

UC Davis

UC Davis Electronic Theses and Dissertations

Title

From Idiosyncrasy to Generalization: The Role of Sociolinguistic Experience on Adaptation to Novel Talkers

Permalink

<https://escholarship.org/uc/item/3gr336q5>

ISBN

9798189237072

Author

Aoki, Nicholas Bamshad

Publication Date

2026-06-25

Peer reviewed|Thesis/dissertation

From Idiosyncrasy to Generalization:
The Role of Sociolinguistic Experience on Adaptation to Novel Talkers

By

NICHOLAS BAMSHAD AOKI
DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Linguistics

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Georgia Zellou

Fernanda Ferreira

Santiago Barreda-Castañón

Committee in Charge

2026

© Copyright

Nicholas Bamshad Aoki, 2026

All rights reserved.

Abstract

Cognitive scientists have long been interested in cross-talker generalization. When listeners are exposed to phonetic variation, when do they apply what they have learned to novel talkers? A debate about generalization has emerged between two ostensibly opposing theories. Some studies find that exposure to multiple talkers is necessary for generalization (a “numerosity account”); others claim that acoustic similarity between the exposure and novel talkers is what drives generalization, not multiple-talker exposure (a “sufficient similarity account”).

But are numerosity and sufficient similarity as different as they seem? This dissertation makes the following insight – multiple-talker exposure only ever affects generalization when the exposure and novel talkers are acoustically similar. This suggests that, rather than being in competition, numerosity and sufficient similarity only differ in the amount of evidence required for generalization. Numerosity relies on two types of evidence (acoustic similarity and multiple-talker exposure), while sufficient similarity relies on just one (acoustic similarity).

This observation leads to a novel “linguistic experience hypothesis” – when listeners possess less experience with variation, more evidence is needed to prove that such variation is truly generalizable and not just an idiosyncrasy (i.e., a numerosity account should be observed

in situations of lower listener familiarity, and vice versa for sufficient similarity). To investigate the linguistic experience hypothesis, this dissertation presents two experiments as a proof of concept. Experiment 1 focused on a test case of relatively high listener familiarity (L1-English listeners were exposed to L2-accented English). Experiment 2 then presented a test case that was specifically designed to be unattested in L1-English (a gendered /p/ to [b] phonetic shift produced by text-to-speech voices). The results were consistent with the linguistic experience hypothesis – less evidence was required for generalization in Experiment 1 (just acoustic similarity); meanwhile, more evidence was needed for generalization in Experiment 2 (both acoustic similarity and exposure to multiple talkers).

The linguistic experience hypothesis makes specific, testable predictions about how social context mediates adaptation to novel talkers, with implications for typology, theories of sound change, and more. In a broader sense, this dissertation highlights the theoretical and practical value of integrating a deeper sociolinguistic perspective into cognitive science.

Acknowledgments

The central finding of this dissertation somehow feels like an apt metaphor for my time in graduate school. In order to successfully adapt to a difficult challenge in an unfamiliar environment, you can't do it with just one person – multiple (and in this case many) people are needed to help you along the way.

First and foremost, thank you to Georgia Zellou for being the most diligent, thoughtful, and understanding advisor ever. I first started working with Georgia in Fall 2021, while she was in Paris on a Fulbright. On top of her own research and all of her professional and personal responsibilities (in addition to navigating a global pandemic), she still strove to create the best advising experience possible for me as a new graduate student. She woke up at 4am to meet with me weekly on Zoom, she pore through dozens of Google Docs with all of my ideas for new experiments (some of which were a bit odd), and she guided me through every step of the publication process. Somehow, I felt at home and cared for, despite Georgia being thousands of miles away.

Now that I've worked with her for the last five years, I can confidently say that Georgia is one of the most hard-working people I've ever met, not just as my advisor but in all aspects

of her professional life. I really don't know how she does it all, and I feel extremely fortunate to have been her graduate student mentee.

I must also acknowledge my two other dissertation committee members. Thank you to Fernanda Ferreira for providing encouragement and feedback on my qualifying paper and for hosting Language Group on Fridays. No matter who presented at Language Group, she always had something incredibly thought-provoking to say, and I appreciated her attentiveness and constructive critique on everyone's work. Thank you also to Santiago Barreda-Castañón for his course about statistics and Bayesianism, and for informing us about his interesting work on theories of normalization.

This dissertation is based on two published journal articles that benefited greatly from the peer-review process. For their insightful comments, all thanks to Meredith Tamminga and three anonymous reviewers at *Glossa Psycholinguistics*, as well as to Meghan Clayards, Cynthia Clopper, and two anonymous reviewers at *Language and Speech*. I would also like to thank Xin Xie and Chigusa Kurumada, who I met at the 2025 Linguistic Society of America Institute – post-publication, they made several points about how the *Glossa Psycholinguistics* article could be improved, and these ideas are woven into the present dissertation.

Attending graduate school at UC Davis has been one of the best choices I've ever made – not only have I learned tremendously, but I've also appreciated the collegiality and professionalism of all the faculty, administrators, and graduate students of the Linguistics Department. I would particularly like to thank three people: (i) Michelle Cohn, for providing valuable guidance on my very first research project as a graduate student and for just being an all-around nice person with a can-do attitude; (ii) Emily Morgan, for teaching my favorite

graduate course on usage-based linguistic theories and for teaching me an important lesson about providing student feedback; (iii) Laurie Lawyer, for her dedication to teaching and for brightening so many of my Tuesdays and Thursdays with her sense of humor and verve for quirky phonological data.

In addition to the members of our department, I’ve also had the pleasure of getting to know many wonderful people at conferences and many motivated undergraduates in the classroom. Thank you to Anne Pycha for meeting with me to discuss my qualifying paper, to Roxanne Elliott and Ana Ortega Pérez for being such terrific and organized co-chairs of the Cluster on Language Research (CLR), and to Nicole Holliday for being a passionate keynote speaker at the CLR Symposium in 2023. An enormous thank you to all of my students – every time I feel despair about the world, I think of Quinn Alarcon, Bayleigh Baldwin, Cal Byrd, Morgan Caldera, Kendall Campbell, Ganyu (“Gloria”) Chen, Griffin (“Fin”) Engels, Delaney Ervin, Macie Gard, Naveed Jones, Annabelle Juanitas, Nandana Koonantavida, Ashlynn Kyaw, Natasha Lee, Grace Macasaet, Audrey Marrero, Kailee (“Lucky”) Meade, Yasmin Moule, Kelsey Sennett, Bálint Szigeti Csúcs, Chuyuan Wang, Samantha Yang, and the hundreds of other curious and driven undergraduates I’ve had the privilege of teaching, and I realize that the future is incredibly bright.

I would be remiss if I did not mention my mentors and professors during my undergraduate education. Thank you to Diane Brentari, whose enthusiasm for linguistics inspired me to become a major, even though I barely knew what “linguistics” was. Thank you to Jackie Yan-Ki Lai for generously providing feedback on my writing sample for graduate school. Thank you to Monica Phillips, for meeting me countless times both at her office hours and

at private appointments to go over my terrible essays line-by-line at the Oriental Institute. Not only did Monica help me become a better writer, but she also showed how a model graduate student teaching assistant should behave, and she inspires me every day. Thank you to Jinghua Ou, for mentoring me as a research assistant for several years on a cue weighting project, and for showing patience despite my overuse of electrode gel. Most of all, thank you to Alan Yu, for graciously letting me into his Phonology Lab, for giving me the opportunity to conduct independent research (even though I had no clue what I was doing), and for still being a kind and caring mentor to this day.

I am very lucky to have a large and supportive extended family. Thank you to Mahin Zarbinian, Zarrin Zarbinian, and Mariko Aoki for being my biggest cheerleaders and for their unconditional love. And of course, how could I forget Marcus Bartholomew Sprout, the Cheerful Gingerbread Man, the Bespectacled Strawberry, and all my other friends at home?

My father, Victor Aoki, is the critic who provides the necessary balance to my cheerleaders, the one who has taught me to never rest on my laurels. Every time I introspect and think about what I could do better for next time, I hear his voice in my head.

My mother, Maryam Bamshad-Alavi, not only inspired this dissertation through her idiosyncratic accent, but she also inspires me in life. It is *remarkable* how much effort she puts in to support the junior faculty of her department, the students in her class and lab, and all of her immediate and extended family. I only wished that she appreciated herself as much as we all appreciate her.

This dissertation is for you, Mom! After I graduate, let's *arriba clemento* and dance, dance, dance for many years to come.

Declarations

1. **Ethical Considerations:** All experiments received approval from the Institutional Review Board (IRB) of the University of California, Davis (reference number: 1328085-2). All procedures adhere to the tenets of the Declaration of Helsinki (World Medical Association, 2025). All participants provided informed consent, and the identity of all participants has been anonymized.

2. **Funding:** The author is grateful for financial support from the Office of Graduate Studies at the University of California, Davis. This includes a Provost's Fellowship in the Arts, Humanities and Social Sciences (Year 1, \$59802), a Graduate Research Mentorship Fellowship (Year 2, \$48900), and a yearly stipend (Years 3-5, approximately \$41370 per year). These funds were partially used to compensate participants, although they were not specifically earmarked for research purposes (other than a \$600 allowance from the Graduate Research Mentorship Fellowship).

3. **Generative Artificial Intelligence (AI) Usage:** Many academic journals now require (or strongly recommend) authors to be transparent about their use of generative AI tools

(e.g., Google Gemini and ChatGPT), given their potential to automate certain aspects of the research and writing process (Tang et al., 2024). However, definitions of “generative AI” can vary and can sometimes be nebulous/circular (Ronge et al., 2025), leading to a lack of clarity about what exactly falls under its purview.

The author has selected the following as a working definition of “generative AI”: “a system of algorithms or computer processes that can create novel output in text, images or other media based on user prompts” (National Library of Medicine, 2024). This description was deemed ideal because it: (i) stems from an official, governmental source; (ii) avoids circularity by not including the words “generative/generate”, “artificial”, or “intelligence”; (iii) encompasses many of the prototypical examples of “generative AI” (e.g., Google Gemini, ChatGPT, etc.).

Based on this definition from the National Library of Medicine, generative AI was only used in the creation of the auditory stimuli in Experiment 2. Generative AI was otherwise not employed in the making of this dissertation, from conceptualization to completion.

4. Data Accessibility: It is widely acknowledged that data accessibility is critical for reproducibility and transparency (Martone et al., 2018). All data and code have been made available as a zip file on the Open Science Framework: <https://doi.org/10.17605/osf.io/n8hws>. This includes: (i) the raw and cleaned data sets; (ii) all R code (for cleaning, analyzing, and visualizing the data); (iii) the R Markdown script used to create this document.

5. Stimulus Accessibility: The experiments in this dissertation use auditory stimuli. The auditory stimuli in Experiment 1 cannot be made available due to lack of explicit permission from the IRB. The auditory stimuli in Experiment 2 cannot be made available either as they

were generated by Amazon Polly and are thus the property of Amazon Web Services (AWS). That said, as of writing this statement (December 27th, 2025), the reader should be able to recreate the Amazon Polly stimuli if they create an AWS account.

6. Prior Publication: Portions of Chapters 2-4 are published in *Glossa Psycholinguistics* (Aoki & Zellou, 2025a) and *Language and Speech* (Aoki & Zellou, 2025b), as alluded to in the Acknowledgments. There are several reasons why the text in this dissertation is not directly copied from the published articles:

- i. Although both articles share considerable theoretical overlap and their conclusions dovetail fairly neatly with each other, they also differ methodologically and were originally designed as stand-alone papers. Therefore, certain sections were either added or deleted to help make the writing flow more smoothly.¹
- ii. Unlike the journal articles, which were subject to certain constraints from their respective journals, the current dissertation has no word limit. The exposition has thus been greatly expanded to more thoroughly motivate the experiments theoretically, to provide greater detail about the methodology, to conduct additional statistical analyses, and to more clearly unpack the broader practical implications of this work.

¹The idea for putting these two papers together actually came about through the peer-review process. It was June 2025 – the *Glossa Psycholinguistics* paper had already been published, and the author was struggling to piece together their dissertation. Then, the first set of reviews from *Language and Speech* arrived, and Reviewer 1 wrote the following: “As a final note, I enjoyed reading ‘When multiple talker exposure is necessary for cross-talker generalization: Insights into the emergence of sociolinguistic perception’ from the same team earlier this year (Aoki & Zellou, 2025)... While of course it is up to the authors and editorial team rather than reviewers to decide the granularity of work submitted, I thought it worth mentioning that as an audience member for this interesting line of findings I would have preferred to read a single manuscript presenting all converging evidence”. Upon reading this comment, this dissertation sprung to life – the author realized that, although the two papers have slightly different vantage points, their overall conclusions complement each other. Thank you to Reviewer 1 from *Language and Speech* for their suggestion!

iii. Science is constantly in flux, and this work is no exception. It is true that, as noted in the Acknowledgments, both papers went through rigorous, thorough, and helpful peer-review processes at *Glossa Psycholinguistics* and *Language and Speech*, facilitated by highly constructive reviewers and editors. At the same time, it is the author's opinion that there is no such thing as a "perfect" paper – even after peer-review.

Since their publication in 2025, the author has realized that both articles could have been better in several respects, due to a combination of self-reflection, having conversations with various researchers, and reading/listening to others' papers and presentations. This dissertation reflects new versions of the original articles that have (hopefully) been improved, and footnotes are included to give credit for specific suggestions made by other researchers.

Despite these changes, the essence of the articles has been maintained – the theoretical grounding and general conclusions remain the same, and no new data have been collected.

7. Competing Interests: The author declares no potential conflicts of interest with respect to the research, authorship, and/or publication of this dissertation. The author takes full responsibility for any errors.

