, and showed that where there were differences between conditions, the mismatch condition showed greater movement. Again, these findings are inconsistent with the view that differences in the mismatch condition can be ascribed to generally slower articulatory movements. Taken together, the analyses provide *prima facie* evidence for an influence of linguistic prediction on the manner, rather than the timing, of articulatory movements when the person making the prediction has to speak.

The time-course analysis additionally reveals that the period during which the frame-to-frame change was greater in the mismatch condition was relatively early in the articulatory gesture, at around 500–300 ms before the onset of the picture name. After this period, the articulatory trajectories in the two conditions converged and were statistically indistinguishable by 280 ms prior to the acoustic release, and for the remainder of the articulation. This reflects the facts that the onset of articulatory movement in the mismatch condition occurred significantly earlier in relation to acoustic release than in the match condition, and that less articulatory movement was required overall to achieve the acoustic target in the match than in the mismatch condition. In other words, articulation was ultimately more efficient in the match than in the mismatch condition. To produce words that mismatch a prediction generated as a listener is not only more demanding at a cognitive level (pace RTs), but also more demanding at a motor-execution level.

Findings from the visuomotor control literature suggest that the “inefficiency” seen in the mismatch condition may have been due to articulation in that condition involving a movement toward the (incorrect) predicted target. Perturbation of movement toward an incorrect target has been observed to be the case when two stimuli “try to control the same speeded motor response” (Schmidt & Schmidt, 2009, p. 595; in their case, of movements with a stylus toward a location that either matched or mismatched the location of a masked prime). Because upcoming predicted lexical items can be specified at least as early as presentation of the preceding word (DeLong et al., 2005; see above), it is possible that in the present study both the predicted item and the item-to-be-named were “trying to control” the motor response. Although the analyses employed in the present study did not

allow us to directly address this possibility, it may be feasible to address the question more directly in future studies, given clarity of ultrasound recordings.

Whatever the specifics of the influences on participants’ articulations, it remains the fact that these articulations were qualitatively affected by the presence of lexical representations that were generated entirely endogenously; the “competing” predicted words were the product of the participants’ semantic prediction systems, having an endogenous rather than an exogenous origin. We were able to observe anticipatory speech–motor consequences associated with predicting from another person’s speech. To that extent, the present study has directly implicated the listener’s speech–motor system in the top-down prediction of upcoming material at the level of communicative resonance.

It appears that anticipatory activation in the speech–motor system is largely outside strategic control: Prediction of the upcoming item was not beneficial to overall performance in the experimental context, and previous work using a similar paradigm has shown that mismatching predictions do not produce temporal inhibition (Drake & Corley, 2014). Although it is likely to be automatic, the activation may be specific to situations in which the listener anticipates his or her own role as a speaker (as one does in dialogue; see Rommers, Meyer, Piai, & Huettig, 2013, for evidence that the neural processing of linguistic material differs, depending on whether or not one expects to be required to speak).

Having considered how the data inform our understanding of the issue that the study was specifically designed to address, we turn briefly to a more general issue: The time course of articulator movements in the present study strongly suggests that stimulus-related lingual movement occurs well before the acoustic response onset, at a point when cognitive processing would be expected to be ongoing. This finding is perhaps surprising, in light of psycholinguistic models of picture naming, which generally involve a sequence of at least four processes prior to the initiation of articulation (for a brief recent review, see Strijkers & Costa, 2011). According to the mappings of the time course of picture name production processes determined via meta-analyses of neuroimaging studies (Indefrey, 2011; Indefrey & Levelt, 2004; see also Laganaro, Python, & Toepel, 2013), motor programming and execution occur only in the final 150 ms prior to the acoustic onset of the target picture name. However, the present experiment indicates that articulation starts much earlier, in line with the electromyographic data presented by Riès, Legou, Burle, Alario, and Malfait (2012), which showed that speech-associated muscular innervation is observable around 380 ms prior to acoustic response onset (see also Schuhmann, Schiller, Goebel, & Sack, 2012). This study confirms the conclusion drawn from Riès et al.’s work that, if we are to further understand the processes involved in speech production, it will be necessary to consider effector activity as an important

observable outcome and time-course marker, in addition to the acoustic onset more typically used as a time-locking point.

