5aSCb16. Acoustic correlates of flaps in North American English

Donald Derrick* and Benjamin Schultz

*Corresponding author's address: University of Western Sydney, MARCS Institute, University of Canterbury, NZILBB, NZILBB, University of Canterbury, Christchurch, 8140, Canterbury, New Zealand, donald.derrick@gmail.com

Using B/M mode ultrasound, Derrick & Gick (2010) identified four categorical variations of flaps in North American English, up-flaps, down-flaps, alveolar taps, and postalveolar taps produced in English. These variants can be used to test hypotheses about constraints on speech articulation, such as local context, gravity and elasticity, speech rate, and longer distance anticipatory coarticulation. This study examines acoustic correlates of flap variations in order to make connections between the results of larger, and easier to collect, acoustic databases and the tongue movements underlying flap productions. Preliminary analyses using smoothing spline ANOVAs of z-score normalized f0, F1, F2, F3, F4 and F5 indicate significant differences in each dependent variable for flaps in non-rhotic vowel contexts. The results for flaps adjacent to rhotic vowels is more complex, requiring more detailed analysis. Based on these results, we are currently planning supervised hierarchical clustering to aid in probabilistic identification of flap variants, with reference to both vocalic context and syllable structure.

Published by the Acoustical Society of America through the American Institute of Physics
INTRODUCTION

Characterizing speech variation has been very important in speech production research because doing so allows researchers to connect motor control to phonological behavior. Categorical speech production variation is especially useful because it provides a missing link between gradient phonetic speech variation and categorical phonological variation.

The most famous example of categorical subphonemic speech production variation is English ‘r’. Delattre and Freeman (1968) identified eight variants, but for our purposes, the most important distinction is between tongue tip-up versions and tongue tip-down versions Hagiwara (1995). However, it was thought to be the only example in English, until Derrick and Gick (2011) uncovered 4 subphonemic, yet categorical, kinematic variants of flaps/taps in North American English.

Flaps and taps are consonants with short contacts between the tongue and the alveolar ridge. Flaps differ from stop consonants in that there is little or no buildup of air pressure at the place of articulation, and as a result there is little or no stop burst at the closure release (Zue and Laferriere, 1979). Ladefoged once thought flaps and taps might differ in that taps strike directly and flaps strike tangentially (Ladefoged, 1968). Ladefoged later recanted this analysis (Ladefoged, 1975, 1982, 1993).

However, Derrick and Gick believed that potential subphonemic differences in flap production would be important to identify because flaps interact with adjacent rhotic vowels, which can themselves categorically vary as described above. As a result, they would represent a region of categorically unstable speech production in English and would therefore be helpful in clearly identifying subtle constrains on speech production such as gravity and elasticity, long-distance anticipatory coproduction, and differences in speakers’ motor skills (Derrick, 2011).

Articulatory measurements

In order to accomplish this task, Derrick and Gick used slow (30 Hz) 2D brightness mode (B-mode) ultrasound images of the midsagittal tongue surface with progressive motion mode (M-mode) scans of 1D intersect lines of a plane roughly parallel to the palate capturing the full speed of the ultrasound recording (60-100 Hz, depending on the ultrasound depth and tongue size) (Derrick and Gick, 2011). The M-mode data captures directional motion of the tongue tip and blade in order to categorize flap/tap variants based on motion.