Vita

Education

2026	Ph.D., Linguistics	University of California (UC), Davis
2021	M.A., Digital Studies of Language, Culture, and History	University of Chicago
2021	B.A., Linguistics, <i>summa cum laude</i>	University of Chicago

Selected Honors and Awards

2026	Outstanding Graduate Student Teaching Award	UC Davis
2022	Graduate Research Mentorship Fellow	UC Davis
2021	Provost's Fellow in the Arts, Humanities, and Social Sciences	UC Davis
2020	Student Marshal	University of Chicago
2020	Phi Beta Kappa (Early Induction)	University of Chicago

Selected Publications

1. Zellou, G., Holliday, N., & **Aoki, N. B.** Introduction to the special issue on the phonetics of human-AI interaction. Forthcoming at *Phonetica*.
2. **Aoki, N. B.**, & Zellou, G. (2025b). Apparent talker variability and speaking style similarity can enhance comprehension of novel L2-accented talkers. *Language and Speech*. <https://doi.org/10.1177/00238309251390505>
3. **Aoki, N. B.**, & Zellou, G. (2025a). When multiple talker exposure is necessary for cross-talk generalization: Insights into the emergence of sociolinguistic perception. *Glossa Psycholinguistics*, 4(1). <https://doi.org/10.5070/g6011.21217>
4. Vonessen, J., **Aoki, N. B.**, Cohn, M., & Zellou, G. (2024). Comparing perception of L1 and L2 English by human listeners and machines: Effect of interlocutor adaptations. *The Journal of the Acoustical Society of America*, 155(5), 3060-3070. <https://doi.org/10.1121/10.0025930>
5. **Aoki, N. B.**, & Zellou, G. (2024). Being *clear* about clear speech: Intelligibility of hard-of-hearing-directed, non-native-directed, and casual speech for L1- and L2-English listeners. *Journal of Phonetics*, 104, 101328. <https://doi.org/10.1016/j.wocn.2024.101328>
6. **Aoki, N. B.**, & Zellou, G. (2023c). Visual information affects adaptation to novel talkers: Ethnicity-specific and ethnicity-independent learning of L2-accented speech. *The Journal of the Acoustical Society of America*, 154(4), 2290-2304. <https://doi.org/10.1121/10.0021289>
7. **Aoki, N. B.**, & Zellou, G. (2023b). When clear speech does not enhance memory: Effects of speaking style, voice naturalness, and listener age. *Proceedings of Meetings on Acoustics*, 51(1), 060002. <https://doi.org/10.1121/2.0001766>
8. **Aoki, N. B.**, & Zellou, G. (2023a). Speakers talk more clearly when they see an East Asian face: Effects of visual guise on speech production. In R. Skarnitzl & J. Volín (Eds.), *Proceedings of the 20th International Congress of Phonetic Sciences* (pp. 2294-2298). Guarant International.
9. **Aoki, N. B.**, Cohn, M., & Zellou, G. (2022). The clear speech intelligibility benefit for text-to-speech voices: Effects of speaking style and visual guise. *JASA Express Letters*, 2(4), 045204. <https://doi.org/10.1121/10.0010274>

Contents

Abstract	ii
Acknowledgments	iv
Declarations	viii
Vita	xii
List of Figures	xv
List of Tables	xvi
1 Prologue	1
2 Introduction	5
2.1 Background: Variability within the speech signal	6
2.1.1 The lack of invariance problem	6
2.1.2 Solution 1: Erasing variation?	9
2.1.3 Solution 2: Leveraging the statistical structure of speech	11
2.1.4 The ideal adapter framework	14
2.2 The problem: To generalize or not to generalize?	17
2.3 Numerosity or sufficient similarity?	18
2.3.1 The case for numerosity	18

2.3.2	The case for sufficient similarity	20
2.3.3	Reconciling the accounts: A linguistic experience hypothesis	22
2.3.4	Reinterpreting prior work through the linguistic experience hypothesis	25
2.3.5	The need for novel test cases and a novel sociolinguistic approach in the adaptation literature	30
2.4	Overview of the dissertation	32
2.4.1	Goals of the dissertation	33
2.4.2	Layout of the dissertation	33
3	Experiment 1: Adapting to Mandarin-Accented English	35
3.1	Introduction	36
3.1.1	Why <i>apparent</i> talker variability?	36
3.1.2	Examining effects of acoustic similarity with speaking style	42
3.1.3	Overview of the study design	44
3.1.4	Predictions	45
3.2	Methods	46
3.2.1	Stimuli	46
3.2.2	Participants	50
3.2.3	Procedure	55
3.2.4	Statistical analysis	57
3.3	Results	59
3.3.1	Exposure phase	59
3.3.2	Test phase	61
3.4	Discussion	65
3.4.1	Ruling out a cognitive difficulty account	67
3.4.2	Why was no generalization observed among listeners tested on hard-of- hearing-directed speech?	68
3.4.3	Conclusion and next steps	70

4	Experiment 2: Adapting to a Novel Phonetic Shift	71
4.1	Introduction	72
4.1.1	Developing a suitable test case	72
4.1.2	Overview of the study design	78
4.1.3	Predictions	79
4.2	Methods	81
4.2.1	Exposure stimuli	81
4.2.2	Test stimuli	86
4.2.3	Participants	90
4.2.4	Procedure	92
4.2.5	Statistical analysis	96
4.3	Main results	98
4.3.1	Exposure to a single talker per gender	98
4.3.2	Exposure to multiple talkers per gender	99
4.3.3	Combined analysis	101
4.4	Interim discussion	102
4.5	Post-hoc analysis	103
4.5.1	Statistical analysis and results	106
4.5.2	Refitting the combined model	108
4.5.3	Interim discussion	109
4.6	Discussion	113
4.6.1	Accounting for the order effect	114
5	General Discussion	117
5.1	The utility of sociolinguistic experience in theories of cognitive science	119
5.2	Future directions and limitations	121
5.2.1	Refining the linguistic experience hypothesis	121

5.2.2	Other ways of testing the linguistic experience hypothesis: Integrating typology and linguistic diversity	124
5.2.3	Individual differences and theories of sound change	125
5.3	Conclusion	128
References		129
Appendix A: Experimental Materials		152
Appendix B: Model Output		158

List of Figures

2.1	Distribution of /s/ and /ʃ/ center of gravity in Hertz (Hz), aggregated across 13 speakers from the same dialect region (“The North”). The vertical line is placed at 4920 Hz, the approximate crossover point between the two distributions.	8
2.2	Distribution of /s/ and /ʃ/ center of gravity in Hertz (Hz) by gender (left panel) and by two talkers (right panel; both female). All distributions stem from the corpus visualized in Figure 2.1. The vertical line at 4920 Hz is also carried over from Figure 2.1.	12
2.3	Perceptual categorization of an /ʃ/-/s/ continuum in a hypothetical experiment, where Step 1 is the most /ʃ/-like and Step 9 is the most /s/-like. In tandem with Figure 2.2: (i) a vertical black line is placed at the most ambiguous point along the continuum; (ii) ambiguous tokens are more likely to be categorized as /s/ for male speakers and Talker 049 (blue) than for female speakers and Talker 072 (red). Note that these are idealized results and not based on any specific computational model.	14
2.4	A flowchart illustrating how the linguistic experience hypothesis unites the sufficient similarity and numerosity accounts under one theoretical framework. When listeners possess ample experience with variation, then single-talker exposure is sufficient for generalization, lending credence to a sufficient similarity account (left side; exposure to an L1-Mandarin talker with a canonical L2-English accent). However, when listeners are unfamiliar with variation, then multiple-talker exposure may become helpful (or even necessary) for generalization, supporting a numerosity account (right side; exposure to the L1-Farsi speaker described in the Prologue; i.e., an individual who <i>seems</i> to have an L1-English accent overall, but produces isolated phonetic shifts that would be more characteristic of an L2-speaker).	29
3.1	A map of the contiguous United States, where each dot represents the estimated location of residence for each participant in Experiment 1. The dots are color-coded by their Rural-Urban Commuting Area (RUCA) code, where lower numbers (more purple) indicate more urban locales.	53
3.2	Exposure phase transcription accuracy across conditions. Individual points reflect by-participant mean performance. Error bars are 95% confidence intervals bootstrapped over by-participant means.	60

3.3	Test phase transcription accuracy across conditions, among participants tested on casual speech. Individual points reflect by-participant mean performance. Error bars are 95% confidence intervals bootstrapped over by-participant means.	62
3.4	Average test phase transcription accuracy across conditions, among participants tested on hard-of-hearing-directed speech. Individual points reflect by-participant mean performance. Error bars are 95% confidence intervals bootstrapped over by-listener means.	63
4.1	A user complaining about the pronunciation of British Siri on an online forum (dated September 12, 2020), with agreement from 18 other users. Siri is apparently producing a “t” to “r” shift (e.g., “twenry” instead of “twenty”), randomly replacing one sound with another: https://discussions.apple.com/thread/251785651 (accessed May 6, 2026).	78
4.2	Distribution of Voice Onset Time for word-initial /p/ in the exposure stimuli. The left panel shows the unshifted female tokens, while the right panel shows the (/p/ to [b]) shifted female tokens. Both panels show identical distributions for the unshifted male tokens.	85
4.3	Distribution of sentence-level mean f0 in the exposure stimuli.	85
4.4	Aggregated results by Step and Speaker in the norming experiment (female = top row, red; male = bottom row, blue).	88
4.5	Aggregated results by Step and Speaker Gender (female = red; male = blue) for the original 9-step continuum (left panel) and the adjusted 5-step continuum used in Experiment 2 (right panel).	88
4.6	Aggregated results in the “single talker per gender” exposure condition by Step, Speaker Gender (female = red, male = blue), Block Order (Female → Male, Male → Female), and Shift Type (Female Shifted, No Shift). Error bars are 95% confidence intervals bootstrapped over by-participant means.	99
4.7	Aggregated results in the “multiple talkers per gender” exposure condition by Step, Speaker Gender (female = red, male = blue), Block Order (Female → Male, Male → Female), and Shift Type (Female Shifted, No Shift). Error bars are 95% confidence intervals bootstrapped over by-participant means.	101
4.8	Aggregated results by Step and Speaker (JOA = Joanna; RUT = Ruth; SAL = Salli) in Experiment 2 for all participants assigned to the No Shift condition (i.e., no /p/ to [b] phonetic shift in the exposure phase). Error bars are 95% confidence intervals bootstrapped over by-participant means.	112

List of Tables

3.1	Results of the burst duration analysis for each speaker. The “Mean Diff” column subtracts the average value for hard-of-hearing-directed speech by the average value for casual speech. The full output of paired t-tests are reported: 95% confidence intervals (CI), t-values, degrees of freedom (df), and p-values. Significance annotations reflect meaningful differences between styles (* = $p < 0.05$; ** = $p < 0.01$; *** = $p < 0.001$).	49
3.2	Results of the speaking rate analysis for each speaker. The “Mean Diff” column subtracts the average value for hard-of-hearing-directed speech by the average value for casual speech. The full output of paired t-tests are reported: 95% confidence intervals (CI), t-values, degrees of freedom (df), and p-values. Significance annotations reflect meaningful differences between styles (* = $p < 0.05$; ** = $p < 0.01$; *** = $p < 0.001$).	50
3.3	Mean RUCA value for each of the ten between-participant conditions in Experiment 1, where the variables “Exposure Style”, “Apparent Exposure Talker Number”, and “Test Style” are abbreviated as “Exposure”, “Variability”, and “Test”, respectively. The term “hard-of-hearing-directed speech” is abbreviated as “HOH-DS”.	54
3.4	Summary statistics for the final model of the test phase analysis, which investigated the relative degree of generalization across all critical conditions (i.e., a comparison of test phase accuracy in the four right-most conditions of Figure 3.3 and the four right-most conditions of Figure 3.4). The predictors “Exposure Style”, “Apparent Exposure Talker Number”, and “Test Style” are abbreviated as “Exposure”, “Variability”, and “Test”, respectively. Credible intervals that do not cross zero are in bold.	65
4.1	Summary of predictions for each theoretical account in Experiment 2. Each prediction (“No difference” or “More ‘pie’ response”) compares the novel female talker to the novel male talker in the test phase of the critical condition. (In the control condition, neither exposure talker produces the /p/ to [b] phonetic shift, so “buy-pie” categorization is expected to be the same for both test talkers).	79
4.2	Voice Onset Time (ms) of at each step along the final version of the continua.	90
4.3	Onset f0 (Hz) of each step along the final version of the continua.	90

4.4	Hypothetical set of results supporting a strong version of a numerosity account (i.e., where multiple-talker exposure is <i>necessary</i> for generalization). The column and row labels stand for the possible exposure and test talkers, respectively. The values within each cell refer to presence (‘Yes’) or absence (‘No’) of generalization in this scenario. Dashes reflect the lack of talker-specific conditions in the current study (i.e., the exposure talkers were never presented in the test phase).	104
4.5	Hypothetical set of results that support a sufficient similarity account <i>and</i> still lead to the aggregated results in Section 4.3. The layout of the table is identical to Table 4.4.	105
4.6	Summary statistics for the interaction between Speaker Gender and Shift Type in each post-hoc model. Credible intervals that do not cross zero are in bold.	107
4.7	Actual results of the post-hoc analysis. The structure of this table is the same as Tables 4.4 and 4.5, such that the column labels represent the exposure talkers and the row labels represent the test talkers. ‘Yes’ and ‘No’ refer to the presence or absence of generalization, respectively.	109
A.1	Sentences presented in the exposure and test phases of Experiment 1. All of the sentences come from Kalikow et al. (1977).	152
A.2	Sentences presented in the exposure phase of Experiment 2. The first 16 sentences are based on Zellou et al. (2023), while the rest were created by the author.	156
B.1	Summary statistics for the original ‘single talker per gender’ exposure model in Experiment 2. Credible intervals that do not cross zero are in bold.	158
B.2	Summary statistics for the original ‘multiple talkers per gender’ exposure model in Experiment 2. Credible intervals that do not cross zero are in bold.	159
B.3	Summary statistics for the combined model in Experiment 2. Credible intervals that do not cross zero are in bold.	160
B.4	Summary statistics for the combined model in Experiment 2 that omits all listeners who heard Joanna at test. Credible intervals that do not cross zero are in bold.	162

Chapter 1

Prologue

It was an ordinary day just like any other – or so it appeared at the time. My father, then a laboratory technician, was preparing equipment for the experiments that afternoon. Wash the test tubes, set up the beakers, repeat. Accustomed to silence within the sterile, white walls, he worked alone.

An opening door shattered the monotony of routine. Instead of the stern countenance of his supervisor, my father was greeted by the warm smile of a new researcher. How could he have known that this very moment would mark the beginning of a great adventure?

As the saying goes, opposites attract. The cuisine and customs emblematic of my father's Japanese-American heritage differed completely from that of my mother's homeland, and they both became curious to learn more about their respective cultures. Date nights alternated – sometimes they would snack on *onigiri* flecked with *furikake*, while other times, my mother would prepare her pomegranate stew or other family recipes. Their conversations lasted into the night, and their relationship blossomed into a lifelong friendship.

My father had met many other L2-English speakers, having attended a very diverse university. Yet nobody had quite the same accent as my mother. On the one hand, she is *almost* indistinguishable from an L1-English speaker, due to her conscious effort to assimilate

within American society. Her English is so fluent that most people initially just assume that she is from the United States and would never guess that her first language was Farsi.

At the same time, her way of talking can be unexpected, especially when she is talking more casually. For example, she sometimes pronounces /w/ as [v] (“You’re *vel*come”) and sometimes exhibits a vowel shift from /ɑ/ to [ʊ] (“*peuts* [pʊts] and pans”). These surprising phonetic shifts are accompanied by occasional transposition errors (*chipoletti* instead of “chipotle”), malapropisms (♫ “At first I was *a-scared*, I was petrified. . .” ♫), and slip-ups with idioms (“It’s raining *deugs* [dʊgz] and cats” instead of “It’s raining cats and dogs”).

My mother’s accent is thus somewhat enigmatic and unlike any other L2-accent that my father had heard before. Even though she is actually an L2-English speaker, she *appears* to sound like an L1-English speaker who makes funny remarks and produces isolated phonetic shifts that slip out every so often. Since my father had never met any other L1-Farsi speakers, the source of these shifts was unclear. Do the pronunciations of /w/ as [v] and /ɑ/ as [ʊ] truly represent the last vestiges of an L1-Farsi accent? Or are these phonetic shifts just random quirks, akin to accidentally saying *chipoletti* rather than “chipotle”?