Before concluding, a note of caution should be sounded: The generalizability of the findings reported here may be impacted by the relatively low number of participants tested. In fact, due to pragmatic difficulties with data collection, this is a common issue with speech–motor studies (comparable numbers of participants were reported by Davidson, 2005; Pouplier, 2008; Pulvermüller et al., 2006; Watkins & Paus, 2004; Watkins, Strafella, & Paus, 2003). In the case of the present study, this concern may be partially mitigated by the fact that the pattern of response latencies was in keeping with that reported by Drake and Corley (2014), whose participant numbers were in line with those typically employed in psycholinguistic research.

Given this caveat, the present study has demonstrated the importance of articulatory measurement in two ways. As we discussed above, the muscle activation and motor movements associated with articulation appear to start much earlier than has been supposed in existing psycholinguistic models. This suggests that the use of articulatory information may be important if we are to develop greater insight into the processes of speech production. For example, in previous work the present authors investigated acoustic onset times to name pictures, in a paradigm very similar to the one employed here. On the basis that no facilitatory or inhibitory effects were observed when the to-be-named picture partially overlapped with the predicted word, we concluded that “prediction during comprehension [did] not appear to occur at a phonological–articulatory level” (Drake & Corley, 2014). The present study indicates that this was far from the final word on the matter: The second consequence of using articulatory measurement is that we are now able to conclude that there clearly *is* an effect of prediction on articulation, if you know where to look.

Author note The research reported in this article was supported in part by an EPSRC Doctoral training studentship awarded to E.D. The authors know of no conflict of interest. We thank Alan Wrench (Articulate Instruments) and Steve Cowen (Queen Margaret University) for technical advice and assistance, and Sonja Schaeffler (Queen Margaret University) for useful discussions throughout the project.

Appendix

Table 3 List of experimental sentence stems by high-cloze target

| Word Predicted | Sentence Stem |
|----------------|---|
| cake | There were five tiers to the wedding . . . Jenny lit the candles on the birthday . . . Would you like a muffin or would you prefer some lemon . . . |

Table 3 (continued)

| Word Predicted | Sentence Stem |
|----------------|--|
| can | The gardener picked up the watering . . . There’s no such word as can’t. You have to believe that you . . . You can drink beer from a glass or straight from the . . . |
| cap | On his head he wore the school . . . A soft flat hat is sometimes known as a . . . Your car wheel has lost its hub . . . |
| cape | You’ll know it’s Dracula if he’s got fangs and is wearing a . . . We made a Superman outfit using blue tights and a red sheet to be the . . . He thinks he can fly when he’s wearing his Superhero . . . |
| coast | He loves sailing so they moved to the south . . . Because Britain is an island it has a very long . . . Plymouth is a lovely city on the south . . . |
| cone | She went to the van and bought an ice-cream . . . Would you like a lolly or would you prefer an ice-cream . . . During the roadworks the central reservation was marked out by . . . |
| take | Some people like to give but others always . . . The secret to a happy marriage is a bit of give and . . . We’re running out of film. We’ll try to film the whole scene in a single . . . |
| tan | To me she looked orange but she thought she had a nice . . . She thinks that if she doesn’t use sunscreen she’ll get a better . . . Before she goes on holiday she goes for one of those spray . . . |
| tap | When we want water we just turn on the . . . Jimmy managed to fix the drip from the old leaky . . . I’d love to have a constant source of beer on . . . |
| tape | The only thing holding it all together was gaffer . . . I’m sure you can fix it with a bit of sticky . . . Before discs came in you used to have record TV programs on to video . . . |
| toast | The fire alarm’s gone off again; someone must have burnt the . . . She likes butter and jam on her . . . He asked them to raise their glasses in a . . . |
| tone | His crass jokes really lower the . . . Her voice has such a lovely . . . Type in the numbers when you hear the dial . . . |

References

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation of verbs: Restricting the domain of subsequent reference. *Cognition*, 73, 247–264. doi:10.1016/S0010-0277(99)00059-1

Articulate Instruments Ltd. (2008). *Ultrasound stabilisation headset users manual: Revision 1.4*. Edinburgh, UK: Articulate Instruments Ltd.

Articulate Instruments Ltd. (2012). *Articulate assistant advanced user guide: Version 2.14*. Edinburgh, UK: Articulate Instruments Ltd.

