Low-Back Merger in California English: Working Paper

Christian Brickhouse

March 27, 2019

Caveat

This draft represents a work in progress and therefore **you should not cite this work for any reason**. The analysis is still ongoing and so this document will change substantially and without notice. Further work may demonstrate significant errors in present or past texts, and so readers should be aware that this analysis is subject to change as more work is done. This document is not intended as a scholarly publication but as living documentation of the ongoing analysis.

Methods

Data were collected as part of the Voices of California project which conducts sociolinguistic interviews with life-long California residents. At the end of the interview, participants are asked to read a wordlist to ensure that vowels of itnerest are captured. This study analyzes that word list data from 570 California English speakers. The selection criteria excluded any person who lived outside of the fieldsite of interest for more than 3(?) years between the ages of 8(?) and 18, or for more than 6(?) years after 18. The field sites, with year of fieldwork, were Merced (2010), Redding (2011), Bakersfield (2012), Sacramento (2014), Salinas (2016), Humboldt Bay (2017), and Redlands (2018).

The wordlists were force aligned using the Penn Forced Aligner (citation), extracted by automated script (see extract_vowels_cj.praat), and analyzed using PraatSauce (citation). Each vowel was measured at 10 equidistant points providing change in values over time. Because each measurement represents one tenth of the vowel, it is time normalized and represents a position in the vowel rather than an absolute time into the vowel. For normalization purposes as well as analyses fo vowels as atemporal points, the multiple measures for each vowel were collapsed into a single measurement. This was done by taking the mean of all 10 measurements from the duration of the vowel, yielding a mean F1 and mean F2 within the vowel. These took the place of what would otherwise have been F1 and F2 measurements from the midpoint of the vowel. These means were used as the inputs to the Nearey normalization (citation) method implemented in the `library(vowels)` package.

Results

The sample used in this analysis comprises 570 speakers from 7 field sites.

```
## [1] "Number of participants per fieldsite:"
```

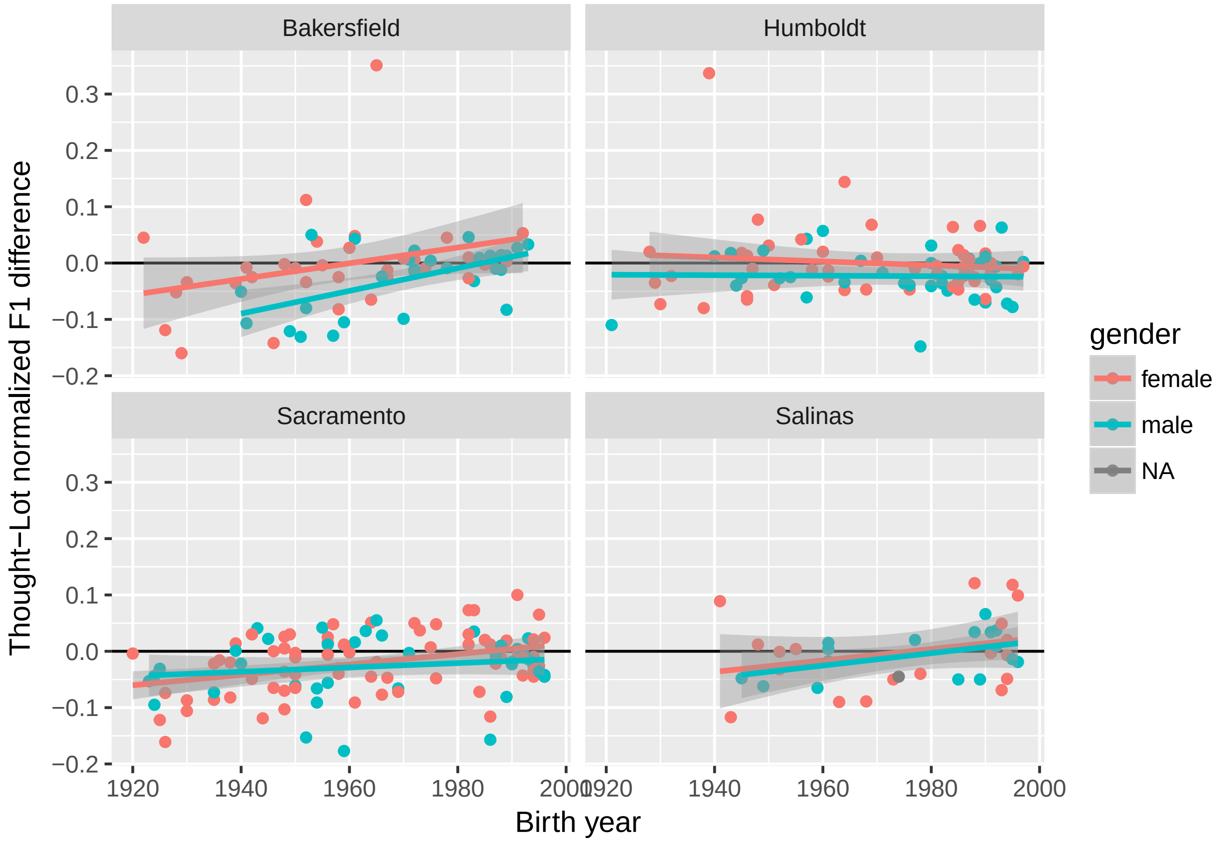
```
##
## BAK HUM MER RDL RED SAC SAL
## 111  96  59  81  97 139  80
```

