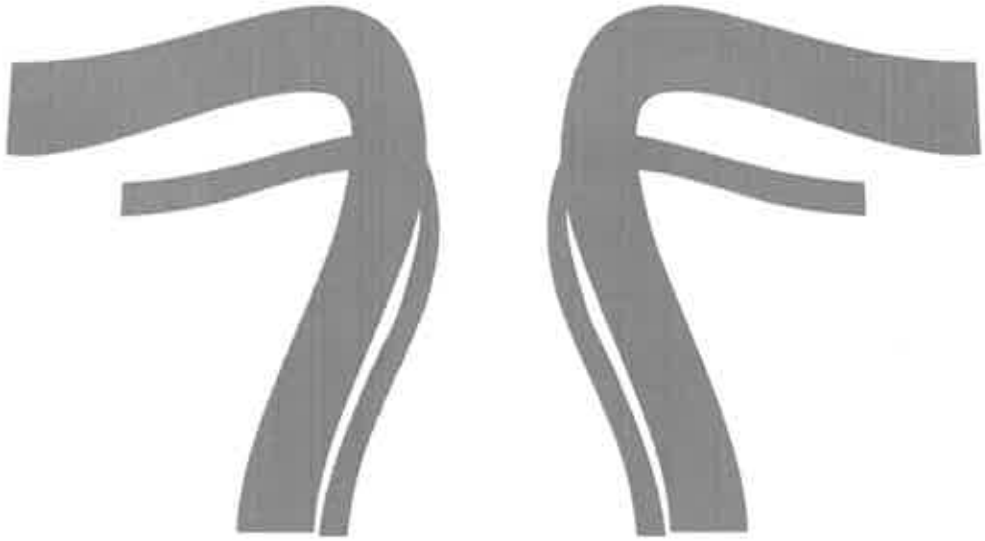
Proceedings of the 10th International Seminar on Speech Production (ISSP)

5 – 8 May 2014
Cologne, Germany



www.issp2014.uni-koeln.de

Edited by

Susanne Fuchs, Martine Grice, Anne Hermes, Leonardo Lancia, Doris Mücke

Articulatory Effects of Prediction During Comprehension: An Ultrasound Tongue Imaging Approach

Eleanor Drake[1], Sonja Schaeffler[2], Martin Corley[1]

[1]*Department of Psychology, University of Edinburgh, Edinburgh, UK*
[2]*Clinical Audiology, Speech and Language Research Centre, Queen Margaret University, Edinburgh, UK*
E.K.E.Drake@sms.ed.ac.uk, sschaeffler@qmu.ac.uk, Martin.Corley@ed.ac.uk

Abstract

We investigated whether effects of prediction during spoken language comprehension are observable in speech-motor output recorded via ultrasound tongue imaging: Predicted words can be specified at a phonological level during reading comprehension, and listening to speech activates speech-motor regions. It has been suggested that speech-motor activation may occur during prediction of upcoming material (Pickering & Garrod, 2007). Speakers model their own upcoming speech, with the effects being observable at an articulatory level in the form of anticipatory co-articulation. We investigated whether the effects of prediction as a listener can also be observed at an articulatory level. We auditorily presented high-cloze sentence-stems, immediately followed by presentation of a picture for naming. Picture names either fully matched the omitted sentence-cloze item or mismatched it at onset (e.g., TAP-"cap"). By-condition differences in picture-name articulation indicated that prediction of upcoming material during speech listening can engage speech-motor processes.

Keywords: ultrasound tongue imaging, prediction, Delta-technique, motor-speech

1. Introduction

Language comprehension involves the prediction of upcoming material, in addition to the processing of perceptually available input (e.g., Altmann & Kamide, 1999; DeLong et al., 2005; Federmeier, 2007). Listening to speech activates neural regions associated with motor-speech planning and execution, and modulates tongue muscle excitability (Fadiga et al. 2002; Pulvermüller et al., 2006; Watkins & Paus, 2004; Wilson et al., 2004). Such communicative resonance may reflect involvement of the speech-motor regions in prediction during comprehension: It has been suggested that the synthesis process involved in predicting another's upcoming speech may recruit mechanisms more typically associated with speech production (e.g., Pickering & Garrod, 2007).

ERP evidence demonstrates that, for written input, predicted linguistic material can be specified at a phonological level during reading comprehension (DeLong et al., 2005). However, it is unclear if, and at what levels, such prediction might involve the infrastructure and mechanisms of speech production. Prediction during language comprehension is performed incrementally. Therefore, when investigating prediction of spoken language, some spoken language must be presented to the listener prior to the critical manipulation point. Under such circumstances it appears not to be possible to distinguish the neural activation posited to be associated with the synthesis of upcoming input (i.e. prediction) from that associated with the analysis of perceptual input.

