An audio-visual account for sound changes towards sibilants merger in Taiwan Mandarin

Baichen Du
baichen@connect.hku.hk
Department of Linguistics, The University of Hong Kong

Apr 15, 2023

Background

- Three-way contrast among sibilants in standard Mandarin, contrasting between place and manner of articulation, as well as aspiration.
  - Alveolars: ts, tsʰ, s
  - Palatoalveolars: tɕ, tɕʰ, ɕ
  - Retroflexes: ʈʂ, ʈʂʰ, ʂ
- Ongoing merger from retroflexes to alveolars in Southern China, especially in Taiwan Mandarin (Y.-H. Chang, 2012; Y.-H. S. Chang & Shih, 2015; Chiu et al., 2020; Chung, 2006).
- The merger is initiated by increased frequency of retroflexes, presumably by more fronted position of tongue contact (Chiu et al., 2020).
  - Articulatorily-motivated? Maybe.

Background

- Production-perception link
  - Bilateral influence between production and perception.
  - Speech accommodation task (Babel, 2009; 2010; 2012; 2014)
  - Second dialect acquisition (Munro et al., 1999; Nycz, 2013; 2011), etc
- The current study:
  - Testing P&P link in the sibilant merger.
  - Experiment 1: (Mechanism) More merged speakers will tolerate more high-COG retroflexes in perception; Significant social factor (i.e., gender).
  - Experiment 2: (Propagation) Significant context factor (i.e., visual information) in convergence towards the merger.

Experiment 1: Design & Procedure

- Phase 1: Production
    - 23 mono- or di-syllabic Mandarin Tone 1 minimal pairs (initial stress)
    - 23 native Mandarin speakers from the University of Hong Kong
    - Read stimuli in carrier phrase three times out loud in a comfortable pace and volume.
- Phase 2: Perception
    - 23 eight-step continuums from retroflexes to alveolars (i.e., low COG to high COG).
    - Two-alternative forced choice task.

Experiment 1: Method

- **Speech synthesis**
  - What are we synthesizing?
    - The frication part of affricates and fricatives.
  - What is manipulated in the synthesis?
    - Center of gravity (i.e., spectral mean) of frication.
  - How to create speech tokens based on that?
    - Splicing synthesized frications onto natural production of closure, bursts, aspiration, and vowels for retroflexes.

Experiment 1: Method

- Synthesizing fricatives at both ends of three groups of continuums.
    - Using a Praat script, white noise is first high- and low-pass filtered at a frequency X, with adjustable slopes on both sides.