There are four variants, which are readily described using the word ‘murder’ as an example: The first variant is an alveolar tap \([\text{r} \uparrow]\). From an initial \([\text{j}]\) position below the alveolar ridge, the tongue moves upwards, makes contact, and moves back down to a final \([\text{j}]\) below the alveolar ridge, giving \([m\text{r} \uparrow \text{j}]\). The second variant is a down-flap \([\text{r} \downarrow\text{]}. From an initial \([\text{j}]\) position above the alveolar ridge, the tongue moves down, makes contact, and continues downward to a final \([\text{j}]\) below the alveolar ridge, giving \([m\text{r} \downarrow \text{]}. The third variant is an up flap \([\text{r} \uparrow\text{]}. From an initial \([\text{j}]\) position below the alveolar ridge, the tongue moves up, makes contact, and continues upward to a final \([\text{j}]\) above the alveolar ridge, giving us \([m\text{r} \uparrow \text{j}]\). The fourth and final variant is a postalveolar tap \([\text{r} \leftrightarrow\text{]}. From an initial \([\text{j}]\) position above the alveolar ridge, the tongue moves across, makes contact, and moves back horizontally to a final \([\text{j}]\) above the alveolar ridge, giving us \([m\text{r} \leftrightarrow \text{j}]\). The four variants can be seen in Figure 1.

Acoustic measurements

The articulatory categorization of flap variants is based on a diagnosis of the direction of tongue front motion, which is a large-scale generalization that would include a high degree of
variability within each articulation. The measurement itself is also centered on a small portion of the tongue along the midsagittal plane. As a result, individual speakers and individual tokens will have a high degree of acoustical variance. Having said that, all English flaps and taps have one thing in common; they are produced by a rapid tongue motion towards and away from the alveolar region of the hard palate. While flaps and taps at one extreme will have a small release burst, and at the other extreme very little tongue contact with the palate, the pattern is consistent. Given that these tongue motions will change the area function of the vocal tract throughout, flaps and taps should have formant transitions and a point of low amplitude marking the flap contact (Zue and Laferriere, 1979).

Therefore a measure of acoustic correlates should examine fundamental frequency and formant transitions for the first five formants from the middle of the proceeding (rhotic or non-rhotic) vowel through to flap contact and to the middle of the following vowel. These are potentially useful because they are the ones that characterize differences in vocal tract shape (F1 - F3) and specific vocal tract characteristics (F4-F5).

In order to compare speakers, a normalization technique is required to make one speaker’s transition data comparable to another. While many techniques for normalization are available, in this case the ideal technique will allow for comparison of speakers of differing means and variabilities in speech production. That is, a simple z-score conversion for each speaker applied to the f0 and F1-F5 will serve the purpose. In order to compare f0 and F1-F5 transitions for the four flap/tap variants, smoothing splines analysis of variance (SSANOVA) meaningfully reduce measurement errors and provide valid comparisons methods for each variant.

SSANOVA comparisons of f0 and F1-F5 will provide preliminary data useful for identifying whether it may be possible to predict the probability that a particular acoustic flap recording was one of the four articulatory flap variants described above. The formula could then be applied to large bodies of acoustically recorded flap/tap data in order to test hypotheses about subtle influences on flap/tap variant production.
METHODS

Here we describe the original data collection technique, the acoustical analysis used in PRAAT used to extract the f0 and F1-F5 contours, and the statistical methods used to analyze the data.

Data collection

Data collection techniques were first reported in (Derrick and Gick, 2011). Recordings from 18 native speakers of North American English between the ages of 18 and 40 were used, ten males and eight females. All participants had normal speech and reported normal hearing. Participants were seated in a customized American Optical Co. model 507-a (1953) ophthalmic chair with a 2-cup rear headrest adjusted to contact the base of the skull just above the neck. A UST-9118 EV 180 electronic curved array ultrasound probe was placed under the chin. The probe has a variable frequency range of 3 to 9.0 MHz with an average (mean) slice thickness of the tissue viewed with this probe of approximately 3 mm (Medicines and Healthcare products Regulatory Agency, 2004). The probe was attached to an Aloka ProSound SSD-5000 ultrasound machine connected via s-video cable to a Canopus ADVC-110 video input adapter.

