Villarreal 2016: Vowel manipulation

These are the sections of my 2016 doctoral dissertation that are most relevant to understanding the acoustic manipulation process used in that study. Section 3.4.3, which starts on page 110 of the dissertation, describes the acoustic manipulation process in broad terms; Appendix E, which starts on page 223 of the dissertation, describes the acoustic manipulation process in greater detail. You can find the full dissertation at https://www.academia.edu/30182487/. If citing the acoustic manipulation process, please cite the dissertation itself.

3.4.3 Implementation of manipulation

Underlying the methods behind vowel resynthesis is the source–filter theory of speech production, which posits that speech sounds originate from a glottal source (i.e., the vibration of vocal folds) and pass through a filter defined by the various articulators of the vocal tract (Johnson 2003). For American English vowels, the filter is defined by the position of the tongue and the rounding of the lips, and these articulators affect the frequencies at which the sound from the glottal source is amplified or dampened (i.e., formant structure). Thus, whereas characteristics like phonation type, pitch, and loudness are properties of the source, the acoustic characteristics that most strongly correlate with the perception of vowel location, formant frequencies, are properties of the filter (Styler 2015).

Vowel resynthesis utilizes source–filter theory (and in particular, the assumption that source and filter are relatively independent) by performing linear predictive coding (LPC) on a vowel to estimate the acoustic effects of the filter (formants), inverse-filtering the vowel to derive the underlying source, modifying the filter, and passing the source through the modified

filter. In essence, this is the basic procedure used in this study. In order to preserve naturalness (and to automate the manipulation process via Praat script), however, I had to make several modifications in order to ensure that this basic procedure worked for the 91 TRAP tokens and 61 GOOSE tokens manipulated in the 24 excerpts. Appendix E gives step-by-step details on how the manipulation process was carried out.

One advantage of using an automated procedure is that the same manipulation steps applied to each token. However, I replaced five tokens (one TRAP, four GOOSE) in the original excerpts because of difficulties manipulating these tokens. Because the problematic tokens were replaced in the original excerpts prior to manipulation, the replaced tokens were manipulated for both guises. Four GOOSE tokens simply resisted being manipulated to one or both of their targets. As a result, these four tokens were replaced by different tokens from elsewhere in the same speaker's retell (but not the same excerpt). For some of these vowels, the source from the original token was retained and passed through a filter derived from a different token; for others, the original, problematic token was replaced completely by splicing in a new token (both source and filter). The pitch and intensity of the new token were adjusted in order to make the new token sound as natural as possible within the larger phrase. In addition, in one excerpt, a speaker's pronunciation of the vowel in *bag* was raised to [e:], a pattern that is typical of the Pacific Northwest (Wassink 2015) but has not been reported in California. In order to avoid this vowel overly influencing listeners' perception of the speaker's origin, this vowel's formants were replaced with that of a different prevelar TRAP token (*back*) from the same speaker.

After the acoustic manipulation process was run for all 24 excerpts, a trained phonetician listened to the manipulated stimuli to gauge naturalness and generalizability, which led to several

small adjustments (e.g., re-splicing a vowel). Stimuli were deemed to be satisfactory after these adjustments.

Figures 3.10 and 3.2 demonstrate the end result of this process for TRAP and GOOSE tokens, with original versions of these tokens compared to the same tokens in the conservative and Californian guises.

