The Early Influence of Phonology on a Phonetic Change

Josef Fruehwald

University of Edinburgh

The Early Influence of Phonology on a Phonetic Change

Abstract

The conventional wisdom regarding the diachronic process whereby phonetic phenomena become phonologized appears to be the "error accumulation" model, so called by Baker et al. (2011). Under this model, biases in the phonetic context result in production or perception errors, which are misapprehended by listeners as target productions, and over time accumulate in new target productions. In this paper, we explore the predictions of the hypocorrection model for one phonetic change (pre-voiceless /ay/ raising) in detail. We argue that properties of the phonetic context under-predict and mischaracterize the contextual conditioning on this phonetic change. Rather, it appears that categorical, phonological conditioning is present from the very onset of this change.[*]

Keywords– phonologization, hypocorrection, phonetic change, Canadian raising, sociolinguistics

[*]Thanks to Bob Ladd, James Kirby and Ricardo Bermúdez-Otero for their comments on this paper, as well as to William Labov, Eugene Buckley, Charles Yang, Tony Kroch, Joel Wallenberg, Márton Sóskuthy, Michael Becker, Karthik Durvasula, Andrew Nevins and Alan Yu for their comments on earlier versions of this work. I would also like to thank audiences at the NWAV, NELS, and mFiL conferences, the Symposium on Historical Phonology, as well as audiences at Newcastle and York Universities. All errors are, of course, our own.

1 Introduction

As a general model of sound change, hypocorrection, proposed by Ohala (1981), appears to be a solution to a number of problems in phonology, phonetics, and sound change, including the Actuation Problem (Weinreich et al., 1968), the Incrementation Problem (Labov, 2001), the naturalness of sound change, and the naturalness of phonological processes (Blevins, 2004). The canonical example of contextual /u/ fronting from Ohala (1981) can be summarized as follows. When adjacent to a coronal, like /t/, a /u/ may be either produced or perceived as further front, perhaps as [ʉ]. This could be either due to physical coarticulation with the tongue body moving towards a fronter target for /u/ than usual, or due to the acoustic effect of a [t] closure on the formant structure of adjacent /u/, perceptually fronting it. Whatever the cause of the perceived [ʉ], hypocorrection occurs when a listener fails to take into account the contextual effects of the speaker's production, and instead reconstructs the speaker's intended production target as [ʉ]. Over the course of many interactions between speakers and listeners, with listeners hypocorrecting, a critical mass of speakers in a speech community may have an underlying distribution of /ʉ/ when adjacent to coronals, and /u/ otherwise, at which point we might say that the language has changed.

The way in which hypocorrection resolves the incrementation and naturalness problems is straightforward. The fundamental mechanism of hypocorrection rests in the physical world and the finite precision of human motor planning. There is a natural and persistent pressure in this model to front a [u] target to a [ʉ] production. The incrementation problem (Why does sound change progress in the same direction over many generations?) is resolved, because both the human articulatory and perceptual systems are remaining constant across the relevant time periods, ensuring a constant bias. An explanation of the naturalness of phonological processes can also be found in hypocorrection. In the case of /u/ fronting, the new distribution of [u] and [ʉ] could described as a phonological rule, like (1).

(1) u $\rightarrow$ [-back] /___[COR]

If we assume that most of the contents of phonological grammars are either the end products of sound changes or the further elaboration and modification of those end products (Blevins, 2004; Bermúdez-Otero, 2007) and that some sound changes are more likely to happen than others because of their articulatory and perceptual grounding (Hale and Reiss, 2008), then it would follow that most phonological processes would be natural, due to their historical origins. The actuation problem is not, however, resolved by the hypocorrection model, as was nicely demonstrated by Baker et al. (2011). Rather than just asking "What actuated a sound change?" the actuation problem really asks "Why did this change occur now, in this dialect, and never before, and not in the neighboring dialects?" Since the motivating factors under hypocorrection are natural and persistent, thus present in all dialects at all times, no good answer to the actuation problem is immediately available; however some approaches, like assuming heterogeneity in the speech community with regards to how often speakers will hypocorrect (Yu, 2013) may go some way towards an answer.

The hypocorrection model is supported by many reasonable assumptions. Experimental work has shown that listeners do hypocorrect, and the way that they do mirrors attested historical changes (Ohala, 1990; Harrington et al., 2007; Yu, 2013, among others). In addition, the outcomes of various simulations of hypocorrection appear similar to the outcomes of attested language changes given similar inputs (Pierrehumbert, 2001; Wedel, 2007; Garrett and Johnson, 2013, among others). One missing strain of evidence for the hypocorrection model of language change is data from actual language change in progress, but this will soon be changing. Variationist sociolinguistics, historically the subfield devoted to the study of language change in progress, is increasing the volume, quality, and time depth of data available to researchers to address questions such as these. Corpora such as the Buckeye Corpus (Pitt et al., 2007), the Origins of New Zealand English (ONZE) corpus, and the Philadelphia Neighborhood Corpus (Labov and Rosenfelder, 2011) are such examples.

In this paper, we investigate the role that hypocorrection plays in conditioned sound changes in the Philadelphia Neighborhood Corpus, specifically the raising of /ay/ before voiceless consonants. This logic of this investigation attempts to evaluate the prediction of the hypocorrection model that phonetic changes are circumscribed by other phonetic properties of speech. In the hypothetical example of /ut/ fronting, the phonetic precursor to /u/ fronting was the coarticulatory pressure and/or acoustic warping from adjacent coronals. Before /u/ fully fronted to [ʉ], this precursor should have been detectable in the speech community.

Importantly, the hypocorrection model predicts that the rate of a hypocorrective change be proportional to the size of the phonetic precursor. For example, suppose in the language, /u/ appeared in two different coronal contexts, pre-alveolar [t] and pre-retroflex [ʈ]. By virtue of its backer place of articulation, [ʈ] would probably exert a weaker coarticulatory/perceptual pressure to front than [t]. As such, when a listener hypocorrects, and incorporates error into their own representation of /u/, that error would be less for /uʈ/ (say [u̟ʈ]) than for /ut/ (say [ʉt]). Over time, and multiple interactions, this would have two results:

i A slower rate of /uʈ/ fronting than /ut/ fronting.

ii A less fronted realization for /uʈ/ than /ut/.

Both of the following would be unexpected under the straightforward hypocorrection account, and it would require further elaboration to account for them.

i An identical rate of fronting for both /uʈ/ and /ut/.

ii Fronting of /uʈ/, but not /ut/.

In addition, if there were a (morpho)phonological process which turned /t/ into /k/ in some contexts, we would expect the coarticulatory/perceptual effect of /u/ fronting to be non-existent before /k/. If a speaker were to hypocorrect, they would update their representation with the error [ʉt], but there would be no such error in the context of [uk]. The result would be:

i /ut/ would front at some rate, and /uk/ would not front at all.

If it were observed that /uk/ were participating in the fronting change to a similar degree as /ut/, then that would be evidence that something a bit more elaborated than hypocorrection is at work.

The example we examine in this paper is directly analogous to the illustrative examples just given. The data is drawn from the Philadelphia Neighborhood Corpus of Ling 560 studies ("the PNC" from here on). We look at the raising of /ay/ from a low nucleus to a mid nucleus before voiceless consonants in Philadelphia ("Canadian" Raising), specifically with how this diachronic change interacts with /t/ and /d/ flapping. We establish that the phonetic precursors which have been suggested for /ay/ raising are either neutralized or strongly mitigated in the flapping context, but despite this fact, the degree to which /ay/ raises or remains low in these flapping contexts appears more or less identical to how it raises or remains low in "faithful," non-flapping contexts. It is the underlying phonological context which best predicts the degree of /ay/ raising, not the phonetic properties of that context.

This case study is an example of a phonetic change progressing at a rate which is disproportionate to the phonetic properties of its context. This should be surprising under the hypocorrection model of phonetic change. We argue that the selection of contexts to undergo /ay/ raising is categorical and phonological. Before concluding, we will examine one alternative explanation (lexical analogy), and will show that insofar as it can be quantitatively operationalized, it does not account for the observed patterns.

2 Data and Methods

The data for this paper is drawn from the Philadelphia Neighborhood Corpus of Ling 560 studies (Labov and Rosenfelder, 2011). It consists of sociolinguistic interviews carried out throughout Philadelphia by graduate students as part of their course work for Ling 560, Researching the Speech Community, at the University of Pennsylvania. The course has run from 1973 to 2012. At the time of writing, 397 of these interviews have been digitized, and approximately 232 hours have been transcribed and force-aligned using the FAVE-suite (Rosenfelder et al., 2014).

In this paper, we will be focusing on the 326 white, Philadelphia-born speakers in the corpus (193 hours of transcribed speech). The earliest date of birth of a speaker in this subsample is 1889, and the most recent is 1998. Figure 1 displays the distribution of ages across each year of the study.

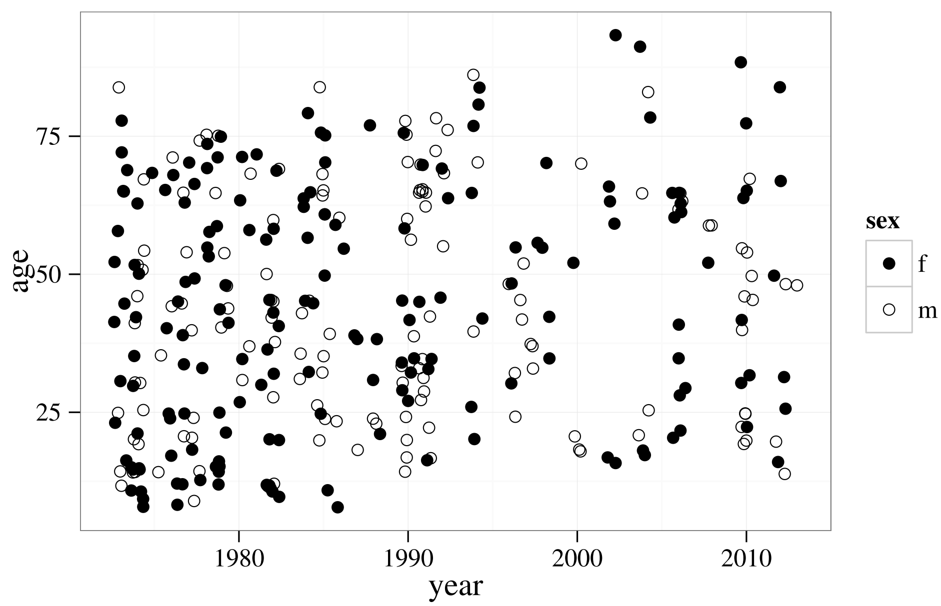


Figure 1: The year of interview and subject's age in the white, Philadelphia born subset of the PNC. A slight jitter has been applied to the year of study.

Point estimates of F1 and F2 were automatically extracted from all vowels with primary lexical stress in this subsample using the Bayesian formant tracking technique from the FAVE-suite (see Labov et al. (2013) for a more complete description), resulting in 615,429 vowel formant estimates. While the FAVE-suite estimates F1 and F2 at 1/3 of the vowel's duration for most vowel classes, they are estimated at F1 maximum for /ay/ and /ey/. Evanini (2009) found that when comparing this automated method of formant estimation to the manual measurements from the Atlas of North American English (Labov et al., 2006), there was a mean absolute difference of about 50 Hz on F1 and 90 Hz on F2 between the two. Within each speaker, F1 and F2 were converted to z-scores (i.e. Lobanov normalization). Adank et al. (2004) found that z-score normalization was the most effective at eliminating physiological differences while preserving social differences, and Rathcke and Stuart-Smith (2014) found that it was most effective at mitigating artifacts caused by poor signal-to-noise ratios and more extreme spectral tilt sometimes found in archival recordings.

3 /ay/ raising: a phonetic change

In his 1944 description of the Philadelphian dialect, Tucker said

> Both the [aɪ]-type diphthong and the [ɑʊ]-type diphthong exist in only one quality, whereas in most American dialects the first element is shortened and modified in quality before a voiceless consonant – the precise sounds vary according to locality. (In my own speech, for example, this short sound, as in *night* or *out*, seems to be identical with the vowel of *but*; contrast with [ɑ] in *ride*, *loud*.) No such distinction is made in the Philadelphia dialect.

It's not clear on which speakers (and importantly, of what generation) Tucker based this statement, but he is very explicit on this point that Philadelphians did not exhibit any raising of pre-voiceless /ay/. His other descriptions of vocalic variation in the same paper are similarly explicit, and concordant with available data, so there is no reason to doubt that his description is in error in this case.

By the 1970s, the Language Change and Variation project at the University of Pennsylvania found that pre-voiceless /ay/ raising was a vigorous change in progress, led by men (Labov, 2001). Subsequent sociolinguistic research on pre-voiceless /ay/ has found that its raising trend has slowed, and that now backer phonetic realizations are associated with masculinity and toughness (Conn, 2005; Wagner, 2007).

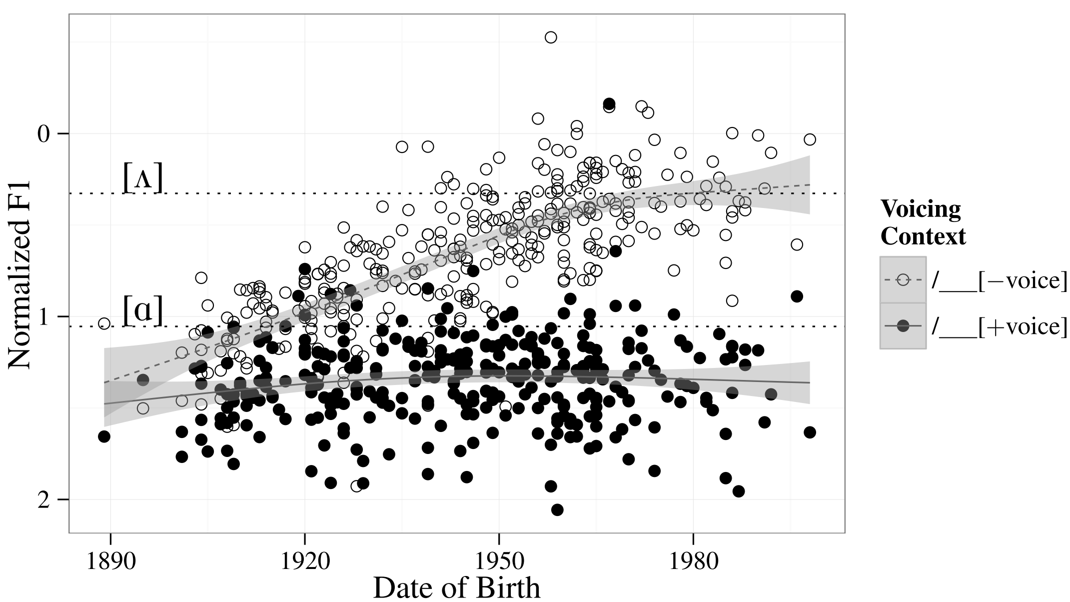


Figure 2: Raising of pre-voiceless /ay/. There are two points for each speaker, representing their mean normalized F1 of /ay/ in the two contexts. The smoothing lines are penalized cubic regression splines. Horizontal dotted lines representing average normalized F1 for [ɑ] and [ʌ] are included for scale.

