Computational model of sound change in Todd, Pierrehumbert, Hay (2019).

Jonathan Harrington

Sound change and word frequency

How to model three different kinds of results.

/t/-glottalisation in Manchester English: no effect of word frequency (Bermúdez-Ortero et al, 2015[1])

/t/-flapping in New Zealand English: sound change takes place first in more frequent words (Hay & Foulkes, 2016[2])

push-chain /a-ɛ/ in NZE: /ɛ/ moves from /a/ occurs in words of low frequency first (Hay et al, 2015[3])

1. Bermúdez-Otero et al (2015). *Paper presented at 2nd Edinburgh symposium on historical phonology*. 2. Hay & Foulkes (2016, *Language*, 92, 298-330. 3. Hay et al (2015) *Cognition*, 139, 83–91.

Similarities in both Todd et al (2019) and Stevens et al (2019, *Glossa*):

- There are words (types), phonemes (categories), and signals (exemplars).
- A word is a generalisation over the signals/exemplars that belong to it. A phoneme is the union of these signals (e.g. /a/ is a generalisation of the signals over map, cat, ...).
- Speaker and listener. A phoneme shift is caused by the listener's absorption (or not) of the signal into memory i.e. sound change is perceptually driven via a production-perception loop.
- The number of signals(exemplars) per word and phoneme remain the same during the ongoing sound change
- Contra a modular feedforward architecture and the Neogrammarian principle: phonemes of words can shift at different rates.
- In contrast to Ohala, in which one allophone is replaced by another (in all words), sound change is incremental.

Differences between Todd et al (2019) and Stevens et al (2019)

In Todd et al (2019) but not Stevens et al (2019):

- Only single phoneme (drift) or two-phoneme interaction.
- Phoneme classes are fixed (no split/merge)
- Monosyllables and no minimal pairs (no bet/bat)
- A single agent or community talking to itself (no multiple agent interaction).
- Acoustics: a single value (a one-dimensional space). No modelling of speech dynamics.
- The data is artificially generated (not based on data from real speakers).
- The sound change is *extrinsically* induced by adding a bias in the direction of sound change - rather than emerging *intrinsically* from phonetic variation as in Stevens et al, 2019.

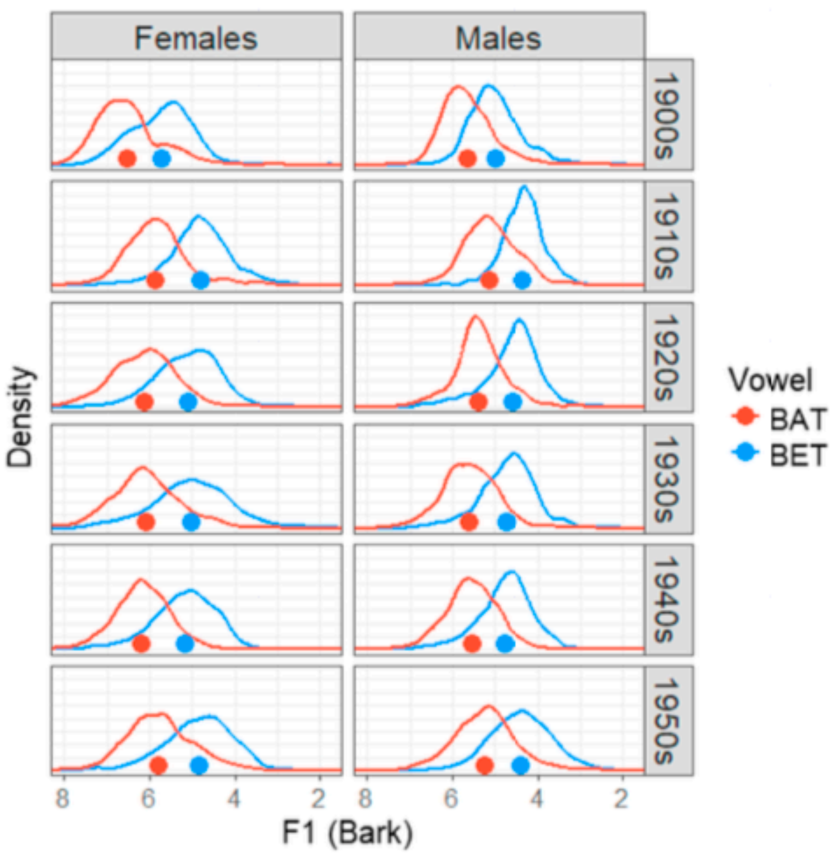
Differences between Todd et al (2019) and Stevens et al (2019)