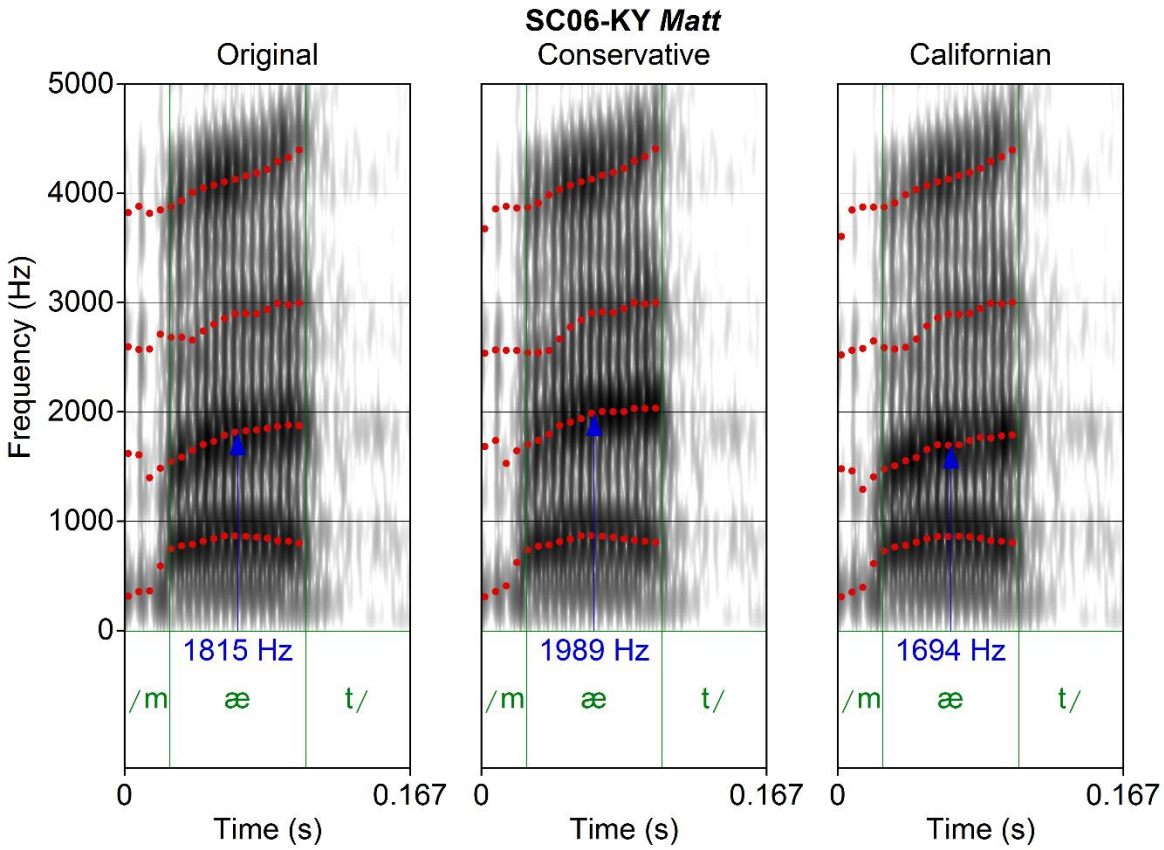

Figure 3.1. Spectrograms and formant tracks of original, conservative, and Californian versions of the token *Matt* by SC06-KY.
Formant tracks are in red. Blue arrows and text indicate F2 value at the vowel midpoint.