In the current study we instead employ an articulatory imaging technique to investigate the effects of prediction as a listener on motor-speech activity itself. When a speaker predicts their own upcoming speech output this is observable at an articulatory level in the phenomenon of anticipatory co-articulation (e.g., Farnetani & Recasens, 1997). Anticipatory co-articulation can be observed via ultrasound tongue imaging (e.g., Zharkova & Hewlett, 2011). We investigated whether ultrasound tongue imaging (henceforth UTI) would reveal articulatory-level effects of prediction during comprehension. We reasoned that if the speech-motor system is activated during prediction as a listener, effects of prediction might be observable in articulation when the listener speaks.

Participants named pictures in three contexts whilst their articulatory movements were recorded via UTI. In the Control condition a visual fixation point was presented for 3 seconds prior to picture presentation; in the Match condition, a sentence-stem predicting the upcoming picture-name was auditorily presented immediately prior to picture presentation; in the Mismatch condition a sentence-stem predicting a rime-partner of the upcoming picture-name was auditorily presented immediately prior to picture presentation. In this way the experimental conditions differ only in whether the auditory linguistic context predicted the target picture-name or an alternative picture-name. Any by-condition differences in picture-name realizations (articulation) must therefore reflect an effect of prediction as a listener.

2. Method

2.1. Participants

Participants (1 male, 7 female) were monolingual speakers of English, had no phonetic training, reported normal hearing and visual acuity, and ranged in age from 22 to 40 years. All gave informed written 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.

2.2. Materials

The picture-name set was created by pairing the consonantal onsets /k/ and /t/ with 6 VC rimes (e.g., /k/ + /æp/ → CAP; /t/ + /æp/ → TAP). Each of the 12 words generated in this way was represented by a colour picture selected from an online database (online pre-test mean picture-name agreement = .76, range = .3 to 1). All picture names were concrete nouns of medium lexical frequency (mean log10CD = 2.93, SD = 0.41, range = 2.07 -3.91; SUBTLEX-US database, Brysbaert & New, 2009). For each picture name, 3 sentence stems were generated that strongly predicted the picture name as their final word (online pre-test minimum cloze probability > .8). The 36 sentence stems were designed each to end in a vowel or semi-vowel in order to allow audio to be cut at a

comparable and non-informative point across all stimuli. Sentence-stems were recorded as spoken by a native female speaker of British English, at a mean rate of 3.92 syllables per second (mean sentence stem duration = 3.10 seconds, range = 1.90 – 5.29 seconds).

2.3. Procedure

The experiment was run at the Ultrasound Tongue Imaging suite at Queen Margaret University. The full experiment was presented on a Dell XPS 1702 laptop using DMDX presentation software (Forster & Forster, 2003). The presentation software fully randomized item presentation within blocks. Participants were familiarized with the 12 picture names prior to the beginning of the experiment, in order to ensure that they would be able to correctly name pictures during the experimental phase. During the familiarization phase each picture was presented once in each of three blocks. All participants used target names for pictures 100% accurately by the third familiarization block.

Following the third familiarization phase participants were fitted with the ultrasound helmet (used to maintain probe position throughout the experimental procedure; Scobbie et al., 2008). Participants then named pictures once more as they had in the third familiarization block, in order to acquire experience of speaking whilst wearing the ultrasound device prior to commencing the experiment proper. The pictures were then presented for naming in 3 conditions: In the Control condition pictures were presented with no auditory context, following presentation of a fixation point. In the experimental conditions (Match and Mismatch) pictures were presented immediately following auditory presentation of a high-cloze sentence-stem: In the Match condition the picture-to-be-named matched the predicted (but missing) sentence cloze word (e.g., "Jimmy fixed the drip from the old leaky" … TAP): In the Mismatch condition the picture-to-be-named differed in onset phoneme from the predicted word (e.g., "On his head he wore the school" … TAP; where the predicted word would be "cap").

Control-trial blocks were presented at the beginning and end of the experiment. Trials in the experimental conditions were presented in between the two Control blocks. Each sentence-stem was presented once in the Match condition and once in the Mismatch condition (i.e. paired once with the picture it predicted and once with the rime-pair of that picture). Each picture was presented twice in the Control condition, three times in the Match condition and three times in the Mismatch condition. The experimental design was therefore fully within-participant and within-items. Whether a given sentence-stem was first heard in a match or a non-match context was balanced across participants.