Setting aside the effect of sex, which for this particular sound change is mitigated in the PNC, the diachronic pattern as described beginning in the 1940s, through to the 1970s and late 2000s is corroborated by the PNC data. Figure 2 plots mean normalized F1 of pre-voiced and pre-voiceless /ay/ for each speaker against their date of birth.[1] We can see how if in 1944 Tucker hadn't based his observations on speakers aged 20 and under (date of birth 1924 and onwards), he probably wouldn't have heard a distinction in vowel quality between *right* and *ride*. Since that time, pre-voiceless /ay/ has changed to an incredible degree. In Figure 2, horizontal lines representing the average normalized F1 of /ɑ/ and /ʌ/ (neither of which exhibit any notable diachronic trend of their own) have been superimposed on the diachronic trends for these /ay/ allophones. For speakers born before the turn of the 20th century, both allophones had a nucleus slightly lower than /ɑ/, but for speakers born around 1970, the nucleus of pre-voiceless /ay/ has risen to be equivalently high as /ʌ/, while elsewhere it has remained at approximately the same lowness for the full century.

At this point, it's important to clarify exactly how pre-voiceless /ay/ has changed. The operating assumption of sociophoneticians when looking at a trend like Figure 2 is that it represents a phonetically gradual change. That is, speakers born in 1900 produced their pre-voiceless /ay/ something like [ɑɪ], speakers born in 1970 produced their pre-voiceless /ay/ something like [ʌi], and speakers born in between produced them in some phonetically intermediate way. However, there are other ways this change could have occurred. For example, there may have been a poorly documented inflow of speakers from Canada (Joos, 1942; Chambers, 1973) or the Northen United States (Dailey-O'Cain, 1997; Vance, 1987) into Philadelphia who brought with them a fully developed pattern of Canadian Raising. The trend in Figure 2 would then be due to population replacement, rather than any linguistic change. If not due to population transfer itself, dialect contact and diffusion (Labov, 2007) could still be the culprit. Perhaps a raised [ʌi] variant entered Philadelphians' repertoire and gradually spread either through the lexicon, or through increasing frequency of use. Categorical variation in vowel quality is certainly possible. For example, Smith et al. (2007) found that categorical variation between [ʌʉ] and [uː] in words like *down* is linked to style shifting in child-caregiver interactions in North-Eastern Scotland. However, insofar as it can be determined from the PNC data, the default sociophonetic interpretation of gradual phonetic change best characterizes the data.

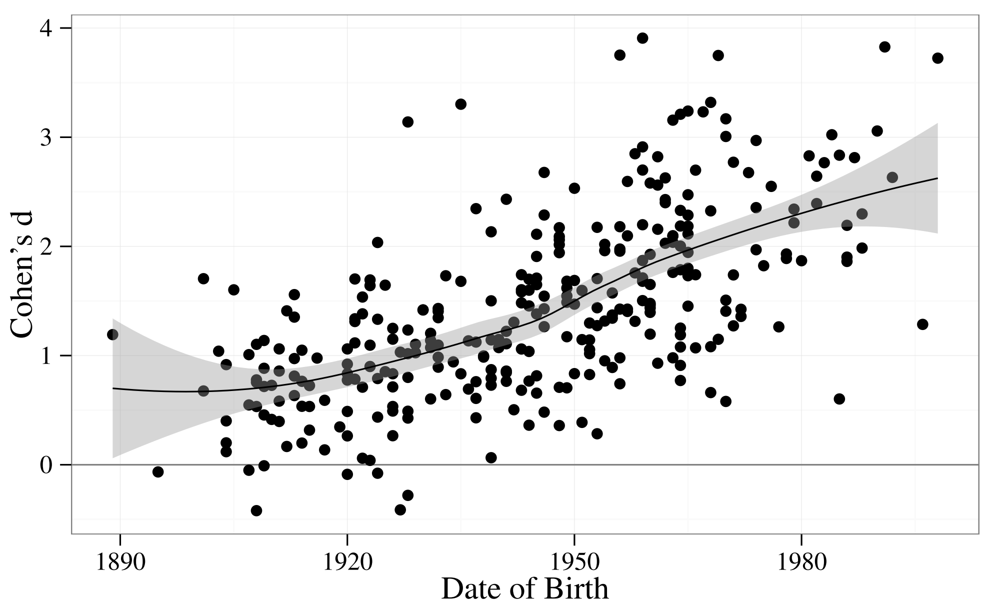


Figure 3: The effect of following voicing (Cohen's d), calculated for every speaker.

[1] Here and throughout, when a figure plots mean formant values for a speaker, they were calculated by first taking the mean formant value for each word for each speaker, and then calculating the mean category value on the basis of these by-word means. This is an attempt to reduce any undue effect of highly frequent words, like *like* and *right* on the estimation of the mean.

To begin with, there does not appear to be two different populations represented in Figure 2. There is no sudden appearance of fully raised [ʌi] speakers, nor a gradual disappearance of [ɑɪ] speakers. In order to visualize this fact more clearly, we calculated the Cohen's d estimate of the voicing effect size for every speaker.[2] If a population with a mature and strong Canadian Raising grammar moved into Philadelphia and replaced the non-raising speakers, we would expect to see the pattern of Cohen's d estimates to form a distribution of two horizontal stripes. The top stripe would represent the incoming speakers for whom following voicing has a strong effect on vowel quality, and the bottom stripe would represent the speakers being replaced for whom following voicing does not have a strong effect. Gradually, the bottom stripe would fade out, and the top stripe would fade in. This hypothetical situation is clearly not the case if we look at Figure 3, which plots every speaker's Cohen's d estimate against their date of birth. Rather than two horizontal stripes representing two different populations of speakers, the pattern instead appears to show one population of speakers which is gradually shifting from having a small effect of following voicing on vowel quality to having a very large effect.

However, neither a plot of speakers' means, nor the plot of Cohen's d would be able to distinguish between continuous and categorical variation within speakers. If speakers born around 1940 produced [ʌi] 50% of the time, or for 50% of their words, and [ɑɪ] the other 50%, their points in Figures 2 and 3 would look approximately same as if they produced something phonetically intermediate between [ɑɪ] and [ʌi] 100% of the time. One way to try to distinguish between categorical and continuous variation is to look at the distributional properties of speakers' data. If, unbeknownst to the researcher, a speaker's pre-voiceless /ay/ productions were drawn from two distributions, one [ɑɪ] and the other [ʌi], the standard deviation of this mixed distribution would be greater than the standard deviation of the two distributions in isolation, and the kurtosis would be less. The kurtosis of a distribution can be thought of either as how sharply peaked a distribution is, how thick its tails are, or, as Darlington (1970) argued, how unimodal it is.

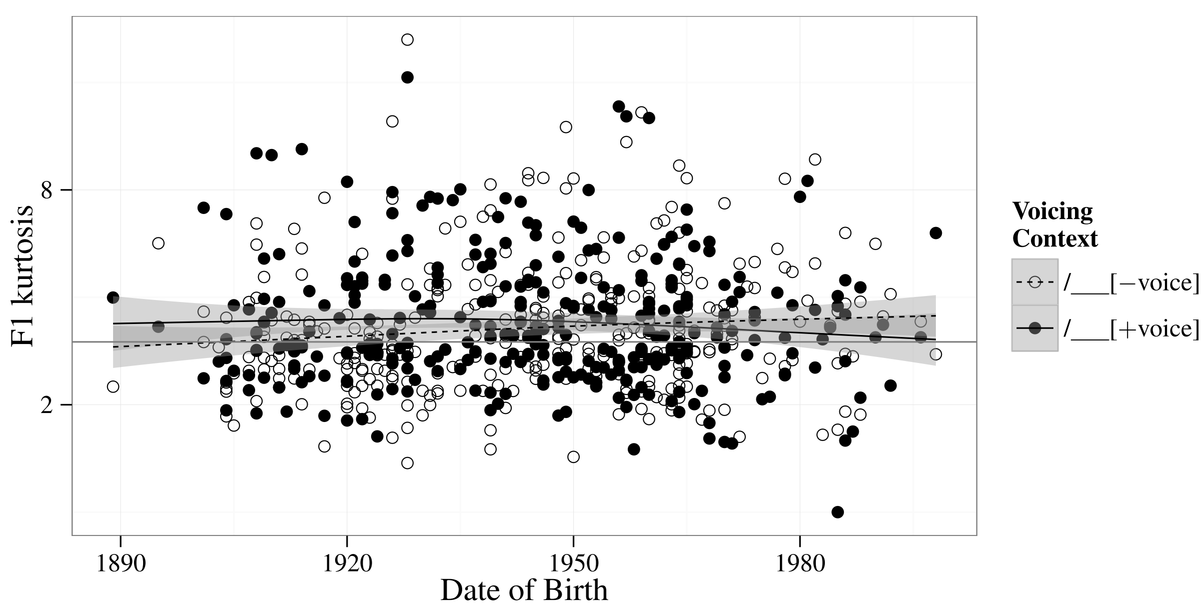


Figure 4: Estimated kurtosis of individual speakers' pre-voiceless /ay/ distributions along F1. The horizontal grey line at y=3 represents the kurtosis of a normal distribution. The y-scale is logarithmic.

[2]Cohen's d was calculated by estimating the mean normalized F1 for pre-voiceless and pre-voiced /ay/, then dividing the difference between these means by the pooled standard deviation.

Figure 4 plots the kurtosis of every speaker's pre-voiceless and pre-voiced /ay/ distributions against their date of birth. A smaller kurtosis corresponds to a flatter, or more bimodal distribution. All normal distributions have the same kurtosis (=3), so a horizontal line at y=3 has been drawn for reference. If the diachronic trend of pre-voiceless /ay/ raising was due to the categorical replacement of an [ɑɪ] distribution with an [ʌi] distribution, then we would expect to see a dip in kurtosis, reaching its minimum at the change's midpoint where the two distributions would be most evenly mixed. From the change's midpoint to its endpoint, kurtosis would begin to increase, reaching its maximum when speakers begin drawing exclusively from an [ʌi] distribution. It should also look markedly different from the kurtosis of pre-voiced /ay/ which didn't undergo any considerable change across this time period. This hypothetical kurtosis profile is not observed for pre-voiceless /ay/ raising. Instead, the trend in kurtosis across the 20th century appears to be more or less flat, with a median value of 3.06: essentially a normal distribution.

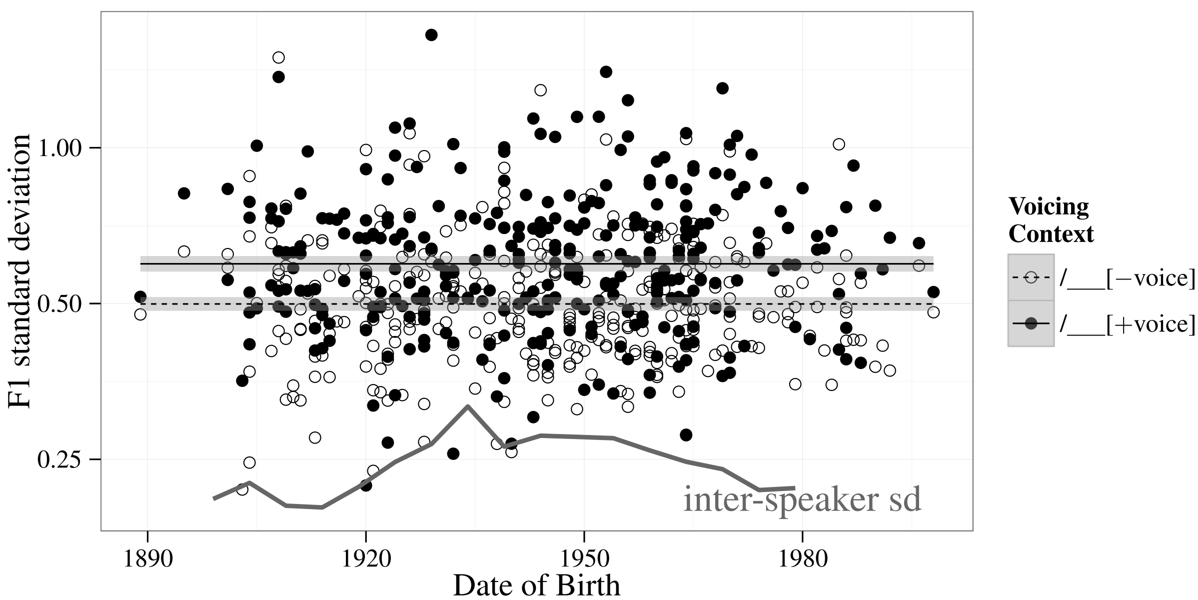


Figure 5: Comparison of within-speaker variation (one point per speaker) to between-speaker variation (grey line). Between-speaker standard deviation calculated based on a moving 20 year date of birth window at 5 year increments. The y-scale is logarithmic.

Turning now to the standard deviation of speaker's pre-voiceless /ay/ distribution, it would exhibit the opposite profile relative to the change from the kurtosis if speakers were drawing from two distinct distributions. The mixture of two normal distributions will always have a standard deviation greater than the the standard deviation of either of the contributing distributions. As such, speakers at the the midpoint of the change, where the mixture of [ɑɪ] and [ʌi] would be most even, would have a greater standard deviation than the beginning or end of the change, where speakers would only be drawing from one or the other distribution. Looking at Figure 5, it appears as if, again, this hypothetical profile of a categorical shift from [ɑɪ] to [ʌi] is not operating. The average standard deviation of pre-voiceless /ay/ appears to be lower than pre-voiced /ay/, probably because it appears in a more restricted set of phonetic environments. This is the opposite of what we would expect to see if pre-voiceless /ay/ raising was a result of mixing [aɪ] and [ʌi] pronunciations.

Included in Figure 5 is a rolling estimate of the inter-speaker standard deviation, based on a window of 20 years, calculated over 5 year increments. This provides some idea of how closely speakers from any given date of birth cohort are clustered together. At any point in time, the degree of between-speaker variation is always less than the degree of within-speaker variation. This information is less relevant to the nature of individuals' behavior, but goes to show that this change wasn't characterized by factions of undergoers and non-undergoers, but rather by relatively strong cohesion within birth cohorts across the entire speech community.