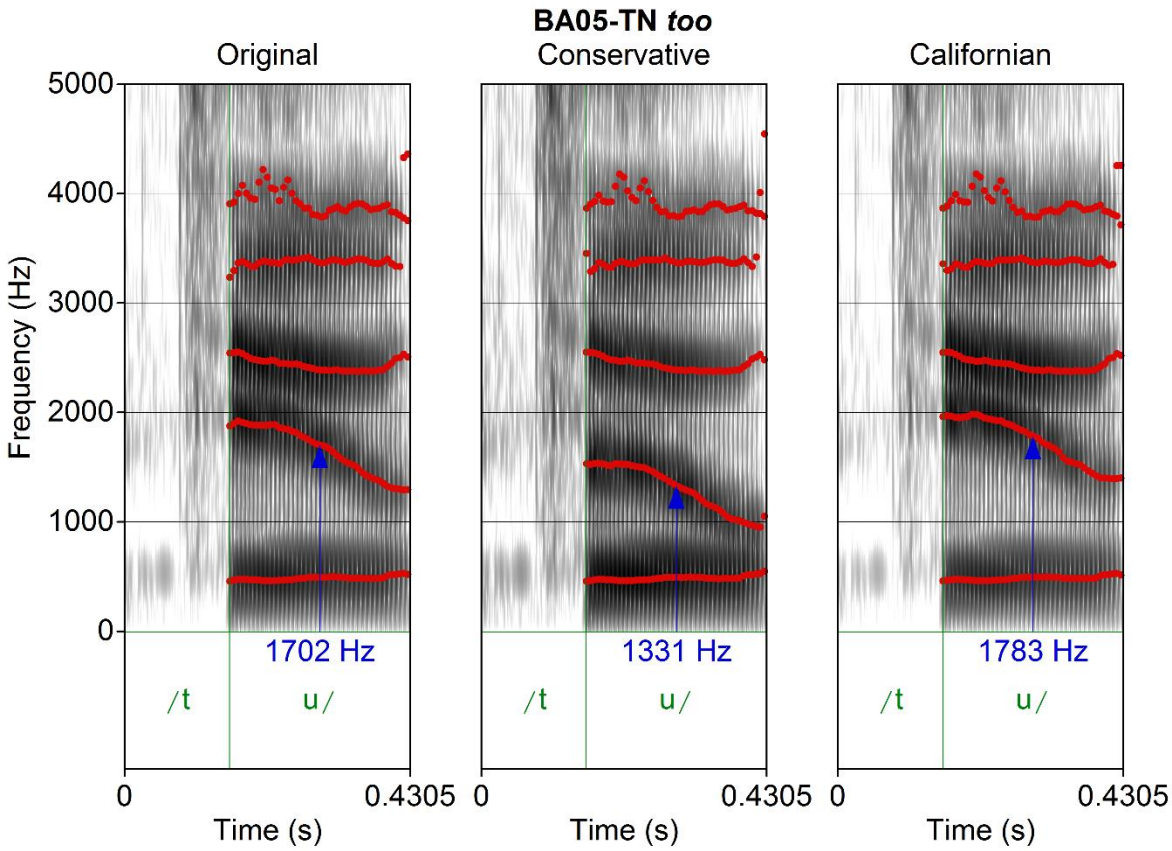

Figure 3.2. Spectrograms and formant tracks of original, conservative, and Californian versions of the token *too* by BA05-TN.
Formant tracks are in red. Blue arrows and text indicate F2 value at the vowel midpoint.

This manipulation procedure largely succeeded in creating manipulated tokens that were acceptably close to F2 target values. I defined a token to be "acceptably close" to the target if the two were separated in F2 by less than 1 just noticeable difference (JND), the minimum difference that human perceivers can detect between two stimuli (in this case, the minimum detectable difference in formant frequency); for TRAP F2, 1 JND is 33.09 Hz, and for GOOSE F2, 1 JND is 21.86 Hz (Kewley-Port & Watson 1994:492). On average, manipulated TRAP tokens were 11.20 Hz off target in F2 and manipulated GOOSE tokens were 8.31 Hz off target in F2, both

well within these vowels' respective ranges of acceptability.[1] All 91 TRAP tokens were within 1 JND of the target for both guises; among the 61 GOOSE tokens, three conservative tokens and four Californian tokens were outside 1 JND of the target. In Figures 3.30 and 3.40, each speaker's original and manipulated TRAP and GOOSE tokens are compared with their conservative and Californian targets. It is clear from these graphs that while there is some variability in the manipulated tokens' F2, in general they are close to their targets irrespective of the location of the original tokens.

[1] For TRAP, conservative tokens (6.58 Hz off target) were on average more accurate than Californian tokens (15.82 Hz off target). For GOOSE, Californian tokens (5.48 Hz off target) were on average more accurate than conservative tokens (11.15 Hz off target).