2.4. Data Capture and Processing

Using AAA software (Scobbie & Wrench, 2008) we recorded acoustic and ultrasound data for each trial: Recording started at the onset of the sentence-stem stimulus and ended once the participant had named the picture. Ultrasound data was captured via an Ultrasonix device used in conjunction with a headmounted micro-convex probe, with depth set at 80mm and angle at 150°, capturing a mid-sagittal tongue image at a rate of 100fps. Data was exported from AAA in AVI format at a rate of 30fps, following which an audio-video synchronization check was performed in VirtualDub (http://www.virtualdub.org/).

2.4.1. Audio data processing

We manually performed acoustic landmark labelling via visual inspection of the spectral signal in Audacity (http://audacity.sourceforge.net/). For each trial we identified the off-set of the sentence-stem audio, the acoustic onset of the picture name, the acoustic onset of the vowel, and the acoustic offset of the steady-state vowel. The time-points of each trial's landmarks were recorded in .csv format, allowing this information to be made available to the ultrasound video-processing software.

2.4.2. Video data processing

Each frame of ultrasound video constituted a 512 x 277 grid of pixels. Pixels ranged in luminance from 000 (black) to 255 (white). In order to achieve data tractability, we processed each frame so that luminance was averaged over blocks of 8 x 8 contiguous pixels (see McMillan & Corley, 2010). A vector was generated from each frame, with each 8 x 8 pixel block assigned a specific position in the vector. Each vector ran from bottom left to top right of the AVI frame and each pixel block was recorded by its luminance (0 to 255). Vectors formed the basis for analyses, which were performed by calculating and comparing "Delta scores"; i.e., the Euclidean distances between individual vectors (frames).

3. Analysis

In order to minimise the effects of noise in the ultrasound images (see Scobbie & Wrench, 2008) we performed a preparatory analysis in order to determine the quality of the data acquired from each participant for each CV onset. This analysis was performed on ultrasound data acquired between the acoustic burst and the end of the steady-state vowel for each token; subsequent analyses were performed on data acquired prior to the acoustic release of the onset consonant. We used multidimensional scaling (Mardia, 1978) to calculate how well the Delta scores distinguished tokens of a given CVC word from tokens of all other CVC words produced by that participant: This was achieved by determining the mean Euclidean distance of a given vector (i.e. articulation) from; (i) all vectors representing different words; (ii) all vectors representing the same word. The Discrimination score for each onset for each participant was equal to (i)/(ii). Therefore the higher the score the better the data discriminated between a given CV onset and others in the picture-name set, and the less "noisy" the data. This information was used to geometrically weight the contribution of each participant's data to subsequent analyses (Carroll & Ruppert, 1988). In this way we were able to avoid arbitrarily discarding "poor quality" data, whilst accounting for the great by-participant variability known to be associated with ultrasound articulatory data.

We used a linear mixed-modelling approach, implemented in R 3.0.2 via the lme4 package, version 0.999999-4 (Bates, Maechler, & Bolker, 2013; R Core Team, 2013). Data were weighted as described above. Condition (Match/ Mismatch) and Onset Consonant (/k//t/) were included as fixed effects, and Participant and Picture-name as random effects. Because this approach provides estimated, rather than exact, effect sizes it was not appropriate to calculate associated p-values exactly. We therefore treat $|t| > 2$ as indicating a statistically significant effect.

3.1. Location of articulation analysis

The first analysis investigated by-condition differences in data topography. Articulatory data acquired between -500 ms and 0 ms of the acoustic burst were collapsed to produce one average-luminance vector per token. This allowed each articulatory token to be compared to a reference vector for that item (picture-name). The reference vector for each item represented the mean of the vectors for that picture-name as produced in the Control condition. Vectors for all individual articulations in the Match and Mismatch conditions were compared to the relevant Control reference vector. This produced a Delta-score for each articulation (token), which indicated the distance in multi-dimensional space between that token and the participant's mean Control articulation of the relevant picture-name.

The Delta-scores were then modelled as the outcome variable in a linear mixed effects model (as detailed above). Inspection of the model indicated that Delta scores in the Mismatch condition were significantly greater than those in the Match condition ($\beta = 10.89$, $t = 2.15$). This indicates that pre-acoustic articulation was less similar to the Control condition in the Mismatch condition than in the Match condition.