My father realized that the answer to these questions was nuanced, once he met his soon-to-be in-laws. All of them spoke with prominent L1-Farsi accents, which included a consistent /w/ to [v] shift. Curiously, however, none of the in-laws possessed an /ɑ/ to [ʊ] vowel shift. Unbeknownst to my father at the time, Farsi already has the /ɑ/ vowel in its phonological inventory (Jam, 2020), making an adjustment from /ɑ/ to [ʊ] unexpected for a Farsi-accented English speaker. To this day, none of us (including my mother) know exactly when, how, or why the vowel shift occurred.¹

¹This is purely speculative, but it seems plausible for the vowel shift to have derived from phonetic imitation (Sancier & Fowler, 1997). My mother learned English in the United Kingdom, having relocated to the London suburbs at 12-years-old (she later moved to the United States where she eventually met my father). In England, she enrolled at a school with many other international students (she recalls encountering immigrants from Nigeria, China, Sweden, El Salvador, and dozens of other countries). English is notorious for its extensive vowel inventory, which often leads to difficulty for L2-learners (Flege & Bohn, 2021; Flege et al., 1999) – perhaps my mother phonetically imitated other students with an /ɑ/ to [ʊ] shift, which somehow persisted even after she moved to the United States.

Regardless of how it came to be, my mother’s accent is interesting because of its ambiguity. Certain aspects of her accent are *idiosyncratic* (the /ɑ/ to [ʊ] shift), while other features are *generalizable* to L1-Farsi speakers (the /w/ to [v] shift). Since my mother’s speech patterns are so different from typical L2-speakers, hearing *multiple* other talkers from the same language background (the in-laws) became necessary for him to tease apart which features were idiosyncratic and which were generalizable.

Partially due to language barriers (and also cultural differences), the interactions between the two halves of our family have not always been straightforward. It was difficult for my father to adapt to Farsi-accented English, and for my mother’s relatives to understand an American English accent. Nevertheless, both sides have always demonstrated a commitment to each other, a quality that the author has grown to appreciate over the years. As my father interacted with more and more friends and extended family members, the challenge of mutual comprehension waned, both linguistically and culturally. With each new person who produced the /w/ to [v] shift and did not produce the /ɑ/ to [ʊ] shift, the clearer it became that the former was a property of Farsi-accented English while the latter was just my mother’s idiosyncrasy.

The wedding arrived only around a year after the first date. Amid the lively chatter, Frank Sinatra’s booming baritone voice (♫ “Strangers in the night...” ♫) and the clinking of silverware, my father ambled around the dance floor, greeting all of the guests. What a difference a year makes! Far from struggling, my father could easily comprehend the Farsi-accented invitees, even the ones he had never talked to before. It was no longer a surprise that, instead of asking “when” my parents first met, they asked *when* they first met. My father was thus engaging in *cross-talker generalization* – having learned the correlation between a social category (my mother’s family) and a phonetic feature (the /w/ to [v] shift) through his long-term experience, he then *generalized* this knowledge to novel speakers from the same social category (Kleinschmidt et al., 2018).

In a broader sense, the story of my parents is a product of modernity – interactions between people of varied linguistic backgrounds are now ubiquitous amid rapid globalization (Coupland, 2010). Above and beyond a personal anecdote, achieving a better understanding of speech adaptation is important for advancing theories of linguistics and cognitive science and for addressing real-world issues of speech communication (e.g., reducing comprehension disparities between speakers of “Mainstream” American English and other speakers from diverse accent/dialect backgrounds; Kang, 2010; Rickford & King, 2016).

This dissertation investigates the mechanisms of cross-talker generalization – under what conditions do listeners generalize an accent or phonetic shift to novel talkers?

Chapter 2

Introduction

Why does this dissertation open with a story, instead of a more customary literature review? The aim of the prologue is to: (i) foreground the question of interest; (ii) highlight a concrete example of adaptation to an atypical L2-accent that a listener has never experienced before; (iii) underscore the potential importance of social context and *numerosity* (the number of exposure talkers) in mediating cross-talker generalization.

In general, the adaptation literature tends to focus heavily on the acoustics of speech, with recent studies being particularly interested in the quantitative modeling of acoustic cue distributions. While not denying the importance of acoustics, this dissertation claims that a more interdisciplinary approach is needed to uncover the mechanisms of cross-talker generalization. More specifically, it will be argued that the listener’s sociolinguistic experience plays a pivotal role in generalization, in addition to speech acoustics.

The first half of the introduction is organized as follows. Section 2.1 reviews prior research on speech variability, highlighting a prominent theory of perceptual adaptation (the ideal adapter framework; Kleinschmidt & Jaeger, 2015). Through the lens of the ideal adapter framework, Section 2.2 more formally delineates the question motivating this dissertation: When do listeners generalize an accent or phonetic shift to a novel talker?

Section 2.3 articulates the gap that this dissertation intends to fill. The first subsections focus on two theoretical accounts (numerosity and sufficient similarity), reviewing recent work that seems to support *sufficient similarity* over numerosity (i.e., acoustic similarity is what really affects generalization, not the number of exposure talkers; Xie et al., 2021; Xie & Myers, 2017). A novel “linguistic experience” hypothesis proposes that sufficient similarity and numerosity are not actually competing accounts – rather, they are mediated by listeners’ sociolinguistic experiences (greater experience should lend *less* support for numerosity and greater support for sufficient similarity, and vice versa).

Section 2.3 then claims that the literature underestimates the importance of social context on generalization, which results in test cases that incidentally disfavor numerosity (e.g., presenting canonical L2-accented speech to L1-English listeners at diverse universities who are already somewhat familiar with L2-accents). Evaluating the linguistic experience hypothesis, this dissertation presents two generalization experiments with contrasting test cases – one where listeners have some degree of familiarity, and another where listeners have no familiarity. The broader aim is to highlight the value of a theoretical approach in cognitive science that is equally informed by both acoustic *and* social considerations.

Section 2.4 concludes by summarizing the main claims of the introduction, providing an overview of the goals and findings of the experiments, and previewing the theoretical and practical implications of this work.

2.1 Background: Variability within the speech signal

2.1.1 The lack of invariance problem

The speech signal abounds with *variation*, with one speaker’s production of a word potentially diverging drastically from another’s. For example, as noted in the Prologue, a Farsi-accented

English speaker might produce /w/ as [v] (e.g., “welcome” as “*velcome*”).¹ The existence of variation is well-known – when comparing accents/dialects, acoustic differences can be highly conspicuous, often stimulating explicit metalinguistic commentary (Bucholtz et al., 2007; Lippi-Green, 1994). Variation, however, is endemic to the speech signal, even among talkers from the same language/dialect background (Hillenbrand et al., 1995) and even in the absence of listener awareness (Johnstone et al., 2006). Understanding how listeners navigate variation has persisted as a longstanding area of research in the cognitive sciences (Liberman et al., 1967; Xie et al., 2023).

Variation has traditionally been framed as an obstacle to successful speech perception (Pisoni, 1981). In particular, the many-to-many relationship between acoustic signals and linguistic categories makes speech ambiguous (Magnuson & Nusbaum, 2007) – the same acoustic signal can map onto different categories, while different acoustic signals can map onto the same category. To provide a clearer definition of “many-to-many relationship” and to elucidate why such a property poses a challenge for the perceptual system, the rest of this section walks through a concrete example of speech variation from a miniature corpus.

Figure 2.1 depicts the center of gravity distributions for the sibilants /s/ and /ʃ/, aggregated across 13 L1-English talkers from the ALLSSTAR corpus (Bradlow, n.d.).² Although /s/ has a higher center of gravity than /ʃ/ on average (Jongman et al., 2000), there is a

¹The usage of slashes and brackets stems from generative phonology – traditionally, symbols within slashes refer to abstract phonemes or underlying categories, while symbols within brackets denote allophones or the acoustic-phonetic realizations of phonemes (Chomsky & Halle, 1968). However, the validity of the generative linguistics paradigm has come under debate in the last several decades (Piantadosi, 2024; Port & Leary, 2005), with certain scholars questioning the psychological reality of the phoneme-allophone distinction (Samuel, 2020). This dissertation remains agnostic about these theoretical issues. Similar to other authors in the adaptation literature (e.g., Xie & Myers, 2017; Zellou et al., 2023), the implementation of slashes and brackets is only adopted as a useful and well-known convention for teasing apart an intended production (/w/) from its actual physical realization ([v]) without implying any commitment about underlying representations. For example, writing /w/ is not meant to insinuate that “w” is “phonemic” or to advocate for one particular phonological theory over another.

²For each talker, there were 37 /s/ and 17 /ʃ/ tokens across 60 “Hearing in Noise Test 1” sentences (702 total tokens and 780 total sentences); center of gravity was measured automatically through a Praat script (DiCanio et al., 2021). These particular 13 talkers were selected because they had all primarily lived in the same dialect region prior to recording (“The North”; Labov et al., 2005). The gender, age range, and ethnicity of the talkers were as follows: 6 women, 7 men; 18-26 years; Asian = 2, Black = 1, White = 10. The specific speaker IDs were: 049, 050, 056, 059, 060, 061, 064, 066, 068, 070, 072, 131, 132.

many-to-many relationship between center of gravity and sibilant category. The same center of gravity value can refer to different categories (e.g., a token with 4920 Hz is equally likely to be /s/ or /ʃ/); meanwhile, different center of gravity values can belong to the same category (e.g., /s/ could be produced with anywhere from 3000 to 9000 Hz). How can listeners reliably identify /s/ and /ʃ/ at all, in the context of such ambiguity? This example illustrates the *lack of invariance problem* (Liberman et al., 1967) – the difficulty of achieving stable perceptual categories from a variable and ambiguous speech signal.

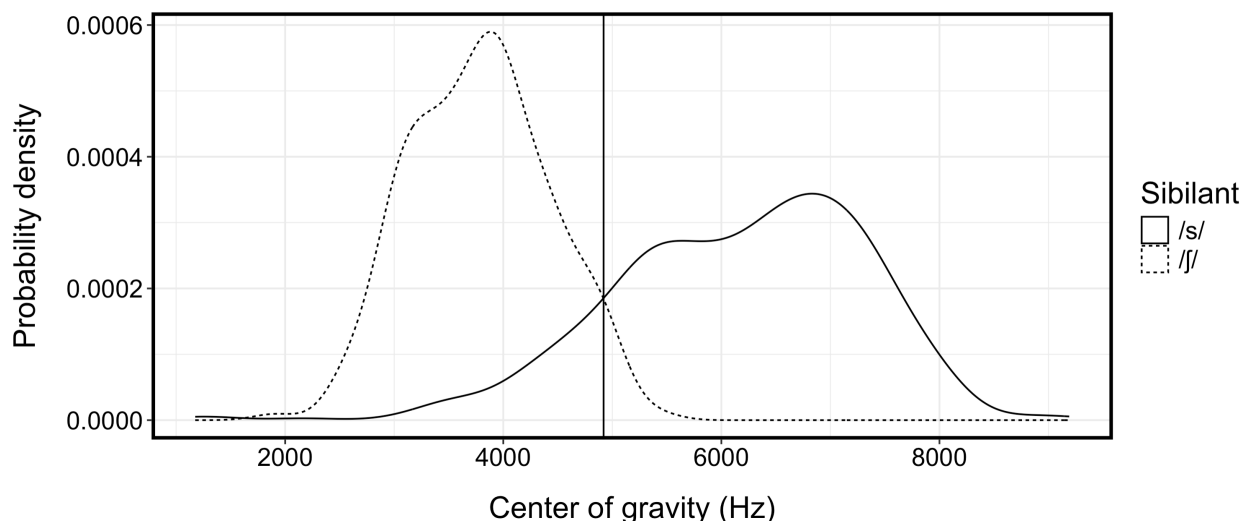


Figure 2.1: Distribution of /s/ and /ʃ/ center of gravity in Hertz (Hz), aggregated across 13 speakers from the same dialect region (“The North”). The vertical line is placed at 4920 Hz, the approximate crossover point between the two distributions.

Note that Figure 2.1 represents an infinitesimal fraction of the lack of invariance problem. Besides the /s/-/ʃ/ distinction in L1-English, many-to-many cue-to-category mappings have been observed for all speech sounds across all dialects and languages, among L1- and L2-speakers alike (Chao & Chen, 2008; Clopper et al., 2005; Idemaru et al., 2012; Xie & Jaeger, 2020; Zellou, 2017). Further compounding its complexity, speech is also acoustically multidimensional, with sounds being differentiated along potentially dozens of cues (Schertz & Clare, 2020). For example, even if there are tokens that are clearly separable by center of gravity in Figure 2.1, these same tokens might be ambiguous in terms of other important

cues (kurtosis, skewness, amplitude, F2 onset, etc.; Jongman et al., 2000). The lack of invariance problem is ever-present, even when interacting with L1-speakers from a matching demographic and linguistic background.

In theory, if the speech signal is a continuous jumble of ambiguity, then speech perception should be unfeasible. Bombarded with ceaseless acoustic variation, listeners should show tremendous difficulty in just recognizing individual segments (and would then be confronted with the additional dizzying challenges of identifying words, parsing syntactic structure, accessing meaning, etc.). Holding any conversation should be a herculean feat, with both interlocutors struggling to understand each other and to formulate coherent responses.

The seeming impossibility of speech perception, however, does not align with many individuals’ everyday experience. Commonly used speech corpora contain hundreds of hours of fluent conversation between interlocutors (Godfrey et al., 1992; Pitt et al., 2005), implying the presence of mutual understanding and successful communication. Even under challenging conditions (e.g., the presence of background noise, listening to L2-accented speech, etc.), listeners generally show a rapid increase in word identification accuracy over time (Banks et al., 2015; Bent et al., 2009). More intuitively, if decoding acoustic variation was truly an insuperable task, it becomes unclear why anyone would actively choose to watch movies, talk with friends, or engage in any type of recreational activity involving speech perception.

In short, the lack of invariance problem must somehow be tractable. How do listeners accommodate and adapt to the speech signal in the face of so much variation?

2.1.2 Solution 1: Erasing variation?

One possible remedy for the lack of invariance problem is to just ignore all variation. For instance, under the assumption of a one-to-one mapping between abstract “phonemes” and neuromotor commands, the motor theory of speech perception claimed that listeners only

perceive the “intended” articulatory gestures of speech, with all acoustic variation being disregarded (Liberman & Mattingly, 1985). This idea would resolve the ambiguity exhibited in Figure 2.1 – if all possible acoustic variants of /s/ were associated with a single, unique neuromotor command (and likewise for /ʃ/), and if these commands were somehow recovered during speech perception, then listeners would have no trouble identifying segments as either /s/ or /ʃ/. There were many other theories of speech perception in the twentieth century besides motor theory, but in general, they all presumed that the listener’s goal was to filter out as much acoustic variation as possible (for a review, see Weatherholtz & Jaeger, 2016).

