Accommodation of End-State Comfort Reveals Subphonemic Planning in Speech

Donald Derrick^a Bryan Gick^b

^aNew Zealand Institute of Language, Brain and Behaviour, University of Canterbury, Christchurch, New Zealand; ^bDepartment of Linguistics, Totem Field Studios, University of British Columbia, Vancouver, B.C., Canada

Abstract

Applying the ‘end-state comfort’ hypothesis of Rosenbaum et al. [J Exp Psych Learn Mem Cogn 1992;18:1058; Acta Psychol (Amst) 1996;94:59] to tongue motion provides evidence of long-distance subphonemic planning in speech. Speakers’ tongue postures may anticipate upcoming speech up to three segments, two syllables, and a morpheme or word boundary later. We used M-mode ultrasound imaging to measure the direction of tongue tip/blade movements for known variants of flap/tap allophones of North American English /t/ and /d/. Results show that speakers produce different flap variants early in words or word sequences so as to facilitate the kinematic needs of flap/tap or other /r/ variants that appear later in the word or word sequence. Similar results were also observed across word boundaries, indicating that this is not a lexical effect.

© 2015 S. Karger AG, Basel

Introduction

Many scientists have been trying to explain low-level speech production, in particular coarticulation, without reference to planning for decades (Joos, 1948; Ohman, 1966, 1967; Fowler, 1980; Saltzman and Munhall, 1989; Boyce, 1990). In this view, speakers may plan larger units of speech such as the phrase (Shattuck-Hufnagel, 2000) or sentence (Butterworth, 1975), but the limited inventory of segments is drawn from stored knowledge in the brain. Other scientists, in contrast, have argued for planning during adjacent anticipatory coarticulation (Whalen, 1990; Roelofs, 1997), anticipatory coarticulation spanning a vowel-consonant-vowel trajectory (Winkler et al., 2011; Barbier, 2013), or they incorporate planning into the structure of speech articulation (Henke, 1966).

Evidence has been found for speech planning at the level of syllables (Levelt, 1994; Hawkins and Nguyen, 2004), with some evidence for planning at lower levels,

D.D. is also at the University of Western Sydney, MARCS Institute, Penrith, N.S.W., Australia.

B.G. is also at Haskins Laboratories, New Haven, Conn., USA.

such as the phoneme (Dell, 1986; Levelt, 1989) or feature (Dell, 1986; Mowrey and MacKay, 1990; Bernhardt and Stemberger, 1998). In addition, Bell-Berti and Harris (1979) argued for a timing-based anticipatory coarticulation, at least when there are no competing constraints, such as with lip-rounding preceded by segments that do not constrain lip position, while Keating (1990) argued for windowed coarticulation of adjacent segments based on subphonemic production variability. Ungrammatical productions in tongue-twisters (Frisch and Wright, 2002) have also been used to argue that spreading activation of more than one competing phoneme or feature can generate subphonemic variation that can only be noticed through careful acoustic (Goldrick and Blumstein, 2006) and articulatory (McMillan and Corley, 2010) analysis. That is, higher level planning can influence the subtlest of speech articulations. Nevertheless, Munhall et al. (2000) have illustrated the difficulty in demonstrating clear cases of planning in speech, as actual speech output may look very similar regardless of whether there is planning or not. As such, a diagnostic allowing identification of low-level planning would be of great use in speech. One such diagnostic is end-state comfort.

Rosenbaum et al. (1992, 1996) and Cohen and Rosenbaum (2004) observe that people grasp objects at the beginning of transport in a way that allows joints to be in a comfortable position at the end of transport. If someone is asked to pick up a glass and put it down the same way, usually the hand is held with the thumb in medial position throughout, but if asked to put the glass down upside down, usually the arm begins twisted so that the thumb is in lateral position when the glass is picked up, and twisted back to the more comfortable thumb-medial position when the cup is put down upside down. Observations of the end-state comfort effect have been used as diagnostics of motor planning in humans, lemurs (Chapman et al., 2010), and cotton-top tamarins (Weiss et al., 2007).

We argue that an analysis relying on the end-state comfort hypothesis also applies to speech motion, provided it is possible to identify a beginning and end-point in a sequence where motion transitions are constrained enough to measure categorical differences. This is not without challenges, as many speech articulators move faster and have more degrees of freedom than skeletal structures. For instance, the human tongue and lips are muscular hydrostats, much like an elephant's trunk or the tentacles of an octopus (Kier and Smith, 1985). The tongue is free to move faster than a skeletal structure, in many more directions and patterns, and with partially independent control of different parts of the tongue (Stone et al., 2004). However, this freedom of motion is in practice constrained in some animals and some situations, as when an octopus mimics the bend-points of an arm when it is trying to grab something (Sumbre et al., 2001, 2005). We argue that in North American English, the interaction of flap/tap variants and surrounding non-rhotic and rhotic vowels provides constraints on tongue motion that work well with testing of the end-state comfort hypothesis in speech planning.

To explain, we must describe subphonemic variability in flaps/taps, rhotic vowels, and non-rhotic vowels: During English /t, d/ ('flap' or 'tap') production, the back of the tongue is braced against the teeth; in contrast the tongue tip must make rapid contact with the front of the hard palate. In previous research, Derrick and Gick (2011) identified four subphonemic categorical kinematic variations for these flap/taps. The first is an alveolar tap ([ɾ¹]), in which the tongue tip moves from below the alveolar ridge upwards, makes contact, then moves back down into position for the following vowel. The second is a down-flap ([ɾ²]), in which the tongue tip moves from above

the alveolar ridge, makes contact, and continues downwards below the alveolar ridge. The third is an up-flap ([r̥]), in which the tongue tip moves from below the alveolar ridge, makes contact, and continues upward into a position above the alveolar ridge. The fourth is a post-alveolar tap ([r̠]), in which the tongue tip moves anteriorly from behind and above the alveolar ridge, makes contact at a point at or above the ridge, then retracts back to a position above and behind the alveolar ridge.

These flap/taps are always produced within a vocalic context, and all begin and end with the tongue tip in either a high position (above the alveolar ridge), or a low position (below the alveolar ridge) based on interaction with their surrounding vowels. These vowels come in two types – rhotic and non-rhotic – and these two types impose significantly different constraints on tongue tip position.