F1 and F2 space

Measurements of the F1 and F2 space have previously been used to argue for an apparent merger and so the data here should show similar patterns of decreasing distance between LOT and THOUGHT vowels in F1-F2 space. For each participant their normalized formant values for LOT were subtracted from the normalized formant values for THOUGHT so that the degree of overlap can be easily quanitified. If the value is 0 then they overlap each other perfectly.

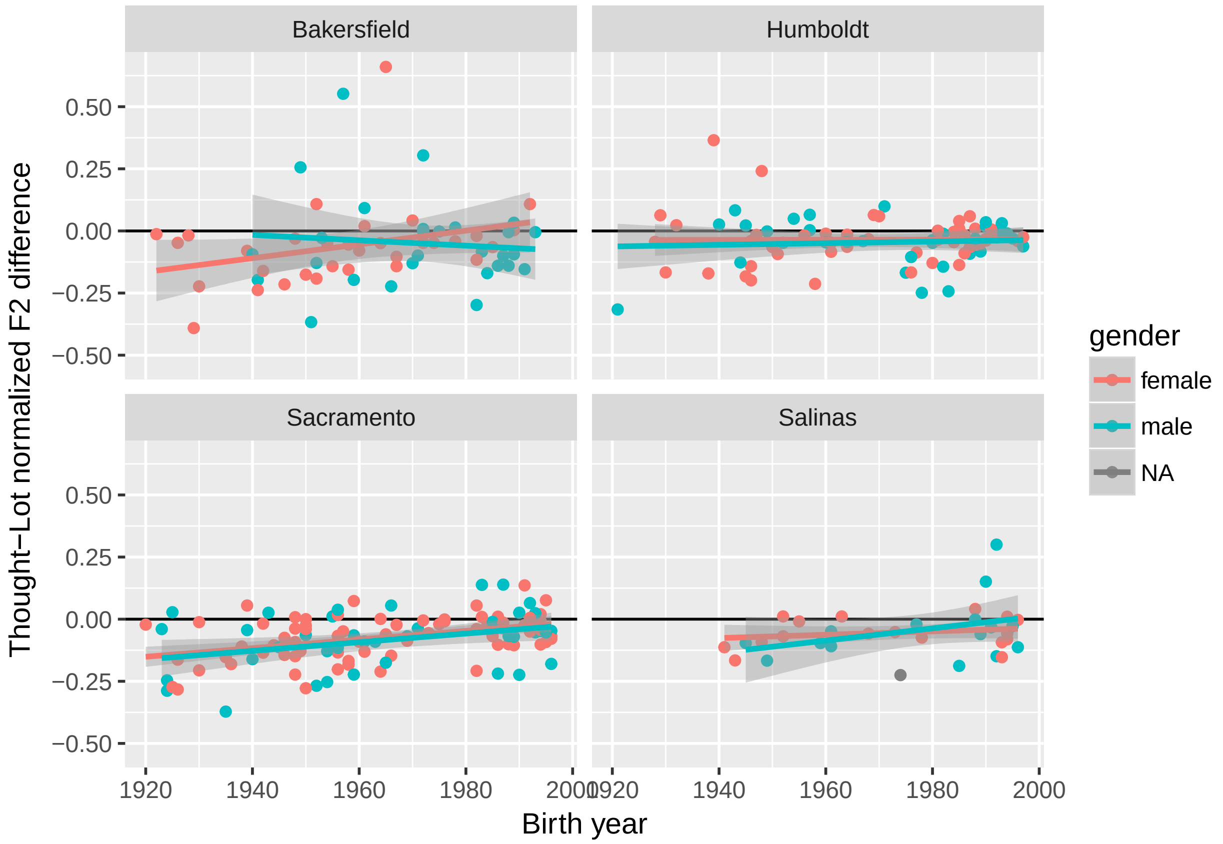
```
## Warning: Removed 33 rows containing non-finite values (stat_smooth).

## Warning: Removed 33 rows containing missing values (geom_point).
```



```
## Warning: Removed 33 rows containing non-finite values (stat_smooth).

## Warning: Removed 33 rows containing missing values (geom_point).
```