Initial findings, such as categorical perception, seemed to provide support for the erasure of variation (Liberman et al., 1957). A classic categorical perception experiment first creates a continuum between two consonant-vowel syllables differing in one segment (e.g., from “ba” to “ga”); participants then hear a single stimulus on each trial and complete a binary, forced-choice task (did you hear “ba” or “ga”?). Interestingly, in spite of the acoustic variation across the continuum, the response curves look highly categorical rather than gradient (e.g., the first few continuum steps are almost exclusively categorized as “ba”; this is followed by a precipitous plummet in “ba”-responses, with the last few continuum steps being almost exclusively categorized as “ga”). In general, categorical perception and other experimental evidence (e.g., Blumstein & Stevens, 1979) purportedly proved that listeners only perceive discrete “phonemes” (e.g., /b/ and /g/) that are abstracted away from the full variation of the speech signal.

The “erasure of variation” hypothesis, however, soon came under question for a wide range of reasons (see Section 1.2 of Liu, 2018). Essentially, the last few decades of research in speech perception have revealed extensive evidence that listeners actually do store acoustic variation (for detailed reviews, see Bent & Holt, 2017; Drager, 2010; Quam & Creel, 2021). Empirical work that ostensibly offered support for the erasure of variation was either not replicated or attributable to task-dependent effects (e.g., perception only *seems* categorical

when listeners complete a *binary*, forced-choice task; once the response variable becomes gradient, as in a visual analog scale task or an eye-tracking study, perceptual responses then appear more gradient; for a thorough critique of categorical perception, see McMurray, 2022).

If variation isn't erased, then how do listeners resolve the lack of invariance problem?

2.1.3 Solution 2: Leveraging the statistical structure of speech

The key insight is that variation is not randomly distributed – rather, speech contains *statistical structure* (Chodroff & Wilson, 2017; Hillenbrand et al., 1995; Newman et al., 2001). Such structure can be informative about the identity of speech segments, aiding listeners in overcoming the rampant ambiguity of the speech signal (Kleinschmidt & Jaeger, 2015).

Two important types of structure are visualized in Figure 2.2. The right panel displays *talker-specific covariation*, or a correlation between speech acoustics and talker identity (e.g., when comparing two female talkers (072 and 049), the former produces a higher center of gravity, on average). The left panel depicts *socio-indexical covariation*, or a correlation between speech acoustics and social category membership (e.g., female talkers possess a higher center of gravity than male talkers, on average; Jongman et al., 2000).

A comparison of the vertical black lines in Figure 2.1 and Figure 2.2 illuminates how structured variation might mitigate the lack of invariance problem. Recall that a black line was situated at 4920 Hz in Figure 2.1, the point of maximal ambiguity between the /s/ and /ʃ/ distributions. When talker-specific and socio-indexical covariation are taken into account (Figure 2.2), the same black line is no longer at a point of maximal ambiguity. A token with a center of gravity of 4920 Hz is much more likely to be /s/ rather than /ʃ/ for male speakers (blue line, left panel) and for Talker 049 (blue line, right panel); the reverse holds for female speakers overall and for Talker 072 as well. If listeners keep track of talker-specific and socio-indexical cue distributions, then the lack of invariance problem becomes tractable.

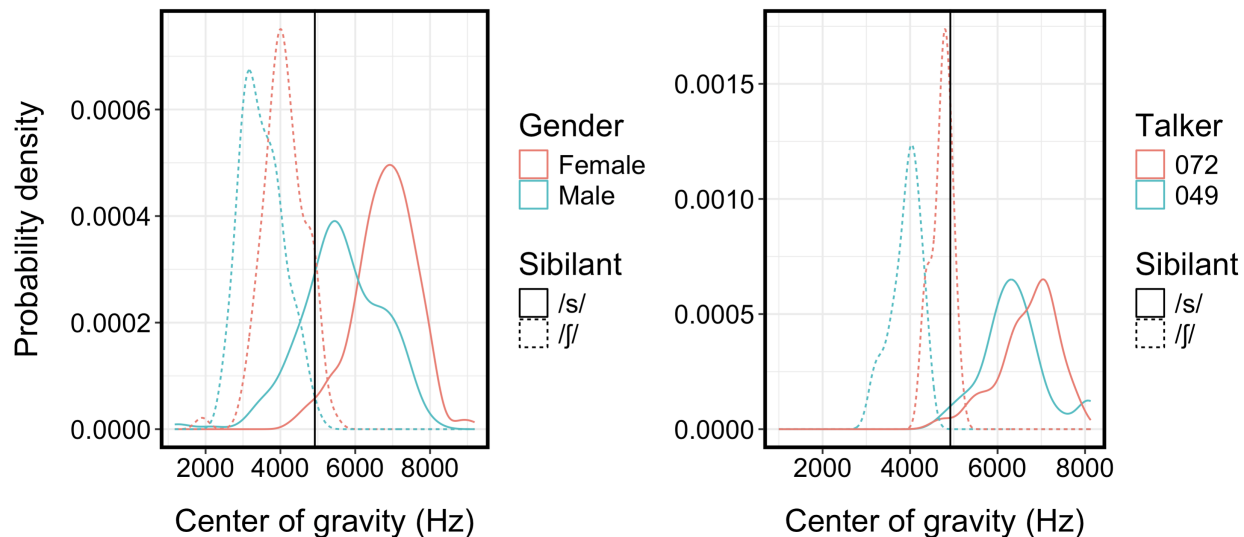


Figure 2.2: Distribution of /s/ and /ʃ/ center of gravity in Hertz (Hz) by gender (left panel) and by two talkers (right panel; both female). All distributions stem from the corpus visualized in Figure 2.1. The vertical line at 4920 Hz is also carried over from Figure 2.1.

Statistical structure is not just helpful in theory – ample empirical evidence demonstrates that listeners leverage both their knowledge of talker-specific and socio-indexical covariation to resolve ambiguity in speech perception.³ In the case of talker-specific covariation, work on lexically-guided perceptual learning provides a case in point (Samuel & Kraljic, 2009). Participants in this type of adaptation experiment complete an exposure phase (often a lexical decision task) followed by a test phase (usually a categorization task). When listeners are placed in a talker-specific condition (i.e., the same speaker is presented at both exposure and test), test phase categorization is altered based on the lexical bias in exposure. For example, consistently replacing /s/ with an ambiguous segment between [s] and [ʃ] in exposure (e.g., ‘hallu?inate’) results in more “asi” responses along an “aʃi”-“asi” continuum at test, while replacing /ʃ/ with the same ambiguous segment (e.g., ‘nego?iate’) leads to a greater “aʃi” response. This finding highlights two points: (i) during exposure, listeners mentally

³It should be emphasized that listeners can rely on many other types of information when categorizing ambiguous speech segments. Visual cues, for example, are critical in speech perception – when listeners hear a “ga” stimulus while looking at a closure of the lips (as in a typical “ba”), many often believe that they’ve heard “da” (i.e., in between “ba” and “ga”, suggesting that auditory and visual information are integrated; McGurk & MacDonald, 1976). This dissertation just concentrates on talker-specific and socio-indexical structure for the sake of brevity.

store idiosyncratic, talker-specific phonetic shifts (they do not just erase variation from memory); (ii) when they later encounter ambiguous speech segments (e.g., along an “afi”-“asi” continuum in a test phase), they transfer what they have learned and stored to this novel context, resolving the lack of invariance problem.

Research in sociolinguistic perception has shown similar findings with socio-indexical covariation. In a seminal experiment, Strand and Johnson (1996) asked listeners to complete an /f/-/s/ categorization task with voices that were gender-ambiguous. The voices were paired with pictures of either stereotypical female or male faces, and the results showed that perception was modulated by apparent gender – consistent with gender patterns in sibilant production (see the left panel of Figure 2.2), a greater /s/ response was provided when male faces were shown (i.e., ambiguous segments were more likely to be categorized as /s/ in the male condition). This suggests that, similar to talker-specific covariation, listeners retain socio-indexical structure in memory and apply their long-term knowledge to new speakers.

For greater clarity, the classic findings of talker-specific adaptation and sociolinguistic perception are visualized in Figure 2.3. The vertical black line in Figure 2.3 is the point of maximal ambiguity along a hypothetical continuum between /f/ and /s/, paralleling the vertical black lines in Figure 2.1 and Figure 2.2. As noted above, a token at this ambiguous point in Figure 2.2 is much more likely to be produced as /s/ by male speakers and Talker 049 than by female speakers and Talker 072; perceptual categorization should match speech production, with listeners showing a greater /s/ response for male speakers and Talker 049.

The introduction has focused heavily on the production/perception of /f/ and /s/ in English as an example, but talker-specific adaptation is a very well-established phenomenon that has been observed across many different contrasts, speakers, and languages (Cummings & Theodore, 2023; Niedzielski, 1999; Norris et al., 2003; Zellou et al., 2023, etc.; for a review, see Samuel & Kraljic, 2009). Effects of social information on speech perception are similarly well-replicated; perceptual categorization has been demonstrated to vary with a

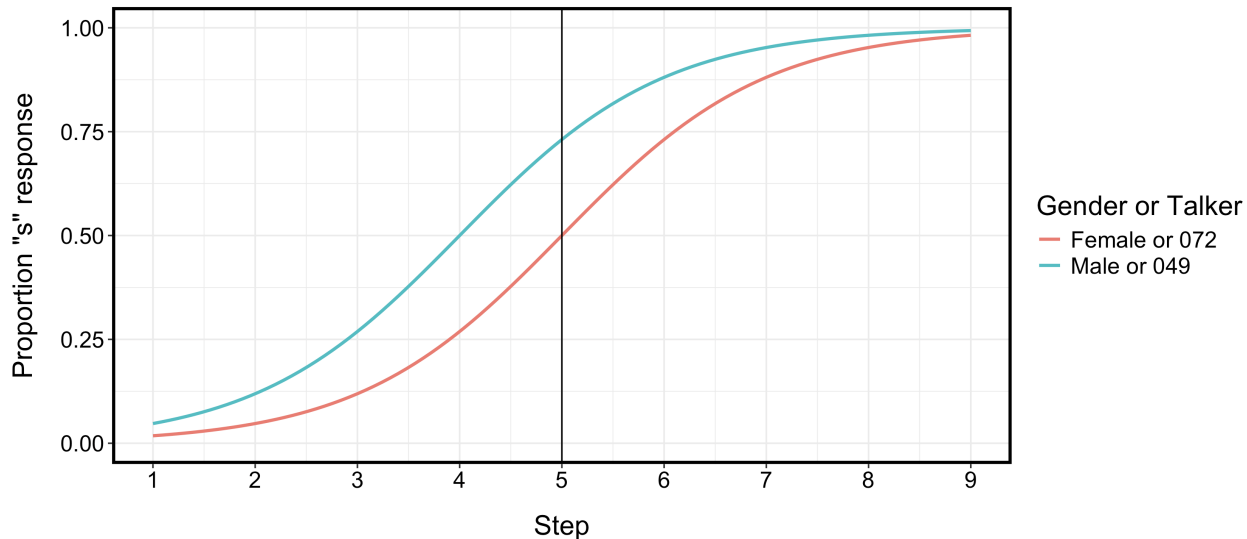


Figure 2.3: Perceptual categorization of an / ʃ /-/s/ continuum in a hypothetical experiment, where Step 1 is the most / ʃ /-like and Step 9 is the most /s/-like. In tandem with Figure 2.2: (i) a vertical black line is placed at the most ambiguous point along the continuum; (ii) ambiguous tokens are more likely to be categorized as /s/ for male speakers and Talker 049 (blue) than for female speakers and Talker 072 (red). Note that these are idealized results and not based on any specific computational model.

talker’s apparent age (Koops et al., 2008), dialect (Niedzielski, 1999), nationality (Hay et al., 2006), and persona (D’Onofrio, 2015), among other social characteristics (for a review, see Campbell-Kibler, 2010).

Before 2015, talker-specific learning and sociolinguistic perception were largely discussed as separate phenomena, even though they seem to be related (both broadly involve listeners storing the statistical structure of language and then applying that knowledge to navigate the variability of the speech signal). What was missing was a clear framework that united the diverse findings of the adaptation literature under one theoretical umbrella.

2.1.4 The ideal adapter framework

The ideal adapter framework is a landmark theory in cognitive science, which frames speech perception as “inference under uncertainty” (Kleinschmidt & Jaeger, 2015, p. 150). Speech is inherently uncertain – the same acoustic cues can refer to different categories, and different

acoustic cues can refer to the same category (i.e., the lack of invariance problem; see Section 2.1.1 for review). Through their knowledge of the statistical structure of speech, an “ideal adapter” develops *generative models* that help make the most accurate inferences or predictions possible about the incoming signal. Importantly, these predictions can be updated through experience, allowing the listener to adjust to novel input in spite of high acoustic variability.⁴

The ideal adapter framework formalizes the ideas in the previous paragraph through Bayes Rule (schematized in Equation 2.1), which states that the *posterior* is proportional to the product of the *prior* and the *likelihood*.⁵ Leveraging the power of Bayes Rule, Kleinschmidt and Jaeger (2015, p. 150) conceive of speech perception as “selecting and adapting the appropriate generative model”, given a combination of both “the structure of listeners’ prior knowledge” and “their ability to learn the statistics of novel situations”. In short, to successfully navigate acoustic variability (i.e., achieve posteriors that accurately reflect the speech signal), listeners must both integrate their prior experiences (the prior) and shift their expectations given new input (an update of the likelihood).

$$\text{Posterior} \propto \text{Prior} * \text{Likelihood} \tag{2.1}$$

As a concrete example, imagine a talker who repeatedly produces /d/ as an ambiguous segment between [d] and [t] (e.g., ‘croco?ile’, ‘legen?ary’, etc.). According to the ideal adapter framework, a listener who encounters this individual will develop a talker-specific mental model, since the talker’s /d/-production strays from their prior expectations. The model would predict that segments with ambiguous acoustic cues (e.g., Voice Onset Time between prototypical /d/ and /t/) are more likely to be classified as /d/ for this talker compared to the general population. Listeners can then store and utilize this talker-specific model to help

⁴Although there are other approaches that could potentially account for perceptual adaptation (e.g., Bayesian exemplar theories; Shi et al. 2010), this dissertation focuses on the ideal adapter framework because it is the one of the most widely cited in the literature.

⁵Note that Equation 2.1 is a mathematically simplified version of Bayes Rule, since this dissertation is not focused on computational modeling.

them make predictions in novel contexts (e.g., an ambiguous token between “ada” and “ata” is more likely to be “ada” when produced by this particular talker; Kraljic & Samuel, 2007).

This exact idea can also account for sociolinguistic perception. For instance, California English talkers tend to produce a backed /æ/ (e.g., “sack” sounds like “sock”; Eckert, 2008). Given sufficient exposure, listeners presumably construct a dialect-specific (talker-general) mental model, predicting that an ambiguous word between “sack” and “sock” is more likely to be “sack” when produced by a California English speaker (D’Onofrio, 2015).