Previous research (Delattre and Freeman, 1968) demonstrated that rhotic vowels (hereafter R in transcription shorthand) could be produced in at least eight categorically different ways. Simplifying these patterns by focusing on anterior tongue constriction, Hagiwara (1995) identified two broad categories, tongue tip-down or ‘bunched’ r ([ɹ]), and tongue tip-up, including but not limited to ‘retroflex’ r ([ɻ]). That is, tip-up productions have the tongue tip and the front of the tongue blade above, and therefore also behind, the alveolar ridge. The tip-down productions have the tongue tip below the alveolar ridge – regardless of the position of the back of the tongue or the angle of the blade. In contrast, non-rhotic vowels (hereafter V in transcription shorthand) are typically produced tip-down only (Chiba and Kajiyama, 1958; Perkell, 1969). This means that a tip-up instance of a non-rhotic vowel would constitute a non-ideal variant of what would normally be a tip-down vowel. Because of this, non-rhotic vowels can be used as a diagnostic of ‘comfort’ when the tongue tip is down, or of ‘non-comfort’ when the tongue tip is up. That is, they can serve as a context in which it is possible to identify end-state comfort effects in speech movements.

Our previous research (Derrick, 2011) suggests that such end-state comfort may play an important role in sequences of these articulations, as evidenced by the observation that in words with one flap/tap, the flap/tap variants are better predicted from the following rhotic vowel variant than from the preceding rhotic vowel variant. That is, for the word ‘Berta’, speakers largely ignored the R variant, producing [r̥] even following an [ɹ]. In these cases, the [ɹ] rapidly transitions to an [ɹ̠] position prior to the [r̥]. The researchers observed few similar transitions into the final rhotic vowel for ‘otter’. That is, the most common production showed anticipatory coarticulation – an [r̥] followed by a word-final [ɹ]. However, this evidence of coarticulation focuses on immediate context, extending no further than the transition from a preceding rhotic vowel into a flap/tap variant in relation to the tongue tip position of the following non-rhotic vowel. This result could potentially be explained via within-word memorized sequences or local assimilation.

In contrast, the present study was designed to test whether speakers take into account tongue movement information from upcoming morphemes and words when producing longer speech sequences, and over non-adjacent spans.

Hypotheses

Our hypotheses for end-state comfort accommodation across morpheme and word boundaries are presented below.

Fig. 1. Schematic of predicted tongue tip height during planned productions of ‘editor/auditor’ vs. ‘edify/audify’, focusing on initial flap/tap. Black line = Tongue tip height; blue circle = flap/tap of interest; green short-dashed circle = ideal state; red long-dashed circle = non-ideal state; grey arrow = predicted influence of ideal end-state on initial flap/tap variant; vertical grey dashed line = morpheme boundary.

Morpheme Boundary

To test for subphonemic planning across morpheme boundaries, we compared two words that end in non-rhotic vowels, ‘edify’ and ‘audify’, to two words ending in rhotic vowels, ‘editor’ and ‘auditor’. Because all of the flap/tap contacts we observed occur at or above the alveolar ridge, in order to produce any of these four words, the tongue tip typically starts low and moves up to produce the initial flap/tap. For ‘edify/audify’, the rest of the sequence involves non-rhotic vowels that are produced with the tongue tip-down, so the tongue tip may be expected to move back down to facilitate their production. In comparison, for ‘editor/auditor’, the sequence ends with a rhotic vowel, which can be either tip up or down. If the words are produced exclusively with word-final [ɹ], we expect a ‘double-tap’ [Vr[↑]Vr[↓]ɹ] sequence regardless of whether the sequence shows anticipatory subphonemic planning or not. However, if those rhotic vowels are tip-up [ɹ[↑]], it is possible to identify the sequences with anticipatory planning. If the speaker plans out the movement sequence in advance, in those cases where a tip-up word-final [ɹ[↑]] is planned, we expect that the sequence is more likely to begin with an initial [r[↖]], anticipating the need to raise the tongue tip later in the word, and overriding the preference for a tip-down vowel in the middle of the sequence. Such a sequence is then more likely to end with an [r[↗]] followed by the [ɹ[↑]]. If the speaker does not plan ahead, but rather engages in purely sequential (left-to-right) production, the tongue tip is more likely to move back down after contacting the alveolar ridge, producing an initial [r[↓]] leading into the second vowel, followed by an [r[↖]] leading into the final [ɹ[↑]]. Our hypothesis generates two predictions:

Prediction 1. If subphonemic planning influences tongue motion sequences, we expect a greater incidence of [r[↖]] for the first flap/tap in ‘editor/auditor’ in comparison to ‘edify/audify’, which should exhibit a greater incidence of [r[↓]]. Figure 1 schematically illustrates prediction 1, as well as how the middle (non-rhotic) vowel would be realized with the tongue tip-up during a planned sequence, illustrating the end-state comfort effect. The black lines illustrate the predicted position of tongue tip height during the given production. The blue circles represent the flap/tap under examination. The green short-dashed circles represent ideal (‘comfortable’) states. The red

Fig. 2. Schematic of predicted tongue tip height during production ‘editor/auditor’ for planned vs. unplanned productions, focusing on final flap/tap. Black line = Tongue tip height; blue circle = flap/tap of interest; green short-dashed circle = ideal state; red long-dashed circle = non-ideal state; grey dashed line = morpheme boundary.

long-dashed circle represents a tongue tip-up non-rhotic vowel, which is a non-ideal state as described above. The grey arrow highlights the relationship between the beginning and end-states. The vertical grey dashed line represents the morpheme boundary.

Prediction 1a. We also note that choice of comparing ‘editor/auditor’ versus ‘edify/audify’ has a potential confound in that ‘editor/auditor’ has two flap/taps in adjacent syllables, whereas ‘edify/audify’ have one flap/tap. We therefore expect that ‘editor/auditor’ would have a greater incidence of $[r^{\cdot}]$ than any sequence with two flap/taps in adjacent syllables that end with a non-rhotic vowel, such as ‘edit/audit a’.