In line with previous work we observe substantial similarity between LOT and THOUGHT vowels. The distance between F1 is NA in normalized formant space. The distance between F2 is NA. There is reason to consider that this is not indicative of a complete merger (i.e., not a near-merger) as both these means are significantly different from 0, however the difference may be caused by some areas which lack the merger and some that do.

```
model.f1 = lm(F1diff ~ site*gender*birthyear,data=data.f1f2)
summary(model.f1)
```

```
##
## Call:
## lm(formula = F1diff ~ site * gender * birthyear, data = data.f1f2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.14799 -0.03469 -0.00154  0.02844  0.34442
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -2.7411801  1.0537179  -2.601  0.00978 **
## siteHUM                        3.4258130  1.3162822   2.603  0.00975 **
## siteSAC                        0.9080507  1.2097392   0.751  0.45352
## siteSAL                        0.7414333  1.6773762   0.442  0.65882
## gendermale                    -1.2651091  1.7170860  -0.737  0.46188
## birthyear                      0.0013983  0.0005383   2.598  0.00988 **
## siteHUM:gendermale             0.6462472  2.1328237   0.303  0.76212
## siteSAC:gendermale             2.3182750  1.9688150   1.177  0.24000
## siteSAL:gendermale            1.0824999  2.7169490   0.398  0.69062
## siteHUM:birthyear            -0.0017462  0.0006710  -2.602  0.00976 **
## siteSAC:birthyear            -0.0004750  0.0006175  -0.769  0.44236
## siteSAL:birthyear            -0.0003864  0.0008525  -0.453  0.65073
## gendermale:birthyear          0.0006205  0.0008733   0.710  0.47801
## siteHUM:gendermale:birthyear -0.0003177  0.0010838  -0.293  0.76967
## siteSAC:gendermale:birthyear -0.0011604  0.0010014  -1.159  0.24751
## siteSAL:gendermale:birthyear -0.0005319  0.0013777  -0.386  0.69973
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05747 on 278 degrees of freedom
##   (34 observations deleted due to missingness)
## Multiple R-squared:  0.1259, Adjusted R-squared:  0.07874
## F-statistic: 2.669 on 15 and 278 DF,  p-value: 0.0008236
```

In order to separate out the effects of demographic and social factors on degree of overlap, a linear model was constructed to predict the normalized F1 difference from field site, gender, and birthyear as well as their interactions. The results of the model show that younger Californians produce LOT and THOUGHT vowels closer together in the F1 dimension. There is a significant main effect of birth year such that younger Californians produce closer first formants. There is also a main effect such that speakers from Humboldt tend to produce LOT vowels higher than THOUGHT vowels. There is also a significant interaction such that the effect of birth year in Humboldt is in the direction of greater overlap over time (it is significant because Humboldt has a very positive difference, meaning their LOT vowel was higher than THOUGHT and so the effect of birth year, which would usually increase scores to bring them closer to zero needs to decrease the scores in Humboldt).

```
model.f2 = lm(F2diff ~ site*gender*birthyear,data=data.f1f2)
summary(model.f2)
```

```
##
## Call:
## lm(formula = F2diff ~ site * gender * birthyear, data = data.f1f2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.33909 -0.05704 -0.00009  0.05057  0.70037
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -5.482851   2.086394  -2.628  0.00907 **
## siteHUM                       5.392502   2.606280   2.069  0.03947 *
## siteSAC                       1.996504   2.395322   0.834  0.40528
## siteSAL                       4.092589   3.321257   1.232  0.21890
## gendermale                    7.560992   3.399884   2.224  0.02696 *
## birthyear                     0.002770   0.001066   2.599  0.00986 **
## siteHUM:gendermale           -8.169731   4.223058  -1.935  0.05406 .
## siteSAC:gendermale           -7.568898   3.898315  -1.942  0.05320 .
## siteSAL:gendermale          -11.087751   5.379644  -2.061  0.04023 *
## siteHUM:birthyear            -0.002742   0.001329  -2.064  0.03999 *
## siteSAC:birthyear            -0.001033   0.001223  -0.845  0.39901
## siteSAL:birthyear            -0.002092   0.001688  -1.240  0.21617
## gendermale:birthyear         -0.003849   0.001729  -2.226  0.02682 *
## siteHUM:gendermale:birthyear  0.004153   0.002146   1.935  0.05398 .
## siteSAC:gendermale:birthyear  0.003848   0.001983   1.941  0.05332 .
## siteSAL:gendermale:birthyear   0.005637   0.002728   2.066  0.03973 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1138 on 278 degrees of freedom
##   (34 observations deleted due to missingness)
## Multiple R-squared:  0.09616,    Adjusted R-squared:  0.04739
## F-statistic: 1.972 on 15 and 278 DF,  p-value: 0.0173
```