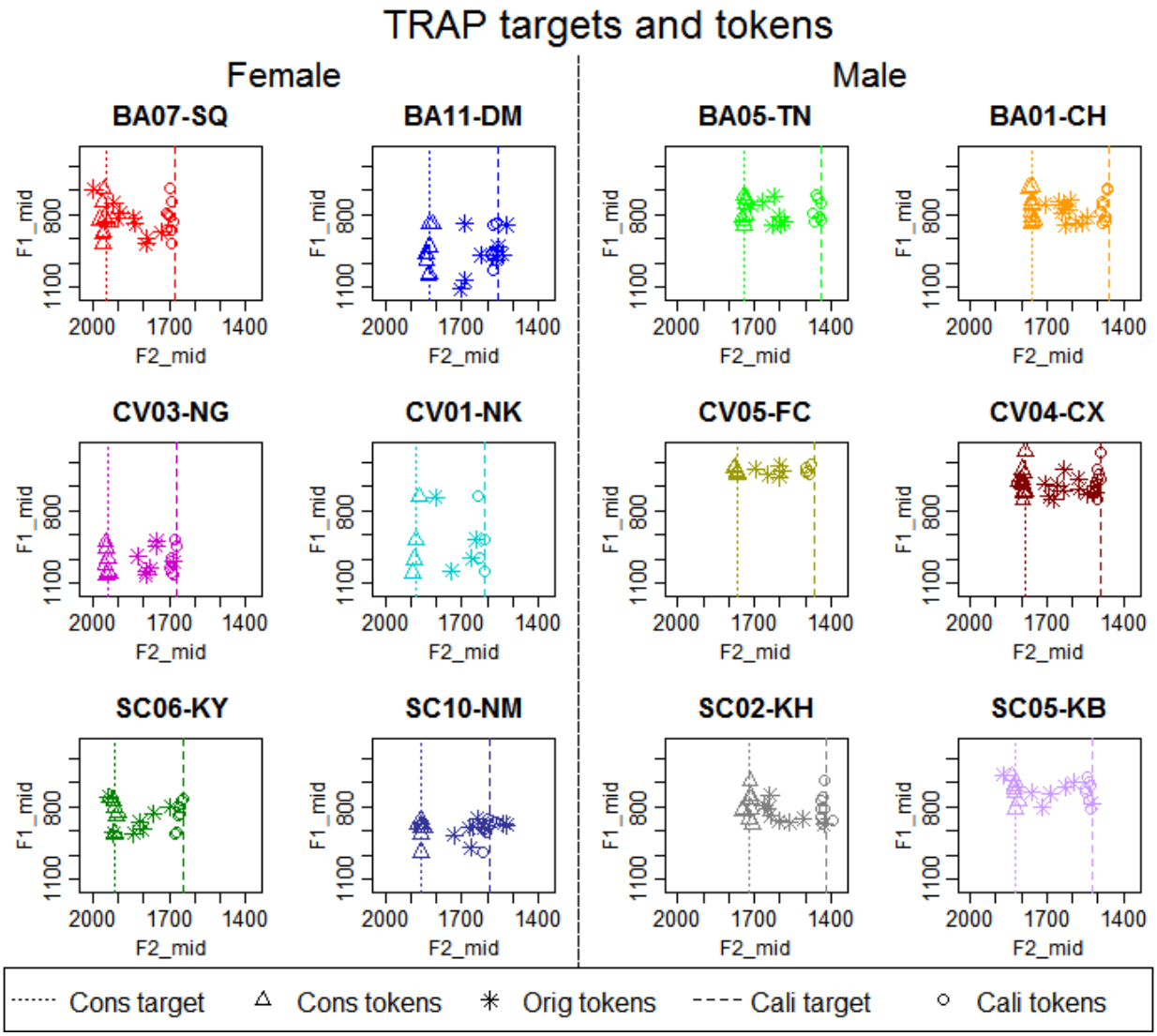

Figure 3.3. TRAP targets and original and manipulated TRAP tokens for each main study stimulus speaker. Targets are denoted by vertical lines since targets were only defined for F2, not F1; tokens are denoted by symbols.