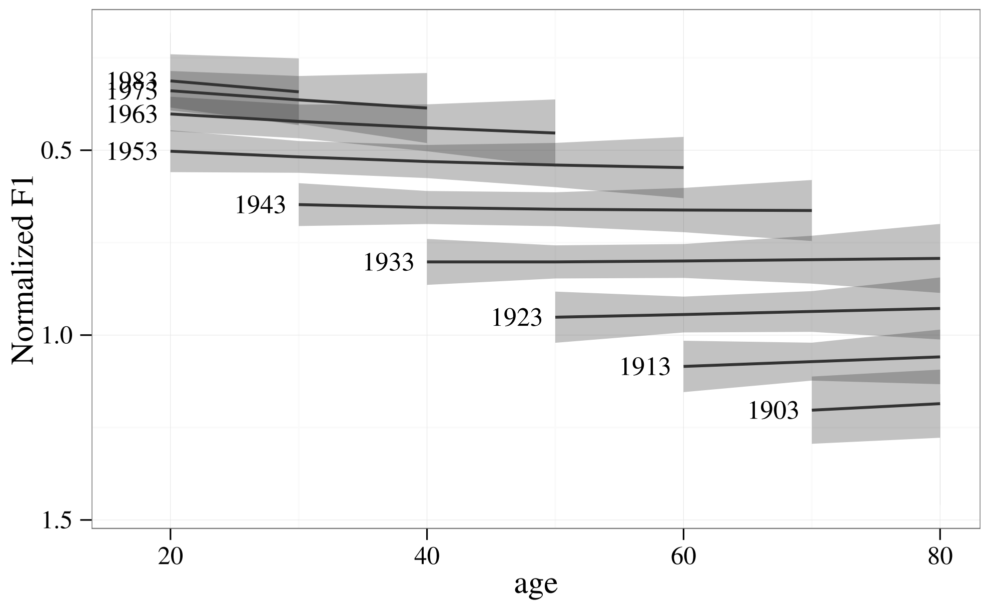


Figure 6: Predicted F1 of pre-voiceless /ay/ for different Date of Birth cohorts at different ages. `gam(NormalizedF1 ~ ti(age) + ti(dob) + ti(age,dob))`

There is also always some concern when relying exclusively on apparent time (Sankoff, 2006) (in this case, speakers' date of birth) that some non-trivial degree of lifespan change is being overlooked. Treating speakers as speech time capsules of the era in which they were born is, of course, a strong idealization, and a number of panel and case studies have found that speakers can, in fact, change their speech well past the critical period (Harrington et al., 2000; Sankoff and Blondeau, 2007). While the PNC does have an important real-time component to it, it is difficult to disentangle the relationship between age and date of birth in its data. By the simple fact that its data was collected between 1973 and 2012, it is impossible for it to contain data from 20 year olds born in 1890, nor 80 year olds born in 1993. Labov et al. (2013) argued that date of birth provided the best explanatory power for the data when compared to age or year of study. In an attempt to improve on that argument here, we fit a generalized additive model, estimating speakers' F1 based on a two dimensional tensor product between age and date of birth using the mgcv package in R (Wood, 2011, 2014). Without getting too deep into the details, this allows us to estimate a non-linear effect of age, date of birth, and their interaction, on pre-voiceless /ay/. Figure 6 plots the estimated lifespan trend for a number of date of birth cohorts. The prediction lines are clipped to the range of what would have been observable within the PNC. While the estimated lifespan trend for speakers born in, say, 1953, is not perfectly flat, it is still not so extremely different from flat that we will be commiting a gross error by only taking into account speakers' date of birth when modeling the change in pre-voiceless /ay/.

Finally, there is always the question of what role word frequency plays in language change in progress (Phillips, 1984; Pierrehumbert, 2001). We'll briefly investigate the effect of word frequency on pre-voiceless /ay/ raising by using the frequency counts from $\text{SUBTLEX}_{\text{US}}$, which contains word counts compiled from the subtitles of U.S. films and television shows. Brysbaert and New (2009) found that these word frequency norms better accounted for participants' behaviors in lexical decision tasks than frequency norms from other sources, meaning they have good psychological validity. Specifically, we'll be utilizing the $\log_2$ transform of the expected word frequency per 1 million words, centered around the median. The $\log_2$ transform lends itself to the intuitively easy interpretation of "the effect of doubling frequency." The full model tried to predict normalized F1 by the three way interaction of date of birth (centered at 1950 and divided by 10, providing the rate of change per decade), $\log_2$ of word frequency, and voicing context (voiced vs voiceless), including random intercepts for speakers and words, with a random slope of date of birth by word, and of voicing context by speaker. The model was fit using the lme4 package, version 1.1-7, in R (Bates et al., 2014), and 95% confidence intervals for the fixed effects were estimated via semiparametric bootstrap replication, using the `bootMer` function. The fixed effects estimates, 95% confidence intervals and density distribution plots of the bootstrap replications are displayed in Table 1.

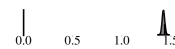
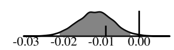
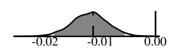
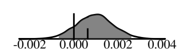
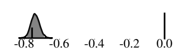
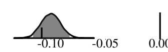
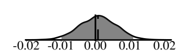
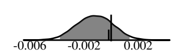
Parameter	Estimate	95% CI	Bootstrap
Intercept*	1.439	[1.402, 1.452]	
Decade*	-0.009	[-0.021, -0.002]	
Doubling Frequency*	-0.011	[-0.018, -0.006]	
Decade × Frequency	0.001	[-0.001, 0.003]	
[-voice]*	-0.755	[-0.784, -0.687]	
Decade × [-voice]*	-0.108	[-0.116, -0.083]	
Doubling Frequency × [-voice]	0.001	[-0.01, 0.011]	
Decade × Frequency × [-voice]	-2×10^{-4}	[-0.004, 0.001]	

Table 1: Parameter estimates and derived values from the mixed effects model `NormalizedF1 ~ Decade * following_voicing * log2Freq + (Decade|Word) + (following_voicing|Speaker)`. `Decade` is the speakers' date of birth, centered at 1950, and divided by 10. `log2freq` is log2(frequency)-median(log2(frequency)), based on frequencies from SUBTLEX. The 95% confidence intervals are based on 5,000 semiparametric bootstrap replicates fitted by `bootMer` from `lme4 v1.1-7`. The density distributions of the bootstraps are provided in the final column.

The fixed effects estimates indicate that there is a main effect of frequency on the F1 of /ay/, but that this effect is not different across the voicing contexts. The effect labeled "Doubling Frequency" corresponds to the effect of doubling frequency on pre-voiced /ay/, and it appears to be reliably different from 0 based on the bootstrap confidence intervals. However, the interaction of "Doubling Frequency × [-voice]" is not reliably different from 0, meaning the frequency effect is more or less the same between pre-voiced and pre-voiceless /ay/. The raising effect of doubling frequency is approximately equal to one year in date of birth for pre-voiceless /ay/.[3] The fact that word frequency affects both pre-voiced and pre-voiceless /ay/ makes sense if we take into account that more frequent words are more likely to undergo phonetic reduction, which in the case of these low vowel nuclei would mean lowering F1. However, word frequency does not appear to reliably interact with the rate of pre-voiceless /ay/ raising. The estimated effect size of the three way interaction "Decade × Frequency × [-voice]" is exceptionally small, and not reliably different

[3] We compared the effect of doubling frequency to the rate of change of pre-voiceless /ay/, which is approximately equal to the effect labeled "Decade × [-voice]" $=-0.108$. A word with double the frequency of another word will have a normalized F1 approximately -0.011 lower, which is about $\frac{1}{10}^{\text{th}}$ the effect of a decade in date of birth, or more simply 1 year.

from 0. These results are very similar to what Zellou and Tamminga (2014) found in the PNC for vowel-nasal coarticulation. More frequent words experienced more reduction/coarticulation, but frequency did not interact with changes in nasal-coarticulation.

If we enter frequency into the model in a stepwise fashion, and compare models using likelihood ratio tests, the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), we can see that the only improvement to the model is made when it is entered as a main effect, without any interactions (Table 2). We can tentatively conclude, then, that the effect of word frequency on the change (i.e. its interaction with date of birth) is marginal at best, and perhaps even non-existent.

	Df	AIC	BIC	logLik	deviance	Chisq	Chi Df	Pr(>Chisq)
no freq	11	54639	54732	-27309	54617			
+freq	12	54634	54735	-27305	54610	6.842	1	0.009
voicing×freq	13	54636	54746	-27305	54610	0.011	1	0.917
decade×voicing×freq	15	54640	54766	-27305	54610	0.351	2	0.839

Table 2: Comparisons of models that differ in terms of how word frequency was included.

Table 3 displays the random effects' standard deviations and correlations from the full model summarized in Table 1. If we compare the standard deviation of [-voice] by speaker, and the residual standard deviation, we can see that these values are broadly similar to the estimates of the within and between speaker standard deviations from Figure 5. These are sensible interpretations of these random effects' standard deviations, and the fact they are so similar to the maximum likelihood estimates from Figure 5 is a sign that the model was sensibly fit.

Group	Name	Std.Dev	Corr
Word	(Intercept)	0.27	
	Decade	0.03	0.39
Speaker	(Intercept)	0.18	
	[-voice]	0.25	−0.72
Residual		0.54	

Table 3: Random effects' standard deviations and correlations from the model described in Table 1

3.1 Preliminary conclusions about pre-voiceless /ay/ raising

On the basis of historical descriptions of the dialect and careful analysis of the data available in the PNC, we can come to a few conclusions about the nature of pre-voiceless /ay/ raising with some certainty. First, it is an innovation that occurred in Philadelphia within the 20th century, and it is not likely that Philadelphians born before 1900 made a distinction in vowel quality between *ride* and *right*. The nature of the innovation also seems clear. It is not the case that speakers with [ʌi] for pre-voiceless /ay/ moved into Philadelphia and replaced the non-raising population, nor does it appear that the speech community was ever divided into a group of speakers who had raising and those who didn't. Nor does it appear that a raised [ʌi] was borrowed into the speech community, nor that there was a period where categorical variation between [ɑɪ] and [ʌi], whether lexically conditioned or otherwise, was the norm. Rather, it looks as if the standard sociophonetic assumption is borne out, summarized in (2)

(2) Pre-voiceless /ay/ changed by a gradual shift in the center of its distribution that propagated in a continuous fashion across generational cohorts.

4 /ay/ raising and /t/, /d/ flapping

Of course, "Canadian" Raising is of great interest to phonologists for its opaque interaction with American /t/ and /d/ flapping (Joos, 1942; Mielke et al., 2003; Idsardi, 2006; Pater, 2014). Even when the voicing contrast is neutralized (or minimized), /ay/ raising still occurs, so that the distinction between *rider* [ɾaɪɾɚ] and *writer* [ɾʌiɾɚ] is maintained in the vowel quality of the preceding /ay/. In contemporary Philadelphia, /ay/ raising does occur before flapped /t/ (this will be demonstrated below) and does not occur before flapped /d/ as a general rule, but there are some lexically conditioned exceptions (Author, 2008). This lexical diffusion pattern appears to be a separate phenomenon overlaid on top of the general raising pattern (Author, 2013) and will not be addressed here.

Given that /ay/ raising in Philadelphia appears to be a phonetically gradual innovation, and that it currently interacts opaquely with flapping, the question arises as to whether /ay/ raising before /t/ flaps is phonetically unexpected at all. Perhaps the phonetic properties of the pre-/t/-flap context are such that we would expect to see raising there. There are three necessary steps to examine this question. First, we need to identify tokens of /ay/ that are almost certainly preceding flaps, and faithful /t/ and /d/. Second, we need to identify the most probable phonetic precursors for /ay/ raising. Third, we need to see in which contexts ([faithful, flapped] × [/t/, /d/]) these precursors were present, and at what strength.

4.1 Identifying Faithful and Flapped /t, d/

For the first step, /ay/ tokens which were most probably preceding faithful and flapped /t/ and /d/ were identified in the PNC. The search definition for /ay/ preceding faithful /t/ and /d/ is given in (3) and the search definition for /ay/ preceding flapped /t/ and /d/ is given in (4).

(3) Faithful (stop realizations):

i. /ay/ is followed by /t/ or /d/.

ii. The /t/ or /d/ is followed by a word boundary.

iii. The /t/ or /d/ is followed by a pause.

(4) Flap:

i. /ay/ is followed by /t/ or /d/.

ii. The /t/ or /d/ is not followed by a word boundary.

iii. The /t/ or /d/ is followed by an unstressed vowel.

Restricting the definition of faithful /t/ and /d/ to be word final and followed by a pause ensures that no tokens of phrase level flaps will be included. Occasionally, the aligner will mis-label a long final closure as a pause, but for these purposes that is a beneficial error, because if the closure was long enough to be mis-labeled a pause, it was certainly not a flap. The resulting numbers of tokens in each context are given in Table 4.

Following Segment	faithful	flap
D	627	245
T	2392	336

Table 4: Number of tokens preceding flapped and faithful /t/ and /d/.

There may be some reasonable concern that this rule based definition of flaps may overapply and label tokens as flaps which are not, in fact, flaps. In order to address this concern, we examined all of the /t/-flap tokens and impressionistically coded them on the basis of the spectrogram. If there was a voicing bar, and/or clear formants in the purported flap, it was coded as being a flap.

After listening to all /t/-flap tokens, 45 were excluded from further analysis. For 17 tokens, the audio was too unclear to accurately code whether the /t/ was flapped, and 2 were excluded because of errors in the transcription. The remaining 26 were excluded because they were actually glottalized tokens, mostly occurring before syllabic nasals. There were no examples of a faithful /t/ in the remaining tokens. The revised numbers of tokens, after these exclusions, are given in Table 5.

Following Segment	faithful	flap
D	627	245
T	2392	285

Table 5: Revised number of tokens preceding flapped and faithful /t/ and /d/.

4.2 Phonetic Precursors

There are two main contenders to be the phonetic precursors of pre-voiceless /ay/ raising. The first, proposed by Moreton and Thomas (2007), is that before voiceless consonants, the offglide of the diphthong is peripheralized. The process of /ay/ raising is thus a result of the diphthong nucleus assimilating to the peripheralized glide. The second, originally proposed by Joos (1942), is that pre-voiceless shortening does not allow enough time for the full gesture from a low-back nucleus to a high front glide, so in reaction, the nucleus raises.

Additional phonetic information beyond point estimates of F1 and F2 are necessary to see how these precursors are distributed across contexts. Specifically, we need full formant tracks to see how the /ay/ glides are affected by flapping, and we need vowel durations. Both of these kinds of data are available from FAVE-extract but they are unfortunately not as high quality as the F1 and F2 point estimates. First of all, FAVE-extract optimizes LPC parameters to arrive at the most likely formant point estimates, but this does not necessarily produce a high quality formant track. This is especially true for a vowel like /ay/, where the high F1 and low F2 at the nucleus would be best estimated with a larger number of poles than the low F1 and high F2 in the glide. The result is that further into the glide, F2 is often poorly tracked. As for vowel duration, the forced-alignments from FAVE-align have not been hand corrected. Alignment errors may be a problem, but Yuan and Liberman (2008) found that most errors based on the acoustic models FAVE-align uses are less than 50ms. However, the aligner only has a precision of 10ms. That is, a phone's duration can increase only by increments of 10ms, and this may pose a problem for finer grained analysis necessary to determine the phonetic precursors of /ay/ raising.

However, we can triangulate between the qualitative generalities of the data from the PNC and results from the literature to arrive at the most likely phonetic situation at the beginning of the change.

4.2.1 Offglide Peripheralization

Using a slightly modified version of FAVE-extract, we extracted F1 from the relevant tokens from 10% to 90% of the duration at 5% intervals. From these full formant tracks, we'll take 80% of the vowel's duration to be indicative of the glide target. Figure 7 plots the height of /ay/ glides over speakers' dates of birth, comparing pre-/t/ and /d/ /ay/ in both flapping and faithful contexts.