3.2. Time-course analysis

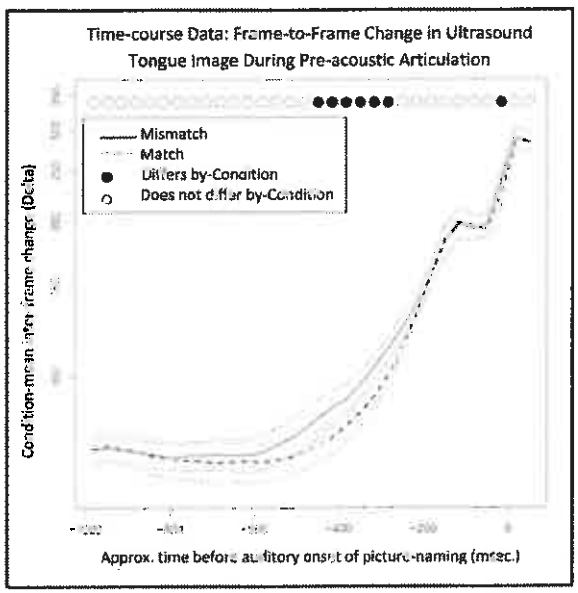


Figure 1: Frame-to-frame change in ultrasound tongue image during pre-acoustic articulation. Faint lines indicate 95% confidence intervals. Circles at top of plot indicate inter-frame intervals. Circles are filled where change differs significantly by condition (Match v. Mismatch).

The second analysis investigated by-condition differences in the articulatory time-course of the pre-acoustic articulations. Articulatory data for each token constituted all ultrasound video frames recorded from 1000 ms prior to the acoustic burst until the acoustic burst (i.e. 31 frames per token). We calculated a Delta score for each inter-frame interval of each token: In this way each Delta score indicated the Euclidean distance between the current frame and that immediately preceding it. Higher Delta scores therefore indicated greater frame-to-frame change, associated with greater change in the configuration of the tongue as indicated by the ultrasound image. Each articulatory token was represented by a series of Delta scores indicating successive frame-to-frame change within that token's data. This output was automatically averaged and plotted by-condition (Match v. Mismatch; see Fig. 1) and by onset-consonant (/k/ v /t/).

We investigated the time-course data by performing a mixed effect model analysis at each time-point (inter-frame interval). Inter-frame change (Delta) was treated as the outcome variable. In order to account for the increased risk of Type I errors associated with the use of multiple comparisons (for discussion see Lage-Castellanos et al., 2009) we treated effects as significant only when they were clustered across three or more consecutive inter-frame intervals Effects of condition were found to be statistically significant (i.e., $|t| > 2$) at all intervals from -483 to -283 ms, and consistently indicated greater frame-to-frame movement in the Mismatch condition than in the Match condition.

4. Discussion and conclusion

We reported a study in which we adapted an automated UTI analysis technique in order to investigate the effect on speech production of prediction during speech comprehension. Participants named pictures in a control condition and in two experimental conditions. The experimental conditions differed only in whether the picture name Matched or Mismatched the predicted word. Predictions were elicited via presentation of a high-cloze auditory sentence stem (i.e., via spoken language comprehension).

We applied two analysis approaches in order to investigate lingual motor-activity in the period immediately prior to the onset of acoustic information associated with picture-naming. The first approach collapsed information about the location of the tongue across time, and compared both experimental conditions to the control condition. The second approach provided information about the degree of movement observable at each inter-frame interval, and compared the two experimental conditions directly. Both approaches revealed by-condition differences in speech-motor activity, indicating that prediction during speech-comprehension produces both spatially and temporally observable effects on motor-speech output. Productions in the Mismatch condition appear to be less "canonical" than those in the Match condition (i.e., differed more from Control productions than did productions in the Match condition).

The current study demonstrates that a Delta-approach to ultrasound tongue image analysis can be adapted to be applicable beyond the paradigm for which it was initially developed (McMillan & Corley, 2010). The automated nature of the approach makes it appropriate for use in psycholinguistically-motivated studies because it reduces demands on researcher time and expertise (compared to a typical tongue-tracing approach), and allows meaningful averaging across differing items. Data-quality weighting provides a non-arbitrary approach to handling between-participant differences in noise-signal ratios, thereby extending the proportion of useable data.

The findings of the current study are novel in that they demonstrate online adaptations to motor-speech realizations arising of prediction during comprehension: Prediction-elicited representations cascaded to directly affect speech-motor production itself, rather than simply affecting the time-point or moment at which motor-execution began. Further investigation will be required to determine more exactly the nature of the information that cascades to a speech motor-execution level: If the predicted onset item itself were

activated at a motor-execution level we might expect to find that tokens produced in the Mismatch condition were more similar to their rime-partner than were tokens produced in the Match condition (e.g., articulation of TAP in the Mismatch condition would be more similar to articulation of CAP in the Control condition than would articulation of TAP in the Match condition). We did not find this to be the case for the time-frame analyzed in this study, although just such an effect has been demonstrated in tongue-twister data when applying the Delta-technique (McMillan & Corley, 2010).