Another linear model was constructed to predict the normalized F2 difference from field site, gender, and birth year. There is a main effect of site such that speakers in Humboldt have overlapping second formants and Redding speakers produce their LOT vowel much lower than speakers at other field sites (though this result should be interpreted with caution due to the very low number of speakers in that sample). There is a main effect of gender such that men seem to have more similar second formants than do women. There is also a small effect of birth year such that younger speakers have closer second formants. The effect of birth year however doesn't hold in Humboldt where similarity of second formants seems to be unaffected by age. There is an interaction such that younger Redding speakers seem to have closer second formants but the same caveat about sample size applies there.

Gender was an important predictor in F2 distance as it interacted with a number of other predictors, so in order to better interpret those interactions a simple effects analysis was conducted.

```
model.f2.simple = lm(F2diff ~ site*gender*birthyear-site,data=data.f1f2)
summary(model.f2.simple)
```

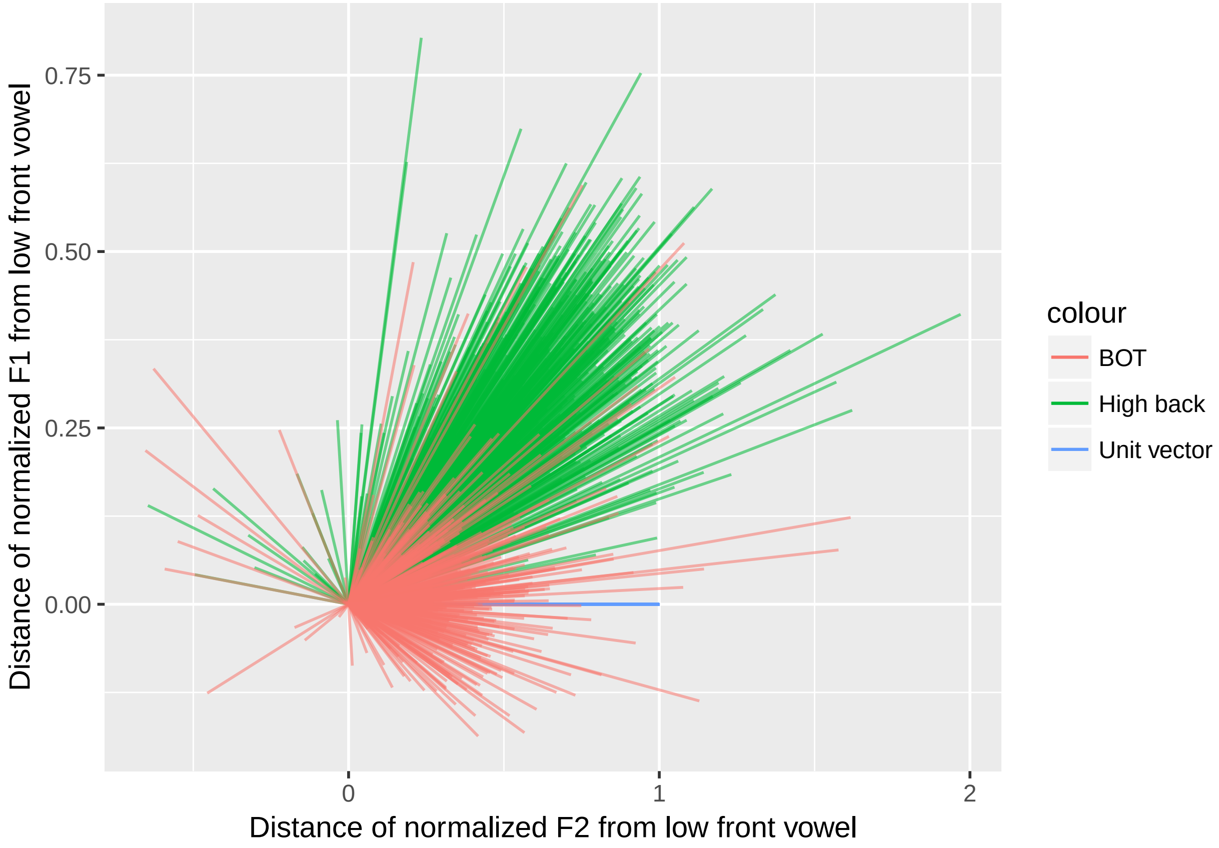
```
##
## Call:
## lm(formula = F2diff ~ site * gender * birthyear - site, data = data.f1f2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.33909 -0.05704 -0.00009  0.05057  0.70037
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -5.482851   2.086394  -2.628  0.00907 **