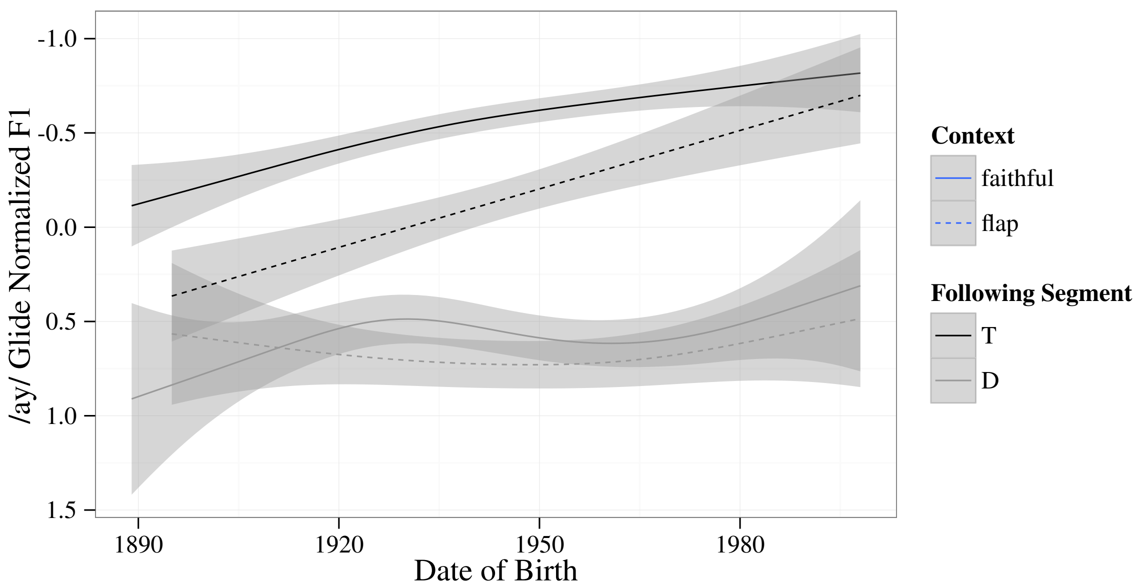


Figure 7: /ay/ glide height across the 20th century.

Looking at the beginning of the 20th century, there is a clear effect of offglide peripheralization for pre-faithful-/t/ only. The height of the /ay/ glide before both variants of /d/ and before flapped /t/ all appear to have roughly the same height. There is a striking raising of the /ay/ glide before flapped /t/ across the 20th century, but this could simply be the result of the raising of the /ay/ nucleus in this context. If the degree of undershoot to the glide remained constant across the 20th century, the mere fact that the nucleus rose would have the knock on effect of also raising the glide.

In an attempt to look at this directly, we calculated the average difference between maximum F1 and F1 at 80% of the vowel's duration for every speaker in these four contexts. This is, in fact, a better measure of the offglide peripheralization precursor, as it directly measures the amount of phonetic space that needs to be traversed in production. Figure 8 plots the diachronic trajectories for these differences.

Briefly looking at the trends for /ay/ before faithful /t/ and /d/ alone, depicted by the solid lines in Figure 8, it would look like the offglide peripheralization hypothesis is strongly confirmed. At the onset of the 20th century, there is a large difference between the nucleus and glide for /ay/ before /t/, and a much smaller one before /d/. There is then a sharp, S-shaped, curve whereby the distance between nucleus and glide is reduced for /ay/ before /t/, such that it comes closer in line with the distance covered by /ay/ before /d/. This distance was reduced primarily by the raising of the /ay/ nucleus towards the glide rather than the other way around. However, the distance from nucleus to glide

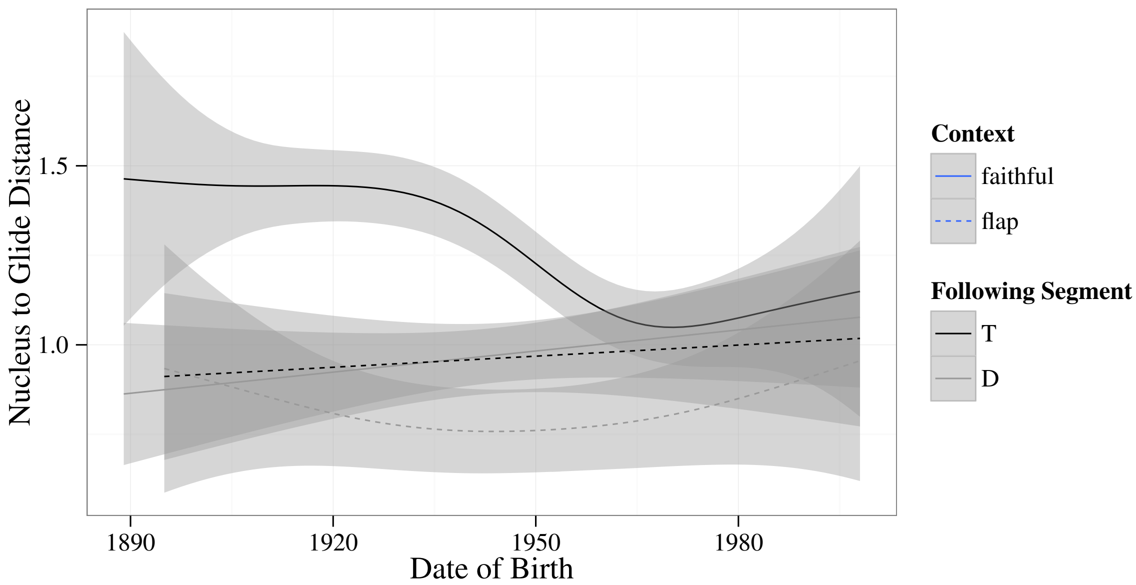


Figure 8: Distance from /ay/ nucleus (maximum F1) to glide (80% of vowel duration) across the 20th century.

for /ay/ before flapped /t/ is indistinguishable from the pre-/d/ context. That is, there doesn't appear to be any offglide peripheralization effect at all for /ay/ when preceding flaps, regardless of their underlying status. This is actually to be expected if the reason for offglide peripheralization is the "spread of facilitation" as Moreton (2004) suggested. Moreton (2004) construes this spread of facilitation as being a low level factor involving "[n]euromuscular coupling between temporally overlapping vowel and consonant articulations." The relationship between any particular [ɾ] and /t/ must be at the cognitive level, relatively far removed from neuromuscular planning, so it is not surprising that there would be no offglide peripheralization before flapped /t/.

These results are concordant with what Rosenfelder (2005) found in Victoria, British Columbia. While she was not similarly focused on the interaction of raising and flapping, she did report the nucleus and glide measurements for /ay/ preceding /t/-flaps separately. Figure 9 plots the means reported in her appendices. We can see that while the glide of /ay/ before /t/-flaps is a bit higher and fronter than the glide before voiced consonants, these glides are more similar to each other than to the glide preceding voiceless consonants. This is even more surprising if we take into account that the nuclei of pre-/t/-flap /ay/ and pre-voiced /ay/ are extremely different.

On the other hand, Kwong and Stevens (1999) did find /ay/ offglide peripheralization was not entirely neutralized before flapped /t/. For most of their speakers, the glide's F1 was lower and its F2 was higher before a /t/-flap than before a /d/-flap. However, they did not report formant estimates for /ay/ preceding faithful /t/ and /d/, making comparison to the data from the PNC and Rosenfelder (2005) difficult. The crux of the matter is not whether there is any offglide peripheralization before /t/-flaps, but rather whether there is enough to drive /ay/ raising before /t/-flaps on phonetic grounds.

Between the data available from the PNC and the data reported by Rosenfelder (2005), it appears as if pre-voiceless glide peripheralization is largely neutralized before flaps. If the contexts in which /ay/ raising took place were defined on the basis of offglide peripheralization, we would expect to see it occurring only before faithful /t/.

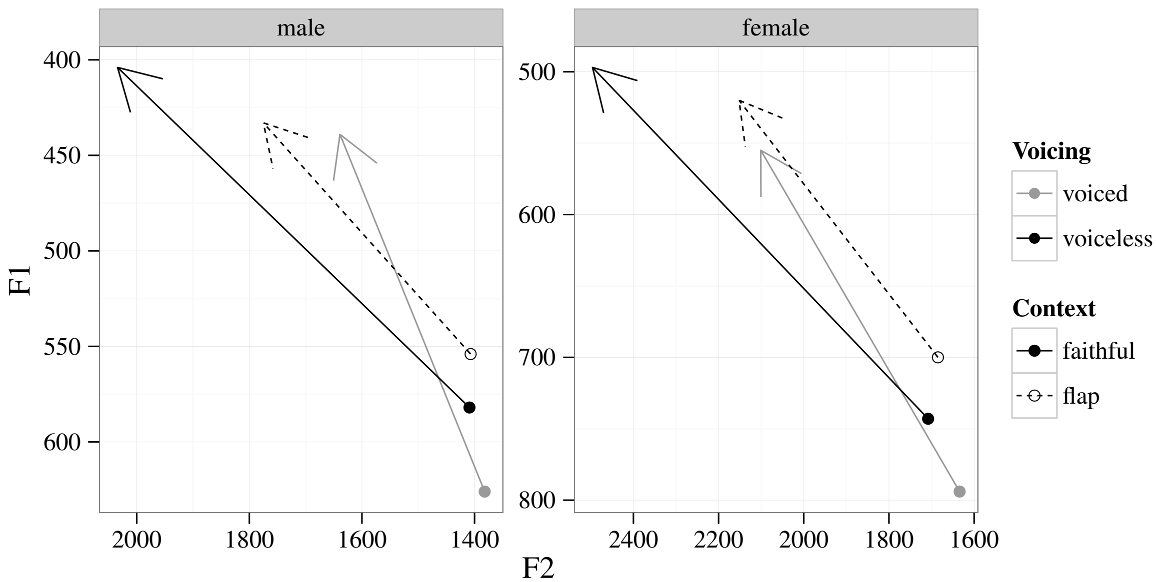


Figure 9: Nucleus and glide measurements of pre-voiced, pre-voiceless, and pre-/t/-flap /ay/ from Victoria, B.C. Data from from Rosenfelder (2005).

4.2.2 Pre-voiceless shortening

Figure 10 plots the diachronic trends for /ay/ durations from the PNC in the four contexts in question. Looking at the early part of the 20th century, we can see that voicing effects on duration are not completely neutralized before flaps, but they are heavily mitigated towards the short end of the spectrum. Table 6 summarizes the mean duration in milliseconds from speakers born before 1920 for /ay/ preceding /t/ and /d/ in flapping and faithful contexts. When arranging contexts from shortest to longest, it looks like when preceding a /d/-flap, the duration of/ay/ is more similar to a faithful /t/ than to a faithful /d/.

Following Segment	Context	Mean Duration (ms)	Difference from next shortest
T	flap	135.4	
T	faithful	146.4	10.95
D	flap	178.3	31.94
D	faithful	227.7	49.42

Table 6: Mean duration of /ay/ preceding both faithful and flapped /t/ and /d/.

context	/ay/ pre-/d/ (ms)	/ay/ pre-/t/ (ms)	difference
faithful	218	135	83
flap	160	124	35

Table 7: Duration differences between /t/ and /d/ in faithful and flapping contexts.

Table 7 displays the same mean durations by following segment and context, emphasizing the duration differences

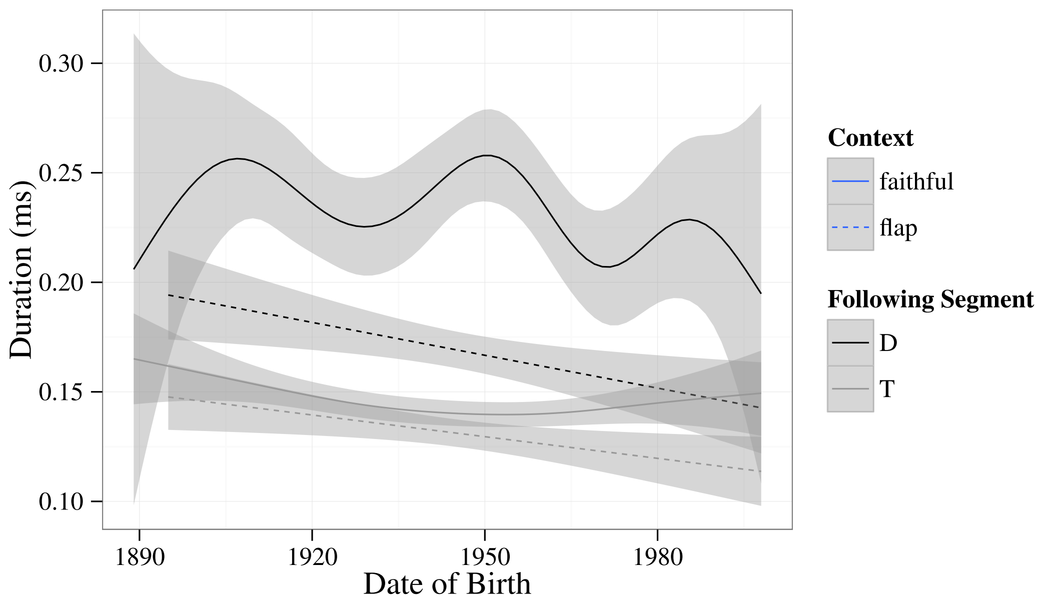


Figure 10: Diachronic trajectory for /ay/ duration across the 20th century

between following /t/ and /d/ within contexts. The duration difference is more than two times greater in the faithful context than in the flapping context.

Braver (in press) recently found an even greater amount of duration neutralization than we have found in the PNC. Using lab speech (and, unfortunately for comparison's sake, not /ay/), Braver found that among his 12 subjects, the largest durational difference before /d/-flaps from /t/-flaps that any of them made was about 15ms. He doesn't report any durations from non-flapped /t/ or /d/, but the vowel durations he reports for pre-/d/-flap vowels are, at most, 140ms. So it appears that in Braver's data, the direction of incomplete neutralization is also towards shorter end of the duration spectrum.

It's not immediately clear why the durational distributions in the PNC should be so different from from Braver (in press). It may be due to dialectal differences between the PNC speakers and Braver's speakers, it may be an /ay/ specific effect, or it may have to do with the methodological and analytic difference between these studies. Regardless, what both sets of results find is that vowels in general, and /ay/ in particular, are shorter before /d/-flaps than before faithful /d/, perhaps nearly (but not statistically) identical to vowels before faithful /t/. This places the predicted participation of /ay/ before /d/-flaps in the raising change in an ambiguous position if it is driven by phonetic length. Before /d/-flaps, /ay/ would either participate to the same degree as /ay/ before faithful /t/, or perhaps would participate a bit more weakly.