Prediction 2. If prediction 1 is true, we expect the number of instances of $[r^{\leftrightarrow}]$ for the final flap/tap in ‘editor/auditor’ to be similar to the number of instances of $[r^{\cdot}]$ tokens for the initial flap/tap in ‘editor/auditor’. Figure 2 schematically illustrates prediction 2.

Word Boundary

To test for subphonemic planning across word boundaries, we compared phrases that contain either one or two flap/taps. The words ‘edit’ and ‘audit’ contain only one flap/tap (the one between vowels). However, the onset of the following word can affect the number of flap/taps in the sequence. For example, the phrases ‘edit the’ and ‘audit the’ contain only one flap/tap, followed by a rapid motion of the tongue tip toward the teeth for the interdental stop + fricative sequence $[t\theta]$ instead (see Browman and Goldstein, 1989, regarding gestural occlusion). In all cases, the tongue tip would move higher than would be preferred for a non-rhotic vowel, but lower than the typical height of contact for any of the four flap/tap variants. In contact, the phrases ‘edit a’ and ‘audit a’ contain two flap/taps, where the second flap/tap in the sequence is conditioned by the vowel following the word boundary. Producing any of these four phrases normally requires starting with the tongue tip low and moving it up to produce the initial flap/tap, again because all flap/tap contact is at or above the alveolar ridge. Thus, a speaker who plans ahead for the double-flap/tap sequence should be more likely to produce an $[r^{\cdot}]$, $[r^{\cdot}]$ sequence for ‘edit/audit a’, but an initial $[r^{\uparrow}]$ for the single flap/tap in ‘edit/audit the’.

Prediction 3. If subphonemic planning influences tongue motion sequences, we expect a greater incidence of $[r^{\uparrow}]$ for the first flap/tap in ‘edit/audit the’ in comparison to ‘edit/audit a’, which would exhibit a greater incidence of $[r^{\cdot}]$. Figure 3 schematically

Fig. 3. Schematic of predicted tongue tip height during planned productions of ‘edit/audit a’ vs. ‘edit/audit the’, focusing on initial flap/tap. Black line = Tongue tip height; blue circle = flap/tap of interest; green short-dashed circle = ideal state; red long-dashed circle = non-ideal state; grey arrow = predicted influence of ideal end-state on initial flap/tap variant; grey dashed line = morpheme boundary.

illustrates prediction 3, as well as how the middle (non-rhotic) vowel would then be retroflected during the planned sequence, illustrating the end-state comfort effect.

Prediction 4. In cases where speakers do not produce an [r^ɹ] for ‘edit/audit the’, we might expect them to produce a greater incidence of [r^ɹ] in comparison to ‘edit/audit a’.

Avoiding Confounds

All of these sequences begin with a non-rhotic vowel at the beginning of a stressed word/phrase, followed by an initial flap/tap and a medial non-rhotic vowel. For each hypothesis, there are tokens with contrasting end-states for analysis.

Since the sequence begins with a non-rhotic vowel, and our analysis begins with the middle of the initial flap/tap motion that follows, initial prosodic strengthening (Fougeron and Keating, 1997) does not interfere with our measurements. In addition, the medial state of the sequence – the vowel in the middle of the sequence – is the one that is expected to show the most deviation from the expected norm. Therefore, end-state comfort analyses do not interfere with observations of initial strengthening or of medial/final reduction or declination (Krakow et al., 1995), and we expect our results to fit within the confines of all of these analyses at the same time.

Methods

Eighteen native speakers of North American English between the ages of 18 and 40 participated in the study. All participants had normal speech and hearing. Participants were seated in a customized American Optical Co. model 507-a (1953) ophthalmic chair with a two-cup rear headrest adjusted to contact the base of the skull just above the neck. This technique is the same one used in Gick et al. (2005); it allows for careful alignment of the ultrasound probe along the midsagittal plane and reduces head motion to about a standard deviation (SD) of 1 mm and angular rotation of about 0.5°.

A UST-9118 EV 180 electronic curved array ultrasound probe was placed under the chin using a mechanical arm. The probe has a variable frequency range of 3–9.0 MHz (set to 6 MHz) with an average (μ) slice thickness of the tissue viewed with this probe of approximately 3 mm (Medicines and Healthcare Products Regulatory Agency, 2004). Therefore head motion in relation to the position

of the ultrasound probe is expected to be less than the thickness of the ultrasound slice. The ultrasound probe was held upright as close as possible to the thyroid notch to allow for effective imaging of most or all of the tongue tip. This is because the anterior portion of the probe is not near the jaw as it would be with a long probe, mitigating occlusion of the tongue tip by the jaw. In addition, the angle between the probe and the tongue tip is maximally well suited to avoiding possible air boundaries under the tongue tip. As a result, for most of our data, the tongue tip was clearly visible, and for the rest, it was similar to results from point-tracking techniques such as X-ray microbeam or electromagnetic articulometry, where the tongue tip sensor is placed 1 cm back from the tongue tip. In addition, the use of three M-mode lines to access tongue tip and blade motion maximized the likelihood of obtaining the most anterior motion information possible.

This probe was attached to an Aloka ProSound SSD-5000 ultrasound machine connected via s-video cable (marked video IN) to a Canopus ADVC-110 advanced digital video recorder. A Sennheiser MKH-416 short shotgun microphone was mounted on a microphone stand and aimed 30 cm away from the participant's mouth. The microphone was plugged into an M-Audio DMP3 pre-amplifier via XLR balanced cable and out with an RCA cable to the Canopus card to ensure time synchronization between the ultrasound and audio output. The Canopus card was connected via FireWire to a MacPro Quad Core 2.8 GHz computer.

An LCD monitor was mounted on the ophthalmic chair's monitor mount in front of the participant. A computer containing the software for experimental stimulus presentation was connected to the LCD monitor so that the participant could easily read the stimuli from the screen.

The ultrasound machine was set up in simultaneous B/M mode and aligned to the acoustic signal. The B-mode ultrasound was used to capture two-dimensional images of the midsagittal plane of the tongue at 30 fps. The M-mode (motion mode) ultrasound provided a progressive scan of three selected one-dimensional lines accessible from an ultrasound probe.