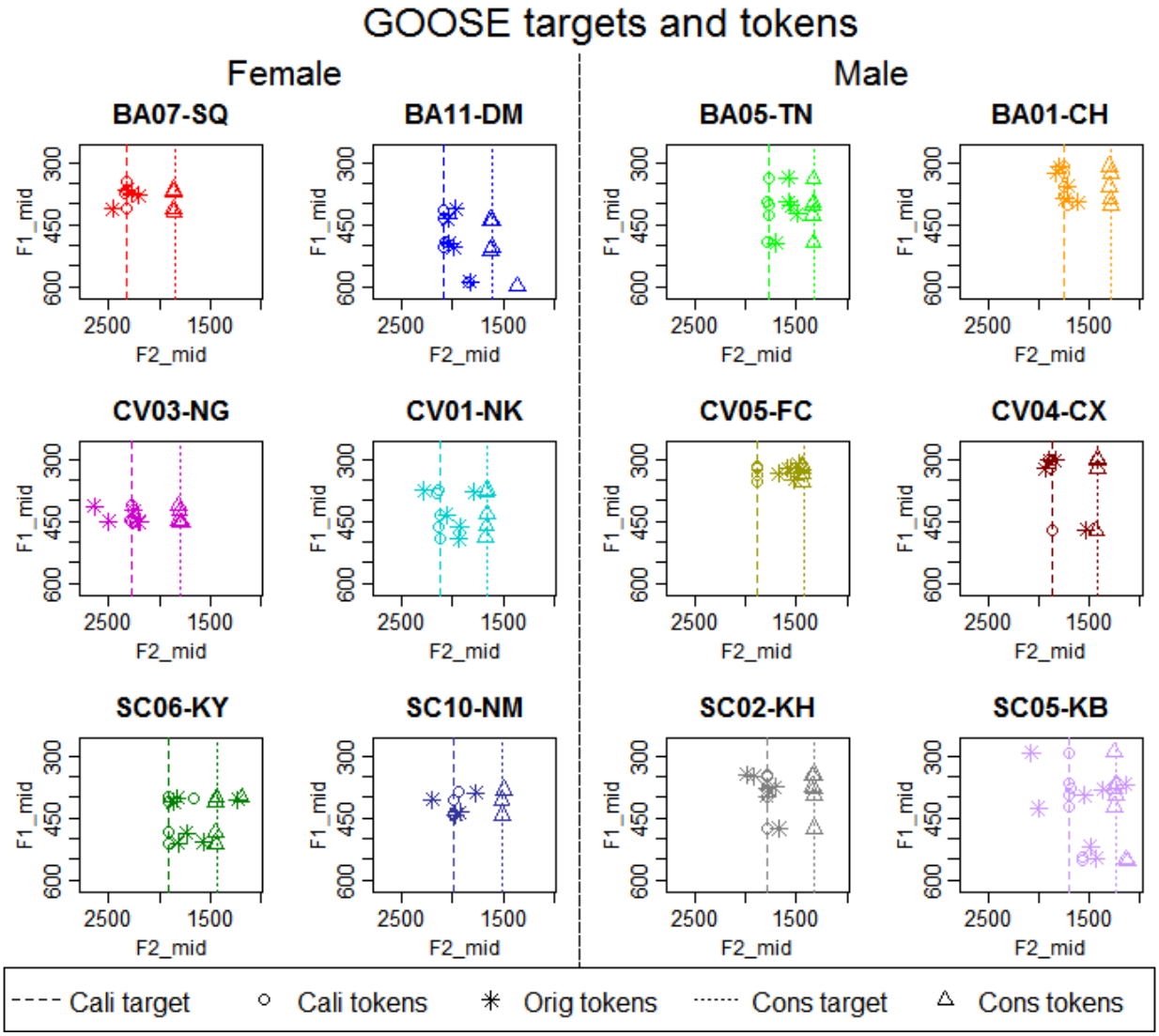

Figure 3.4. GOOSE targets and original and manipulated GOOSE tokens for each main study stimulus speaker. Targets are denoted by vertical lines since targets were only defined for F2, not F1; tokens are denoted by symbols. Four manipulated tokens (one for BA11-DM, one for SC06-KY, and two for SC05-KB) appear substantially offset from the targets. These tokens were non-postcoronal tokens, and as such had both their conservative and Californian targets adjusted to account for the fact that the postcoronal environment promotes GOOSE fronting (see **Error! Reference source not found.**. **Error! Reference source not found.**, above); that is, these tokens are not actually off-target, as their targets are different from other tokens'.

Figures 3.30 and 3.40 reveal an unintended side effect of the manipulation process: F1 changed for many tokens, in some cases dramatically, even though F1 was not manipulated. For example, BA01-CH's lowest original TRAP F1 is 744 Hz, but two of his manipulated TRAP tokens are under 700 Hz for both guises. On average, F1 changed by 16.05 Hz for TRAP tokens and 3.24 Hz for GOOSE tokens. In general, F1 changed to a similar degree for both guises of the same

token (though this was not the case for all tokens, as with BA07-SQ's stray raised Californian GOOSE token). On average, the difference between the same token's conservative and Californian F1 was 5.46 Hz for TRAP and 3.25 Hz for GOOSE, well below Kewley-Port and Watson's (1994:492) JND values of 19.55 Hz for LOT F1 (TRAP F1 is not given) and 10.45 Hz for GOOSE F1. F3 values were slightly more off target, as the average difference between the same token's conservative and Californian F3 was 12.15 Hz for TRAP and 14.16 Hz for GOOSE (not counting tokens where F3 was intentionally raised). It is not clear what produced these changes in F1 and F3, but it is likely a consequence of the difficulties in decoupling the filter from the source (see Appendix E).