4.2.3 Phonetic Precursors Summary

The two main contenders for phonetic precursors of /ay/ raising actually produce different predictions for how /ay/ should behave before flaps. If the precursor were offglide peripheralization, as suggested by Moreton and Thomas (2007), then on the basis of the PNC data and Rosenfelder (2005), we would expect only the context of following faithful /t/ to condition the change, since the glide does not appear to be peripheralized before flaps, nor before faithful /d/. If the precursor were phonetic duration, we would unambiguously expect both /ay/ before faithful /t/ and before /t/-flaps to undergo the change. Our expectations for /ay/ before /d/-flaps are a bit more ambiguous. On the basis of Braver's (in press) study, we would expect /ay/ raising to occur at more or less the same rate before /d/-flaps as before faithful /t/ and /t/-flaps. In the PNC data, /ay/ before /d/-flaps is a bit longer than /ay/ before faithful /t/, but still much shorter than /ay/ before faithful /d/. For the purposes of further investigation, we'll categorize /ay/ before /d/-flaps as a "weak undergoer" context under the duration precursor model. The categorization of contexts into "undergoer" and "non-undergoer" on the basis of these hypothesized precursors are summarized in (5).

(5) faithful /t/   /t/-flap   /d/-flap   faithful /d/

- Above: faithful /t/ = undergoer - periph; /t/-flap, /d/-flap, faithful /d/ = non-undergoer - periph
- Below: faithful /t/, /t/-flap = undergoer-dur; /d/-flap = weaker undergoer - dur; faithful /d/ = non-undergoer - dur

It should be noted that placing contexts into "undergoer" and "non-undergoer" categories is not really consistent with the hypocorrection model of phonetic change. Rather, we would expect the rate of change across contexts to vary continuously in a way proportional to the strength of the phonetic precursor in those contexts. Based on the available precursor data from the PNC, bearing in mind the necessary caveats about its quality, we can try to arrive at more specific, quantitative predictions about how /ay raising ought to interact with flapping.

Taking offglide peripheralizaton first, we'll use the difference between F1 at 80% of the vowel's duration and maximum F1 as our quantitative measure of the peripheralization precursor. We know that /ay/ raising did occur before faithful /t/, and did not occur before faithful /d/. If we take the size of the precursor before faithful /t/ as being at 100% strength, and the size before faithful /d/ as being at 0% strength, we can calculate the relative strength in the remaining contexts, which should be proportional to the degree of participation of /ay/ raising in these contexts. The estimated participation rates based on the peripheralization precursor are given in Table 8. If /ay/ were to raise before flapped /t/ and /d/, we would expect it to do so somewhere between 10% as much as /ay/ before faithful /t/. Taking the same logic and applying it to the duration precursor (Table 9), we find that /ay/ before /t/-flaps ought to participate in the change at a higher rate than /ay/ before faithful /t/, and /ay/ before /d/-flaps ought to participate at about 60% the rate of /ay/ before faithful /t/.

Following Segment	Context	Nucleus to glide distance	participation
T	faithful	1.44	1.00
T	flap	0.91	0.11
D	flap	0.91	0.11
D	faithful	0.84	0.00

Table 8: Estimated participation rates of /ay/ raising on the basis of offglide-peripheralization.

Following Segment	Context	Mean Duration (ms)	participation
T	flap	135	1.13
T	faithful	146	1.00
D	flap	178	0.61
D	faithful	228	0.00

Table 9: Estimated participation rates of /ay/ raising on the basis of phonetic duration.

Section 4.3 is devoted to describing the outcome of a non-linear model of /ay/ raising, but we can briefly prefigure the results in that section by examining a plot of speaker means of /ay/ F1 across these contexts. In Figure 11, we can see that the pattern of raising before /t,d/-flaps looks very similar to the pattern before faithful /t,d/. That is, the observable pattern of raising looks neither like the coarse categorization laid out in (5), nor like the quantitative predictions in Tables 8 and 9. Rather, it appears that /ay/ raising has always been conditioned by the underlying phonological status of the following segment, not the phonetic properties of the context.

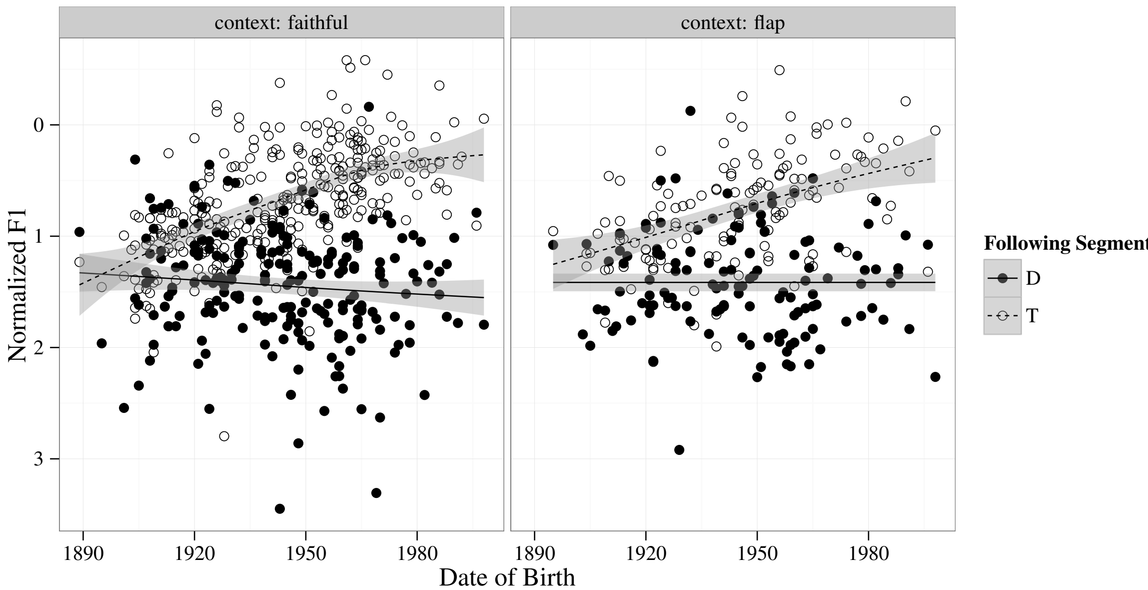


Figure 11: Speaker F1 means for /ay/ preceding different /t/ and /d/ realizations.

4.3 Non-linear Bayesian Modelling

At this point, the qualitative impression from Figure 11 requires quantitative support from statistical modelling. However, standard linear-mixed effects models will be insufficient to address the questions at hand. For example, it might be possible that /ay/ preceding flaps behaved as one of the two precursor hypotheses predict at the beginning of the change, but then underwent reanalysis to be conditioned by underlying voicing at a later point. That is the “initiation” of the change may have been governed by phonetic factors, but the “propagation” of the change by phonological factors (Janda and Joseph, 2003; Ohala, 2012). Simply fitting a linear model to the data would wash out any time dependent effects like that, so a modelling approach with allows for a non-linear relationship between the change and date of birth will be pursued. There are a number of non-linear modelling methods to choose from, including smoothing-spline ANOVAs (Davidson, 2006) and generalized additive models (Wieling et al., 2011). In this paper, however, we will be using a Bayesian method based on Ghitza and Gelman (2014) for a number of reasons. First, the mathematical description of the model is simpler than many other non-linear modelling techniques, even if the way the parameters are estimated is more complex. Second, random effects for speakers and words are more straightforwardly integrated into the model. Finally, it is easier to generate 95% Bayesian credible intervals for all parameters and generated quantities in the model than it is using other methods, and the credible intervals are more intuitively understandable. A full mathematical description of the model is given in Appendix A. This description is probably longer than the descriptions of most statistical models in the literature, but that wouldn’t be the case if the full mathematical description of, say, a smoothing-spline ANOVA, or a tensor product smooth had to be included in the full text.

The single sentence description of the model is it's a first order autogregressive model over the rate of change. The model will produce non-linear estimates for year-over-year changes, which we'll label as δ_j, where j is an index for the date of birth. If normalized F1 of /ay/ before faithful /t/ lowered by 0.01 between 1899 and 1900, we'd say that $\delta_{1900} = -0.01$. The δ_j of any given date of birth is constrained to be similar to δ_{j-1}, i.e. the rate of change of the previous date of birth. Exactly how similar δ_j and δ_{j-1} ought to be is a parameter of the model itself, so the smoothness or wobbliness of the non-linear aspect of the model is optimized on the basis of the data. The expected normalized F1 of /ay/ for a given year, which we'll label as μ_j, is calculated by summing up all of the year over year changes up to the year in question. The first observable date of birth is 1889, so the estimated F1 in 1892 (μ_{1892}) is equal to $\delta_{1889} + \delta_{1890} + \delta_{1891} + \delta_{1892}$. Separate δ_j and μ_j values were estimated for each of the four contexts ([/t/, /d/]$\times$[faithful, flap]), so that there would be no bias in the model for any of the differences between curves to shrink towards 0. Random intercepts of speaker, word, and street were also included in the model.

The parameters of the statistical model described in Appendix A were estimated using the No U-Turn Sampler (NUTS) (Hoffman and Gelman, 2011) as implemented in Stan (Stan Development Team, 2014). It is a form of Markov Chain Monte Carlo, which takes an iterative approach to estimating the probability of parameter values given the data (a.k.a. the posterior), which is proportional the the probability of the data given the parameters (the likelihood) times the probability of the parameters (the prior).[4] All of the priors of this model were either non-informative or weakly-informative, meaning that the values of parameter estimates are driven most strongly by the data. To ensure that the sampler settled on a stable distribution, we fit 4 chains with 4000 iterations. The first 2000 iterations were discarded as a burn-in, and the Rubin-Gelman diagnostic, ($\hat{R}$) was used to determine convergence. For all of the parameters reported here, the $\hat{R}$ was sufficiently close to 1 to consider them converged.

4.3.1 Results

To begin with, Figure 12 plots the estimated scale parameters (σs) from the model as a sanity check. These values can be compared against the maximum likelihood estimates from Figure 5, and the random effects standard deviations from Table 3. The σ labelled "within speaker variation" is estimated to be about 0.51. This is slightly smaller than the residual deviance from the mixed effects model described in Table 3, but very similar to the average within-speaker standard deviation from Figure 5. The estimated between-speakers standard deviation is 0.21, which is very similar to both the rolling estimate of the inter-speaker standard deviation from Figure 5 and the standard deviation of the by-speaker random effects from Table 3. Finally, the between words standard deviation is estimated to be 0.13, which is actually a bit smaller than the standard deviation of the by-word random intercepts from Table 3, but this should perhaps not be too surprising since there are fewer word types and fewer phonological contexts represented in this model than in the full model. These scale parameters appear to be generally reasonable when compared to other similar estimates from the data, giving us some confidence that the model, as described in Appendix A, was reasonably specified.

The first model parameter which varies across time is $\delta_{[j,k]}$, which can be understood as the year-over-year differences in F1, or the rate of change. Figure 13 plots the model estimates along with a ribbon indicating the 95% Bayesian credible interval. The interpretation of these credible intervals is different from frequentist confidence intervals. They indicate that there is a 95% probability that the value of $\delta_{[j,k]}$ lies within the interval. Rather than a tabular representation of coefficient estimates and p-values, these graphical intervals should be understood as indicating the reliability of the effect.

[4] See Kruschke (2011) for an accessible introduction to MCMC and Bayesian Modelling.

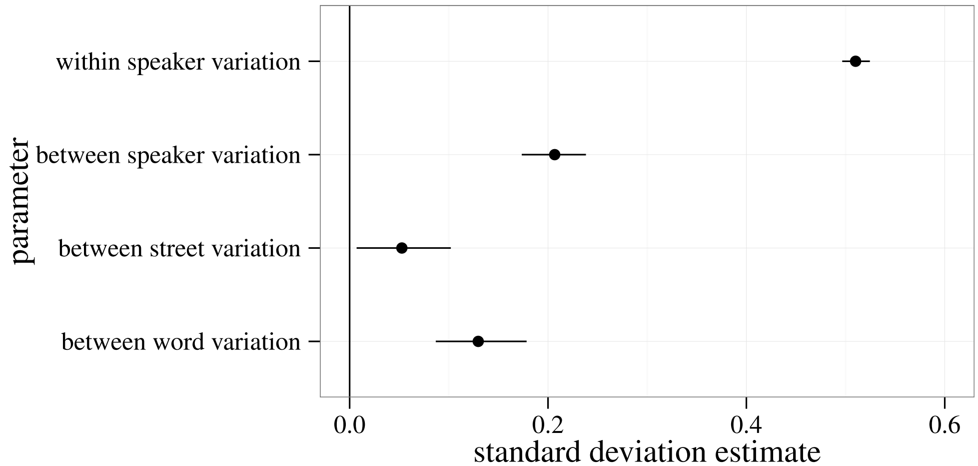


Figure 12: Estimated standard deviations from the Stan model.

For about the first two-thirds of the time course of the change, the estimated year-over-year differences for /ay/ before /t/ (both faithful and flapped) hovers around -0.01, although it doesn't appear to be reliably different from 0 until 1920 for /ay/ preceding /t/-flaps. This is relatively identical to the estimated slope from the full model described in Table 1, which was -0.108 per decade, or -0.0108 per year. So far, it does not look as if there was some time period where pre-flap /ay/s were either both participating, or both not participating in /ay/ raising. Rather, /ay/ preceding /d/-flaps appears to not be undergoing any change, and /ay/ before /t/-flaps appears to be undergoing the same change as /ay/ before faithful /t/. The fact the credible interval for pre-/t/-flap /ay/ doesn't exclude 0 for about 20 years after it first does for pre-faithful-/t/ is almost certainly because there is an order of magnitude more data for /ay/ before faithful /t/.

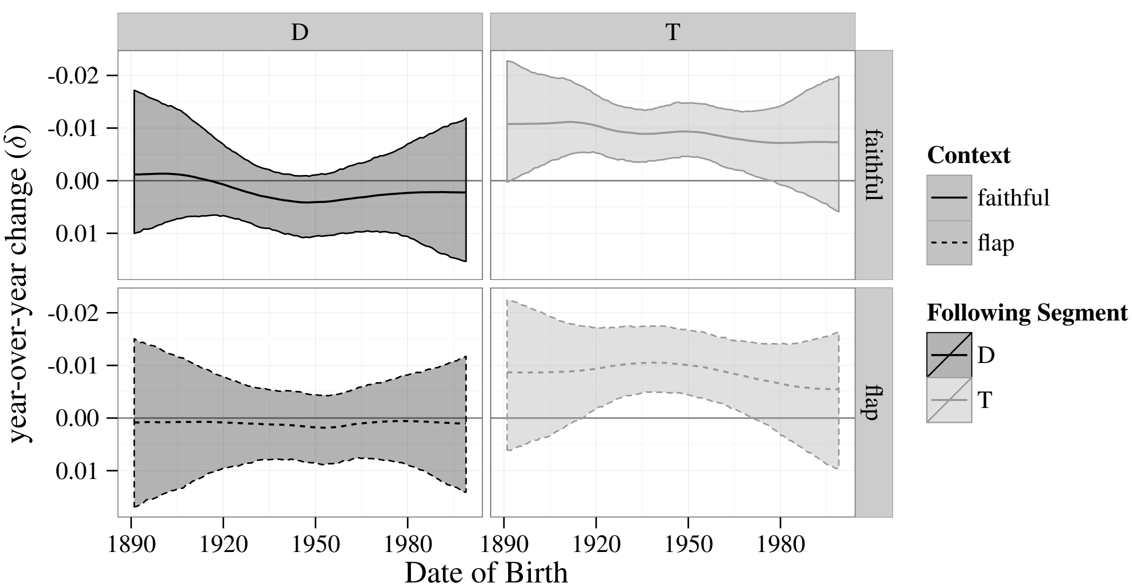


Figure 13: Model estimates and 95% credible intervals for the year-over-year differences (δ) i.e. the rate of change.