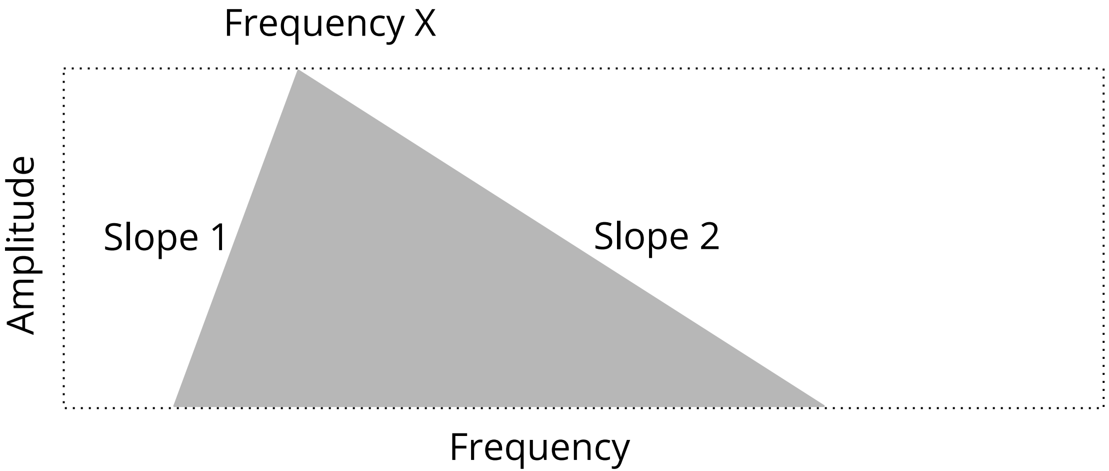


Figure 1. Spectral representation of frication synthesis

Experiment 1: Method

- Synthesizing fricatives
  - The process is repeated for frequency Y…

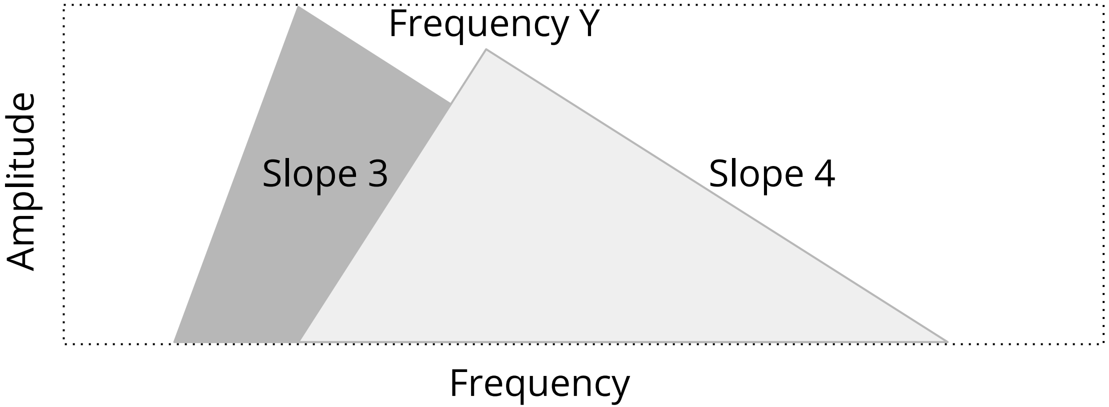


Figure 2. Spectral representation of frication synthesis

Experiment 1: Method

- Synthesizing fricatives
  - And Z.

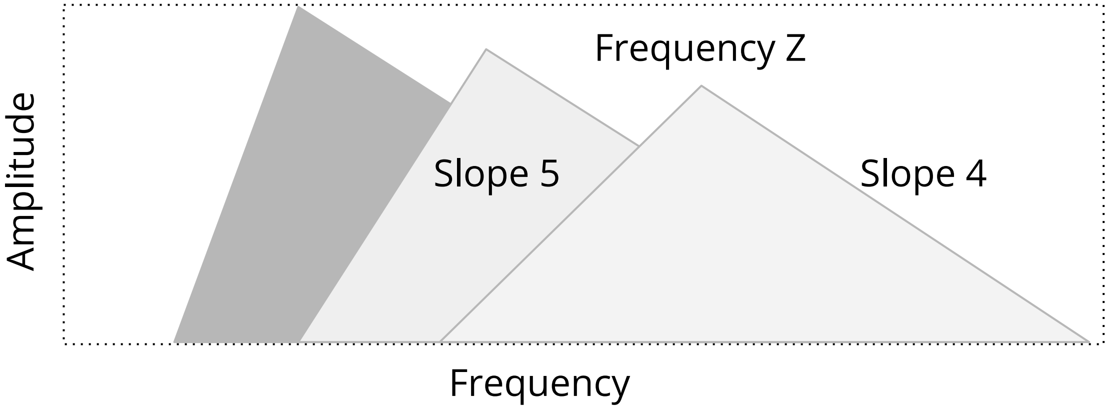


Figure 3. Spectral representation of frication synthesis

Experiment 1: Method

- Synthesizing fricatives
  - All parameters shown above were manually obtained from natural productions of Mandarin in Praat.
  - This, although used elsewhere (e.g., Schiller, 2009), might not be the perfect way of synthesizing.
  - Other option: Klatt + TANDAM-STRAIGHT (Kawahara, 2008)
    - Automatic morphing between two speech tokens on every possible parameters.
    - However, no control over dimensions other than centre of frequency, including “formants” in frication part.

Experiment 1: Method

- Synthesizing fricatives
    - Both ends were interpolated on Bark scale into 8 steps using a Praat script.
    - Waveform envelope shape (Van Tasell et al., 1987) and second formant transitions in the frication part (Soli, 1981) were disregarded.
    - F2 transition was not controlled. But it can be a cue in both production and perception (e.g., Hauser, 2023).