This method also proved to be relatively speedy. In total, manipulating 152 tokens twice across 24 excerpts took slightly under three minutes, about a second per token. Among these 304 manipulations, the median number of iterations of the main manipulation step (see Appendix E) was seven, although it was not uncommon for some tokens to see 20 or more iterations.

Appendix E. Implementation of acoustic manipulation

The basic procedure for vowel resynthesis, which relies upon the source–filter theory of speech production, is as follows: linear predictive coding (LPC) is performed on a vowel to estimate the acoustic effects of the filter (formants), the vowel is inverse-filtered to derive the underlying source, the filter is modified, and the source is passed through the modified filter. This basic procedure required several modifications to be implemented in this study in order to preserve naturalness, however, as it is relatively easy to produce obviously computer-like sounds with this process. This is especially the case given that the process had to account for 152 tokens (each manipulated twice), some of which were rather far from their target values, produced in spontaneous speech by 12 speakers with differing vocal tract characteristics and embedded within carrier phrases. In order to ensure that the process was replicable and that it applied evenly to all 152 tokens, I coded several Praat scripts (available upon request) that took the 24 excerpts as input (as well as tabular data on the timing, formant measurements, and targets of each token) and produced two versions of each excerpt as output: one conservative guise and one Californian guise.

The acoustic manipulation proceeded according to the following steps for each token:

1. The token was extracted from the excerpt with a buffer of 50 ms at the left edge and the right edge (later manipulation steps would trim these buffers). For the purposes of acoustic manipulation, tokens were defined as the TRAP/GOOSE vowel plus any preceding and/or following sonorant (except /r/) within the same word. These neighboring sonorants were included in the manipulation given that sonorant consonants have formant structure (Johnson 2003) in order to avoid abrupt "jumps" between the F2 of the unmodified sonorant and the modified vowel. For some

vowels, a neighboring sonorant reacted poorly to the manipulation; in these cases, the neighboring sonorant was simply excluded from the token (i.e., not manipulated), with no appreciable negative effect on naturalness. Neighboring /r/s were excluded from all tokens because /r/ regularly resisted being manipulated cleanly.

2. The token was separated into a high-frequency component (via high-pass filtering) and a low-frequency component (via downsampling). Because LPC calculations find the number of formants within a given frequency range (and because most spectral information about vowel location is relatively low in frequency), it is easier to perform an LPC over a smaller low-frequency range than to attempt to count formants in higher frequencies, where their low amplitude makes them difficult to detect. As a result, the original token was high-pass filtered above a specified maximum frequency to create the high-frequency component and the token was downsampled from 44100 Hz to twice the maximum frequency (which effectively low-pass filtered it below the maximum frequency). The maximum frequency was specified by token (for reasons that will be discussed below); the maximum frequency was 5000 Hz for most tokens and 5500 Hz for others.
3. Praat computed an LPC on the downsampled token, with a prediction order (i.e., the number of coefficients to use) of twice the number of formants. Like the maximum frequency, the number of formants was found manually for each token; most TRAP tokens were best modeled with 4.5 formants (i.e., 9 LPC coefficients) below the maximum frequency and most GOOSE tokens were best modeled with 5 formants, but some tokens used 4 or 5.5 formants. Praat's standard values were used for the other

input parameters: 25 ms window length, 5 ms time step, and 50 Hz pre-emphasis frequency.

4. The glottal source was derived by inverse-filtering the downsampled token through the LPC (the filter).
5. The LPC was converted to a Formant object, a type of Praat object that stores the frequency and bandwidth of each formant at each time point. (In Praat, Formant objects store the same basic data as LPC objects, but Formant objects are far easier to manipulate directly than LPC objects.)
6. **Main manipulation step:** The Formant object was adjusted iteratively so that the midpoint F2 of the *vowel* (not the token) matched the target value within an acceptable margin of error.[2] This step utilized a Praat function by which a constant can be added to the value of a formant at all timesteps (e.g., "increase F2 by 70 Hz along the entire trajectory").[3] One issue with this approach is that once the source is filtered through the adjusted Formant object, there is no guarantee that the new formants will fall exactly where they are specified to fall; for example, if the Formant

[2] In an earlier implementation of the acoustic manipulation process, this adjustment step was performed non-iteratively (e.g., if a given vowel had an original F2 of 1582 Hz and its target value was 1322 Hz, the function was used once to subtract 260 Hz from F2), but the result was sometimes wildly off target, especially when the adjustment was large (200 Hz or more). Hence, the main manipulation step was performed iteratively in the final version of this manipulation process.