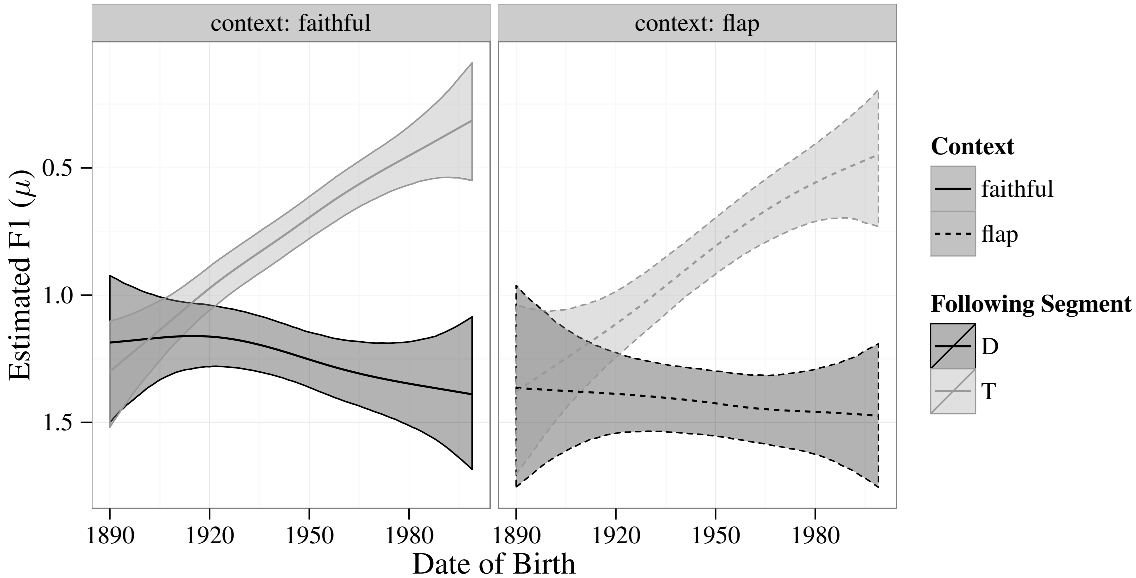


Figure 14: Model estimates and 95% credible intervals for normalized /ay/ F1.

Figure 14 plots the actual expected F1 and 95% credible intervals for /ay/ across the four contexts. Modulo the wider credible intervals in the flapping facet, which again are almost certainly due to the sparser data for pre-flap /ay/, the profile of the change is largely identical between the two contexts: /ay/ raises before underlying /t/, and does not before underlying /d/, and flapping does not seem to perturb that change.

One thing not immediately clear from Figure 14 is that there appears to be a weak main effect where /ay/ before flaps is slightly lower than before faithful /t,d/. This effect is clearer in Figure 15, which plots the same estimates from Figure 14, but this time emphasizes the difference between faithful and flapping contexts. It's not immediately clear why /ay/ preceding flaps should be slightly lower than preceding faithful /t,d/, but two things should be noted. First, it appears to affect /ay/ before /d/-flaps and /t/-flaps to a similar degree. Second, neither of the phonetic precursors considered above would predict an effect like this. If anything, the duration precursor would predict that /ay/ should be higher before flaps than before faithful realizations, and the peripheralization precursor would predict that /ay/ before /d/-flaps would be higher and /ay/ before /t/-flaps would be lower than before faithful realizations.

Something else that is not immediately clear from a visual inspection of Figure 14 is whether the way /ay/ differentiates itself between pre-/t/ and pre-/d/ contexts is the same between faithful and flapping realizations. It could be the case that /ay/ before /t/-flaps does differentiate from /ay/ before /d/-flaps, but at a slower rate, or at a later time than it does before faithful-/t/ and faithful-/d/. The height difference between pre-/t/ and pre-/d/ contexts was not an actual parameter of the model itself, but it is possible to generate estimated differences and 95% credible intervals for those differences from the model parameters, and Figure 16 plots these estimates. It appears that the trend over time for /ay/ differentiation is nearly identical between flapping and faithful context. In fact, the 95% credible interval for the height difference begins to exclude 0 at the same time for both contexts (approximately 1915).

It is worth re-iterating at this point that the curves for each of the four contexts investigated here were estimated separately in the model. That is, the model did not assume that there should be any similarity in the diachronic curves

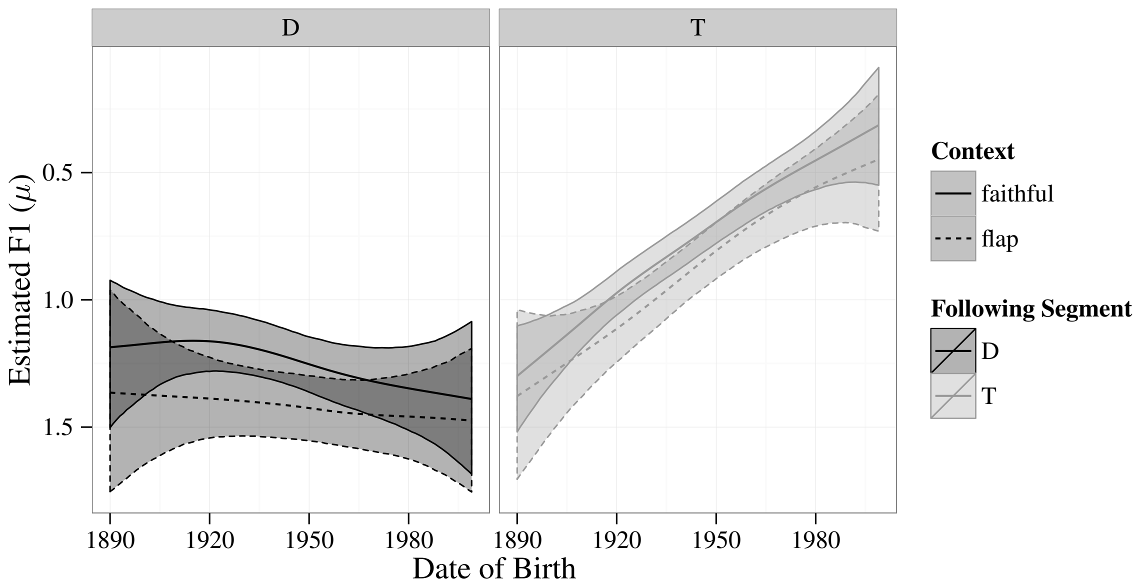


Figure 15: Model estimates and 95% credible intervals for normalized /ay/ F1.

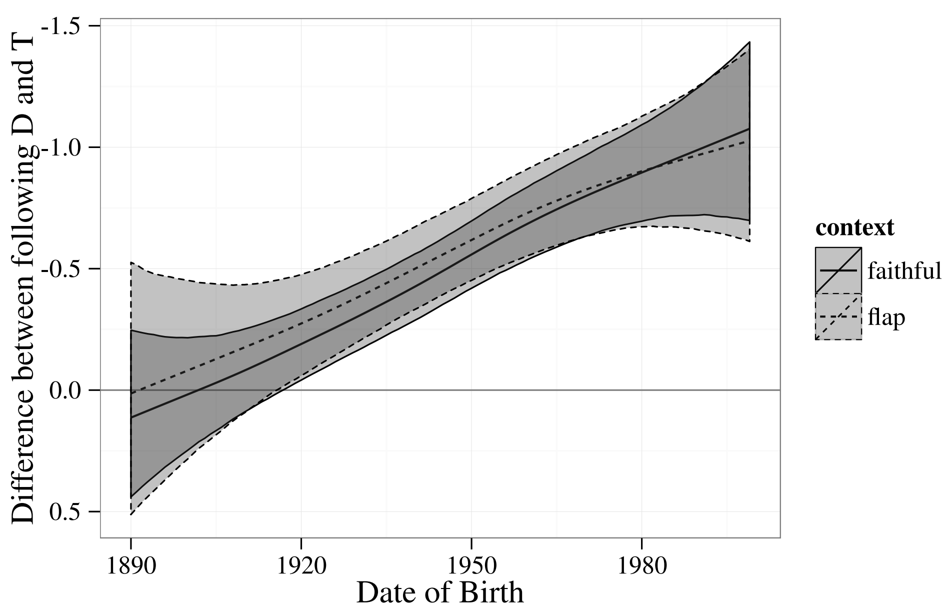


Figure 16: Estimated /ay/ height difference between pre-/t/ and pre-/d/ contexts.

we've plotted here. The fact that /ay/ height diverges identically in faithful and flapping contexts is a property of the data, not modelling assumptions.

4.3.2 Non-linear Modelling Summary

If there was one goal of this non-linear Bayesian modelling strategy, it was to produce Figure 16. Its takeaway point is twofold. First, as soon as there was a detectable difference in height for /ay/ in pre-faithful-/t/ position and pre-faithful-/d/ position, there was also a detectable difference of the same magnitude in pre-/t/-flap and pre-/d/-flap contexts. Secondly, the overall way in which /ay/ differentiated in height across the 20th century before faithful-/t/ and faithful-/d/ appears to be the same before /t/-flaps and /d/-flaps.

4.4 Analogy

The primary thesis of this paper is that pre-voiceless /ay/ raising has always been conditioned by the phonological properties of its context, not the phonetic properties. However, there is one confound to the results from §4.3 that may call into question whether the effect demonstrated there was truly phonological. Almost all of words in which /ay/ appears before /t/-flaps are morphologically complex (e.g. *fighting*, *united*, *writing*). For very few roots (38 tokens in total) does /ay/ appear exclusively before /t/-flaps, and they are forms of *title*, *vitamin*, and the suffix *-itis*. It could be that /ay/ raising began in the phonetically predicted context (before faithful /t/), and then the vowel quality analogized to derived and inflected forms of the root without ever making reference to other phonological properties of the root. Such a process doesn't seem very likely in view of the results from §4.3, especially looking at Figure 16. The analogy would have had to be nearly instantaneous for all roots involved. Moreover, while this analogy would be very different from phonology as it is traditionally understood, it is still quite a few steps removed from the continuous properties of speech upon which the hypocorrection model is based.

However, in an effort to appropriately address this open question, we coded every word root in the flapping data for how frequently it occurs in flapping and faithful contexts according to the word frequency norms from $\text{SUBTLEX}_{\text{US}}$. Table 10 illustrates what this looked like for the root *unite*. For each individual word which contained the root *unite*, its frequency per 1,000 words was collected, and it was coded for whether or not the /t/ would be flapped. Then, the frequencies for each context were summed, and a ratio of flapping to faithful realizations was calculated (see Table 11). For the root *unite*, it appeared in flapping contexts about 12 times more often than in faithful contexts.

root	word	freq	context
unite	unite	3.02	faithful
unite	reunite	0.71	faithful
unite	unites	0.53	faithful
unite	reunites	0.04	faithful
unite	united	50.27	flap
unite	reunited	1.78	flap
unite	uniting	0.29	flap
unite	reuniting	0.27	flap

Table 10: $\text{SUBTLEX}_{\text{US}}$ frequency norms for the root *unite*.

The hypothesis being pursued here is that the more frequently a root appears with /ay/ in a faithful context, where the

root	faithful	flap	flap:faithful
unite	4.3	52.61	12.23

Table 11: Summed frequency norms for the root *unite*, and the ratio of flapping frequency to faithful frequency

raising change is phonetically natural, compared to a flapping context, the more likely the vowel quality is to analogize to other realizations of the root. Of course this is a relatively simplistic approach to quantifying the likelihood of analogy, but if analogy is playing a powerful enough role in producing the appearance of a phonologically conditioned phonetic change, then even an imperfect measure like this should indicate some hint of an effect.

As a first pass at the question, Figure 17 plots F1 means for /ay/ in flapping contexts. The data was split into two categories: roots which occur more often in flapping contexts, and roots which occur more often in faithful contexts. The impression from Figure 17 is that if there is an effect of the flapping to faithful ratio, it is a weak one.

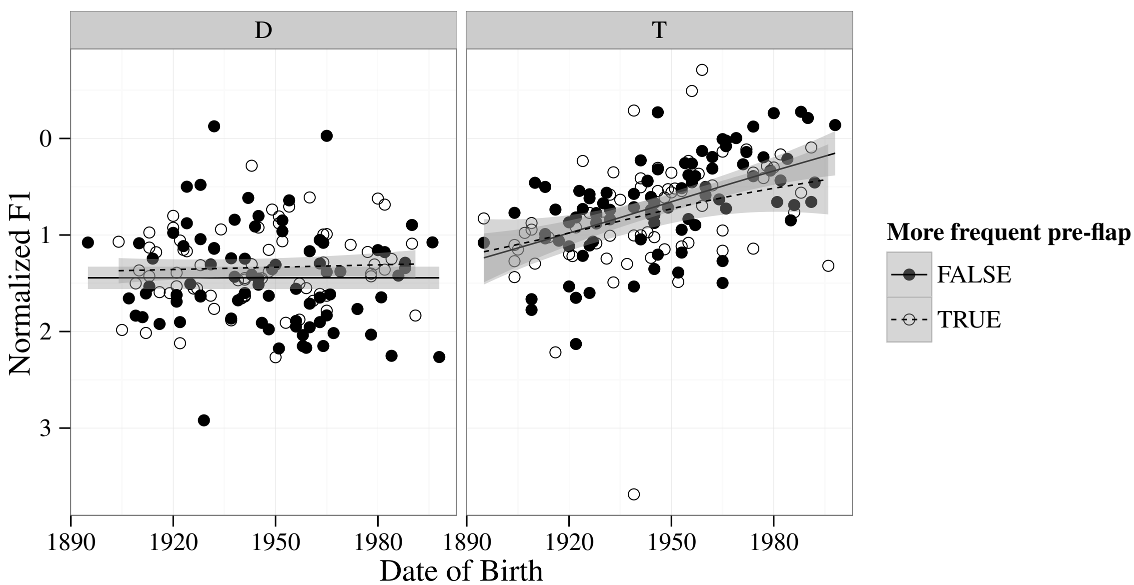


Figure 17: Estimated F1 of /ay/ in flapping contexts, divided by roots which appear more often in flapping contexts, and those which appear more often in faithful contexts.