## gendermale                    7.560992   3.399884   2.224  0.02696 *
## birthyear                     0.002770   0.001066   2.599  0.00986 **
## siteHUM:genderfemale          5.392502   2.606280   2.069  0.03947 *
## siteSAC:genderfemale          1.996504   2.395322   0.834  0.40528
## siteSAL:genderfemale          4.092589   3.321257   1.232  0.21890
## siteHUM:gendermale           -2.777228   3.322879  -0.836  0.40399
## siteSAC:gendermale           -5.572393   3.075597  -1.812  0.07109 .
## siteSAL:gendermale           -6.995162   4.232000  -1.653  0.09948 .
## siteHUM:birthyear            -0.002742   0.001329  -2.064  0.03999 *
## siteSAC:birthyear            -0.001033   0.001223  -0.845  0.39901
## siteSAL:birthyear            -0.002092   0.001688  -1.240  0.21617
## gendermale:birthyear         -0.003849   0.001729  -2.226  0.02682 *
## siteHUM:gendermale:birthyear  0.004153   0.002146   1.935  0.05398 .
## siteSAC:gendermale:birthyear  0.003848   0.001983   1.941  0.05332 .
## siteSAL:gendermale:birthyear  0.005637   0.002728   2.066  0.03973 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1138 on 278 degrees of freedom
##   (34 observations deleted due to missingness)
## Multiple R-squared:  0.09616,    Adjusted R-squared:  0.04739
## F-statistic: 1.972 on 15 and 278 DF,  p-value: 0.0173
```

The marginal interaction between Humboldt and gender seems to be driven by women who have a great deal of overlap in their second formants. The interaction between Salinas and gender on the other hand seems to be driven by men who produce more distinct second formants.