These three one-dimensional M-mode lines were set visually to about 45° left-up, and adjusted to capture anterior (tip and blade) tongue motion during flap/tap production. Because M-mode ultrasound is a progressive scan, it presents the motion data at the full capture rate of the ultrasound probe, ranging from 60 to 100 Hz depending on the depth of the scan. This arrangement of M-mode lines allows capture of the general direction of motion of the tongue tip and blade, which is ideal for identifying the flap/tap variants described above. At the same time, the B-mode ultrasound allows examination of the midsagittal plane of the tongue surface at 30 fps (NTSC), which along with the M-mode data allowed identification of the rhotic vowel variants described above. The three M-mode intersect lines were placed parallel to each other in order to provide redundancy in case one or two of the intersects fell outside of the range of tongue motion for the participants. Too close to the palate and part of the tongue motion could be missing, too far and the motion would be less pronounced as the tongue body does not move much during flap production.

Tokens were selected to contain single flap/taps or sequences of flap/taps in consecutive syllables. Data were collected on 17 control sentences, 9 sentences with 1 flap/tap, 10 sentences with double flap/tap sequences, and 2 sentences with triple flap/tap sequences, for a total of 38 unique sequences. The sentences were randomized within each of 12 blocks, giving a total of 456 stimulus sentences. The stimuli were presented using PXLabRT (Irtel, 2007) such that each sentence was displayed on an LCD screen for 2.2 s. The software automatically paused the experiment after the first 6 blocks to allow participants to swallow some water or take a short break if needed. Each set of 6 blocks took 9 min, for a total of 18 min recording time. The present study is based on a subset of the phrases collected, as shown in table 1.

The acoustic signal was labelled and transcribed in Praat (Boersma, 2001). Vowels were identified using the standard techniques of listening and aligning to waveform and spectrogram. Vowels were labelled as rhotic [R] or non-rhotic [V]. Flap contact duration could not be measured precisely using ultrasound, and doing so was not required for our analysis. However, flap/tap contacts could be and were identified temporally as the point of lowest amplitude in the acoustic signal, as defined by Zue and Laferriere (1979). See figure 4 for an example from a token of 'editor'. Flap/tap boundaries were identified by their attenuation of vocalic energy in the waveform and spectrograms. Flap/taps were labelled as [T], and were expected to be from 10 to 40 ms in length (Zue and Laferriere, 1979; Fukaya and Bird, 2005). However, the duration from the middle of the preceding vowel to the middle of the following vowel contains the articulatory transition information used to identify flap/tap variants, and is expected to be much longer, at over 100 ms.

Fig. 4. Praat transcription of a token of the word ‘editor’. V = Non-rhotic vowel; T = flap/tap; R = rhotic vowel.

Table 1. Subset of phrases used for analysis of subphonemic planning

Token	Word	Carrier phrase	Flap/tap count
1	editor	We have editor books	2
2	auditor	We have auditor books	2
3	edify	We have him edify a book	1
4	audify	We have him audify a book	1
5	audit the	We have him audit the books	1
6	edit the	We have him edit the books	1
7	edit a	We have him edit a book	2
8	audit a	We have him audit a book	2

Praat transcriptions were then imported into ELAN (Sloetjes and Wittenburg, 2008) and the flap/tap variants were identified from the B/M-mode ultrasound image. The B-mode data allowed categorical identification of the rhotic vowels as either tip-down [ɹ] or tip-up [ɹ̥] based on tongue shape and position.

The M-mode data allowed tracking of tongue tip/blade motion for identifying which of the four flap/tap variants was uttered. The top of figure 5 illustrates the position of M-mode intersect lines.

Fig. 5. Schematic of B/M-mode ultrasound with visualization of the technique for identifying T variants through M-mode. (Image from Derrick et al., 2015)

As the surface of the tongue moves through the intersect lines, the white surface appears in the progressive scans produced from the data captured along the intercept lines. When the tongue tip/blade moves high and back the white lines move higher, and when the tongue tip/blade is low and front the white lines are lower. The three parallel intersect lines typically move in unison, though one or two may be absent due to motion of the tongue outside the intersect lines and/or occlusion of that part of the tongue. The three together ensure that the most anterior portion of the tongue possible was always tracked, allowing for accurate observation of the movement patterns of the tongue tip.

The flap/tap variants were identified by first examining the B-mode video just before, during, and after flap/tap contact. The identification was confirmed by examining the three M-mode progressive scans, starting from a 2–3 frames ahead of the flap contact, and focusing on the M-mode data adjacent to the leading edge, as identified by the thick black lines, and highlighted as the ‘area of interest’ in figure 5. Within the M-mode data, there are four ‘patterns of interest’ illustrated in figure 5: alveolar taps ($[r^1]$) are identified by a white up-down loop centred on the acoustically identified time of contact. Down-flaps ($[r^2]$) are identified by a downward motion of the white air boundary. Up-flaps ($[r^3]$) are identified by an upward motion of the white air boundary. Lastly, postalveolar taps ($[r^4]$) are identified by a flat or slightly wavy horizontal white air boundary, higher than the typical up-down loop of an $[r^1]$ (Derrick et al., 2015). In figure 5, the three M-mode intersect scans each show the pattern of interest highlighted.

Data were analysed using Wilcoxon signed-rank and rank-sum tests (Wilcoxon, 1945) in R (R Core Team, 2013). The Wilcoxon signed-rank test is a non-parametric alternative to the paired t test based on an ordered ranking of the measured observations. For example, for prediction 2, to test whether ‘edify/audify’ are produced with more [r^l] than ‘editor/auditor’, the percentages of flap/taps produced as [r^l], for each subject and for each group, were computed. The results were paired by participant, and the statistical test completed. Wilcoxon rank-sum tests were selected because they are the most conservative of the statistical tests and make no assumptions about normality. They are therefore highly appropriate for data that varies by participant as much as this data does, though they are prone to type II errors, so even when results are not statistically significant, the descriptive statistics may be informative as to behavioural trends.

In contrast, the Wilcoxon rank-sum test, or Mann-Whiney U test, is a non-parametric alternative to the two-sample t test. This was used as a test of significance for the behaviour of individual participants for data where there was particularly high between-subject variability, so as to avoid type II errors that might be introduced by using the signed-rank test. Because the rank-sum test does not require paired samples, but simply two groups, it can be used to examine differences in contextual behaviour for each participant individually.