A Sennheiser MKH-416 short shotgun microphone was mounted on a microphone stand and aimed at the participant approximately 30 cm away from the participant's mouth. The microphone was plugged into an M-Audio DMP3 preamplifier via XLR balanced cable and out with an unbalanced RCA cable to the Canopus card to guarantee 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 facing the front of the participant. A computer running the experiment stimuli presentation software 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 B/M mode and aligned to the acoustic signal. B-mode ultrasound was used to capture 2-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 one-dimensional M-mode lines follow the line of the palate, in the region of intercept with the blade/tip of the tongue. Because M-mode ultrasound is a progressive scan, it presents the motion data at the full capture rate of the ultrasound probe, which ranged from 60 to 100 Hz depending on the depth of the scan. At the same time, the B-mode ultrasound allows examination of the midsagittal plane of the tongue surface at 30 fps for tongue shape and position verification.

Stimulus tokens were selected to contain a single flap/tap 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 (Derrick 2011).

Stimuli were carefully chosen to control for vowel/rhotic contexts and syllable count. However, due to the limited number of relevant tokens in English, tight control of word frequency was not possible. Each stimulus was embedded in a carrier phrase. The purpose of the carrier phrase was to be as repetitive as possible and therefore help place focus on the stimulus instead of the carrier phrase. For this reason, all stimuli begin with “We have (him)”. The phrase “We have (him)” was also chosen because it contains only labial or glottal consonants and so leaves the tongue free for other articulations. Similarly, carrier phrases end in words that have coronal/velar consonants only at the end (e.g., books), if at all.

The stimuli were presented using PXlabRT set to present stimuli such that each sentence
was displayed on an LCD screen for 2.2 seconds, for a total of 12 blocks. The software automatically paused the experiment after the first six blocks (nine minutes) to allow participants to swallow some water or take a short break if needed. The 12 blocks were presented in set order, but the entire set of 38 sentences was randomized for each block.

Participants were asked to repeat “ta” at least 10 times rapidly in order to record tongue motion speed and to provide data for audio synchronization. The experiment software was then activated and experiment data were recorded as described above. Participants were then asked to say 38 stimuli, repeated six times for each of two blocks, for a total of 456 stimuli. Each block took nine minutes, for a total of 18 minutes recording time.

Data were recorded directly onto a Macbook via the Canopus card, and the audio was extracted from the DV recordings. Audio and video synchronization were confirmed using the sequences of acoustic transients from the alveolar stop releases in the spoken sequences of “ta” with tongue-dropping gestures associated with them. The Canopus card’s audio and video synchronization were consistently within one frame, requiring no special post-production synchronization.

The acoustic signals were labeled and transcribed in PRAAT, with attention to identifying segment boundaries and the acoustic low amplitude point (center) of each flap/tap. Data were then imported into ELAN (a tool for annotating audio and video recordings simultaneously) and the kinematic variants of each flap/tap were identified.

**Acoustical analysis in PRAAT**

The acoustic signals were then processed in PRAAT. The fundamental frequency was extracted using PRAAT’s pitch tracker with a 75 to 500 Hz window. The first 5 formants were extracted using the LPC formant tracker, with a window of 0-5500 Hz, a time-step of 10 ms, a window-length of 25 ms, and pre-emphasis from 50 Hz. Data for each flap/tap were extracted using the percentage duration from the hand-labeled mid-point of the preceding vowel to the flap contact point as measured by the lowest point of amplitude in the waveform, to the mid-point of the following vowel at 10% increments. Data were saved to text files for processing in R, with annotations for relevant contextual information such as source word, flap/tap variant, and surrounding vowel context.

**Statistical analysis**

The acoustic data were loaded into R, and, for each subject, f0 and F1-F5 were normalized to their respective z-scores. For each of f0 and F1-F5, and for each of the four possible contexts of rhotic or non-rhotic vowels preceding and following the given flap/tap, SSANOVA graphs were produced to allow visual comparison of the transitions. The SSANOVA graphs contain the transition lines with shading to 95% confidence intervals. Therefore white space between the transition lines for each of the four flap/tap variants represents a statistically significant difference between the given acoustic measures.

**RESULTS**

The results for flap/taps with non-rhotic vowels before and after the flap show significant differences in (intrinsic) f0 between all four flap/tap variants, as seen in Subfigure 2(a), representing a difference in intrinsic fundamental frequency between the flap/tap variants.