All of the examples in the introduction have only considered individual phonetic shifts, but the ideal adapter framework can theoretically scale up to more complex, naturalistic contexts, such as L2-accent adaptation. L2-accent consists of a constellation of acoustic features that differ from L1 norms – the ideal adapter can create an accent-specific (talker-general) mental model with predictions for all of these features at once (e.g., a model for Spanish-accented English might simultaneously include an epenthetic vowel before word-initial /s/, a mapping from /p/ to an acoustic [b], etc.; Magen, 1998).

The contexts in this section can be grouped into one of two categories: *talker-specific adaptation* and *cross-talker generalization*. Talker-specific adaptation is when a listener is exposed to a talker and then adjusts their expectations for that same talker (Kraljic & Samuel, 2007). Meanwhile, cross-talker generalization is when a listener applies a pre-existing generative model to a *novel* talker. Sociolinguistic perception can be considered a kind of long-term generalization – even before the undergraduate listeners in Strand and Johnson (1996) completed the study, they already possessed extensive long-term experience with gendered variation of /f/ and /s/; they then applied their well-developed generative models of female and male speech to the novel talkers in the experiment, shifting their categorization of /f/ and /s/ by apparent speaker gender.

Sociolinguistic perception and generalization have largely been taken for granted so far, but an understudied question is how they emerge in the first place. As a thought experiment,

imagine an alien who resides in a parallel universe that is identical to Earth in every possible way, except that there is no gender differentiation between /j/ and /s/ in English. Instead of teleporting to their home planet, the alien mistakenly appears on Earth, and they bump into a human woman and a human man; the woman has the same center of gravity as everyone else on the alien’s planet, while the man has a lower-than-expected center of gravity. Would the alien mentally store this difference as talker-specific covariation (i.e., this one man is the “odd-one-out”), or would the alien assume that this difference is gendered and socially meaningful? Only the latter would predict generalization to occur, with the alien applying what they have learned in exposure to novel talkers (i.e., that all male speakers on Earth possess a lower center of gravity than all female speakers).

In general, although it is well-established that listeners can generalize based on their long-term experience, what is unclear is how/whether listeners take their recent experiences and generalize in the short term.

2.2 The problem: To generalize or not to generalize?

This dissertation is centered on cross-talker generalization. When a listener is exposed to an accent or variant, when do they apply what they have learned to a novel talker?

Generalization is interesting theoretically because it involves an inherent trade-off between efficiency and accuracy. On the one hand, under the assumption that developing a generative model is cognitively taxing (Xie et al., 2023), listeners have an incentive to generalize as often as they possibly can (repurposing the same mental model for many different talkers eliminates the need to expend any extra effort in creating a new model). Overgeneralizing, however, can lead to incorrect predictions. As an extreme example, it would be quite odd for listeners to develop a mental model for Spanish-accented English and to indiscriminately transfer the model’s predictions to all novel talkers, including L1-English speakers.

Navigating this tradeoff is made difficult by the uncertainty of the speech signal (Kleinschmidt & Jaeger, 2015). Talker-specific and socio-indexical covariation coexist, which means that listeners must *infer* whether variation is applicable to novel talkers. If an exposure talker produces a phonetic shift that seems idiosyncratic, then it makes sense to be conservative and not generalize; if a novel talker is likely to produce the shift as well, then generalization is the rational choice. But how do listeners determine whether or not to generalize?

This dissertation reconsiders a debate between two accounts (*numerosity* and *sufficient similarity*), which differ in the types of evidence listeners use to generalize – whereas numerosity stresses the importance of talker variability in exposure, sufficient similarity focuses on the degree of acoustic similarity between the exposure and novel talkers. The novel theoretical contribution of this work is to introduce the *linguistic experience hypothesis*, which relies on a sociolinguistic perspective to reconcile the two accounts.

2.3 Numerosity or sufficient similarity?

After discussing numerosity and sufficient similarity (Sections 2.3.1-2.3.2), the linguistic experience hypothesis is defined (Section 2.3.3) as a way of reinterpreting prior work (Section 2.3.4). Section 2.3.5 highlights the need for both novel test cases and a novel sociolinguistic approach in the adaptation literature.

2.3.1 The case for numerosity

Bradlow and Bent (2008) came out with a bold claim – multiple-talker exposure is necessary for cross-talker generalization. The idea that talker variability is helpful (or even potentially necessary) for generalization will be termed a *numerosity account*.⁶

⁶“Numerosity” is being framed here in a moderate way (talker variability is helpful *or* necessary for generalization). Bradlow and Bent (2008) will be thought of as an extreme version of a numerosity account.

“[N]ative American English listeners recruited from the Northwestern community” completed a speech-transcription-in-noise task for a single Mandarin-accented English speaker in a test phase (Bradlow & Bent, 2008, p. 719). They were placed in one of five training conditions that varied the type of exposure they received prior to test:

1. Untrained (No exposure)
2. Task control (Five L1-English speakers)
3. Single-talker (One L1-Mandarin speaker that differed from the test talker)
4. Talker-specific (The same talker presented at test)
5. Multiple-talker (Five L1-Mandarin speakers, all different from the test talker)

Generalization was defined as greater test phase accuracy in the critical conditions (single-talker, multiple-talker) relative to control (untrained and task control). In theory, if listeners develop a mental model of Mandarin-accented English in exposure and apply that model to a novel talker, comprehension should increase compared to listeners with either no exposure or no chance to develop an appropriate mental model. More concretely – L1-Mandarin speakers often devoice word-final stop consonants, such that “seed” sounds like “seat”; if a novel talker says “overloat”, generalization will help listeners more quickly recognize that the talker intended “overload”, thus raising comprehension and accuracy).⁷

Bradlow and Bent (2008) found that test phase accuracy was equally highest in the multiple-talker and talker-specific conditions, equally intermediate in the single-talker and task control condition, and lowest in the untrained condition. Among the two critical conditions (single-talker and multiple-talker), only the multiple-talker condition increased accuracy compared to the control conditions, suggesting that multiple-talker exposure is necessary for cross-talker generalization.

⁷Transcription tasks have been critiqued because they are not necessarily transparent as to what listeners are learning (Baese-Berk et al., 2023). However, since so much of the literature has relied on this paradigm (Baese-Berk et al., 2013; Bradlow & Bent, 2008; McLaughlin & Van Engen, 2023; etc.), this dissertation assumes for the sake of simplicity that boosts in accuracy truly reflect adaptation.

Kleinschmidt and Jaeger (2015) attributed this effect to the distinct mental models that listeners develop following single- and multi-talker exposure. On the one hand, participants who are exposed to just one talker might assume that the particularities of the accent are idiosyncrasies, thereby resulting in an ungeneralizable, talker-specific model. However, listeners who are exposed to multiple talkers with the same L2-accent can more readily recognize the covariation in the speech signal between category membership (Mandarin-accented English speaker) and acoustic variation (e.g., devoicing of word-final stop consonants). This recognition ostensibly leads to the development of a talker-general mental model for Mandarin-accented English, which then gives rise to successful generalization (i.e., enhanced comprehension of a novel talker with the same, Mandarin-accented English accent).

Bradlow and Bent (2008) became an extremely influential paper (1174 citations as of May 2026).⁸ There are at least two possible reasons for why this is so: (i) the benefit of talker variability aligned neatly with previous work (e.g., Bradlow et al., 1997; Clopper & Pisoni, 2004; Wang et al., 1999); (ii) generalization of L2-accent is not only of theoretical interest to cognitive scientists and linguists, but also of practical interest for improving intelligibility in a diverse world (Derwing & Munro, 2014). The central finding of Bradlow and Bent (2008) was not substantively questioned for the next decade, and it seemed clear that multiple-talker exposure was necessary for cross-talker generalization.

2.3.2 The case for sufficient similarity

Contrary to Bradlow and Bent (2008), Xie and Myers (2017) observed no effect of the number of exposure talkers on generalization. They instead found evidence for a *sufficient similarity* account, which claims that generalization is facilitated when the exposure and novel talkers are acoustically similar.

⁸<https://scholar.google.co.uk/scholar?cites=8740483286481257637>

Participants in Xie and Myers (2017) consisted of monolingual L1-English speakers at the University of Connecticut with “no or minimal prior experience with Mandarin-accented English or the Mandarin language” (p. 33). All listeners completed an exposure phase (auditory lexical decision) followed by a test phase (cross-modal priming); everyone was tested on the same (novel) Mandarin-accented English speaker and exposed to Mandarin-accented English. There were three types of critical conditions: (i) exposure to multiple talkers; (ii) exposure to a single talker who was acoustically similar to the novel test talker; (iii) exposure to a single talker who was acoustically different from the novel test talker.

Three major findings emerged. First, both the multiple-talker and the “single-talker, acoustically similar” conditions resulted in generalization, with no meaningful difference observed between the two conditions. Second, generalization was blocked in the “single-talker, acoustically different” condition. Third, listeners’ subjective judgments about talker and accent similarity across exposure and test had no effect on the results. Taken together, Xie and Myers (2017) concluded that the bottom-up acoustics are what influence adaptation, not simply the number of exposure talkers or top-down expectations. Multiple-talker exposure only incidentally leads to generalization when at least one of the exposure talkers happens to be acoustically similar to the novel speaker.⁹

As further evidence against the primacy of talker variability, Xie et al. (2021) conducted a replication study of Bradlow and Bent (2008) and did not find a meaningful difference in transcription accuracy between the single- and multiple-talker exposure conditions.

Xie et al. (2021) pointed to a variety of methodological issues in Bradlow and Bent (2008) that may have compromised the results. Not only was statistical power low (e.g., the multiple-

⁹In contrast to the summary of Bradlow and Bent (2008) in Section 2.3.1, a specific operationalization of generalization is not provided here, mainly because defining “generalization” is much more complicated in the context of a cross-modal priming task (Xie & Myers, 2017), as opposed to a transcription task (Bradlow & Bent, 2008). Essentially, Xie and Myers (2017) exposed listeners to devoiced word-final stop consonants (e.g., “seed” as “seat”) and found that priming effects were only shifted when listeners were exposed to multiple talkers or to a single talker who was similar in terms of vowel, burst, and closure duration (i.e., the acoustic cues differentiating word-final voiced and voiceless stops). For details, please refer to Xie and Myers (2017).

talker condition only had ten listeners), but there was a lack of talker counterbalancing, and the single- and multiple-talker conditions employed different exposure talkers. The last point is particularly troublesome for a numerosity account – the absence of generalization in the single-talker condition of Bradlow and Bent (2008) may have merely occurred because the authors coincidentally selected exposure talkers that were all acoustically different from the test talker (and conversely, generalization in the multiple-talker exposure condition might have been blocked if none of the exposure talkers were acoustically similar to the test talker).

Coupled with similar findings in lexically-guided perceptual learning (e.g., Kraljic & Samuel, 2005; Reinisch & Holt, 2014), it seems that acoustic similarity is what mediates generalization, not the number of exposure talkers. But has sufficient similarity truly won out over numerosity, or is there a possible way of harmonizing the two accounts?

2.3.3 Reconciling the accounts: A linguistic experience hypothesis

The seed for the linguistic experience hypothesis is that sufficient similarity and numerosity are not quite as different as they seem. In the initial study by Bradlow and Bent (2008), there were two multiple-talker exposure conditions (five Mandarin-accented English speakers, five L1-English speakers), but only the former group generalized to a novel Mandarin-accented English speaker (i.e., when there was a match in language background) – in other words, the benefit of numerosity was only observed when the exposure and test talkers were acoustically similar. This suggests that, although sufficient similarity and numerosity are sometimes framed as competing accounts (Xie et al., 2021; Xie & Myers, 2017), they only differ in the amount of evidence required to generalize. Whereas sufficient similarity relies on just one type (acoustic similarity), numerosity is based on two types (acoustic similarity *and* the number of exposure talkers).

What factors might mediate the amount of evidence required for generalization? Thinking back to sociolinguistic perception (see Sections 2.1.3-2.1.4), listeners required no evidence for

generalization, ostensibly because of their extensive *experience* with sociolinguistic variation. For example, the L1-English listeners in Strand and Johnson (1996) did not need to learn in an exposure phase that female speakers often produce a higher center of gravity for /s/ and /ʃ/ relative to male speakers – they already possessed knowledge of such gender-mediated covariation, and they generalized their long-term experience to novel talkers in an experiment, adjusting their categorization of /s/ and /ʃ/ by perceived gender.

If greater experience reduces the amount of evidence needed for generalization, then conversely, less experience should increase the amount of evidence needed for generalization. Less experience should make listeners more uncertain about whether an accent or phonetic variant is talker-specific or applicable to novel talkers; this uncertainty should then lead listeners to respond more conservatively, demanding more evidence for generalization.

Taken together, all of the ideas in this section coalesce into a *linguistic experience hypothesis*, which is formally defined below (along with several corollaries that unpack the particular wording of the hypothesis). This sets the stage for the next subsection (2.3.4), which reinterprets Xie and Myers (2017) and also brings back the story from the Prologue.

The Linguistic Experience Hypothesis: The less experience a listener has with variation, the more evidence will likely be helpful/needed for generalization, above and beyond acoustic similarity.

1. The linguistic experience hypothesis suggests that sufficient similarity and numerosity can be placed within the same theoretical framework. When listeners have less experience with linguistic variation, there should be greater support for numerosity (i.e., the number of exposure talkers should become helpful or even required for generalization).
2. The first clause (“the less experience a listener has with variation”) explicitly situates adaptation within sociolinguistic context by making reference to the listener’s experi-

ence. This is notable, as many prominent theories of adaptation focus heavily on the computational modeling of acoustic cues (i.e., the influence of the *speaker* rather than the listener; e.g., Feldman et al., 2009; McMurray & Jongman, 2011; Xie et al., 2023).

3. The second clause (“the more evidence will likely be helpful/needed for generalization”) makes a few important points. The first is that “evidence” is kept intentionally vague. Acoustic similarity and the number of exposure talkers are the two pieces of evidence that this dissertation happens to be most interested in, but other cues could be relevant (e.g., visual similarity; if an exposure and novel talker are acoustically similar *and* are wearing the same types of clothing, generalization might be more robust).

In alignment with Bayesian models of cognition (Kleinschmidt & Jaeger, 2015), the word “likely” frames the linguistic experience hypothesis as inherently probabilistic, not as a rigid rule or law. The listener’s experience is one factor among potentially many that might affect adaptation, and there could easily be individual differences between listeners in how strongly their experiences mediate generalization. However, if a study has a sufficiently large sample size, and all of the listeners have a similar demographic background (and thus are more likely to share similar experiences with linguistic variation), then the predictions made by the linguistic experience hypothesis should tend to hold.

Similar to the term “likely”, the inclusion of “helpful/needed” (as opposed to just “needed”), avoids framing generalization as a binary, “all-or-nothing” phenomenon where a piece of evidence is either always necessary or never necessary.

It should be clarified that “more evidence” means both a greater number of tokens *and* more types of evidence. For example, if a listener is loathe to generalize an L2-accent, then they may need a greater number of tokens (exposure to many sentences) *and* more types of evidence (e.g., the sentences should be acoustically similar to the novel talker *and* they should be divided evenly among multiple exposure talkers).