Results

The results of the experiment are presented below, including descriptive statistics and logistic regression tests for each of the predictions listed above.

Descriptive Acoustic Analysis

Flap/tap duration, as measured by the boundaries of attenuation of vocalic energy seen in the waveform and spectrograms, were an average (mean) of 32.7 ms (SD = 10.9 ms) for initial flap/taps, and a mean of 56.8 ms (SD = 19.4 ms) for the second flap/taps for ‘editor/auditor’ and ‘edit/audit a’. The transition times from the middle of the preceding vowel to the middle of the following vowel were a mean of 106.1 ms (SD = 19.8 ms) for initial flap/taps, and a mean of 123.2 ms (SD = 24.3 ms) for second flap/taps for ‘editor/auditor’ and ‘edit/audit a’. These transition durations provide lines of progressive M-mode data based upon the frame rate of the ultrasound machine (between 60 and 100 Hz) and duration of the transitions. That is, an average of 6.4 (at 60 Hz) to 10.6 (at 100 Hz) ultrasound cycles for initial flap/taps, and an average of 7.4 (at 60 Hz) to 12.3 (at 100 Hz) ultrasound cycles for second flap/taps. These correspond to the M-mode ‘area of interest’ windows shown in figure 5 and are quite sufficient for flap/tap identification.

Hypothesis 1: Subphonemic Planning across Morpheme Boundaries

We begin with the tests for subphonemic planning across morpheme boundaries, as described in predictions 1 and 2.

Prediction 1. Recall that we predicted more [r^l] for ‘edify/audify’, and more initial [r[^]] for ‘editor/auditor’ depending upon differences in tongue tip position for the final vowel. Of the 399 measurable rhotic vowels in our dataset, for ‘editor/auditor’ 314 (78.7%) of them were tip-up [ɹ], allowing a meaningful comparison with tip-down non-rhotic vowels at the end of ‘edify/audify’. All instances of flap/tap variants produced during the word ‘edify/audify’ versus the initial flap/taps from ‘editor/auditor’ are plotted in figure 6. Overall, the participants produced 238 [r[^]] out of 399 successfully measured tokens for ‘editor/auditor’, compared to only 73 out of 414 tokens for ‘edify/audify’.

Fig. 6. Comparison of the flap/tap variants in ‘edify/audify’ vs. the first flap/tap in ‘editor/auditor’. Left side = By-subject comparison. Right side = All subject combined comparison. Each box contains four quadrants sized proportionally to the counts of the relevant flap/tap variant. Upper-left quadrant = Alveolar tap [r^l]; upper-right quadrant = down-flap [r^r]; lower-left quadrant = up-flap [r^u]; lower-right quadrant = postalveolar tap [r^w].

Three participants (8, 15 and 23) almost always produced [r^l] in all four phrases, violating the assumption that would make them testable for hypothesis 1. However, the rest did not. Five participants (2, 4, 5, 6 and 26) produced exclusively or almost exclusively [r^l] during the production of the flap/tap in ‘edify/audify’, but exclusively or almost exclusively [r^r] during the production of the first flap/tap in ‘editor/auditor’. Seven more participants followed the same pattern, but not as consistently (9, 10, 12, 14, 16, 17 and 18). One participant (13) followed the opposite pattern, producing mostly [r^l] during ‘editor/auditor’, and [r^r] during ‘edify/audify’. One participant (3) produced mostly [r^r] in all four phrases. The last participant (21) produced a mixed bag of [r^l], [r^r] and [r^u].

Wilcoxon signed-rank tests were performed on the data summarized in figure 6. For each of the four flap/tap variants, the percentage of productions matching that variant were compared based on ‘edify/audify’ versus ‘editor/auditor’. As expected from the descriptive statistics in figure 6, the results show that there is a significantly higher ratio of [r^l] in ‘edify/audify’ than the first flap/tap of ‘editor/auditor’ ($V = 127$, $p = 0.003$), and a significantly lower ratio of [r^r] ‘edify/audify’ than the first flap/tap of ‘editor/auditor’ ($V = 8$, $p = 0.002$).

Prediction 1a. Wilcoxon signed-rank tests were performed on the data for ‘editor/auditor’ summarized in figure 6, and ‘edit/audit a’ summarized in figure 7. For each of the

Fig. 7. Flap/tap sequences for 'editor/auditor'. Y-axis = Initial flap/tap; X-axis = final flap/tap.

four flap/tap variants, the percentage of productions matching that variant were compared based on 'edit/audit a' versus 'editor/auditor'. As expected from the descriptive statistics in figures 6 and 7, the results show that there is a significantly higher ratio of [r[↕]] in 'edit/audit a' than the first flap/tap of 'editor/auditor' ($V = 148$, $p < 0.001$), and a significantly lower ratio of [r[↕]] 'edit/audit a' than the first flap/tap of 'editor/auditor' ($V = 0$, $p < 0.001$).

Prediction 2. We predicted predominantly [r[↕]Vr[↔]] final sequences for 'editor/auditor'. The full sequence of flap/tap variants produced for 'editor/auditor' is shown in figure 7. Note from figure 7 that of the 238 [r[↕]] tokens produced, 234 were followed by an [r[↔]].

Hypothesis 2: Subphonemic Planning across Word Boundary

The following section tests for subphonemic planning across word boundaries, as described in prediction 3.

Prediction 3. We predicted more [r[↕]] for phrases 'edit/audit the', and more initial [r[↕]] for phrases 'edit/audit a'. Figure 8 plots all instances of flap/tap variants produced during 'We have him edit/audit the books' versus 'We have him edit/audit a book'. Note that several participants, particularly participant 5, sometimes did not produce a second flap/tap in 'edit/audit a' because they slowed down and produced the flap/tap as a stop. These individual tokens were excluded from statistical analysis. Figure 9 plots the flap/tap sequences in 'We have him edit/audit a book' in order to show sequence context not otherwise visible in figure 8.