- Audacity. (2014). Audacity: Free audio editor and recorder (Version 2.0.0) [Computer program]. Retrieved April 20, 2014, from <http://audacity.sourceforge.net/>
- Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*, 255–278. doi:10.1016/j.jml.2012.11.001
- Carroll, R. J., & Ruppert, D. (1988). *Transformation and weighting in regression* (Vol. 30). New York, NY: Chapman & Hall.
- Costa, A., Alario, F.-X., & Caramazza, A. (2005). On the categorical nature of the semantic interference effect in the picture–word interference paradigm. *Psychonomic Bulletin & Review*, *12*, 125–131. doi:10.3758/BF03196357
- Damian, M. F., & Dumay, N. (2007). Time pressure and phonological advance planning in spoken production. *Journal of Memory and Language*, *57*, 195–209.
- Dang, J., Wei, J., Suzuki, T., Honda, K., Perrier, P., & Honda, M. (2004). Investigation and modeling of coarticulation in speech production. In *2004 International Symposium on Chinese Spoken Language Processing* (pp. 25–28). Piscataway, NJ: IEEE Press.
- Davidson, L. (2005). Addressing phonological questions with ultrasound. *Clinical Linguistics and Phonetics*, *19*, 619–633.
- DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, *8*, 1117–1121. doi:10.1038/nl1504
- Dikker, S., & Pytkänen, L. (2013). Predicting language: MEG evidence for lexical preactivation. *Brain and Language*, *127*, 55–64.
- Drake, E., & Corley, M. (2014). Effects in production of word pre-activation during listening: Are listener-generated predictions specified at a speech-sound level? *Memory & Cognition*, *43*, 111–120. doi:10.3758/s13421-014-0451-9
- Fadiga, L., Craighero, L., Buccino, G., & Rizzolatti, G. (2002). Speech listening specifically modulates the excitability of tongue muscles: A TMS study. *European Journal of Neuroscience*, *15*, 399–402.
- Federmeier, K. D. (2007). Thinking ahead: The role and roots of prediction in language comprehension. *Psychophysiology*, *44*, 491–505. doi:10.1111/j.1469-8986.2007.00531.x
- Federmeier, K. D., & Kutas, M. (1999). A rose by any other name: Long-term memory structure and sentence processing. *Journal of Memory and Language*, *41*, 469–495. doi:10.1006/jmla.1999.2660
- Federmeier, K. D., McLennan, D. B., De Ochoa, E., & Kutas, M. (2002). The impact of semantic memory organization and sentence context information on spoken language processing by younger and older adults: An ERP study. *Psychophysiology*, *39*, 133–146.
- Fischer, M. H., & Zwaan, R. A. (2008). Embodied language: A review of the role of the motor system in language comprehension. *Quarterly Journal of Experimental Psychology*, *61*, 825–850. doi:10.1080/17470210701623605
- Forster, K. I., & Forster, J. C. (2003). DMDX: A Windows display program with millisecond accuracy. *Behavior Research Methods, Instruments, & Computers*, *35*, 116–124. doi:10.3758/BF03195503
- Fowler, C. A. (2014). Talking as doing: Language forms and public language. *New Ideas in Psychology*, *32*, 174–182.
- Gambi, C., & Pickering, M. J. (2013). Talking to each other and talking together: Joint language tasks and degrees of interactivity. *Behavioral and Brain Sciences*, *36*, 423–424.
- Garnier, M., & Henrich, N. (2014). Speaking in noise: How does the Lombard effect improve acoustic contrasts between speech and ambient noise? *Computer Speech and Language*, *28*, 580–597.
- Goffman, L., Smith, A., Heisler, L., & Ho, M. (2008). The breadth of coarticulatory units in children and adults. *Journal of Speech, Language, and Hearing Research*, *51*, 1424–1437. doi:10.1044/1092-4388(2008/07-0020)
- Gower, J. C. (1966). A Q-technique for the calculation of canonical variates. *Biometrika*, 588–590.