Experiment 1: Method

- Synthesized frications:

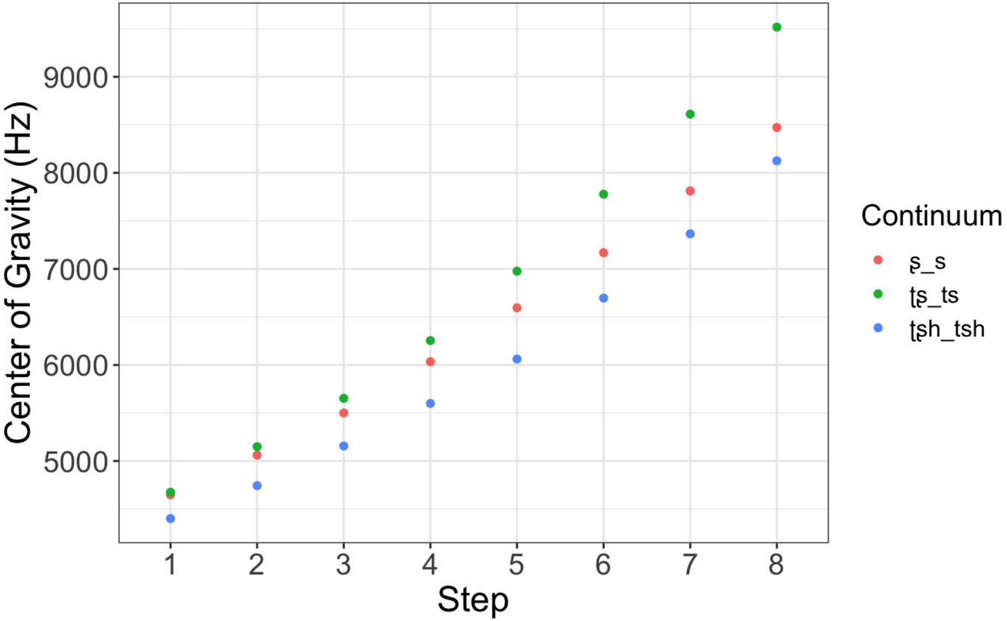


Figure 4. Frequencies of synthesized steps

Analysis: Experiment 1

- All recordings were annotated in Praat. Frication was hand-marked.
- Four spectral moments were obtained using a Praat script.
- Production of minimal pairs were paired together and COG differences were calculated in R.
- Perception results were analyzed in R.
- Production and perception results were analyzed in R:
    - Frequentist and Bayesian Logistic mixed-effects regression:
        - Response ~ Step + Continuum * Merger + (1|Participant)

Results & Discussion: Experiment 1

- Global spectral property of sibilants:
    - One thing to notice: Compared to previous reports of frequency (i.e., around 10000Hz and 5000Hz for alveolars and retroflexes, see Chang & Shih, 2015; Kallay & Holliday, 2012; Li et al., 2016; Li et al., 2014), our participants produced sibilants relatively lower.

Pair	Retroflexes	Alveolars	Difference	N
ʂ_s	3156	6657	3501	420
ʈʂ_ts	2649	5998	3349	540
ʈʂh_tsh	2413	5232	2724	418

Table 1. Production results

Results & Discussion: Experiment 1

- Gender effect: Female leading the merger.