4. The final part (“above and beyond acoustic similarity”) implies a hierarchy of “types of evidence” where acoustic similarity is ranked highest – no matter how many exposure talkers there are, listeners are very unlikely to generalize if none of them are acoustically similar to the novel talker. The linguistic experience hypothesis therefore complements, rather than contradicts, prior work on acoustic similarity in adaptation (Kraljic & Samuel, 2005; Reinisch & Holt, 2014; Xie et al., 2021; Xie & Myers, 2017).

2.3.4 Reinterpreting prior work through the linguistic experience hypothesis

As a proof of concept, this section reinterprets Xie and Myers (2017) by unpacking its social context and explaining why this particular social context may have incidentally favored sufficient similarity over numerosity. This is followed by a counterpoint (the example of adaptation from the Prologue), which illustrates a scenario that would ostensibly support numerosity over sufficient similarity. There are two underlying goals here: (i) to reframe the relationship between sufficient similarity and numerosity (from being competing accounts to being understood within the same theoretical framework of the linguistic experience hypothesis); (ii) to broadly illustrate the potential value of integrating an explicit sociolinguistic perspective into theories of perceptual adaptation and cognitive science.

As discussed in Section 2.3.2, Xie and Myers (2017) investigated generalization of word-final devoicing in Mandarin-accented English (e.g., “seed” as “seat”). All listeners were “monolingual English speakers” from the University of Connecticut who self-reported “no or minimal prior experience with Mandarin-accented English or the Mandarin language” (p. 33). The speakers were “six male native-Mandarin speakers”, all of whom were “undergraduates or graduate students from the University of Connecticut” and “were L2 learners of English [who] acquired English in mainland China” (p. 33). Their age-of-acquisition was 7-12 years old, their age of arrival in the United States was 18-22 years old, and their length of residence

in the United States was 6 months to 5 years (Xie & Myers, 2017, Table A.1). There is an implication that the speakers have a strong L2-accent – stimulus transcription accuracy was around 50% on average (Xie & Myers, 2017, Table A.1).

Unlike the listeners in Strand and Johnson (1996), who presumably possessed extensive experience with gendered variation of /s/ and /ʃ/ in English, the listeners in Xie and Myers (2017) had comparatively less experience with the test case at hand (this implies that some degree of exposure is necessary for generalization, contra Strand and Johnson (1996)). However, self-reports of “no or minimal prior experience with Mandarin-accented English or the Mandarin language” do not necessarily imply a lack of experience altogether. There are two reasons why the social context in Xie and Myers (2017) is one of *intermediate* familiarity.

The first reason involves word-final devoicing. There are various articulatory explanations for why utterance/word-final devoicing is less challenging than voicing in the same position (e.g., “seat” with a word-final voiceless [t] is easier to produce than “seed” with a word-final voiced [d]; for details, see Jatteau et al., 2025). Theories of phonetically-driven phonology expect less articulatorily complicated structures to be more typologically common (Hayes, 1999), and this prediction is often borne out – for example, word-final devoicing is a very widespread phonological process across unrelated languages (Dutch: Eisner et al., 2013; Turkish: Sanyar & Tunçer, 2022; Mandarin: Xie & Fowler, 2013) and across various English dialects (Chicano English: Bayley & Holland, 2014; Midwest English: Purnell et al., 2005; British English: Verhoeven et al., 2011). Moreover, given that casual speaking styles are characterized by lower articulatory effort (Smiljanić & Bradlow, 2009), L1-English speakers should regularly encounter at least some kind of partial word-final devoicing in more relaxed conversational settings (as opposed to a scenario eliciting more effortful speech, e.g., when communicating with someone who is hard-of-hearing; Picheny et al., 1985). Based on these frequency effects, listeners likely possess implicit knowledge from their everyday experience that word-final devoicing is highly generalizable across contexts.

As a second point, a lack of exposure to Mandarin-accented English does not entail a lack of exposure to L2-accented English. In general, regardless of language background, L2-English speakers are acoustically similar through accent-independent properties (e.g., slower speaking rate, Baese-Berk & Morrill, 2015; reduced usage of spectral cues when producing tense/lax vowel contrasts, Feng & Wang, 2024; etc.). L1-English listeners are often very effective at identifying whether a speaker has an L1- or an L2-accent (Atagi & Bent, 2017; Flege, 1984), which suggests an awareness of accent-independent features (and thus implicit knowledge that the acoustics of L2-accented speech are often generalizable). Such awareness is likely increasing, given that the number of L2-English speakers is rising in the United States (Graddol, 2003; ShareAmerica, 2023), and that public institutions like the University of Connecticut are becoming more diverse over time (University of Connecticut, n.d.).

According to the linguistic experience hypothesis, the bar for generalization should be relatively low in Xie and Myers (2017) – given their intermediate familiarity with the test case in question, fairly little evidence should be necessary for generalization. Since the speakers had strong accents, all of the listeners presumably recognized that they had some L2-accent; based on their previous experience with L2-accented speech and word-final devoicing, participants only needed to rely on acoustic similarity to generalize, supporting sufficient similarity over numerosity (multiple-talker exposure is unnecessary because listeners can already deduce that word-final devoicing in a single L2-accented talker is likely not an idiosyncrasy). Overall, even when listeners in an adaptation study are unfamiliar with the specific accent in question (in this case, Mandarin-accented English), the experiment is still embedded within a rich sociolinguistic context that potentially informs perceptual behavior.

The context of Xie and Myers (2017) marks a stark contrast to the Prologue. The Prologue described an L2-English (L1-Farsi) speaker who has taken pains to eliminate their accent due to societal pressure, but still produces certain phonetic shifts, particularly in their casual speech (e.g., /w/ to [v], “welcome” as *velcome*; /ɑ/ to [ʊ], “pots and pans” as “*peuts* [pʊts]

and pans”). This results in a highly ambiguous situation that the listener in the Prologue had little to no experience with – communicating with an individual who is perfectly intelligible and essentially cosplays as an L1-English speaker, yet produces isolated shifts that would otherwise be characteristic of an L2 speaker. In accordance with the linguistic experience hypothesis, the listener had to rely on more evidence beyond acoustic similarity to determine whether the shifts were idiosyncratic or generalizable (i.e., by socializing with his L1-Farsi in-laws, who had much more typical L2-accents). This piece of anecdotal evidence now seems to support numerosity over sufficient similarity.

The takeaway from this section is that, through the lens of the linguistic experience hypothesis, numerosity and sufficient similarity are not actually competing accounts – rather, both can be sheltered under the same theoretical umbrella. This idea is illustrated in the flowchart of Figure 2.4, which shows how the listener’s prior experiences might mediate the helpfulness (or necessity) of multiple-talker exposure on generalization.

The broader contribution of the linguistic experience hypothesis is that, rather than appealing solely or primarily to acoustic/methodological explanations to reconcile competing findings (e.g., Xie et al., 2021), sociolinguistic considerations are now placed at the front and center. For example, in Figure 2.4, the flowchart begins with the social context (“a diverse American university”), and then directly weaves that context into making predictions. Overall, the claim of this work is not that acoustics and methodology are not important, but that the listener’s perspective and sociolinguistic theory are equally critical.

So far, the introduction has only listed two concrete examples where the number of exposure talkers was helpful or necessary for generalization: (i) Bradlow and Bent (2008), which was shown to have a variety of methodological issues, see Section 2.3.2; (ii) the anecdote from the Prologue, which does not constitute strong proof. If a numerosity account is truly viable, then why is there little published work that validates such an account empirically?

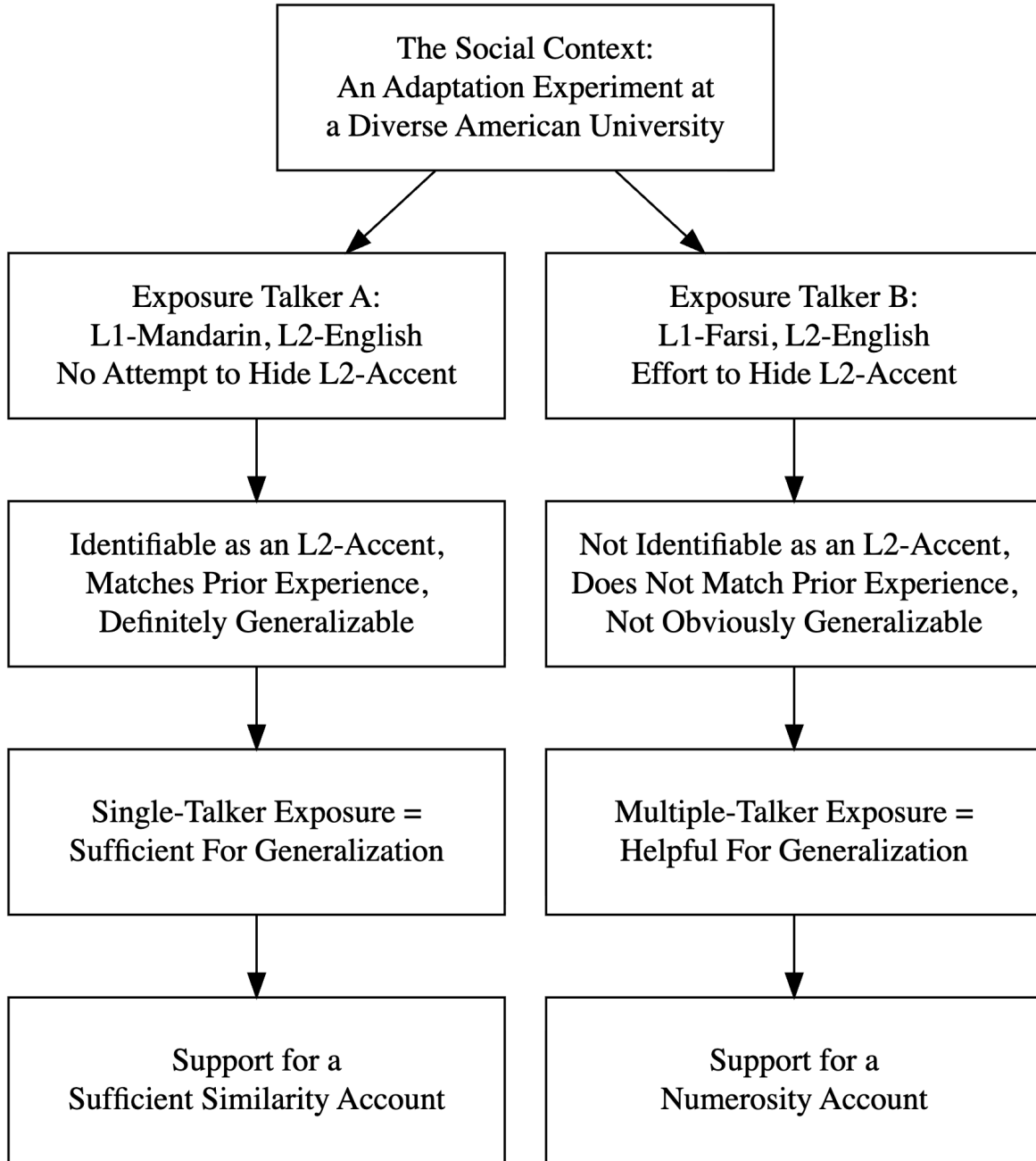


Figure 2.4: A flowchart illustrating how the linguistic experience hypothesis unites the sufficient similarity and numerosity accounts under one theoretical framework. When listeners possess ample experience with variation, then single-talker exposure is sufficient for generalization, lending credence to a sufficient similarity account (left side; exposure to an L1-Mandarin talker with a canonical L2-English accent). However, when listeners are unfamiliar with variation, then multiple-talker exposure may become helpful (or even necessary) for generalization, supporting a numerosity account (right side; exposure to the L1-Farsi speaker described in the Prologue; i.e., an individual who *seems* to have an L1-English accent overall, but produces isolated phonetic shifts that would be more characteristic of an L2-speaker).

2.3.5 The need for novel test cases and a novel sociolinguistic approach in the adaptation literature

The absence of strong evidence for a numerosity account could have at least two sources: (i) the full range of possible test cases has not yet been explored; (ii) the theoretical approach is not sufficiently broad to find appropriate test cases.

In terms of the first point, many experiments on perceptual adaptation fall into one of a few set paradigms. This is exemplified by a comparison between Samuel and Kraljic (2009) and Xie et al. (2023) – although both are published over ten years apart, their literature reviews essentially present the same two-part structure, where one section focuses entirely or predominantly on L2-accented speech and another section concentrates on perceptual recalibration (i.e., exposure to ambiguous segments between /t/ and /d/, or some other contrast). This is not to claim that there is no variation in the literature (cf. Charoy & Samuel, 2023), but by and large, many studies tend to investigate similar test cases and recruit listeners from similar demographic backgrounds (e.g., undergraduates at institutions in North America or Western Europe, L1 American English speakers from crowdsourcing platforms; see the citations within Samuel and Kraljic (2009) and Xie et al. (2023)).

There can be no question that the field has made tremendous progress overall. However, if the test cases are not varied enough, then all studies may incidentally lead to the same conclusions, leaving adaptation undertheorized (for a comparable type of argument, see Blasi et al. (2022) and Henrich et al. (2010)). For example, Kraljic and Samuel (2005), Reinisch and Holt (2014), and Xie and Myers (2017) specifically present either L2-accented speech or ambiguous segments to L1-English undergraduates at large American universities with diverse student bodies (Stony Brook University, Carnegie Mellon and University of Pittsburgh, and University of Connecticut, respectively). All three found that single-talker exposure facilitated generalization, which is expected by the linguistic experience hypothesis laid out in Sections 2.3.3-2.3.4 – since the listeners likely have extensive experience with

linguistic variation, they are more likely to generalize with just single-talker exposure. But let's say that listeners were presented with a completely different type of test case that they have no familiarity with and that has never been examined before in the literature – would multiple-talker exposure become necessary for generalization, supporting a more extreme version of a numerosity account? In general, it remains unclear whether the conclusions of current adaptation experiments reflect deep principles of language and cognition or are at least partially artefacts of the particular test cases that are being employed.

A second issue may lie with decontextualization in theoretical accounts of adaptation. As discussed in Section 2.1.4, Kleinschmidt and Jaeger (2015, p. 150) rely on Bayes Rule to postulate two important ingredients for successful adaptation: “the structure of listeners’ prior knowledge” and the “ability to learn the statistics of novel situations” (an updating mechanism). The prior might technically be part of the theory, but by focusing primarily on the quantitative modeling of acoustic cue distributions and the updating mechanism (e.g., Feldman et al., 2009; McMurray & Jongman, 2011; Xie et al., 2023; etc.), the literature does not often focus on the listener’s prior sociolinguistic experience in practice. Since social context is not considered often enough in perceptual adaptation studies (cf. Section 2.3.2 and Sections 2.3.3-2.3.4), listener familiarity is not often manipulated as a variable, which means that the full range of possible test cases are not being implemented (including types of variation that might find stronger support for a numerosity account).