Vowel Space Shape

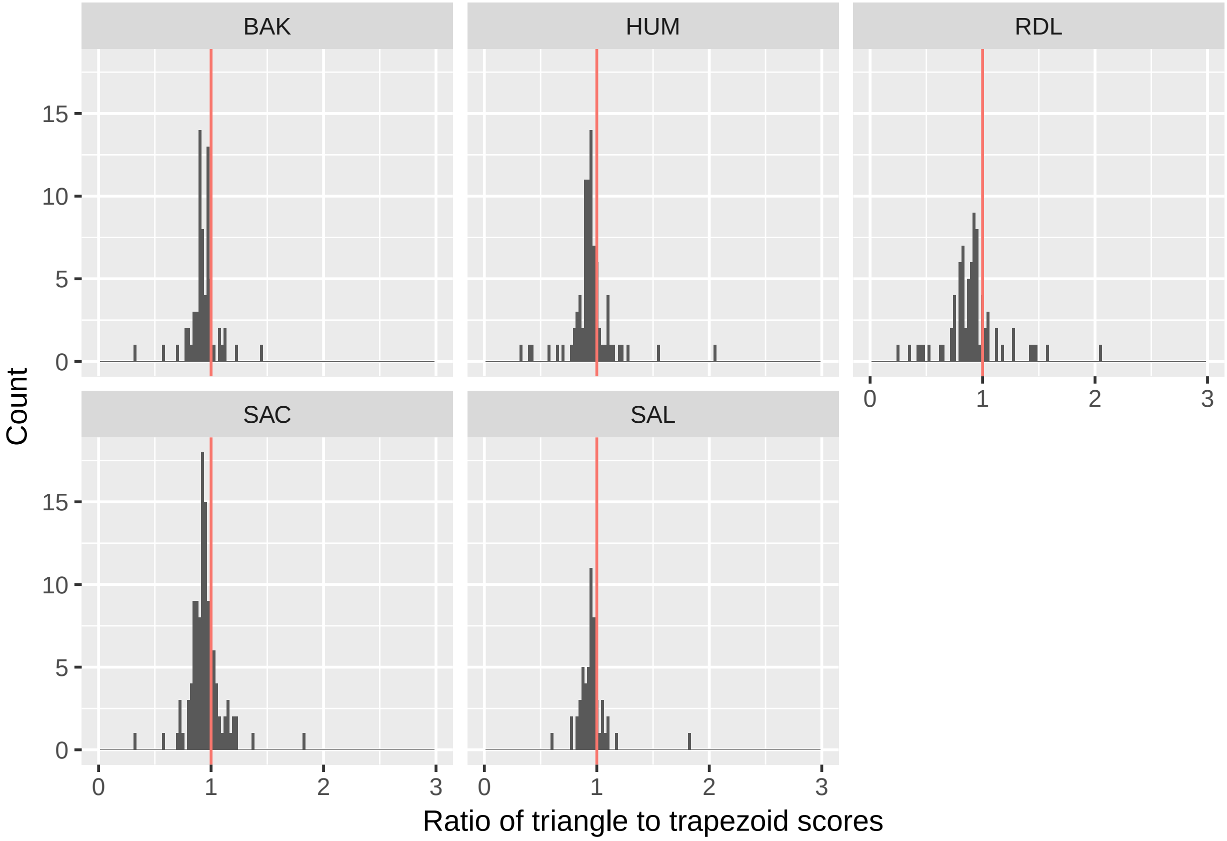
To investigate whether the phonetic movement of the LOT and THOUGHT vowels is causing a shift in the shape of the vowel space, the vowels were treated as vectors and their angles measured. If the vowel space were triangular then the LOT-THOUGHT vowels should lie on the line between the lowest vowel and the highest backest vowel. For each speaker an algorithm identified both the lowest vowel that was not LOT or THOUGHT and the highest backest vowel. The high back vowel was defined as a vowel in the vowel classes of BOAT or POOL that had the lowest Euclidean distance from the origin in Hz space (i.e., closest to the origin).

The high back vowel and the LOT vowel were then converted to vectors from the low front vowel in order to test the hypothesis that LOT lies along the line between the low vowel and high back vowel. If this hypothesis were true then the angle between the two vectors should be zero. The alternative hypothesis—that the vowel space is trapezoidal—would predict that the LOT vector would have an angle of zero with the horizontal (F2) unit vector. As the plot below shows, the LOT vector appears to have minimal overlap with the high back vector, suggesting that there is not a consensus for triangular vowel spaces.