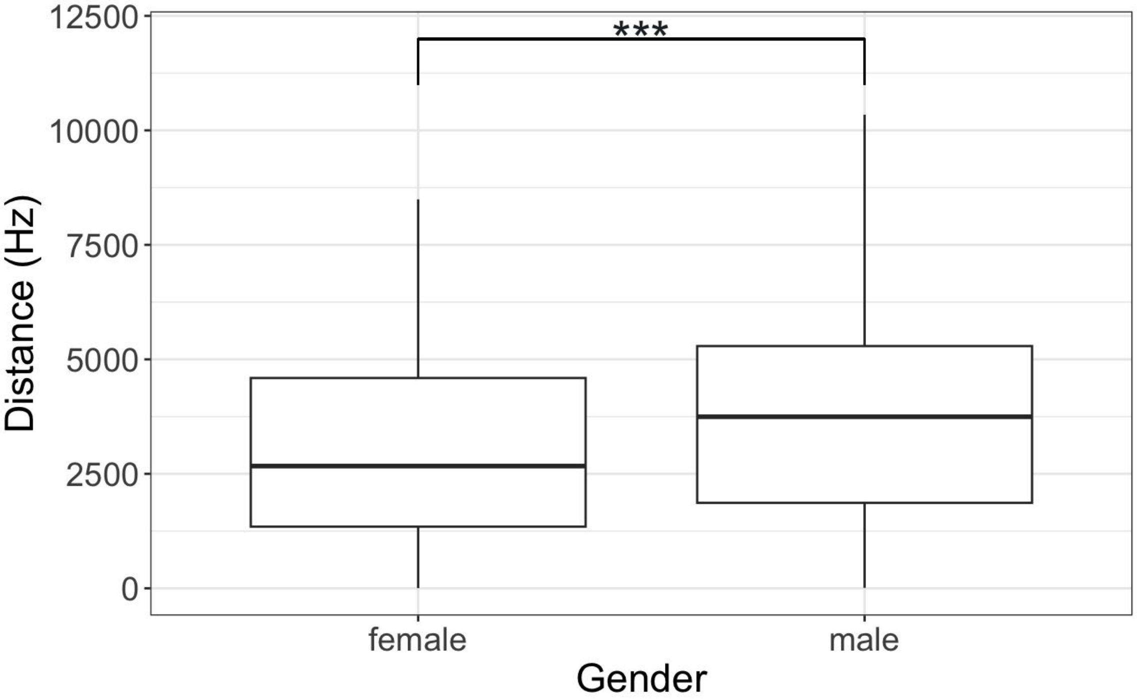


Figure 4. Distance between alveolars and retroflexes

Results & Discussion: Experiment 1

- Global perceptual boundary between retroflexes and alveolars:

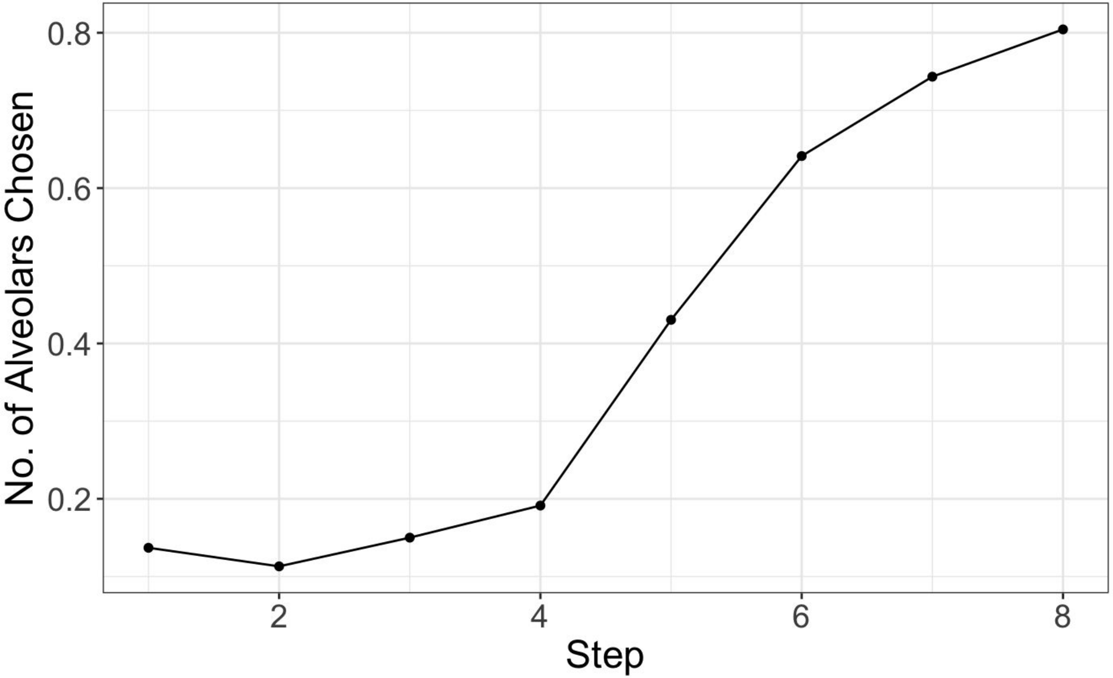
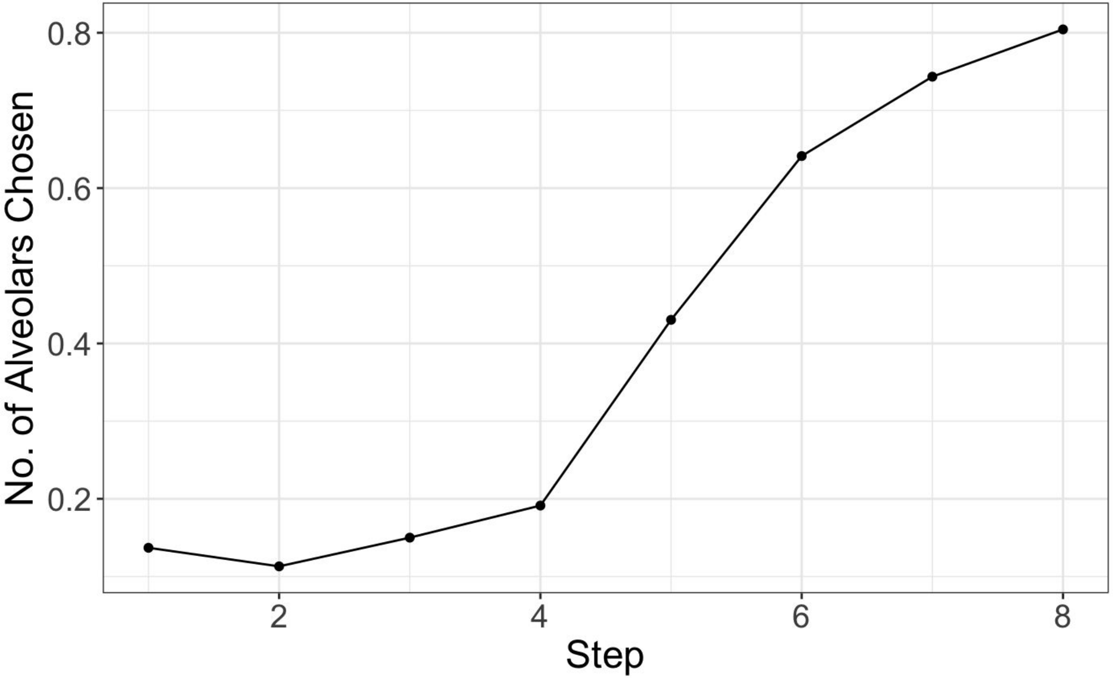