The descriptive statistics seen in figure 8 show that 1 participant (2) consistently used [r[↕]] for 'edit/audit a', and [r[↕]] for 'edit/audit the', whereas 6 participants (10, 12, 15, 16, 17 and 18) used [r[↕]] in 'edit/audit a' sequences more often than in 'edit/audit the' sequences. Two other participants (3 and 21) have a mixture of [r[↕]], [r[↕]] and [r[↕]]. Four participants (8, 9, 14 and 23) have predominantly [r[↕]] in all conditions. One participant (21) produced only [r[↕]] and [r[↕]] tokens. Three participants (13, 15 and 26) have mostly [r[↕]] in 'edit/audit a' and [r[↕]] in 'edit/audit the', and one (3), had more [r[↕]] for 'edit/audit the', which is opposite the predicted pattern.

Fig. 8. Comparison of first flap/tap variants in 'edit/audit a' vs. the flap/tap variants in 'edit/audit the'. Left side = By-subject comparison; Right side = All subject combined comparison. Each box contains four quadrants sized proportionally to the counts of the relevant flap/tap variant. Upper-left quadrant = Alveolar tap [r^l]; upper-right quadrant = down-flap [r^v]; lower-left quadrant = up-flap [$r^{\sim}$]; lower-right quadrant = postalveolar tap [r^{**}]. * p = Significant ($\alpha = 0.05$); + p = marginally significant ($\alpha = 0.1$).

Fig. 9. Count by flap/tap sequences in 'edit/audit a'. Y-axis = Initial flap/tap; X-axis = final flap/tap.

Table 2. Wilcoxon rank-sum tests comparing prevalence of final flap/tap variants in ‘edit/audit a’ versus ‘edit/audit the’

Participant	Flap variant					
	[r ^ɹ]		[r ^ʌ]		[r ^{ɹ̥}]	
	W	p	W	p	W	p
2	25.5	*p < 0.001	144.5	NA	263.5	*p < 0.001
3	90	p = 0.601	94.5	p = 0.353	58.5	p = 0.184
4	126	p = 0.194	112	NA	98	p = 0.194
5	NA	NA	NA	NA	NA	NA
6	168	p = 0.541	153	NA	138	p = 0.541
8	162	NA	162	NA	162	NA
9	36	NA	36	NA	36	NA
10	60	*p < 0.001	153	NA	246	*p < 0.001
12	76.5	*p = 0.007	144.5	NA	212.5	*p = 0.007
13	198	p = 0.119	162	NA	126	p = 0.119
14	136	p = 0.347	144.5	NA	153	p = 0.347
15	198	*p = 0.015	153	NA	108	*p = 0.015
16	87.5	*p = 0.069	117	NA	146.5	+p = 0.068
17	96.5	*p = 0.047	136	NA	175.5	*p = 0.047
18	108	*p = 0.015	153	NA	198	*p = 0.015
21	153	p = 0.412	117	p = 0.412	135	NA
23	20	p = 0.716	18	NA	16	p = 0.716
26	129	p = 0.211	96	p = 0.260	99	p = 0.541

* p = Significant ($\alpha = 0.05$); + p = marginally significant ($\alpha = 0.1$). NA = Not available/did not converge. All comparisons are against [r^{ɹ̥}].

Because it is clear that many individuals appeared to produce different flap/tap variants for ‘edit/audit a’ versus ‘edit/audit the’, Wilcoxon rank-sum tests were performed on each of the participants separately. These tests show significantly more tokens with [r^ʌ] for the first T of ‘edit/audit a’ and more tokens with [r^ɹ] for ‘edit/audit the’ for 5 of 18 participants (2, 10, 12, and 15, 17 and 18), and marginal significance for 1 more (participant 16), as seen in table 2. Note, however, that for 1 participant (15), the test showed a significant result is in the opposite direction.

In addition, Wilcoxon signed-rank tests were performed for between-subject analysis. The results of between-subject Wilcoxon signed-rank tests were not statistically significant.

Prediction 4. While all 4 participants who produced [r^ʌ] (3, 21, 23 and 26) produced more [r^ʌ] in ‘edit/audit the’ than in ‘edit/audit a’, as predicted, the results were not statistically significant using either Wilcoxon test.

Discussion

The results show that, as per prediction 1, participants produce more instances of [r^ɹ] for ‘edify/audify’ and more instances of [r^ʌ] for the first flap/tap in ‘editor/auditor’. Similarly, as per prediction 2, the results show that participants produce a number of [r^{ɹ̥}] for the second flap/tap in ‘editor/auditor’ consistent with the number of [r^ʌ] for

the first flap/tap. These results support the hypothesis that speakers can plan tongue tip motion sequences across morpheme boundaries, and that the selection of the initial flap/tap variant in 'editor/auditor' is related to the final posture of the rhotic/vowel segment in accord with the end-state comfort effect, as described in figures 1 and 2. In addition, supporting prediction 1a, speakers produce more instances of [r[^]] for the first flap/tap in 'editor/auditor' than for that in 'edit/audit a', supporting the hypothesis that the rhotic vowel at the end of the sequence had more of an influence than the absence of a second flap/tap for 'edify/audify'.

For prediction 3, and as seen in figure 9, most of the participants produced [r[↑]] and [r[↓]] sequences for 'edit/audit a'. While this result is consistent with an end-state comfort analysis of the influence of sequence-final rhotic vowels on preceding flap/taps, it was prevalent enough to make prediction 3 difficult to test. This behaviour was impossible to exclude from between-subject analysis and attenuated the possibility of identifying an end-state comfort effect through less-than-ideal production of initial or subsequent non-rhotic vowels when comparing 'edit/audit a' with 'edit/audit the'.

However, the descriptive statistics show that, overall, participants were still more likely to produce [r[↑]] for the 'audit/edit the' phrases, and [r[^]] for 'audit/edit a' phrases, and that these results were significant for 6 participants, and marginally significant ($\alpha = 0.1$) for 1 other. This provides some support for prediction 3.

For prediction 4, there were simply too few examples of [r[^]] for any of the statistical tests to function. For the 4 relevant participants, their behaviour trended in the direction hypothesized – but the results were not statistically significant.

Our conclusion is that there was a tendency to strategize tongue tip motion to accommodate end-states across word boundaries, but there were too many differences between the speakers, and possibly too small an effect of flap sequences in comparison to the influence of surrounding vowels, to demonstrate that the effect was statistically significant for the group as a whole.