[3] Although Praat's function for altering Formant objects allows for more complex manipulations of formants than simply adding or subtracting a constant (e.g., the linear function used in formant transition smoothing in step 11 below), I determined that this simple additive method would be best for the main manipulation step. One advantage of this method is that it preserves the original trajectory of the formants, which is important because formant transitions serve as an acoustic cue to neighboring stops' place of articulation (Johnson 2003). More importantly, this method could be easily applied to all of the tokens in the excerpts, rather than attempting to come up with different manipulation formulae for a variety of cases. Moreover, the tokens produced by this method seemed to satisfy the criteria of naturalness and generalizability upon auditory inspection; similarly, in synthesizing vowel continua based on actual speakers' vowels, Fridland et al. (2004:7) found that "the most natural sounding tokens resulted from the same degree of formant change (ΔF) along the entire vowel trajectory." It should be noted, however, that because of the 'messiness' of formant manipulation (see fn. 4), it was not always the case that formants moved by the same amount along their trajectory.

object is adjusted so that F2 is increased by 70 Hz, the resultant sound may have an F2 that is 50 Hz higher, 68 Hz higher, or 84 Hz higher than the original. In short, the 'messy' nature of speech sounds (especially those produced spontaneously)—not to mention the messiness of Praat's estimates of formant frequencies—precludes a hyper-precise adjustment of formants in resynthesized vowels.[4] Fortunately, hyper-precision is not necessary because there is a lower limit to which human perceivers can detect differences in formant frequencies (or other types of sensory stimuli): the just noticeable difference (JND). Manipulated vowels were deemed acceptably close to the target value if they fell within 1 JND of the target. This JND standard was derived from Kewley-Port and Watson's (1994:492) empirical JND values of 33.09 Hz and 21.86 Hz for TRAP F2 and GOOSE F2, respectively.

7. After each iteration of the main manipulation step, the distance between the current and target F2 was measured and compared to the JND for that vowel. If the distance exceeded 1 JND, the Formant object's F2 values were adjusted by 1 JND. The source was then filtered through the adjusted Formant, F2 was measured for the resultant sound, and this F2 value served as the input to the next iteration of the main manipulation step.[5] If the distance between the current and target F2 was less than 1

[4] This 'messiness' of the manipulation is due not to shortcomings in Praat's implementation but because it is impossible to completely decouple the filter from the source (Titze 2008). In addition, LPCs are imperfect representations of the vocal tract filter. The calculation of LPCs assumes that there are no antiresonances (zeroes) in the signal, so zeroes are not accounted for in the filter and thus remain in the source after inverse-filtering (Ladefoged 1996). As a result, if a formant is manipulated to be near the frequency of a zero that remains in the source, its amplitude will be greatly diminished.

[5] The resultant sound was used only for measuring the change in F2, as this was seldom equal to the JND (for reasons of 'messiness' discussed in fn. 4). The alternative would have been to use this sound as the basis for the following iteration of changing F2, but an earlier implementation of this step that did so showed that this alternative was unworkable. After numerous iterations, the eventual sound was often highly degraded, as the noise introduced by the process of calculating LPCs, inverse filtering to derive the source, and passing the source through the modified filter was multiplied. As a result, the final version of this process modified only the filter iteratively, thus minimizing the amount of noise introduced by filtering.

JND, the Formant object's F2 values were adjusted by the remaining distance. At times, the actual F2 value of the sound resulting from this final iteration was further from the target than that of the penultimate iteration; in these cases, the penultimate version of the filter was accepted as the final version.