Figure 5.

Results & Discussion: Experiment 1

- Global perceptual boundary between retroflexes and alveolars:



Not reaching 100%?

Figure 5.

Results & Discussion: Experiment 1

- By-onset boundary:

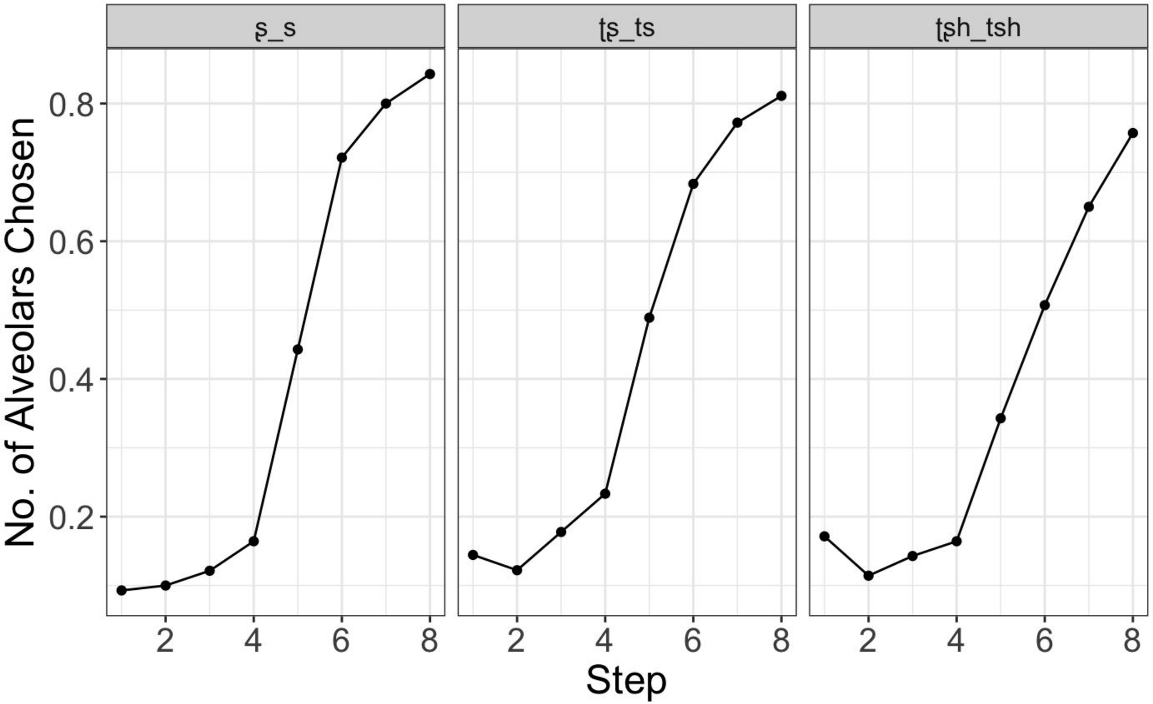


Figure 6.