Goldrick and Blumstein (2006) and McMillan and Corley (2010) argue that overlapping activations of word and phoneme plans generate gradient effects on articulation in speech production. They therefore do not argue for planning at the subphonemic level. However, the present findings support the idea that desired articulations at the end of a complex word or phrase sequence can influence tongue tip trajectories from the very beginning of the sequence. Instead of competition between overlapping activations in a context-free system, our data suggest that there is an interaction between an upcoming speech plan and earlier tongue tip motion trajectories. Within an overlapping activation system, this is evidence not just for subphonemic planning, but for look-ahead planning in the form of anticipatory accommodation of end-state comfort.

Alternate analyses of these observations, wherein a wide range of possible outcomes are stored rather than planned, could also help account for these results. For instance, usage-based grammars (e.g. Bybee, 1995; Tomasello, 2005) where commonly used chunks such as 'edit/audit a' versus 'edit/audit the' are stored in memory could explain differences in flap/tap selection. Similarly, exemplar theories (e.g. Pierrehumbert, 2001) in which multiple variants of the whole word 'edit' and 'audit' are stored based on individual experience, with different contextual variants used before vowels and consonants, could explain different flap/tap selection between phrase contexts. Lastly, models that allow complex representation of each word with stored choices at specific points in the word (e.g. Hudson, 1980) could also account for the observed trends. Note that all three of these memory-based analyses would need

to encode instances of groups of words spanning a phrase boundary (as in ‘edit/audit a/the’), and not just single words or even phrases. Further, these theories must still involve subphonemic planning at some stage, as the production output is optimized to anticipate a comfortable end-state. Memory-based explanations based on phonological context must account for the realization that the observed behaviours are themselves behavioural trends, not behavioural certainties. Lastly, they do not account for the remarkable degree of variation we have already seen within speaker and context (Derrick and Gick, 2011). Our results thus demonstrate look-ahead planning across both morpheme and word boundaries, necessitating a theory of subphonemic planning in models of the motor control of speech production.

Nevertheless, the high degree of variability suggests that individuals speaking the same language likely generate substantially different subphonemic plans from each other and may additionally be influenced by intrinsic speaker and speech act conditions (e.g. motor ability, speech rate) combined with ongoing optimization in speech production. In addition, our previous research has shown that extrinsic effects such as gravity and elasticity can influence interactions between flap/taps and rhotics during the word ‘Saturday’, leading to a strong preponderance of up-flap, down-flap sequences (i.e. [sær`ɹ`ei]; Derrick et al., 2015).

Future Work

In order to understand the relationship between speech rate and subphonemic planning, we are currently conducting follow-up research where speakers produce similar word/phrase sequences at different speech rates. This follow-up study combines ultrasound and electromagnetic articulometry to look for both categorical and gradient evidence for end-state comfort effects on flap/tap sequences across morpheme/word boundaries. We believe that the results will uncover a relationship between speech rate and optimization of subphonemic planning that can lead to a speaker employing distinct but stable speech plans under different speaking conditions, hearkening to multiple stabilities observed in other motor subsystems, such as locomotion (e.g. Dominici et al., 2011).

Acknowledgements

This research was funded by a Discovery Grant from the Natural Sciences and Engineering Council of Canada (NSERC) to the second author, and by National Institutes of Health (NIH) grant DC-02717 to Haskins Laboratories. Thanks to Cathi Best and Jason Shaw for help with argumentation. Special thanks to Aislin Stott for labelling and segmenting the acoustic data.

References

- Barbier G, Perrier P, Ménard L, Payan Y, Tiede MK, Perkell JS (2013): Speech planning as an index of speech motor control maturity. Proc 14th Annu Conf of the International Speech Communication Association, INTERSPEECH 2013, Lyon, pp 1278–1282.
- Bell-Berti F, Harris KS (1979): Anticipatory coarticulation: some implication from a study of lip rounding. J Acoust Soc Am 65:1268–1270.
- Bernhardt BH, Stemberger JP (1998): Handbook of Phonological Development from a Nonlinear Constraints-Biased Perspective. San Diego, Academic Press.
- Boersma P (2001): Praat, a system for doing phonetics by computer. Glot Int 5:341–345.
- Boyce SE (1990): Coarticulatory organization for lip rounding in Turkish and English. J Acoust Soc Am 88:2584–2595.