The ratio of flapping-to-faithful frequency is a continuous factor, so it was entered into a linear mixed-effects model using a $\log_2$ transform.[5] Table 12 displays the fixed-effects estimates from the model, along with 95% confidence intervals based on 10,000 semiparametric bootstrap replicates obtained using `bootMer`. The only parameters where the confidence interval excludes 0 are the Intercept, the main effect of a following /t/, and the interaction of /t/ with date of birth. The confidence interval for three way interaction of Decade$\times$Ratio$\times$[-voice] doesn't exclude 0, but just barely. Assuming, briefly, that the effect is actually non-zero, we can visualize the size of the effect by calculating the fitted values from each of the bootstrap replicates, and plotting from the 2.5th percentile to the 97.5th percentile,

[5]Using the $\log_2$ transform means that a flap:faithful ratio of 2:1 has a value of 1, a 1:1 ratio has a value of 0, and a 1:2 ratio has a value of -1.

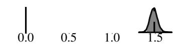
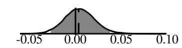
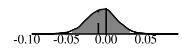
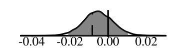
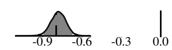
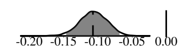
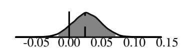
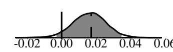
Parameter	Estimate	95% CI	Bootstrap
Intercept*	1.491	[1.404, 1.575]	
Decade	0.006	[−0.027, 0.038]	
$\log_2$ Flap Ratio	−0.009	[−0.043, 0.03]	
Decade × Ratio	−0.008	[−0.021, 0.011]	
[-voice]*	−0.813	[−0.893, −0.672]	
Decade × [-voice]*	−0.11	[−0.146, −0.068]	
Ratio × [-voice]	0.02	[−0.017, 0.071]	
Decade × Ratio × [-voice]	0.016	[−0.003, 0.035]	

Table 12: Parameter estimates and derived values from the mixed effects model `NormalizedF1 ~ Decade * following_voicing * log2Ratio + (Decade|Word) + (following_voicing|Speaker)`. `Decade` is the speakers' date of birth, centered at 1950, and divided by 10. `log2ratio` is log2(flap/faithful), based on the frequencies described in §4.4. The 95% confidence intervals are based on 10,000 semiparametric bootstrap replicates fitted by `bootMer` from `lme4 v1.1-7`. The density distributions of the bootstrapped parameters are provided in the final column.

illustrating a 95% range of the fitted values. The result is Figure 18, which is admittedly not a victory of data visualization, but it is very difficult to successfully visualize an effect as small as the this one may be.

A model comparison route to evaluating the effect of the flapping:faithful ratio suggests that there is not even a main effect. Table 13 displays the results of likelihood ratio tests, as well as the AIC and BIC for models which differ only in the way the ratio was added to the model. None of the likelihood ratio tests are significant, and the the model with the smallest AIC and BIC does not include the ratio as a predictor.

	Df	AIC	BIC	logLik	deviance	Chisq	Chi Df	Pr(>Chisq)
no ratio	11	796.4	843.6	-387.2	774.4			
+ratio	12	798.4	849.9	-387.2	774.4	0.017	1	0.895
voicing×ratio	13	800.4	856.1	-387.2	774.4	0.018	1	0.893
decade×voicing×ratio	15	801.9	866.3	-385.9	771.9	2.467	2	0.291

Table 13: Comparisons of models that differ in terms of how the ratio of flaps:faithful was included.

While the ratio of the frequency with which a root appeared in flapping contexts to its frequency in faithful contexts may not be the most sophisticated operationalization for analogy, any more sophisticated approach will face the challenge of successfully analogizing the vowel quality for /ay/ raising while simultaneously failing to analogize the vowel quality for /ey/ raising, which is another phonetic change that occurred in Philadelphia in the 20th century. Following consonants conditioned /ey/ raising, but this conditioning applied completely transparently. Figure 19a is

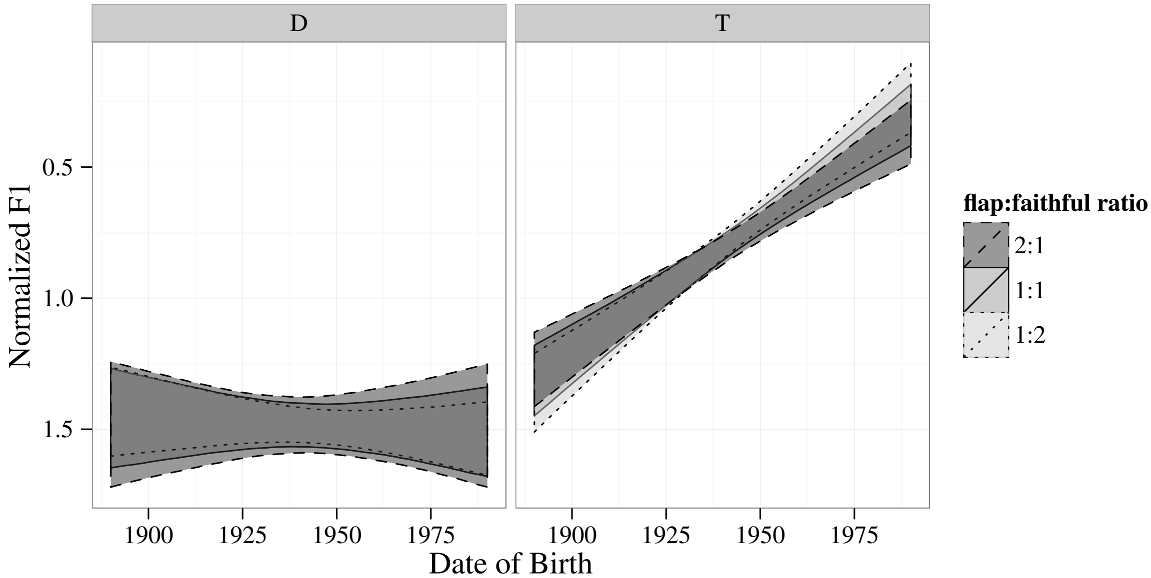
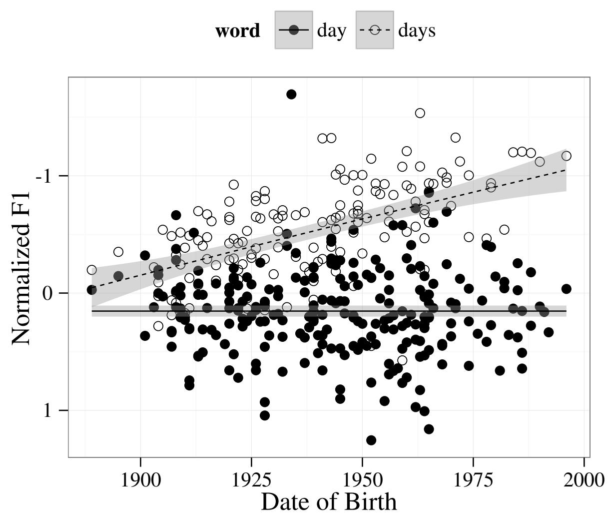


Figure 18: 95% percentiles of the fitted values from 10,000 bootstrap replicates.

an illustrative example, plotting the mean F1 for just the lexical items *day* and *days*. *Days* undergoes a phonetic change which is just slightly smaller in magnitude than pre-voiceless /ay/ raising, and no analogical force appears to drag *day* along for the ride. For the sake of comparing apples to apples, Figure 19b plots a similar illustrative plot for the lexical items *fight* and *fighting*. Both of these lexical items appear to undergo their raising change in lockstep.

(a)



(b)

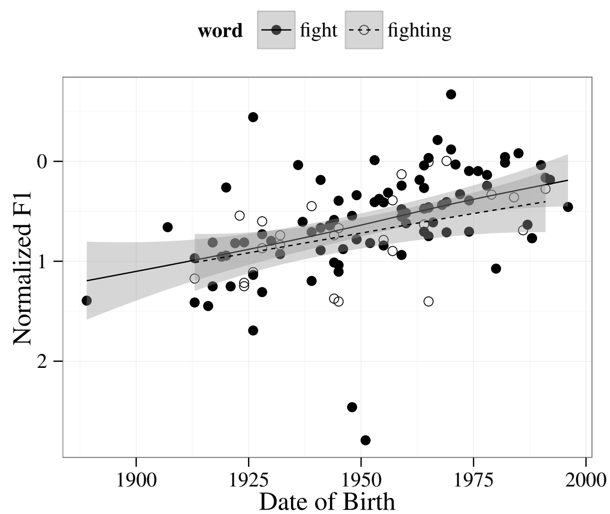


Figure 19: Mean F1 over date of birth for different lexical items.

5 Discussion

Most of the analysis in this paper has, so far, been devoted to being certain about what didn't happen to /ay/. Pre-voiceless raising didn't first begin in the contexts with the strongest phonetic precursors, then subsequently generalize or analogize along phonological or lexical dimensions. In different terms, the evidence suggests that the set of environments where this change was initiated were defined on phonological grounds, rather than in terms of the phonetic properties of those environments.

The appropriate next steps forward when faced with this result depends greatly on one's theoretical commitments. If our commitment to the phonetic precursors model of phonetic change extended beyond apparent counter-evidence, then the next appropriate step would be to search for new phonetic precursors that could accurately predict how this change was circumscribed. This route could prove to be long and fruitful, as the set of possible precursors is large and possibly non-finite. However, the same argument could be levied against the set of possible phonological explanations for phonetic changes. This underlines a central problem in comparing phonetic and phonological theories for how this change occurred, as most explanations grounded in either categorical phonology or continuous phonetics will be post-hoc and highly flexible. A deductive approach to settling the question is therefore simply not open to us on the basis of current theory since the premises are not fixed. However, just because there may be infinite explanations on phonetic and phonological grounds does not mean that all explanations are equally probable. If we take pre-voiceless shortening and offglide peripheralization to be the most likely phonetic precursors simply because they have been proposed in the literature, then the results presented here are highly unlikely on the basis of the most likely phonetic predictions. On the other hand, these results are exactly what would be expected if /ay/ raising has always been conditioned by the underlying phonological status of the following segment.

So what did happen to /ay/ in Philadelphia? These results support part of a larger argument we would like to make that in order for two contextual variants of a speech sound to diverge in their phonetics over time, they must, all else being equal, be treated as being qualitatively different categories by speakers from the moment they begin to diverge. That is, a categorical split of /ay/ into two new allophones or phonemes is not the reanalysis of a longer term phonetic change. Rather, the longer term phonetic change is only possible because /ay/ split into two new allophones or phonemes either previous to or concurrent with the onset of the phonetic change. The split allowed for their phonetic targets to be learned separately, and to change independently.

We propose that very early in the change, there were two categorically distinct, but phonetically similar variants of /ay/, distributed according to the voicing specification of the following segment, and that one of them underwent a phonetically gradual change in height, and the other remained low. This is a more extreme version of the Big Bang theory of sound change put forward by Janda and Joseph (2003). They propose that purely phonetic factors guide sound changes very briefly, and are eventually overridden by phonological conditioning. There is, in fact, no detectable period where the pattern of /ay/ raising aligned with what would be predicted on purely phonetic grounds. The conclusion we draw is that either the period of purely phonetic conditioning was too brief to be identified, or was non-existent. If the situation is the former, then that means there is a greater challenge than perhaps has been appreciated in identifying phonologization *in vivo*. If the situation is the latter, then the question arises as to how these two categorically different variants came to exist.

Under our proposal, there must have been a categorical difference between pre-voiced and pre-voiceless /ay/ at the onset of the change, but the nature of that difference is, unfortunately, not specified by our model. We can't differentiate between proposals that place the distinction in the underlying representation (Mielke et al., 2003) and those that generate it in the phonological grammar (Idsardi, 2006; Pater, 2014). However, Bermúdez-Otero (2013) makes a compelling argument that a categorical distinction between pre-voiced and pre-voiceless /ay/ is not necessarily a new development. Rather, he argues that a long-standing categorical process of pre-fortis clipping is responsible for producing two allophones of /ay/. As was already demonstrated, the selection of contexts to undergo /ay/ raising is not proportional to phonetic duration, but pre-fortis clipping is construed here as a categorical, phonological process, even if its primary phonetic consequence is a shorter vowel duration. It was the clipped allophone which underwent the /ay/ raising change under this proposal. Whether that necessitates a reorganization in the phonological grammar of Philadelphians depends on how much or how little one wants to make their phonology dictate phonetics. For example, there may have only been one phonological process in the grammar across the entire 20th century, which could be given as (6).

(6) CLIPPING ay → ăy/___−voice

The substantive change observed in this paper would thus be a shift in the phonetic realization of [ăy], which used to just be realized with a shorter duration, but then began to also exhibit a change in its height. Alternatively, a context-free process could have been introduced to the grammar which altered the phonological specification of height for [ăy], which could be given as (7).

(7) CLIPPING ay → ăy/___−voice
RAISING ăy → -low

This change in the phonological specification of [ăy] resulted in the observed gradual phonetic shift. An additional possibility is that these two phonological grammars, in addition to any others which might result in similar phonetic outcomes, were all covertly being used in a mixture, such as Mielke et al. (forthcoming) have found for the distribution of bunched and retroflex /r/ (which are largely acoustically indistinguishable) in American English. Additional diagnostics to differentiate between these possibilities, like other phonological processes which interact with vowel height, are not forthcoming, so at the moment this discussion will have to be set aside.

Regardless of how different phonological categories are being represented or generated, our conclusion that they are not products of the phonetic change, but rather necessary ingredients for the change to happen, reopens most of the questions that hypocorrection was supposed to resolve. Gradual assimilation of the vowel's nucleus to a peripheralized glide, for example, would explain why the change happened at all. Without this explanation, we are left with the mystery of why [āy] and [ăy] didn't just both maintain their low nuclei forever. Perhaps a maximal dispersion theory could be turned to salvage the situation. For example, Boersma and Hamann (2008) explicitly model maximal acoustic dispersion as being a product of cross generational change. However, we are again left with the same, fundamentally difficult Actuation Problem: Why now? Why never before? Why here? Why not everywhere? Since a definitive answer to the Actuation Problem has not been provided in the nearly 50 years since it was first given a name, we can hopefully be forgiven for not settling the issue here.

As for the larger argument we want to make regarding the early influence of phonology on phonetic changes, a definitive case for it will require more examples of phonetic changes from more dialects investigated in similar depth as we have done here. We have focused exclusively on /ay/ raising in order to provide it the thorough treatment necessary to establish with relative certainty that for this change, phonological conditioning was present at its outset. Obviously we should avoid generalizing too broadly from one specific case, but we hope to have at least laid the groundwork for establishing a direction of inquiry, and sufficiently problematized the widely held conventional wisdom regarding changes of this sort.

6 Conclusion

The hypocorrection model of conditioned phonetic change relies crucially on the presence of phonetic precursors that drive the change. In the case of pre-voiceless /ay/ raising, we established how the two phonetic precursors which have been proposed for it (pre-voiceless shortening, and offglide peripheralization) predict /ay/ raising ought to interact with /t, d/ flapping. We found that in 20th century Philadelphia, neither set of predictions are borne out. Rather, it appears that /ay/ raising has always been conditioned by the phonological voicing of the following segment, not the phonetic properties of the context. We argued that this calls into question the hypocorrective model of phonetic change, and that the possibility of early involvement of phonology ought to be explored in more sound changes.