Results & Discussion: Experiment 1

- By-onset boundary:

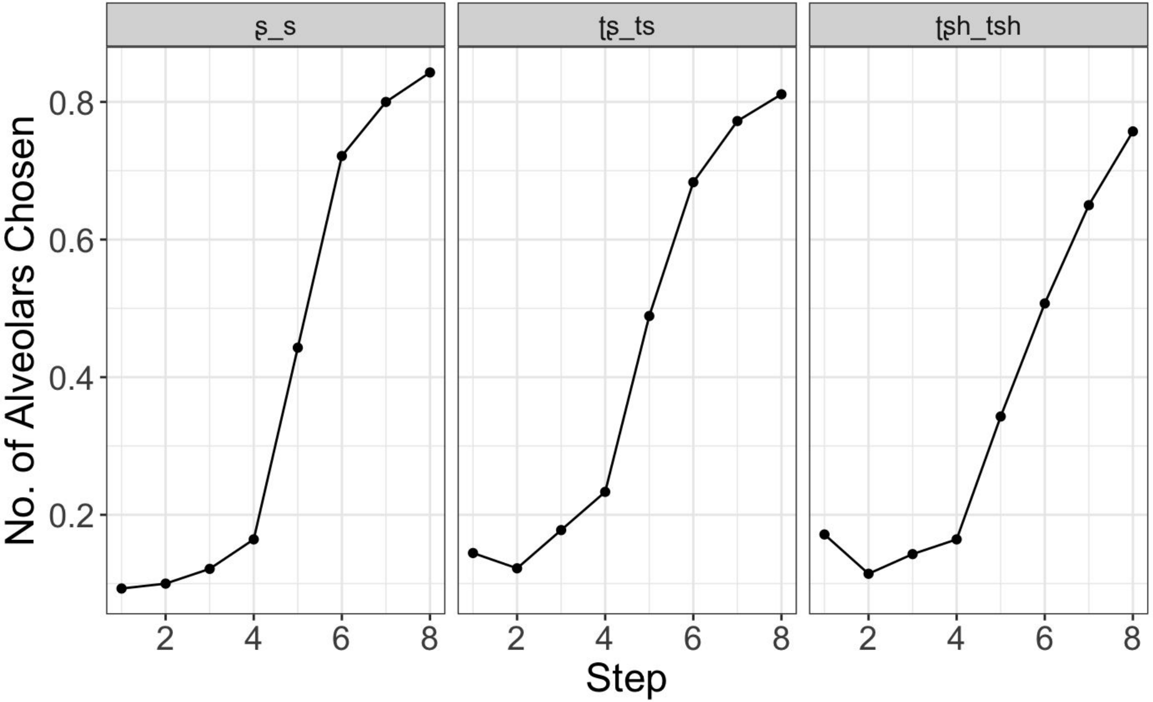


Figure 6.

Not reaching 100%?

Results & Discussion: Experiment 1

- Mergers versus non-mergers: Kolmogorov-Smirnov Score in production.