- Hardcastle, W. J., & Hewlett, N. (Eds.). (1999). *Coarticulation: Theory, data and techniques*. Cambridge, MA: Cambridge University Press.
- Hauk, O., Johnsrude, I., & Pulvermüller, F. (2004). Somatotopic representation of action words in human motor and premotor cortex. *Neuron*, *41*, 301–307. doi:10.1016/S0896-6273(03)00838-9
- Huber, S., & Krist, H. (2004). When is the ball going to hit the ground? Duration estimates, eye movements, and mental imagery of object motion. *Journal of Experimental Psychology: Human Perception and Performance*, *30*, 431–444. doi:10.1037/0096-1523.30.3.431
- Indefrey, P. (2011). The spatial and temporal signatures of word production components: A critical update. *Frontiers in Psychology*, *2*, 255. doi:10.3389/fpsyg.2011.00255
- Indefrey, P., & Levelt, W. J. (2004). The spatial and temporal signatures of word production components. *Cognition*, *92*, 101–144. doi:10.1016/j.cognition.2002.06.001
- Kamide, Y., Altmann, G. T. M., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language*, *49*, 133–156. doi:10.1016/S0749-596X(03)00023-8
- Katz, W. F. (2000). Anticipatory coarticulation and aphasia: Implications for phonetic theories. *Journal of Phonetics*, *28*, 313–334.
- Laganaro, M., Python, G., & Toepel, U. (2013). Dynamics of phonological–phonetic encoding in word production: Evidence from diverging ERPs between stroke patients and controls. *Brain and Language*, *126*, 123–132. doi:10.1016/j.bandl.2013.03.004
- Loerts, H., Stowe, L. A., & Schmidt, M. S. (2013). Predictability speeds up the re-analysis process: An ERP investigation of gender agreement and cloze probability. *Journal of Neurolinguistics*, *26*, 561–580.
- Lubker, J. (1981). Temporal aspects of speech production: Anticipatory labial coarticulation. *Phonetica*, *38*, 51–65.
- Lupker, S. J. (1982). The role of phonetic and orthographic similarity in picture–word interference. *Canadian Journal of Psychology*, *36*, 349–367. doi:10.1037/h0080652
- Mardia, K. V. (1978). Some properties of classical multi-dimensional scaling. *Communications in Statistics: Theory and Methods*, *7*, 1233–1241.
- McMillan, C. T., & Corley, M. (2010). Cascading influences on the production of speech: Evidence from articulation. *Cognition*, *117*, 243–260.
- Meyer, A. S., & Schriefers, H. (1991). Phonological facilitation in picture–word interference experiments: Effects of stimulus onset asynchrony and types of interfering stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*, 1146–1160. doi:10.1037/0278-7393.17.6.1146
- Miall, R. C., & Reckess, G. Z. (2002). The cerebellum and the timing of coordinated eye and hand tracking. *Brain and Cognition*, *48*, 212–226.
- Miall, R. C., & Wolpert, D. M. (1996). Forward models for physiological motor control. *Neural Networks*, *9*, 1265–1279.
- Mitra, V., Nam, H., Espy-Wilson, C. Y., Saltzman, E., & Goldstein, L. (2011). Articulatory information for noise robust speech recognition. *IEEE Transactions on Audio, Speech and Language Processing*, *19*, 1913–1924.
- Neiberg, D., Ananthakrishnan, G., & Engwall, O. (2008). The acoustic to articulation mapping: Non-linear or non-unique? In *INTERSPEECH 2008: 9th Annual Conference of the International Speech Communication Association* (pp. 1485–1488). Retrieved from <http://dblp1.uni-trier.de/db/conf/interspeech/interspeech2008.html>
- Pianesi, F. (2007). Temporal reference. In M. Everaert & H. van Riemsdijk (Eds.), *The Blackwell companion to syntax* (Vol. 5, pp. 94–136). Malden MA: Blackwell. doi:10.1002/9780470996591.ch72
- Pickering, M. J., & Garrod, S. (2007). Do people use language production to make predictions during comprehension? *Trends in Cognitive Sciences*, *11*, 105–110. doi:10.1016/j.tics.2006.12.002

- Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, *36*, 329–347. doi:10.1017/S0140525X12001495
- Poupplier, M. (2008). The role of a coda consonant as error trigger in repetition tasks. *Journal of Phonetics*, *36*, 114–140.
- Pulvermüller, F., Huss, M., Kherif, F., del Prado Martin, F. M., Hauk, O., & Shtyrov, Y. (2006). Motor cortex maps articulatory features of speech sounds. *Proceedings of the National Academy of Sciences*, *103*, 7865–7870.
- Recasens, D., Pallarès, M. D., & Fontdevila, J. (1997). A model of lingual coarticulation based on articulatory constraints. *Journal of the Acoustical Society of America*, *102*, 544–561. doi:10.1121/1.419727
- Riès, S., Legou, T., Burle, B., Alario, F.-X., & Malfait, N. (2012). Why does picture naming take longer than word reading? The contribution of articulatory processes. *Psychonomic Bulletin & Review*, *19*, 955–961. doi:10.3758/s13423-012-0287-x
- Rommers, J., Meyer, A. S., Piai, V., & Huettig, F. (2013). *Constraining the involvement of language production in comprehension: A comparison of object naming and object viewing in sentence context*. Paper presented at the 19th Annual Conference on Architectures and Mechanisms for Language Processing [AMLaP 2013], Marseille, France.
- Rommers, J., Meyer, A. S., Praamstra, P., & Huettig, F. (2013). The contents of predictions in sentence comprehension: Activation of the shape of objects before they are referred to. *Neuropsychologia*, *51*, 437–447.
- Rothermich, K., & Kotz, S. A. (2013). Predictions in speech comprehension: fMRI evidence on the meter–semantic interface. *NeuroImage*, *70*, 89–100.
- Schiller, N. O., Horemans, I., Ganushchak, L., & Koester, D. (2009). Event-related brain potentials during the monitoring of speech errors. *NeuroImage*, *44*, 520–530.
- Schmidt, T., & Schmidt, F. (2009). Processing of natural images is feedforward: A simple behavioral test. *Attention, Perception, & Psychophysics*, *71*, 594–606. doi:10.3758/APP.71.3.594
- Schuhmann, T., Schiller, N. O., Goebel, R., & Sack, A. T. (2012). Speaking of which: Dissecting the neurocognitive network of language production in picture naming. *Cerebral Cortex*, *22*, 701–709.
- Scott, S. K., McGettigan, C., & Eisner, F. (2009). A little more conversation, a little less action—Candidate roles for the motor cortex in speech perception. *Nature Reviews Neuroscience*, *10*, 295–302.
- Severens, E., Ratinckx, E., Ferreira, V. S., & Hartsuiker, R. J. (2008). Are phonological influences on lexical (mis)selection the result of a monitoring bias? *Quarterly Journal of Experimental Psychology*, *61*, 1687–1709. doi:10.1080/17470210701647422
- Stone, M. (2005). A guide to analysing tongue motion from ultrasound images. *Clinical Linguistics and Phonetics*, *19*, 455–501.
- Strijkers, K., & Costa, A. (2011). Riding the lexical speedway: A critical review on the time course of lexical selection in speech production. *Frontiers in Psychology*, *2*, 356. doi:10.3389/fpsyg.2011.00356
- Tettamanti, M., Buccino, G., Saccuman, M. C., Gallese, V., Danna, M., Scifo, P., ... Perani, D. (2005). Listening to action-related sentences activates fronto-parietal motor circuits. *Journal of Cognitive Neuroscience*, *17*, 273–281. doi:10.1162/0898929053124965
- Tilsen, S. (2007). Vowel-to-vowel coarticulation and dissimilation in phonemic-response priming. In *UC Berkeley Phonology Lab 2007 Annual Report*, 416–458. Retrieved from <https://escholarship.org/uc/item/8971k261>
- Verfaillie, K., & Daems, A. (2002). Representing and anticipating human actions in vision. *Visual Cognition*, *9*, 217–232.
- Watkins, K., & Paus, T. (2004). Modulation of motor excitability during speech perception: The role of Broca's area. *Journal of Cognitive Neuroscience*, *16*, 978–987. doi:10.1162/0898929041502616
- Watkins, K. E., Strafella, A. P., & Paus, T. (2003). Seeing and hearing speech excites the motor system involved in speech production. *Neuropsychologia*, *41*, 989–994.
- Whalen, D. H. (1990). Coarticulation is largely planned. *Journal of Phonetics*, *18*, 3–35.
- Willems, R. M., & Hagoort, P. (2007). Neural evidence for the interplay between language, gesture, and action: A review. *Brain Language*, *101*, 278–289.
- 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. doi:10.1038/nn1263
- Wilson, M., & Wilson, T. P. (2005). An oscillator model of the timing of turn-taking. *Psychonomic Bulletin & Review*, *12*, 957–968. doi:10.3758/BF03206432
- Wrench, A. A., & Scobbie, J. M. (2008). *High-speed cine-loop ultrasound vs. video ultrasound tongue imaging: Comparison of front and back lingual gesture location and relative timing*. Paper presented at the Eighth International Seminar on Speech Production (ISSP), Strasbourg, France.