In Todd et al (2019) but not Stevens et al (2019):

- there is explicit modelling of word frequency
- decisions about whether or not to absorb new items are based on exemplar strength (Gaussian probability in Stevens et al, 2019).
- there is category (phoneme) overlap but no mechanism for actual merger. Todd et al (2019) explicitly argue against the mechanism in Stevens et al (2019) in which A is progressively absorbed into B in an A->B merger (A, B are two phonemes). Todd et al (2019): B must not 'leech' A.

Differences between Todd et al (2019) and Stevens et al (2019)

In the starting conditions and in the ongoing sound change in Todd et al (2019), the distribution of phonemes remains more or less symmetrical i.e. no skew. In Stevens et al (2019), and indeed in many other models of sound change (Beddor, 2009[1]; Garrett & Johnson, 2013[2], Ohala 1993[3]), skew is one of the factors that drives sound change.



'As vowel distributions moved in the New Zealand short front vowel shift, they maintained their distance from one another, their shapes (width and skewness), and their degree of overlap with one another. At all times, they exhibited little skewness and sub- stantial overlap relative to their width'

Starting conditions in Todd et al (2019): words

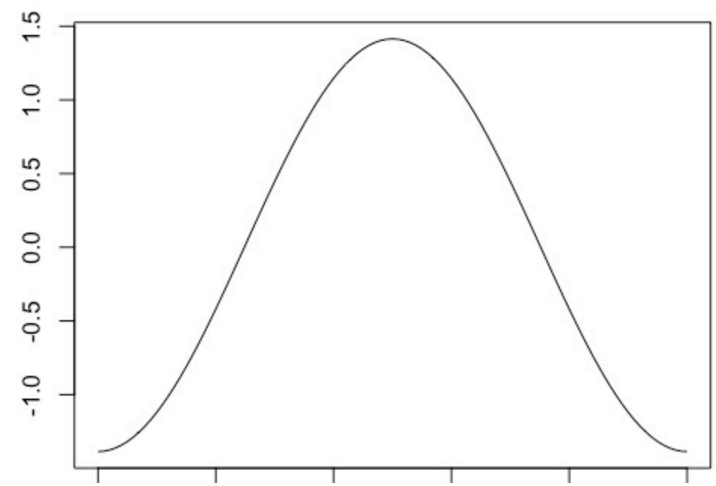
Each word is associated with a lexical (log.) frequency.

More frequent words have more exemplars than low frequency words. 'We follow the multiple-trace hypothesis (Hintzman & Block, 1971) in assuming that the number of exemplars for a given type (word) represents that type's frequency'.

There were fewer high frequency than low frequency words in the simulation in order to ensure that the total number of exemplars in high and low frequency words (and in e.g. in /a, e/ phonemes was the same).

Starting conditions in Todd et al (2019): acoustics

Each exemplar consists of a single value generated at random from a raised cosine distribution



Interaction: Production

A word, *W* is retrieved in relation to its lexical frequency (high frequency words are more likely to be selected).

One of the *W*'s exemplars, *E*, is chosen at random.

A bias is added to any *E* whose phoneme is in the direction of the sound (in an A->B shift, a bias is added to any exemplar that is a member of /A/, not /B/).

Noise is added to *E* regardless of its phonemic association.

W and *E* are sent to the listener...

Interaction: Perception

E activates all of the listener's stored exemplars. The degree of activation depends on a (Gaussian weighted) distance to *E*.

E is then absorbed depending on discriminability and typicality

Perception and Discriminability

"compare the ratio of category activations (intended category activation, Ai, divided by other category activation, Ao) to the discriminability threshold, δ (a parameter). When the ratio is equal to the threshold, the probability of passing the evaluation is 0.5"

$$P(\text{pass discriminability evaluation} | A_i, A_o) = \frac{\frac{A_i}{A_o}}{\frac{A_i}{A_o} + \delta}$$

Todd pers. com: the evaluation is probabilistic. So it can be passed (with low probability) even in cases where the activation of the other category is greater than the activation of the intended category. Note: this is a point of difference from your model, where the exemplar passes the equivalent test only if the probability of the intended category is greater than the probability of the other category. In your model, the strict nature of this test means that categories cannot overlap. In our model, the probabilistic nature of the test means that overlap is permitted (provided there is some force encouraging the two categories to stay near one another).