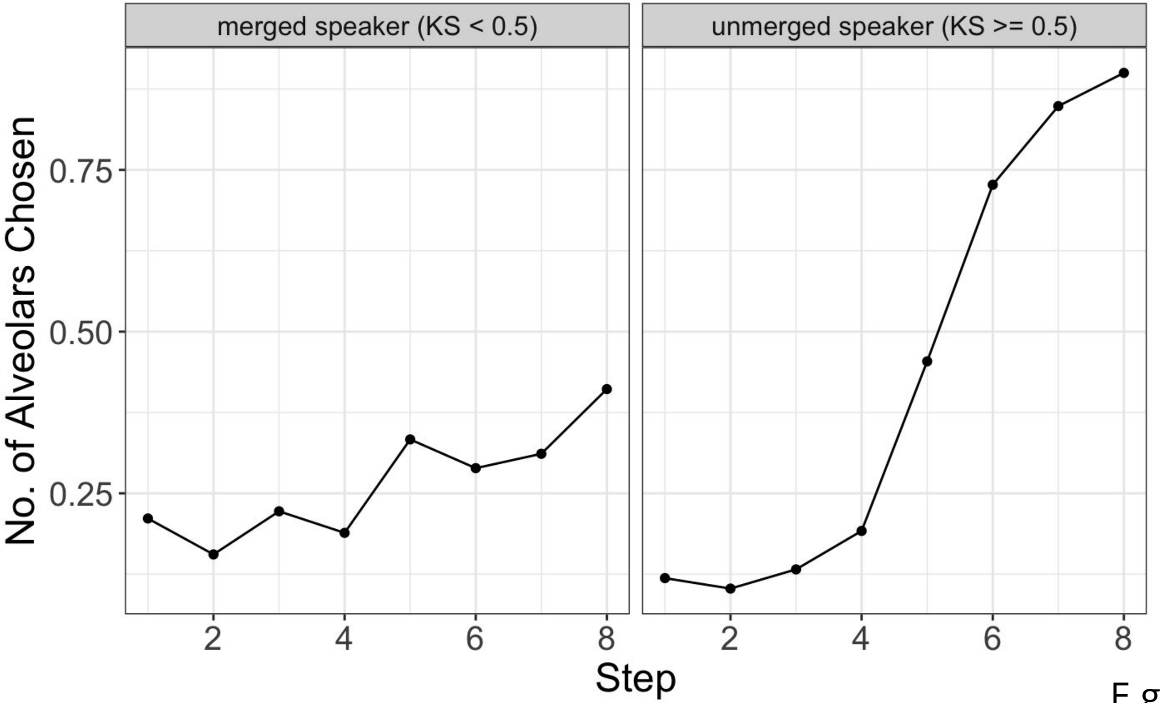


Figure 7.

E.g., /r/ vs /l/ in ESL

Results & Discussion: Experiment 1

- Mergers versus non-mergers: Kolmogorov-Smirnov Score in production.

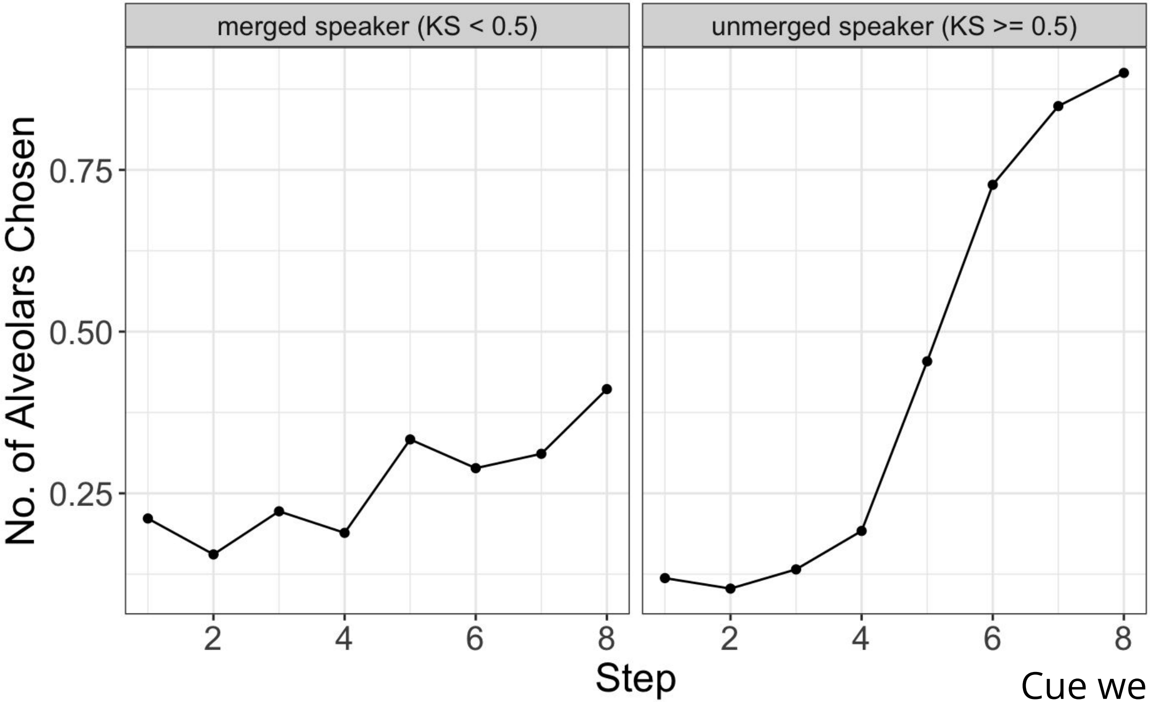


Figure 7.

Cue weighting variation?

Results & Discussion: Experiment 1

- Testing: Closer distance between retroflexes and alveolars in production correlates with more tolerance of high-frequency retroflexes in perception.