The results are also significant (as indicated by white-space separation of the lines) for F1 for the transition through the preceding vowel, as seen in Subfigure 2(b), and for the following
vowel differentiation $r^\uparrow$ from the other flap/taps, suggesting a greater degree of tongue height displacement before flap contact than after.

![SSANOVA of $f_0$ for VɾV](image1.png)

**Figure 2:** SSANOVA comparisons of flap variants for VɾV contexts, $f_0$ and $F_1$. Each band represents the 95% confidence interval

The results for $F_2$ show a significant difference between all four flap/tap types in the transition to flap contact, with the $F_2$ merging for all the flap types except $r^\uparrow$, which diverges from the rest during the transition to the following vowel, as seen in Subfigure 3(a). This result suggests that tongue frontness differs before flap/tap contact more than it does after.

The results for $F_3$ show a significant difference between 1) $r\downarrow$, 2) $r^\uparrow$ and $r\downarrow$, and 3) $r^\uparrow$, especially at the time of flap contact, as seen in Subfigure 3(b). These results suggest that these three flap groups have a different place of contact within the alveolar region of the palate.

![SSANOVA of $F_2$ for VɾV](image2.png)

![SSANOVA of $F_3$ for VɾV](image3.png)

**Figure 3:** SSANOVA comparisons of flap variants for VɾV contexts, $F_2$ and $F_3.
The results for F4 show significant differences between all four flap/tap types at various places throughout the transitions, as seen in Subfigure 4(a). The results suggest a greater degree of motion relevant to F4 in $\ddot{r}$, as compared to $\dddot{r}$ and $\grave{r}$, with the least amount of motion for $\grave{r}$. The results for F5 show some significant differences between flap types, as seen in Subfigure 4(b), but these differences are limited to a small portion of the transition just prior to tongue-tip contact, and are therefore difficult to interpret.

Results for flaps in rhotic contexts are more complicated to interpret, and space precludes their discussion here.

**FIGURE 4:** SSANOVA comparisons of flap variants for VrV contexts, F4 and F5

**DISCUSSION**

The results show that, despite a high degree of variability in f0 and F1-F5 measurements within flap/tap variants, there are highly significant differences between the variants. The intrinsic f0 of each flap variant differs throughout the transition from preceding vowel to following vowel, with consistent distinctions between $\ddot{r}$ with the lowest f0, $\grave{r}$ with median f0, and $\dddot{r}$ with the highest f0 in comparison to the rest of the speaker’s speech. This indicates that the different flap/tap variants use the whole tongue differently such that there are different loads placed on the larynx.

In addition, the results show that the different flap/tap types all begin with different tongue height (F1) and backness (F2), even from the midpoint of the proceeding vowel. This difference is less pronounced after the flap/tap, which indicates that the tongue travels less after a flap/tap than before.

The results for F3 indicate that different flap variants tend to have different contact points along the alveolar ridge. It is likely that $\grave{r}$ have the most forward/lowest contact point, $\dot{r}$ have a shared contact point, and $\dddot{r}$ have the highest contact point. This is consistent with simulations from research on the effects of gravity and elasticity on flap/tap variant production (Derrick and Gick, 2010).

Having said that, there are still many questions to answer. These analyses do not
distinguish flap/taps based on syllable counts of words, or whether the flap/taps are in isolation or in flap/tap sequences.

The results are also preliminary in that they do not yet allow us to predict how likely a given flap/tap is to be one of the four recorded variants based on acoustic data alone. This will require cluster analysis of the transition data; work which is currently in progress.

ACKNOWLEDGEMENTS

The authors acknowledge the valuable contributions of Aislin Stott of the UBC Interdisciplinary Speech Research Laboratory for segmenting the bulk of the acoustic data. Original data collection funded by an NSERC Discovery Grant to Bryan Gick, and by an NIH Grant to Haskins Laboratories. Continued statistical analysis was funded by the University of Western Sydney and the MARCS institute.

REFERENCES