Perception and Typicality

The probability of passing the typicality evaluation is determined by comparing the activation of the intended category, Ai (normalized for the number of exemplars of the category,), to the typicality threshold, τ (a parameter). When the activation is equal to the threshold, the probability of passing the evaluation is 0.5

$$P(\text{pass typicality evaluation}|A_i) = 1 - \exp\left(-\ln 2 \cdot \frac{A_i}{N_i \tau}\right)$$

Todd pers. comm: "If you just have the choice between the two categories, then the probability would not fall as you move outward from the mode, in the direction away from the other category – i.e. you would say that anything that doesn't seem like the other category would pass the typicality evaluation, even if it doesn't much seem like the intended category either. We characterize using activation; the activation of the intended category must be sufficiently high. As for the discriminability evaluation, the typicality evaluation is probabilistic "

Perception and absorption into memory

if *E* is absorbed into memory (having passed the discriminability and typicality tests), it replaces one of the exemplars from the same word selected at random

Modelling single phoneme drift: one phoneme, no competitor phonemes

to model cases such as /t/-Glottalisation in Manchester English in which high and low frequency words change at the same rate

Suppose two words W_H, W_L (high and low frequency) such that the frequency of W_H is 4 times that of W_L

Then W_H has 4 times as many exemplars as W_L.
Suppose 5 exemplars (4 in W_H, 1 in W_L).

Then in 40 selections, each exemplar is likely to advance by 40/5 = 8 units. i.e. **each of the exemplars** of W_H will advance by 8 units, and **the single exemplar in** W_L will also advance by 8 units.

Generalising: the bias in the direction of sound change is distributed across a larger number of exemplars in W_H which is offset by W_H being selected more often. These effects cancel each other out, so W_H and W_L change at the same rate.

Modelling A in the sound change A->B

e.g. A = /t/-flapping in NZE which encroaches on the space for B = /d/.

Hay & Foulkes (2016). A->B is faster in high frequency words.

Psycholinguistic evidence used in computational model: e.g. high frequency words are more robustly identified in noise (Howe, 1957) are in lexical decision are more likely to be classified as real words (Luce & Pisoni, 1998), and faster (Forster & Chambers, 1973).

In the computational model, Todd et al (2019) make high frequency words more discriminable. Therefore, in the region of overlap, where A encroaches on B, high frequency words are more likely to be retained than low frequency words (hence high frequency words change faster).

Modelling B in the sound change where B moves away from A

bat -> bet and 'bet' moves away from 'bat'. It is this movement away of 'bet' words from 'bat' that is being modelled.

Presumably a bias is applied to 'bet' words to move them away from 'bat' words.

high frequency 'bet' words in the region of overlap with 'bat' are more likely to pass the discriminability threshold. Hence the change (involving 'bet' moving away from 'bat') is more likely in low frequency words.

Modelling B in the sound change where B moves away from A

JMH: HF-words change faster than LF-words in /a/ for the same reason as for /t/-flapping

Todd: "In the model, yes, but this isn’t actually a model of /ae/ in NZE (in which LF words change fastest). The simplifying assumptions of the model only allow us to develop a treatment of /E/, as discussed in Appendix A.1."

"In essence, it’s the reverse of the situation for /t/-tapping. /E/ is moving away from another category, i.e. away from a region of acoustic ambiguity (of course, this sense of movement “away” is only local, because /ae/ is also moving). The discriminability asymmetry means that HF words are more robust to acoustic ambiguity than LF words, so they can remain in the ambiguous region longer than LF words. That is, LF words are repelled by the ambiguous region more, so they retreat from it faster.

The fact that LF words are also subject to the typicality force less than HF words (due to the interaction between production and storage described in Appendix B) helps them to extend further in the direction away from /ae/. So it helps the effect, but it doesn’t drive the effect"