Note that this section does not contradict prior work, nor does it devalue the test cases and approaches of the literature. The claim is that, in addition to commonly used test cases and to acoustic modeling, it could be important to look at novel test cases and integrate a sociolinguistic point of view. Harmonizing what we already have with new ideas would achieve both a better understanding of cross-talker generalization in a narrow sense, but also deepen theories of cognition and language in a broader sense.

2.4 Overview of the dissertation

Before discussing the goals and layout of this dissertation, it seems helpful to summarize the introduction, so that the most important points are concisely written in one place.

1. The Theoretical Question: This dissertation is concerned with *cross-talker generalization*. After listeners are exposed to a phonetic shift or accent, when do they apply (or generalize) what they have learned to a novel talker?
2. The Theoretical Debate: Two prominent accounts of generalization are *numerosity* and *sufficient similarity*. A numerosity account emphasizes the importance of talker variability, with multiple-talker exposure being helpful (or possibly even necessary) for generalization, as long as the exposure talkers are acoustically similar to the novel talker (Bradlow & Bent, 2008). Meanwhile, a sufficient similarity account proposes that it is just acoustic similarity that predicts generalization, not the number of exposure talkers (Xie et al., 2021; Xie & Myers, 2017).
3. The Potential Resolution: This dissertation proposes a novel theory to unite numerosity and sufficient similarity – the *linguistic experience hypothesis*. The idea is that, the less experience a listener has with variation, the more evidence will be helpful/needed for proving that the variant is truly generalizable and not just an idiosyncrasy. Resolving the debate between numerosity and sufficient similarity, the linguistic experience hypothesis predicts that listeners should: (i) rely on less evidence in situations of relative familiarity (e.g., just acoustic similarity); (ii) count on more types of evidence in unfamiliar contexts (e.g., acoustic similarity *and* the number of exposure talkers). Numerosity and sufficiently similarity are thus mediated by the listeners’ sociolinguistic experiences, and both are built into the same underlying theory.
4. The Broader Implications: Adopting a distinct sociolinguistic and listener-focused perspective departs from the current adaptation literature, which has a tendency to

focus on the quantitative modeling of acoustic cues (i.e., a speaker-focused point of view). This dissertation claims that *both* approaches are equally valuable and must be more closely intertwined in theories of language and cognition.

2.4.1 Goals of the dissertation

This dissertation has two goals:

1. The first aim is to provide initial empirical support for the linguistic experience hypothesis, investigating how social context mediates effects of acoustic similarity and talker variability on cross-talker generalization. Although the amount of evidence needed for generalization should vary gradiently with experience, this dissertation makes a binary comparison between two test cases as a first pass – one where listeners have relative familiarity and another where listeners have a complete lack of familiarity.

Since the intention is to lay the foundation for future work (see the next goal), it is not necessary to report many different studies (in fact, too many experiments would likely decrease concision). Instead, two experiments are reported and described here in detail.

2. The second aim is to highlight the value of integrating a sociolinguistic perspective into cognitive science, and to showcase the ways that such an approach can both generate novel predictions and yield new theoretical and practical insights.

2.4.2 Layout of the dissertation

This dissertation is structured as follows. Chapters 3 and 4 present Experiments 1 and 2, respectively. The general discussion in Chapter 5 concludes by delving into the implications of this work for cognitive science and sociolinguistics.

Experiment 1 revisits a very well-studied test case that listeners are relatively familiar with. More specifically, L1-English listeners in the United States are first exposed to Mandarin-accented English and then given a comprehension test on a novel Mandarin-accented English talker. Experiment 1 generally replicates prior work – although multiple-talker exposure affords a small perceptual benefit, single-talker exposure is sufficient for generalization, consistent with a sufficient similarity account (Xie et al., 2021; Xie & Myers, 2017).

Experiment 2 presents listeners with a synthetic, gendered phonetic shift that they have never encountered before. Female and male text-to-speech voices are heard in exposure, where the female talkers produce /p/ as [b] (e.g., “paint” as *baint*) and the male talkers are unshifted; a novel female and a novel male talker are heard in a test phase. In support of a numerosity account (Bradlow & Bent, 2008), listeners only generalize the phonetic shift to the novel female talker when they hear the shift in multiple female exposure talkers.

Taken together, the results are consistent with the linguistic experience hypothesis. Less evidence is needed for generalization in contexts of intermediate listener familiarity (just acoustic similarity in Experiment 1), and more evidence is needed in contexts of no listener familiarity (both acoustic similarity and multiple-talker exposure in Experiment 2).

The General Discussion first underscores the necessity of a sociolinguistically-informed approach to cognitive science. The novel predictions and implications of the linguistic experience hypothesis are then discussed, linking perceptual adaptation to typology, theories of sound change, and linguistic anthropology, among other fields. While not discounting the value of current work, the General Discussion highlights how a more interdisciplinary perspective can plant the seed for new avenues of inquiry to bloom.

Chapter 3

Experiment 1: Adapting to Mandarin-Accented English

Through the speech-transcription-in-noise paradigm (Bradlow & Bent, 2008), Experiment 1 first presented Mandarin-accented English talker(s) in an exposure phase and then a novel talker from the same language background in a test phase. The goal was to investigate effects of acoustic similarity and talker variability on cross-talker generalization (defined in this paradigm as a boost in comprehension of the novel talker relative to some control condition).

According to the linguistic experience hypothesis (see Section 2.3.3), single-talker exposure should be sufficient for generalization in Experiment 1, supporting a sufficient similarity account (Xie et al., 2021; Xie & Myers, 2017). All participants were online workers living in the United States who necessarily had access to a computer and a stable Internet connection (otherwise, they would not be able to complete studies on a crowdsourcing platform). Given the linguistic diversity of the United States (Graddol, 2003) and widespread exposure to dialects/accents through the media (Nycz, 2019), all listeners: (i) were presumably able to identify the speakers as L2-accented (even without much exposure to Mandarin-accented English in particular); (ii) already knew implicitly that the acoustic properties of L2-accents

are often systematic and therefore generalizable. In this sociolinguistic context (decent listener familiarity with L2-accent), multiple-talker exposure should not be necessary for generalization, contra a stringent numerosity account (Bradlow & Bent, 2008). Single-talker exposure should be enough to generalize, as long as the exposure and novel talkers are acoustically similar (Xie et al., 2021; Xie & Myers, 2017).

One caveat is in order before charging on to the introduction. As noted in Point 6 of the Declarations above, this dissertation is essentially a concatenation of two published papers, with the present experiment being based on an article in *Language and Speech* (Aoki & Zellou, 2025b). The original *Language and Speech* paper is broadly concerned with generalization (and is thus pertinent to this dissertation), but there were also a few additional theoretical questions that are not the central focus here, as they are not germane to Experiment 2. Since these additional theoretical questions led to certain methodological choices, they must now be motivated in order for Experiment 1 to make sense.

The introduction of this chapter opens by explaining why talker variability was *simulated* in the original article (Section 3.1.1). The following section then justifies speaking style as a way of testing effects of acoustic similarity (Section 3.1.2). Finally, a brief overview of the experimental design and the predictions are provided (Section 3.1.3 and Section 3.1.4).

3.1 Introduction

3.1.1 Why *apparent* talker variability?

The author was conflicted about whether this section should be included at all. On the one hand, considering that the theoretical motivation for simulating talker variability is not relevant to the later experiments, inserting a lengthy explanation about apparent talker variability feels somewhat disruptive to the flow of the overall dissertation. On the other

hand, the complete omission of such detail does not seem to uphold the integrity of the original article (relegating this content to an appendix or asking readers to just navigate to the original article both seem equally suspect).

This section strikes a balance between these two concerns by including a “short” and a “long” explanation. The reader can now choose their own adventure by either: (i) just getting the gist of apparent talker variability and skipping to Section 3.1.2 of the introduction; (ii) perusing the long explanation that remains more faithful to the original work, but goes on somewhat of a theoretical tangent.

The short explanation for simulating talker variability

The crux is that single- and multiple-talker exposure not only differ in the number of talkers, but also in acoustic variability. Due to individual differences between speakers (Xie & Jaeger, 2020), multiple-talker exposure will always have more acoustic variability than single-talker exposure. Since increasing acoustic variability is also theorized to enhance adaptation (Raviv et al., 2022), this leads to a potential confound – are observed benefits of multiple-talker exposure really due to the number of talkers or to incidental acoustic properties?

Simulating talker variability through the “Change Gender” function in Praat (Boersma & Weenink, 2023) is intended to resolve these competing explanations. Passing stimuli through the Change Gender function can successfully alter both the apparent gender and identity of a talker (Luthra et al., 2021), while keeping acoustic changes to a minimum (i.e., by only adjusting f_0 and formant frequencies; Merritt & Bent, 2022). There are theoretical reasons why these minimal acoustic changes are not expected to influence adaptation (see the “long explanation” below for details) – as such, simulating talker variability boosts confidence that the multiple-talker exposure benefits genuinely stem from the heightened number of talkers, not from a concomitant increase in acoustic variability.

The long explanation for simulating talker variability

Up to this point, only two theoretical accounts (numerosity and sufficient similarity) have been discussed as potential mechanisms for generalization. A third account is possible, however, in the context of L2-accent adaptation: *heterogeneity*. As discussed in detail below, numerosity and heterogeneity are often confounded, and introducing apparent talker variability is intended to resolve this confound.

What is meant by “heterogeneity”, and how does this differ from “numerosity”? Raviv et al. (2022) note that multiple-talker exposure differs from single-talker exposure in at least two respects: (i) an increase in the number of talkers (numerosity); (ii) an amplification of phonological variability (heterogeneity). Both numerosity and heterogeneity accounts expect talker variability to boost generalization, but attribute the effect to different properties of multiple-talker exposure.

As a review from Section 2.3.1, a numerosity account claims that the higher number of talkers is what facilitates generalization, not the higher amount of phonological variation. In theory, if listeners only hear a single L2-accented talker initially, there might be some uncertainty about whether the phonological properties of exposure are idiosyncratic or characteristic of the overall accent, thus weakening generalization. On the other hand, those exposed to multiple L2-accented speakers could notice more shared attributes (e.g., devoicing of word-final stop consonants, such that “seat” is intended as “seed”; Xie & Myers, 2017), implying that the phonological properties of the input are broad accent features, not merely talker-specific particularities. The development of an accent-specific (talker-general) mental model might promote stronger generalization, increasing comprehension for novel talkers with the same L2-accent (e.g., applying what they have learned from exposure, listeners might recognize that a stimulus produced as “beet” is intended as “bead”).

A heterogeneity account proposes that the greater phonological variation of multiple-talker exposure is what encourages generalization, not the larger number of talkers. L1- and L2-

accents often diverge in cue weighting, and exposure to a more variegated stimulus set boosts the likelihood of learning which dimensions are relevant for distinguishing speech sounds (Apfelbaum & McMurray, 2011). L1-English and Catalan-accented English speakers, for instance, differ in their production of tense/lax vowel contrasts (e.g., /i/ in “seat” versus /ɪ/ in “sit”) – whereas L1-English speakers rely more heavily on formant frequencies than vowel duration (Hillenbrand et al., 2000), Catalan-accented English speakers show the opposite trend (Cebrian, 2007). As an extreme case, if there was no phonological variation in exposure (e.g., presentation of the same tokens of “seat” and “sit” from one talker), then it would be unclear whether the formant frequency and duration patterns were specific to those particular tokens or reflective of a broader cue weighting strategy. By contrast, exposure to a diverse set of Catalan-accented tokens would enable listeners to recognize a consistent lack of differentiation in formant frequencies compared to vowel duration. Multiple-talker exposure magnifies phonological variation, given the ubiquity of talker-specific differences in speech production (e.g., due to physiological and/or social factors, individual speakers tend to produce vowels with their own unique pattern of formant frequencies; Xie & Jaeger, 2020). Leveraging this greater variation, listeners exposed to multiple L2-accented talkers are more likely to successfully shift their cue weighting, which can then facilitate understanding of a novel talker from the same language background.

Numerosity and heterogeneity accounts are not intractably confounded, but teasing them apart is complicated in practice, especially in experiments on L2-accent adaptation that employ naturalistic, sentence-length stimuli (Bradlow & Bent, 2008; Xie et al., 2021). For one, speakers do not possess control over their exact degree of variation across utterances (Smith & Kenney, 1994), so researchers often cannot investigate numerosity while keeping heterogeneity constant (i.e., it is usually infeasible to present identical phonological variation across single-talker and multiple-talker exposure conditions). Meanwhile, even if the number of exposure talkers is stable (thus controlling for numerosity), probing heterogeneity accounts also brings its own challenges. L2-accented and L1-accented sentences potentially differ on

dozens of phonological contrasts (e.g., word-final stop consonants, tense/lax vowels, etc.; Wang et al., 2025), with each contrast being highly multidimensional (e.g., word-final stop consonants can be cued by vowel duration, burst duration, and/or closure duration; Xie & Myers, 2017). In attempting to adjust the amount of phonological variation, it becomes difficult to determine which contrasts and which dimensions to manipulate, as well as how much to alter them to achieve a robust effect. Although disentangling numerosity and heterogeneity accounts could be simplified by focusing on single contrasts (e.g., /b/ versus /p/; Clayards et al., 2008), this would sacrifice the ecological validity of the sentence-length stimuli used in prior work (Bradlow & Bent, 2008; Xie et al., 2021).

For researchers examining L2-accent adaptation through more naturalistic experimental paradigms (e.g., speech-transcription-in-noise), resolving the numerosity-heterogeneity confound is a difficult task without a straightforward solution. If the confound cannot be entirely controlled, one might ask the following as a starting point: Is there a way of adjusting one factor (numerosity or heterogeneity) while holding the other as constant as possible? This experiment introduces *apparent talker variability* (the simulation of multiple-talker exposure) as a potential method of elevating numerosity while minimizing changes in heterogeneity.

Talker variability can be simulated via the Change Gender function (Boersma & Weenink, 2023), a Praat algorithm that “uses pitch-synchronous overlap-and-add (PSOLA) to allow independent manipulation of formant frequencies, median f0, f0 variation, and signal duration while maintaining all of the other acoustic parameters” (Merritt & Bent, 2022; p. 487). Passing stimuli through the Change Gender function can successfully alter both the apparent gender and identity of a talker, based on anecdotal evidence (e.g., sufficiently decreasing the formant frequencies and f0 of a female speaker can make listeners think they are hearing a novel “male” talker; Luthra et al., 2021). Thus, even if all stimuli originate from one speaker in reality, the *illusion* of multiple talkers can be created with most acoustic properties remaining constant (i.e., an increase in numerosity with minor changes in heterogeneity).