It should be noted that in our analyses all data was time-locked to the acoustic onset of speech production. We adopted this approach in order to avoid finding by-condition differences simply as a function of when articulation commenced. That situation might occur under a stimulus-locked approach if motor-execution were identical across conditions but commenced later (i.e., longer after stimulus presentation) in the Mismatch condition. However, we agree with a reviewer who commented that, when exploring effects of prediction on articulation, it would be valuable to study speech-motor behaviour at the point of stimulus presentation (i.e., time-locked to picture presentation, in the case of the current study). This is an area for further development of the Delta-technique, and a spatial analysis of data time-locked to stimulus presentation might well provide valuable information regarding the exact nature of the prediction-effect demonstrated in the current study.

For the purposes of the current paper we note that although we do not report stimulus-locked data, the time-frame used in the second analysis approach includes data acquired at the point of stimulus presentation, and indeed extends further back in time to include articulatory data acquired whilst listening to the auditory material. Differences in speech-motor activity become observable as a consequence of whether or not a comprehension-elicited prediction is met. Given the nature of our stimuli, this confirms that; (i) listeners produce predictions during comprehension of spoken language presented at a typical conversational speech-rate, and; (ii) the effect of such predictions is observable in the listeners' own speech productions.

5. Acknowledgements

We thank Professor Alan Wrench (Articulate Instruments) and Steve Cowen (QMU) for technical advice and assistance. We would also like to thank two anonymous reviewers.

6. References

Altmann, G., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition, 73*(3), 247-264.

Bates, D., Maechler, M., & Dai, B. (2008). lme4: Linear mixed-effects models using S4 classes (R package version 0.999375-27). Retrieved from: *http://www. r-proiect. org.*

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*(4), 977-990.

Carroll, R. J., & Ruppert, D. (1988). "Discussion". *Technometrics, 30*(1), 30-31.

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). "Probabilistic word pre-activation during language comprehension inferred from electrical brain activity". *Nature neuroscience, 8*(8), 1117-1121.

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(2)*, 399-402.

Farnetani, E., & Recasens, D. (1997). Coarticulation and connected speech processes. In W. Hardcastle (Ed.) *The handbook of phonetic sciences*, 371-404.

Federmeier, K. D. (2007). "Thinking ahead: The role and roots of prediction in language comprehension". *Psychophysiology, 44(4)*, 491-505.

Fischer, M. H., & Zwaan, R. A. (2008). "Embodied language: a review of the role of the motor system in language comprehension". *The Quarterly Journal of Experimental Psychology, 61*(6), 825-850.

Forster, K. I., & Forster, J. C. (2003). "DMDX: A Windows display program with millisecond accuracy". *Behavior Research Methods, Instruments, & Computers, 35*(1), 116-124.

Lage-Castellanos, A., Martínez-Montes, E., Hernández-Cabrera, J. A., & Galán, L. (2010). False discovery rate and permutation test: an evaluation in ERP data analysis. *Statistics in medicine, 29*(1), 63-74.

Mardia, K. V. (1978). "Some properties of classical multi-dimesional scaling". *Communications in Statistics-Theory and Methods, 7*(13), 1233-1241.

McMillan, C. T., & Corley, M. (2010). "Cascading influences on the production of speech: Evidence from articulation". *Cognition, 117(3)*, 243-260.

Pickering, M. J., & Garrod, S. (2007). "Do people use language production to make predictions during comprehension?". *Trends in cognitive sciences, 11*(3), 105-110.

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*(20), 7865-7870.

Wrench, A. A., & Scobbie, J. M. (2008). High-speed Cineloop Ultrasound vs. Video Ultrasound Tongue Imaging: Comparison of Front and Back Lingual Gesture Location and Relative Timing. In Proceedings of the Eighth International Seminar on Speech Production (ISSP).

Scobbie, J. M., Wrench, A. A., & van der Linden, M. (2008). "Head-probe stabilisation in ultrasound tongue imaging using a headset to permit natural head movement". In *Proceedings of the 8th International Seminar on Speech Production* (pp. 373-376).

Watkins, K., & Paus, T. (2004). 'Modulation of motor excitability during speech perception: the role of Broca's area". *Journal of Cognitive Neuroscience, 16*(6), 978-987.

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*(7), 701-702.

Zharkova, N., Hewlett, N., & Hardcastle, W. J. (2011). Coarticulation as an indicator of speech motor control development in children: an ultrasound study. *Motor Control, 15*(1).