- Browman CP, Goldstein L (1989): Articulatory gestures as phonological units. *Phonology* 6:201–251.
- Butterworth B (1975): Hesitation and semantic planning in speech. *J Psycholinguist Res* 4:75–87.
- Bybee J (1995): Regular morphology and the lexicon. *Lang Cognit Process* 10:425–455.
- Chapman KM, Weiss DJ, Rosenbaum DA (2010): Evolutionary roots of motor planning: the end-state comfort effect in lemurs. *J Comp Psychol* 124:229–232.
- Chiba T, Kajiyama M (1958): *The Vowel: Its Nature and Structure*. Tokyo, Phonetic Society of Japan.
- Cohen RG, Rosenbaum DA (2004): Where grasps are made reveals how grasps are planned: generation and recall of motor plans. *Exp Brain Res* 157:487–495.
- Delattre P, Freeman D (1968): A dialect study of American Rs by X-ray motion picture. *Linguistics* 44:29–68.
- Dell GS (1986): A spreading-activation theory of retrieval in sentence production. *Psychol Rev* 93:283–321.
- Derrick D (2011): *Kinematic Patterning of Flaps, Taps and Rhotics in English*; PhD thesis University of British Columbia.
- Derrick D, Gick B (2011): Individual variation in English flaps and taps: a case of categorical phonetics. *Can J Linguist* 56:307–319.
- Derrick D, Stavness I, Gick B (2015): Three speech sounds, one motor action: evidence for speech-motor disparity from English flap production. *J Acoust Soc Am* 137.
- Dominici N, Ivanenko YP, Cappellini G, D'Avella A, Mondì V, Cicchese M, Fabiano A, Silei T, Di Paolo A, Giannini C, Poppele RE, Lacquaniti F (2011): Locomotor primitives in newborn babies and their development. *Science* 334:997–999.
- Fougeron C, Keating P (1997): Articulatory strengthening at the edges of prosodic domains. *J Acoust Soc Am* 101:3728–3740.
- Fowler CA (1980): Coarticulation and theories of extrinsic timing. *J Phon* 8:113–133.
- Frisch SA, Wright R (2002): The phonetics of phonological speech errors: an acoustic analysis of slips of the tongue. *J Phon* 30:139–162.
- Fukaya T, Bird D (2005): An articulatory examination of word-final flapping at phrase edges and interiors. *J Int Phonet Assoc* 35:45–58.
- Gick B, Bird S, Wilson I (2005): Techniques for field application of lingual ultrasound imaging. *Clin Linguist Phon* 19:503–514.
- Goldrick M, Blumstein SE (2006): Cascading activation from phonological planning to articulatory processes: evidence from tongue twisters. *Lang Cognit Processes* 21:649–683.
- Hagiwara R (1995): *Acoustic Realizations of American /r/ as Produced by Women and Men*; PhD thesis University of California.
- Hawkins S, Nguyen N (2004): Influence of syllable-coda voicing on the acoustic properties of syllable-onset /l/ in English. *J Phon* 32:199–231.
- Henke W (1966): *Dynamic Articulatory Model of Speech Production using Computer Simulation*; PhD thesis Massachusetts Institute of Technology.
- Hudson G (1980): Automatic alternations in non-transformational phonology. *Language* 51:94–125.
- Irtel H (2007): *PXLab: The Psychological Experiments Laboratory*; version 2.1.11. Mannheim, University of Mannheim.
- Joos M (1948): Acoustic phonetics. *Language* 24:5–136.
- Keating PA (1990): The window model of coarticulation: articulatory evidence; in Kingston J, Beckman M (eds): *Papers in Laboratory Phonology*. Cambridge, Cambridge University Press, vol I, pp 451–470.
- Kier WM, Smith KK (1985): Tongues, tentacles and trunks: the biometrics of movement in muscular-hydrostats. *Zool J Linnean Soc* 83:307–324.
- Krakow RA, Bell-Berti F, Wang QE (1995): Supralaryngeal declination: evidence from the velum; in Bell-Berti F, Raphael LJ (eds): *Producing Speech: A Festschrift for Katherine Safford Harris*. Woodbury, AIP Press, pp 333–353.
- Levelt WJM (1989): *Speaking: From Intention to Articulation*. Cambridge, MIT Press.
- Levelt WJM (1994): Do speakers have access to a mental syllabary? *Cognition* 50:239–269.
- McMillan CT, Corley M (2010): Cascading influences on the production of speech: evidence from articulation. *Cognition* 117:243–260.
- Medicines and Healthcare Products Regulatory Agency Evaluation report MHRA 03107 (2004).
- Mowrey RA, MacKay IRA (1990): Phonological primitives: electromyographic speech error evidence. *J Acoust Soc Am* 88:1299–1312.
- Munhall KG, Kawato M, Vatikiotis-Bateson E (2000): Coarticulation and physical models of speech production; in Broe MB, Pierrehumbert JB (eds): *Papers in Laboratory Phonology*. Cambridge, Cambridge University Press, vol V: Acquisition and the Lexicon, chap 1, pp 9–28.
- Ohman S (1966): Coarticulation in VCV utterances: spectrographic measurements. *J Acoust Soc Am* 39:151–168.
- Ohman S (1967): Numerical model of coarticulation. *J Acoust Soc Am* 41:310–320.
- Perkell JS (1969): *Physiology of Speech Production: Results and Implications of a Quantitative Cineradiographic Study*. Research Monograph No 53. Cambridge, MIT Press.
- Pierrehumbert J (2001): Exemplar dynamics: word frequency, lenition and contrast; in Bybee J, Hopper P (eds): *Frequency Effects and Emergent Grammar*. Amsterdam, Benjamins, pp 137–157.

- R Core Team (2013): R: A Language and Environment for Statistical Computing. Vienna, R Foundation for Statistical Computing.
- Roelofs A (1997): The WEAVER model of word-form encoding in speech production. *Cognition* 64:249–284.
- Rosenbaum DA, van Heugten CM, Caldwell GE (1996): From cognition to biomechanics and back: the end-state comfort effect and the middle-is-faster effect. *Acta Psychol (Amst)* 94:59–85.
- Rosenbaum DA, Vaughan J, Barnes HJ, Jorgensen MJ (1992): Time course of movement planning: selection of handgrips for object manipulation. *J Exp Psych Learn Mem Cogn* 18:1058–1073.
- Saltzman E, Munhall G (1989): A dynamical approach to gestural patterning in speech production. *Ecol Psychol* 1:333–382.
- Shattuck-Hufnagel S (2000): Phrase-level phonology in speech production planning: evidence for the role of prosodic structure; in Horne M (ed): *Prosody: Theory and Experiment*. New York, Springer.
- Sloetjes H, Wittenburg P (2008): Annotation by category – ELAN and ISO DCR. *Proc 6th Int Conf on Language Resources and Evaluation (LREC)*.
- Stone M, Epstein MA, Iskarous K (2004): Functional segments in tongue movement. *Clin Linguist Phon* 18:507–521.
- Sumbre G, Graziano F, Flash T, Binyamin H (2005): Neurobiology: motor control of flexible octopus arms. *Nature* 433:595–596.
- Sumbre G, Gutfreund Y, Fiorito G, Flash T, Hochner B (2001): Control of octopus arm extension by a peripheral motor program. *Science* 293:1845–1848.
- Tomasello M (2005): *Constructing a language: a usage-based theory of language acquisition*. Cambridge, Harvard University Press.
- Weiss DJ, Wark JD, Rosenbaum DA (2007): Monkey see, monkey plan, monkey do: the end-state comfort effect in cotton-top tamarins (*Saguinus oedipus*). *Psychol Sci* 18:1063–1068.
- Whalen DH (1990): Coarticulation is largely planned. *J Phon* 18:3–35.
- Wilcoxon F (1945): Individual comparisons by ranking methods. *Biometrics Bull* 1:80–83.
- Winkler R, Ma L, Perrier P (2011): A model of optimal speech production planning integrating dynamical constraints to achieve appropriate articulatory timing. *Proc of the 9th ISSP, Montreal*, pp 235–236.
- Zue VW, Laferriere M (1979): Acoustic study of medial /t,d/ in American English. *J Acoust Soc Am* 66:1039–1050.