Appendix A Non-linear Model Definition

We'll call the difference in F1 for any given year from the year before $\delta_{[j]}$, where j is an index for the date of birth, assigned the value 1 for the first observable date of birth. We'll be estimating this $\delta_{[j]}$ value for all four contexts (/t, d/ $\times$ [faithful, flapped]), so we'll assign each context an index from 1 to 4 called k, and call the year-over-year change for a given year for a given context $\delta_{[j,k]}$.

$$\delta_{[j,k]} \begin{cases} = 0 & \text{if } j = 1 \\ \sim \mathcal{N}(0, 100) & \text{if } j = 2 \\ \sim \mathcal{N}(\delta_{[j-1,k]}, \sigma_\delta) & \text{if } j > 2 \end{cases} \tag{8}$$

For every $\delta_{[j,k]}$ for $j > 2$, these are treated as being drawn from a normal distribution centered around the previous year's $\delta_{[j,k]}$. σ_δ is sampled from a uniform prior between 0 and 100, as are all other free scale parameters in the model. The first observable date of birth, $\delta_{[1,k]}$, has been fixed to be 0. This was done to allow for the identifiability of an intercept and contextual effects. Finally, $\delta_{[2,k]}$ has been treated specially, being drawn from a normal distribution with a standard deviation of 100. Since $\delta_{[2,k]}$ is the first non-zero $\delta_{[j,k]}$ value to be estimated, it might be substantially different from 0, but the differences between every subsequent $\delta_{[j,k]}$ might be much smaller. If the same prior distribution for $\delta_{[2,k]}$ was used for all of the other $\delta_{[j,k]}$, this would either have a detrimental effect on estimating the smoothing parameter σ_δ, or $\delta_{[2,k]}$ would be estimated to be closer to 0 than it actually is. For this reason, $\delta_{[2,k]}$ was given a less potentially restrictive prior than the rest of $\delta_{[j,k]}$.

We'll call the expected normalized F1 for a particular date of birth $\mu_{[j]}$, and since we're estimating $\mu_{[j]}$ for four contexts, we'll also index $\mu_{[j]}$ by context, using k again.

$$\mu_{[j,k]} = \beta_0 + \beta_{\text{context}[k]} + \sum_{1}^{j} \delta_{[j,k]} \tag{9}$$

$$\beta_0 \sim \mathcal{N}(0, 100) \tag{10}$$

$$\beta_{\text{context}[k]} \begin{cases} = 0 & \text{if } k = 1 \\ \sim \mathcal{N}(0, 100) & \text{if } k > 1 \end{cases} \tag{11}$$

For each speaker (s), a random intercept will be estimated for each context (k), drawn from a normal distribution with a standard deviation σ_s. The same goes for each street (r). Since context is a between-words factor in this model, simply a random intercept effect for each word is estimated.

$$\text{speaker.random}_{[s,k]} \sim \mathcal{N}(0, \sigma_s) \tag{12}$$

$$\text{street.random}_{[r,k]} \sim \mathcal{N}(0, \sigma_r) \tag{13}$$

$$\text{word.random}_{[l]} \sim \mathcal{N}(0, \sigma_l) \tag{14}$$

Finally, the observed normalized F1 of each token is understood to be the sum of the expected value for the speaker's date of birth, plus the random effects of speaker, street and word, an effect of duration of the token, plus some random error. Both duration and random error, ϵ are drawn from normal distributions. σ represents the residual deviation.

$$\begin{aligned} y_{[i]} =& \mu_{[\text{speaker.dob}[i],\text{context}[i]]} + \\ & (\beta_{\text{duration}} \times \text{duration}[i]) + \\ & \text{speaker.random}_{[\text{speaker}[i],\text{context}[i]]} + \\ & \text{street.random}_{[\text{street}[i],\text{context}[i]]} + \\ & \text{word.random}_{[\text{word}[i]]} + \epsilon_{[i]} \end{aligned} \tag{15}$$

$$\beta_{\text{duration}} \sim \mathcal{N}(0, 100) \tag{16}$$

$$\epsilon \sim \mathcal{N}(0, \sigma) \tag{17}$$

The duration data passed to the model is log-transformed and centered at the median. This should have the effect of largely factoring out effects strictly due to phonetic duration, and means that the curves represent the estimates for /ay/s of median duration.

References

Adank, Patti, Roel Smits, and Roeland van Hout. 2004. A comparison of vowel normalization procedures for language variation research. *The Journal of the Acoustical Society of America* 116:3099. URL `http://link.aip.org/link/JASMAN/v116/i5/p3099/s1&Agg=doi`.

Baker, Adam, Diana Archangeli, and Jeff Mielke. 2011. Variability in American English s-retraction suggests a solution to the actuation problem. *Language Variation and Change* 23:347–374. URL `http://www.journals.cambridge.org/abstract_S0954394511000135`.

Bates, Douglas, M Maechler, B Bolker, and S Walker. 2014. lme4: Linear mixed-effects models using Eigen and S4 using Eigen and S4. URL `http://cran.r-project.org/package=lme4`.

Bermúdez-Otero, R. 2007. Diachronic Phonology. In *The Cambridge Handbook of Phonology*, ed. Paul de Lacy, chapter 21, 497–517. Cambridge: Cambridge University Press.

Bermúdez-Otero, Ricardo. 2013. Philadelphia /aɪ/-raising without rule insertion. Paper Presented at the Symposium on Historical Phonology.

Blevins, Juliette. 2004. *Evolutionary phonology: the emergence of sound patterns*. New York: Cambridge University Press.

Boersma, Paul, and Silke Hamann. 2008. The evolution of auditory dispersion in bidirectional constraint grammars. *Phonology* 25:217–270.

Braver, Aaron. in press. Imperceptible Incomplete Neutralization : Production , Non-Identifiability, and Non-Discriminability in American English Flapping. *Lingua* .

Brysbaert, Marc, and Boris New. 2009. Moving beyond Kucera and Francis: a critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior research methods* 41:977–90. URL `http://www.ncbi.nlm.nih.gov/pubmed/19897807`.

Chambers, J K. 1973. Canadian Raising. *Canadian Journal of Linguistics* 18:131–135.

Conn, Jeffery. 2005. Of 'moice' and men: The evolution of a male-led sound change. Ph.D., University of Pennsylvania.

Dailey-O'Cain, Jennifer. 1997. Canadian raising in a midwestern U.S. city. *Language Variation and Change* 9:107–120.

Darlington, Richard B. 1970. Is Kurtosis Really "Peakedness?". *The American Statistician* 24:19–22.

Davidson, Lisa. 2006. Comparing tongue shapes from ultrasound imaging using smoothing spline analysis of variance. *The Journal of the Acoustical Society of America* 120:407. URL `http://scitation.aip.org/content/asa/journal/jasa/120/1/10.1121/1.2205133`.

Evanini, Keelan. 2009. The permeability of dialect boundaries: A case study of the region surrounding Erie, Pennsylvania. Ph.D., University of Pennsylvania.

Garrett, Andrew, and Keith Johnson. 2013. Phonetic bias in sound change. In *Origins of sound change: Approaches to phonologization.*, ed. Alan C. L. Yu, September 2014. Oxford University Press.

Ghitza, Yair, and Andrew Gelman. 2014. The Great Society, Reagan's Revolution, and Generations of Presidential Voting. ms.

Hale, Mark, and Charles Reiss. 2008. *The Phonological Enterprise*. New York: Oxford University Press.

Harrington, John, S Palethorpe, and C I Watson. 2000. Does the Queen speak the Queen's English? *Nature* 408:927–8. URL `http://www.ncbi.nlm.nih.gov/pubmed/11140668`.

Harrington, Jonathan, Felicitas Kleber, and Ulrich Reubold. 2007. /u/-fronting in RP: A link between sound change and diminished perceptual compensation for coarticulation. *ICPhS XVI* 1473–1476.

Hoffman, Matthew D, and Andrew Gelman. 2011. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *arXiv.org* URL `http://arxiv.org/abs/1111.4246`.

Idsardi, William. 2006. Canadian Raising, Opacity, and Rephonemicization. *The Canadian Journal of Linguistics / La revue canadienne de linguistique* 51:119–126.

Janda, Richard D, and Brian D Joseph. 2003. Reconsidering the Canons of Sound-Change: Towards a "Big Bang" Theory. In *Historical Linguistics 2001. Selected Papers from the 15th International Conference on Historical Linguistics, Melbourne*, ed. Barry Blake and Kate Burridge, 205–219. John Benjamins.

Joos, Martin. 1942. A phonological dilemma in Canadian English. *Language* 18:141–144.

Kruschke, John K. 2011. *Doing Bayesian Data Analysis: A Tutorial with R and BUGS*. Oxford: Elsevier.

Kwong, Katherine, and Kenneth Stevens. 1999. On the voiced-voiceless distinction for writer/rider. *Speech Communication Group Working Papers (Research Laboratory of Electronics at MIT)* 11:1–20.

Labov, William. 2001. *Principles of linguistic change. Volume 2: Social Factors. Language in Society*. Oxford: Blackwell.

Labov, William. 2007. Transmission and Diffusion. *Langauge* 83:344–387.

Labov, William, Sherry Ash, and Charles Boberg. 2006. *The Atlas of North American English: Phonetics, Phonology and Sound Change*. New York: Mouton de Gruyter.

Labov, William, and Ingrid Rosenfelder. 2011. The Philadelphia Neighborhood Corpus.

Labov, William, Ingrid Rosenfelder, and Josef Fruehwald. 2013. One hundred years of sound change in Philadelphia: Linear Incrementaion, Reversal, and Reanalysis. *Language* 89:30–65.

Mielke, Jeff, Mike Armstrong, and Elizabeth Hume. 2003. Looking through opacity. *Theoretical Linguistics* 29:123–139.

Mielke, Jeff, Adam Baker, and Diana Archangeli. forthcoming. A typology of American English /*r/ allophony patterns.

Moreton, Eliott, and Erik R Thomas. 2007. Origins of Canadian Raising in voiceless-coda effects: a case study in phonologization. In *Papers in laboratory phonology 9*, ed. J Cole and J I Hualde, 37–64. Cambridge University Press.

Moreton, Elliott. 2004. Realization of the English postvocalic [voice] contrast in F1 and F2. *Journal of Phonetics* 32:1–33. URL http://linkinghub.elsevier.com/retrieve/pii/S0095447003000044.

Ohala, John J. 1981. The listener as a source of language change. In *Papers from the parasession on language and behavior*, ed. C S Masek, R A Hendrick, and M F Miller. Chicago Linguistics Society.

Ohala, John J. 1990. The phonetics and phonology of aspects of assimilation. In *Papers in laboratory phonology 1: Between the grammar and physics of speech*, ed. John Kingston and Mary E Beckman, 258–282. Cambridge: Cambridge University Press.

Ohala, John J. 2012. The listerner as a source of sound change: An update. In *The initiation of sound change: Perception, production, and social factors*, ed. Maria-Josep Solé and Daniel Recasens, 21–35. Philadelphia: John Benjamins.

Pater, Joe. 2014. Canadian raising with language-specific weighted constraints. *Language* 90:230–240. URL http://muse.jhu.edu/content/crossref/journals/language/v090/90.1.pater.html.

Phillips, Betty S. 1984. Word Frequency and the Actuation of Sound Change Betty S. Phillips. *Language* 60:320–342.

Pierrehumbert, Janet B. 2001. Exemplar dynamics: Word frequency, lenition and contrast. In *Frequency and the emergence of linguistic structure*, ed. Joan Bybee and Paul Hopper, volume 45 of *Typological Studies in Language*, 137–157. John Benjamins.

Pitt, M A, L Dilley, Keith Johnson, S Kiesling, W Raymond, Elizabeth Hume, and E Fosler-Lussier. 2007. Buckeye Corpus of Conversational Speech (2nd release). URL www.buckeyecorpus.osu.edu.

Rathcke, Tamara, and Jane Stuart-Smith. 2014. On the impact of noise on vowel formant measures. Paper presented at Methods in Dialectology XV.

Rosenfelder, Ingrid. 2005. Canadian Raising in Victoria , B. C.: an Acoustic Analysis. MA, Ruprecht Karl University of Heidelberg.

Rosenfelder, Ingrid, Josef Fruehwald, Keelan Evanini, Scott Seyfarth, Kyle Gorman, Hilary Prichard, and Jiahong Yuan. 2014. FAVE (Forced Alignment and Vowel Extraction) 1.1.3. URL http://dx.doi.org/10.5281/zenodo.9846.

Sankoff, Gillian. 2006. Gillian Sankoff. Age: Apparent time and real time. In *Encyclopedia of language and linguistics*. Oxford: Elsevier.

Sankoff, Gillian, and Hélène Blondeau. 2007. Language Change across the Lifespan: /r/ in Montral French. *Language* 83:560–588.

Smith, Jennifer, Mercedes Durham, and Liane Fortune. 2007. "Mam, my trousers is fa'in doon!": Community, caregiver, and child in the acquisition of variation in a Scottish dialect. *Language Variation and Change* 19:63–99.

Stan Development Team. 2014. Stan: A C++ Library for Probability and Sampling. URL http://mc-stan.org/.

Tucker, R Whitney. 1944. Notes on the Philadelphia dialect. *American Speech* 19:37–42.

Vance, Timothy. 1987. 'Canadian Raising' in some dialects of the northern United States. *American Speech* 62:195–210.

Wagner, Suzanne E. 2007. We act like girls and we don't act like men: The use of the male-associated variable (ay0) in South Philadelphia. In *Penn Working Papers in Linguistics*, ed. Tatjana Scheffler, Joshua Tauberer, Aviad Eilam, and Laia Mayol, volume 13.1, 393–406. Philadelphia: University of Pennsylvania Press.

Wedel, AB. 2007. Feedback and Regularity in the Lexicon. *Phonology* 24:147–185.

Weinreich, Uriel, William Labov, and Marvin Herzog. 1968. Empirical foundations for a theory of language change. In *Directions for historical linguistics*, ed. W Lehmann and Y Malkiel. U. of Texas Press.

Wieling, Martijn, John Nerbonne, and R Harald Baayen. 2011. Quantitative social dialectology: explaining linguistic variation geographically and socially. *PloS one* 6:e23613.

Wood, Simon. 2014. mgcv. URL `http://cran.r-project.org/web/packages/mgcv/`.

Wood, SN. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)* 73:3–36. URL `http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9868.2010.00749.x/full`.

Yu, Alan C. L. 2013. Individual differences in soci-cognitive processing and the actuation of sound change. In *Origins of sound change: Approaches to phonologization.*, ed. Alan C. L. Yu, 201–227. Oxford University Press.

Yuan, Jiahong, and Mark Liberman. 2008. Speaker identification on the SCOTUS corpus. In *Proceedings of Acoustics*.

Zellou, Georgia, and Meredith Tamminga. 2014. Nasal coarticulation changes over time in Philadelphia English. *Journal of Phonetics* 47:18–35. URL `http://linkinghub.elsevier.com/retrieve/pii/S0095447014000734`.