8. Once the resultant vowel was acceptably close to the target value (within 1 JND, as determined iteratively), the source was passed through the modified filter.
9. As mentioned in the section on manipulation target calculation,[6] eight tokens of GOOSE and one token of TRAP were set to have their F3 values raised in a particular guise, in order to avoid F2 'colliding' with F3. For these tokens, steps 6–8 were performed on F3 *prior* to F2.
10. After F2 was adjusted and the source was passed through the modified filter (step 8), the new, manipulated low-frequency component was resampled back to 44100 Hz and combined with the (unchanged) high-frequency component.
11. The manipulated token was matched in intensity (amplitude) with the original token. Manipulating the formants also affected the intensity of the token, in some cases causing abrupt rises or drops in amplitude that would be articulatorily impossible. To mitigate this effect, the token was subjected to a script that copied the intensity contour of the original token onto the manipulated token, then scaled the manipulated token to the mean intensity of the original.[7]

[6] From pp. 108–9 of the dissertation, "For eight GOOSE tokens, F2 was raised in the Californian guise to an extent that it got too close to F3, producing odd effects and non-speech-like chirping sounds. In order to avoid this effect, F3 was set to be raised in only the Californian guise by 100, 200, or 300 Hz for these tokens. (F3 was not manipulated for any other tokens.) After this adjustment, the chirping sounds were eliminated and there was no appreciable difference between the quality of Californian GOOSE tokens with raised F3 or non-raised F3. One TRAP token sounded highly nasalized in its conservative guise (but not in its Californian guise), so F3 was set to be raised by 100 Hz in only the conservative guise; after this adjustment, the percept of nasality was mitigated."

[7] Pitch was identical in the original and manipulated tokens, given that pitch is a property of the glottal source and the source was not modified.

12. The manipulated token was stripped of the left and right 50-ms buffers and spliced back into either the conservative or Californian guise.
13. Issues of discontinuous formant transitions between manipulated vowels and non-manipulated neighboring sonorants were mitigated by smoothing these transitions. In 14 cases, manipulating a GOOSE token that was adjacent to a non-manipulated sonorant (e.g., *Stu walks*) resulted in a discontinuous 'jump' in F2. To eliminate these jumps, F2 was measured at the start and end points of a 40-ms clip centering on the token–sonorant boundary; F2 was then re-defined as a linear function connecting these start and end points. For example, if F2 was 1552 Hz at the start point and 992 Hz at the end point, a change of −560 Hz, F2 was set to start at 1552 Hz and decrease by a slope of 14 Hz/ms. This clip was subjected to the same process outlined in steps 1–12, except that the main manipulation step (step 6) was performed just once, not iteratively. (Note that this smoothing procedure, which only affected the first or final 20 ms of the token, had no effect on the targets, which were measured at the vowel midpoint.) Although this step bolsters the naturalness of stimuli in theory, in actuality I was only able to detect an increase in naturalness for smoothed over non-smoothed F2 transitions in one out of the 14 cases; in the others, the F2 transitions sounded sufficiently natural regardless of smoothing.
14. Once all tokens were manipulated for a given stimulus, the average intensity of the stimulus was scaled to 65 dB so all stimuli would have the same loudness.

Prior to manipulation, five problematic tokens (one TRAP, four GOOSE) were replaced in their original excerpts. After the acoustic manipulation process was run for all 24 excerpts, a

trained phonetician listened to the manipulated stimuli to gauge naturalness and generalizability. Stimuli were deemed to be satisfactory after several small adjustments.

References

Fridland, Valerie, Kathryn Bartlett, and Roger Kreuz. 2004. Do you hear what I hear? Experimental measurement of the perceptual salience of acoustically manipulated vowel variants by Southern speakers in Memphis, TN. *Language Variation and Change* 16.1-16.

Johnson, Keith. 2003. *Acoustic & auditory phonetics*, 2nd edition. Malden, MA: Blackwell.

Kewley-Port, Diane, and Charles S. Watson. 1994. Formant-frequency discrimination for isolated English vowels. *Journal of the Acoustical Society of America* 95.485-96.

Ladefoged, Peter. 1996. *Elements of acoustic phonetics*, 2nd edition. Chicago: University of Chicago Press.

Styler, Will. 2015. Using Praat for linguistic research. Online: http://savethevowels.org/praat/UsingPraatforLinguisticResearchLatest.pdf.

Titze, Ingo R. 2008. Nonlinear source–filter coupling in phonation: Theory. *The Journal of the Acoustical Society of America* 123.2733-49.

Wassink, Alicia. 2015. Dialect evolution in the Pacific Northwest: Reanalysis and conventionalization of a universal phonetic pattern. Paper presented at the Annual Meeting of the Linguistic Society of America, Portland, OR.