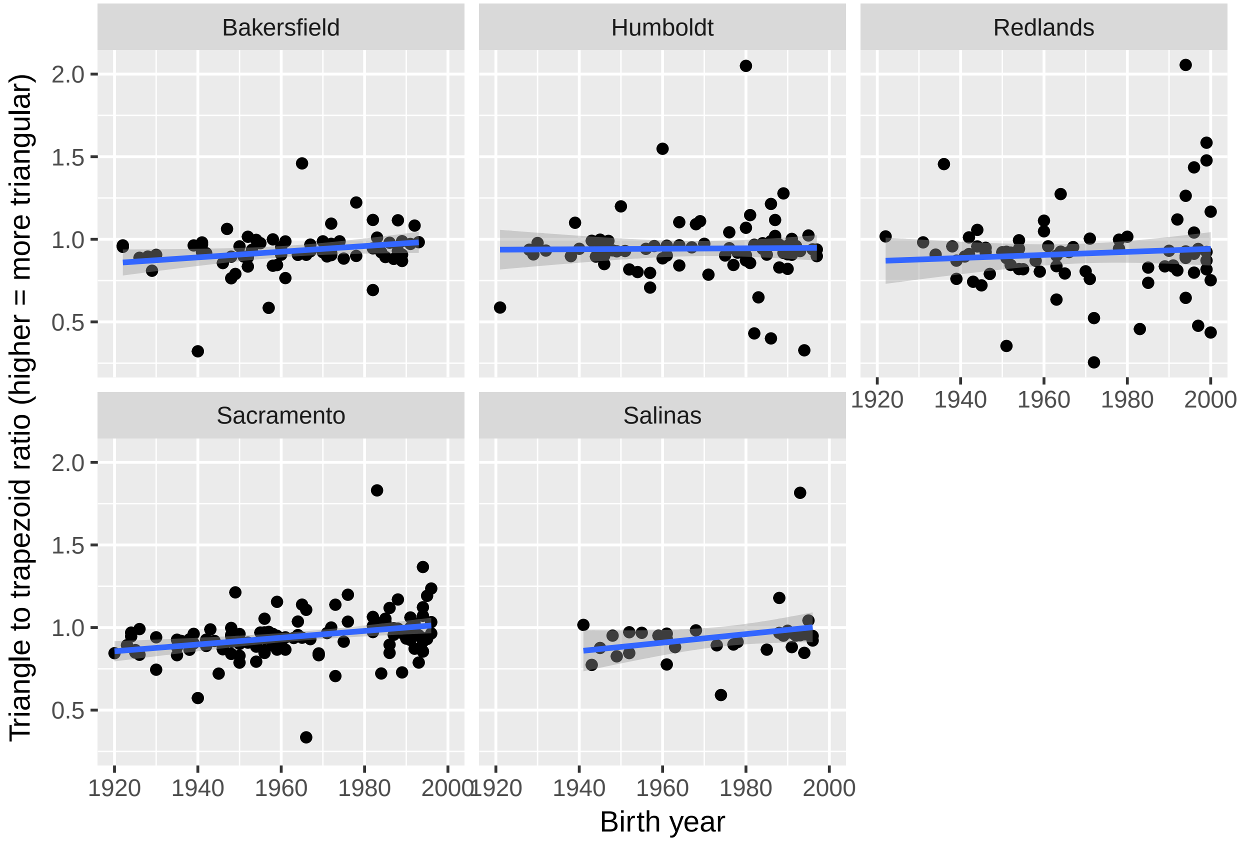


To more precisely test this hypothesis, a metric for how triangular or trapezoidal a speaker's vowel space is was computed. The triangularity of a vowel space was defined as the angle between the LOT vector and the high back vector. The trapezoidality of a vowel space was defined as the angle between the horizontal unit vector and the LOT vector. The ratio of triangularity to trapezoidality allows for the comparison of which state, triangle or trapezoid, best describes a speaker's vowel space shape. If this ratio has a value of 1 then the LOT vowel lies perfectly between triangular and trapezoidal states. If the value is greater than 1 then the vowel space is more triangular than it is trapezoidal supporting the hypothesis that the LOT-THOUGHT movement is causing a more triangular vowel space. However if the ratio is less than 1 then the vowel space is more trapezoidal supporting the alternative hypothesis that the LOT-THOUGHT movement is not causing a change in the vowel space shape. The plot below corroborates the intuitions from the previous figure as for each field site (Merced and Redding were not included due to small sample size) the bulk of speakers have a more trapezoidal than triangular space, though a notable few have more triangular vowel space.

```
## Warning: Removed 21 rows containing non-finite values (stat_bin).
```



However because the California Vowel Shift seems to be a change in progress, this analysis should consider how the vowel space has been changing over time. The figure below shows the triangle to trapezoid ratio for a participant by their birth year, and there appears to be a trend in apparent time whereby younger participants have a more triangular vowel space. For this analysis the data was further subsetted to remove those whose birthyear was unknown and any participant with negative ratio values as they were few and (from the first figure) likely to be measurement errors. The pattern seems strongest in Bakersfield and Sacramento. Redlands looks to have an increasing trend however younger speakers there seem to have a wider envelope of variation which may indicate an interesting social patterning of the low back vowels.



Observations

Given the pattern of data in both analyses it seems that over time and across California the LOT and THOUGHT vowels are moving closer together and perhaps upwards or inwards.

Humboldt also seems to be a very interesting location as it appears the merger in production took place before the other field sites.