	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**
(Intercept)	-2.4063	0.2397	-10.038	< 2e-16 ***
step 2	-0.2221	0.2015	-1.102	0.27028
step 3	0.1078	0.1897	0.568	0.56976
step 4	0.4062	0.1817	2.236	0.02536 *
step 5	1.6042	0.1672	9.594	< 2e-16 ***
step 6	2.5092	0.1699	14.772	< 2e-16 ***
step 7	3.019	0.1763	17.126	< 2e-16 ***
step 8	3.3866	0.1836	18.442	< 2e-16 ***
ʈʂ_ts	-0.294	0.2668	-1.102	0.27055
ʈʂh_tsh	**-0.6429**	**0.2482**	**-2.59**	**0.00959 ** **
unmerger	**0.664**	**0.214**	**3.103**	**0.00192 ** **
ʈʂ_tsh:unmerger	0.4726	0.2874	1.644	0.10008
ʈʂh_tsh:unmerger	0.5461	0.2754	1.983	0.04738 *

Table 2. Logistic Regression on responses

Results & Discussion: Experiment 1

- Adding random intercept for each subject: Generalized Linear Mixed-Effects Model failed to converge. Bayesian model was used instead.

	Estimate	**Est.Error**	**L-95% CI**	**U-95% CI**	**Rhat**	**Bulk_ESS**	**Tail_ESS**
Intercept	-2.41	0.28	-2.99	-1.88	1	1466	1766
step2	-0.22	0.2	-0.61	0.17	1	1947	2614
step3	0.11	0.2	-0.27	0.49	1	1764	2744
step4	0.42	0.19	0.07	0.79	1	1631	2454
step5	1.65	0.17	1.31	2	1	1533	2215
step6	2.58	0.18	2.24	2.92	1	1497	2373
step7	3.1	0.18	2.75	3.46	1	1611	2195
step8	3.48	0.19	3.12	3.86	1	1610	2766
ʈʂ_ts	-0.26	0.28	-0.79	0.27	1	1603	2729
ʈʂh_tsh	**-0.77**	**0.26**	**-1.27**	**-0.25**	**1**	**1642**	**2385**
unmerger	**0.62**	**0.26**	**0.12**	**1.14**	**1**	**1611**	**2548**
ʈʂ_ts : unmerger	0.44	0.3	-0.15	1.03	1	1557	2449
ʈʂh_tsh : unmerger	0.7	0.29	0.14	1.26	1	1598	2322

Table 3. Bayesian LMER on responses with by-subject random intercepts.

Experiment 2: Method & Procedure

- Speech accommodation task towards a merged talker.
  - “Repeat after the model talker you hear or see.”
  - Audio-visual condition: Video of talking face shown to shadowers
  - Audio only condition: Audio stimuli only
- Measurement: Center of gravity of frications.
  - Phase 1: natural production of stimuli (baseline)
  - Phase 2: shadowing task using speech with more advanced merger (exposure)
  - Phase 3: reproduction of the same stimuli (post-exposure)

Experiment 2: Analysis

- Difference-in-difference:
  - Difference in advancedness of merger between model talker and shadower in each phase.
- Advancedness of merger:
  - Distance between alveolars and retroflexes measured in raw frequency and K-S score.
- Predicted result:
  - DID: Phase 1 > Phase 3 >= Phase 2 (Baseline DID > post-shadowing >= shadowing)
  - More accommodation in A-V condition than in A-only condition.
  - Rejection of H1 if ranked otherwise.

Results & Discussion: Experiment 2

- If the results are indeed what we predicted:
  - Visual information might coordinate sound change on a higher level, providing more contexts or social pressure for convergence.
  - However, according to our existing study design, even if our hypothesis is supported, we do not know if this is due to synchronized articulatory gesture or the mere presence of human face.
    - Blurred or partially masked face?

Conclusion

1. There is indeed a merger in the participants we tested.
2. The merger is led by female speakers, consistent with previous sound change studies.
3. Compared to unmerged speakers, merged speakers were more likely to perceive more retroflexes along the same continua.
4. The merger might be conditioned by onsets or vowels, or cue weighting differences among the population.
5. We predict that audio-visual condition induces more accommodation to the merger, by offering more social contexts and/or visible articulatory cues.

Future Studies

- If it fits Ohala’s listener based sound change model (i.e., misperception), other phonological factors might also come into play:
    - Following vowel? Aspiration? Other psycholinguistic effects?
- Cue weighting variation?
- Merger by expanding versus by shifting? Important but unclear.
- For the exact role played by visual information, an ongoing project is looking at the question by manipulation of visual stimuli.

Acknowledgement

- This study was made possible by TransDisciplinary Undergraduate Research Initiative Grant from the University of Hong Kong.
- I would like to express my special thanks to Dr. Jonathan Havenhill and Dr. Youngah Do at the Language Development Lab of HKU for offering feedback and comments of the project, and Dr. Alexandra Pfiffner at UC Berkeley Linguistics for offering suggestions on the presentation.
- I also thank all subjects who participated in this study.