Although the presentation of multiple apparent talkers would increase variability in terms of f_0 and formant frequencies, this caveat is tempered by a useful feature of the Change Gender function: the adjustment of “the spectral envelope of speech sounds up/down according to *uniform scaling*” (Barreda & Silbert, 2023, Section 13.2.2; emphasis added). Recent work in the normalization literature demonstrates that uniform scaling does not affect vowel identification – when all formants of a particular vowel are scaled up or down by the same ratio, the stimulus is still heard as the same vowel with the same quality (Barreda, 2021). According to Barreda (2020), listeners only associate shifts in uniform scaling with anatomical and/or social differences (e.g., vocal tract length, speaker size, and speaker gender), not phonological differences. Phonological differences are only perceived when non-uniform scaling occurs (e.g., when the proportional increase in the first formant (F1) is greater than the other formants, a vowel should be judged as relatively lowered; Barreda & Nearey, 2012). In short, even if *acoustic* variability in f_0 and formant frequencies is heightened, this type of variation should be largely “removed by the ‘normalizing ear’ of the listener” (Barreda, 2021, p. 29).

Despite the maintenance of most phonological features, certain increases in variability are likely unavoidable. For instance, after running stimuli from a female talker through the Change Gender function to generate a “novel male” talker, Cummings and Theodore (2022) reported inadvertent alterations in fricative acoustics:

... though the Change Gender parameters do not suggest nor explicitly allow control over changes in voiceless energy [of /s/ and /ʃ/], this function does introduce a scaling of center of gravity in line with the change in fundamental frequency. Specifically, center of gravity for [the “novel male”] tokens is shifted towards lower frequencies compared with [the original female] tokens. (p. 2340)

Employing the Change Gender function therefore does not entirely eliminate the numerosity-heterogeneity confound – rather, the proposed method is intended to increase numerosity while keeping changes in heterogeneity to a minimum.

The Change Gender function allows for two types of conditions to be compared – single talker exposure (where all stimuli come from one talker and are unmanipulated) and multiple apparent talker exposure (where all stimuli come from one talker, but some are manipulated to alter apparent identity). A numerosity account predicts that multiple apparent talker exposure should enhance generalization. By contrast, a null effect of apparent talker variability seems more likely for a heterogeneity account, given the minimized change in phonological variation across exposure conditions.

In the original article that the current chapter is based on, the goal was to examine numerosity, heterogeneity, and sufficient similarity in one experiment (Aoki & Zellou, 2025a). This dissertation, however, is not focused on heterogeneity (partly for concision, but also because cue weighting is not relevant for the next two experiments; see “Rationale for *not* presenting apparent talker variability” in Section 4.2.4 of Experiment 2). Other than Section 4.2.4, this dissertation will not mention the term “heterogeneity” again. The results will be interpreted with respect to the two accounts of interest (numerosity and sufficient similarity).

3.1.2 Examining effects of acoustic similarity with speaking style

Even after controlling for acoustic variability in exposure, acoustic similarity effects present a further quandary. A sufficient similarity account predicts generalization to occur when at least one exposure talker is sufficiently similar phonologically to the novel talker (e.g., an alignment in the use of vowel duration, burst duration, and closure duration when distinguishing word-final stop consonants; Xie & Myers, 2017). Merely increasing the number of exposure talkers thus does not guarantee a boost in generalization if the additional talkers are not similar enough to the novel talker.

Akin to numerosity, similarity accounts can be difficult to evaluate, especially when working with sentence-length stimuli. Numerous phonological contrasts can be heard within a set of sentences, and it becomes challenging to determine which features are important

for assessing between-talker similarity. Another issue is that individual differences in cue weighting are well-documented (Clayards, 2018), meaning that even for the same phonological contrast, listeners might diverge in their perceptual judgments of similarity. For example, if the word-final stop consonants of an exposure and novel talker are only alike in burst duration, but not in closure or vowel duration, then generalization might be facilitated for listeners attuned to burst duration, but blocked for those who weight closure or vowel duration more heavily. Defining “similarity” is ultimately not straightforward, and adaptation studies with sentence-length stimuli must often resort to exploratory, post-hoc assessments of similarity for a limited subset of features (Bradlow et al., 2023; Sidaras et al., 2009).

One way of alleviating these issues is to examine *speaking style*. Speaking style is a ubiquitous type of within-talker variation, referring to systematic shifts in production based on the social context and/or interlocutor (Aoki & Zellou, 2023a; Smiljanić, 2021; Vonessen et al., 2024). Both L1- and L2-accented speakers adjust their productions based on listener identity – for instance, when asked to imagine talking to someone who is hard-of-hearing, speakers make exaggerated acoustic-phonetic modifications (e.g., expanded vowels, greater f_0 variation, slower speaking rate) relative to their default, “casual” mode of speaking (Kato & Baese-Berk, 2022; Scarborough & Zellou, 2013). Perceptual effects are usually consistent, with hyper-articulated styles like hard-of-hearing-directed speech tending to enhance intelligibility in noise compared to casual speech (Aoki & Zellou, 2024; Picheny et al., 1985).

Introducing speaking style into L2-accent adaptation experiments does not completely eliminate the aforementioned difficulties about measuring similarity. Speaking styles are highly multidimensional acoustically and phonologically (Smiljanić, 2021), and perceptual effects can vary across listeners (Ferguson & Kewley-Port, 2002). Nevertheless, when employing sentence-length stimuli, speaking style proffers an a priori hypothesis about similarity effects without relying on post-hoc analyses. In particular, a similarity account predicts that generalization should be facilitated if the styles of the exposure and novel talkers are congruent.

3.1.3 Overview of the study design

The current study employed a speech-transcription-in-noise task to investigate how (apparent) talker variability and speaking style impact cross-talker generalization of L2-accented speech. There were three main reasons for selecting this particular paradigm: (i) it facilitates a more direct comparison with prior work relying on the same task (Bradlow & Bent, 2008; Xie et al., 2021); (ii) transcribing sentences in noise is fairly ecologically valid; (iii) obscuring the stimuli with noise may render the manipulation of apparent talker variability less transparent.

Listeners were either assigned to a critical condition (an exposure phase followed by a test phase) or to a control condition (no exposure). All of the critical conditions presented one Mandarin-accented English talker in exposure and one novel talker from the same language background at test. The experiment had a 5x2 between-participant design, with five exposure conditions and two test conditions:

- E1. No exposure (the same control in Bradlow et al. (2023))
- E2. A single talker producing casual speech
- E3. A single talker producing hard-of-hearing-directed speech
- E4. Multiple (apparent) talkers producing casual speech
- E5. Multiple (apparent) talkers producing hard-of-hearing-directed speech
- T1. A single novel talker producing casual speech
- T2. A single novel talker producing hard-of-hearing-directed speech

Note that the exposure stimuli were originally produced by a single talker across all critical conditions – the number of *apparent* talkers was manipulated by uniformly scaling f_0 and formant frequencies (see Section 3.1.1 for why talker variability was simulated).

The broad goal is to evaluate cross-talker generalization – which types of exposure conditions facilitate adaptation to the novel test talker?

3.1.4 Predictions

Generalization is defined as increased test phase transcription accuracy compared to the no-exposure control condition (Bradlow et al., 2023). As noted in the footnotes within Section 2.3.1, the issue with all speech-transcription-in-noise experiments is that they are opaque as to what exactly is being learned (Bradlow & Bent, 2008; Xie et al., 2021). The current study takes the following as working assumptions: (i) since the exposure and novel talkers share the same language background, they are acoustically similar (e.g., both likely devoiced word-final /d/, producing “seat” instead of “seed”); (ii) when listeners generalize what they have learned in exposure to a novel talker, such generalization ameliorates transcription accuracy (e.g., if the intended keyword is “bad”, listeners who have successfully generalized word-final devoicing are more likely to correctly transcribe the keyword as “bad”, rather than “bat”, boosting accuracy).

According to a *sufficient similarity account*, generalization should be facilitated when the stimuli in the exposure and test phases are acoustically similar (Xie & Myers, 2017). Two predictions fall out from this idea: (i) single-talker exposure should be sufficient for generalization due to the shared language background between the exposure and novel talkers; (ii) generalization should be boosted when there is an alignment in speaking style across phases. To make the second prediction more concrete, listeners exposed to casual speech should show an extra boost in test phase accuracy when tested on casual speech; meanwhile, participants presented with hard-of-hearing-directed speech in exposure should benefit even more perceptually when tested on hard-of-hearing-directed speech.

A *numerosity account* states that increasing the number of perceived talkers should enhance generalization – therefore, exposure to multiple apparent talkers is expected to

augment generalization beyond exposure to a single talker (in an extreme case, multiple apparent talkers might even be necessary for generalization, in a replication of Bradlow and Bent (2008)). No clear predictions are made about speaking style by a numerosity account.

3.2 Methods

3.2.1 Stimuli

Development

The speakers were four undergraduate international students from China who were recruited from the University of California (UC) Davis Psychology participant pool and compensated with course credit.¹ All talkers were 20-22 year-old Asian men who had been studying English for 7-15 years, reported that their first and strongest language was Mandarin, and did not self-report knowledge of any other language besides Mandarin and English.

The recording procedure was identical to Aoki and Zellou (2024). Participants were first seated in front of a computer within a sound-attenuated booth. After giving informed consent, speakers completed a three-block sentence production task and a demographic questionnaire. Each block was preceded by a written prompt designed to elicit one of three speaking styles to an imagined interlocutor: hard-of-hearing-directed speech, non-native-directed speech, or casual speech. Within each block, speakers produced the same 78 low-predictability sentences, all of which ended in a keyword (e.g., “Mr. Black knew about the *pad*”; Kalikow et al., 1977;

¹These speakers were part of a corpus of 96 talkers (half L1-English, half Mandarin-accented English), all of whom provided informed consent and completed the same recording procedure. The four specific talkers in this study were chosen because they all showed evidence of hyper-articulation in their hard-of-hearing-directed productions (see the acoustic analysis in this same section for details). The only reason for selecting male speakers is to be consistent with previous studies on L2-accent adaptation (e.g., Baese-Berk et al., 2013; Bradlow & Bent, 2008; Xie et al., 2021).

emphasis added).² The entire study was conducted through a self-paced Qualtrics survey, and recordings were made with a Shure WH20XLR head-mounted microphone at a digital sampling rate of 44.1-kHz. Details about block and sentence counterbalancing, the wording of the speaking style prompts, and other aspects of the recording procedure can be found in Section 2.1.3 of Aoki and Zellou (2024).

Note that the speakers technically produced sentences in three speaking styles, but only two were presented to listeners in the current experiment (hard-of-hearing-directed speech and casual speech). The reason for excluding non-native-directed speech has to do with the author’s recent work (Aoki & Zellou, 2024) – between hard-of-hearing-directed and non-native-directed speech, only the former seems to be effective at enhancing transcription accuracy in noise, despite both being ostensibly intended to facilitate comprehension (Piazza et al., 2022; Picheny et al., 1986). To ensure a robust perceptual difference across conditions, the present study focused on a binary comparison between hard-of-hearing-directed speech and casual speech. Future research should explore how non-native-directed speech and other speaking styles might influence cross-talker generalization.

For each of the four male speakers, two apparently female talkers were created using the Change Gender function in Praat. Both the f0 median and formant frequencies of the original stimuli were manipulated, given evidence that adjusting only f0 or only formants is insufficient for shifting apparent gender judgments (Hillenbrand & Clark, 2009). The specific parameter adjustments were based on the “h95” data set in the *phonTools* R package, which contains “[f]ormant frequency, f0 and duration information for vowels collected from 139 speakers in the Hillenbrand et al. (1995) data” (Barreda, 2015, p. 17). For the first set of

²The types of stimuli used in speech-transcription-in-noise experiments can vary from one study to the next. For instance, unlike the use of low-predictability sentences in the present work, certain studies employ semantically anomalous (Vaughn, 2019) or high-predictability sentences (Bradlow & Bent, 2008). In some cases, participants are asked to transcribe as many words as they can within each sentence (Van Engen et al., 2012), whereas other experiments instruct listeners to focus on sentence-final keywords (Clopper, 2011). The particular set of stimuli in this study were selected to align with the previous work of the authors (e.g., Aoki et al., 2022; Aoki & Zellou, 2023b; Cohn et al., 2021). There is no immediately apparent theoretical reason for why the results reported here would not replicate with other kinds of sentences, although this must be examined through future research.

apparently different talkers, the f0 median was set to 220 Hz (the average f0 of the adult female speakers in h95), while the formant shift ratio equaled 1.15 (roughly the average scale factor between the adult female and adult male speakers in h95 across the first three formants). The second set of apparently different talkers was generated by setting the f0 median and formant shift ratio to 259 Hz and 1.23, respectively, approximating the female talker in h95 with the highest f0. In summary, there were three apparent talkers per speaker (Original Male, Apparent Female 1, Apparent Female 2) and 1872 total stimuli (4 talkers * 2 styles * 78 sentences * 3 apparent talkers).

To prepare the sentences for the speech-transcription-in-noise experiment, all stimuli were set to a presentation level of 65 dB SPL using Praat and mixed with speech-shaped noise at a +5 dB signal-to-noise ratio (McCloy, 2015), following Bradlow and Bent (2008). The noise, which was developed from the long-term spectrum of the concatenation of all sentences (Winn, 2019), commenced 500 milliseconds before sentence onset and ended 500 milliseconds after sentence offset.

Acoustic analysis

Although the present experiment is focused on speech perception, an acoustic analysis was conducted to confirm that the hard-of-hearing-directed stimuli are hyper-articulated, in alignment with prior work (Aoki & Zellou, 2024; Picheny et al., 1986). The present analysis is limited in scope – since hyper-articulation is often characterized by simultaneous adjustments on many acoustic-phonetic parameters (e.g., speaking rate, f0, vowel space area, etc.; Smiljanić & Bradlow, 2009), it can be challenging to determine which specific cues are responsible for intelligibility enhancements (Kato & Baese-Berk, 2024).

Two features were examined, both of which are modulated by L1-Mandarin speakers in English: speaking rate (Kato & Baese-Berk, 2024) and burst duration of word-final voiceless stop consonants (Xie & Myers, 2017). For all sentences, speaking rate was calculated with a

custom-made Praat script and defined as the number of syllables divided by sentence duration (Cohn et al., 2021). Burst duration was only examined for sentence-final keywords that ended in a voiceless stop consonant (e.g., /p/ in “rope”) and that were released in both styles (160 total tokens across all speakers). If speakers are hyper-articulating relative to casual speech (Uchanski, 2005), then hard-of-hearing-directed speech should exhibit a slower speaking rate (fewer syllables per second) and a longer burst duration (i.e., an exaggerated contrast relative to voiced stop consonants, which tend to have shorter burst durations).

Paired t-tests were conducted comparing hard-of-hearing-directed and casual speech for each speaker on each acoustic cue. The results are shown for burst duration and speaking rate in Tables 3.1 and 3.2, respectively.³ Despite some variation in effect size, all speakers were acoustically similar, producing a significantly slower speaking rate and longer burst duration in hard-of-hearing-directed speech than casual speech. These results verify that the hard-of-hearing-directed stimuli are hyper-articulated and thus consistent with prior work.

Table 3.1: Results of the burst duration analysis for each speaker. The “Mean Diff” column subtracts the average value for hard-of-hearing-directed speech by the average value for casual speech. The full output of paired t-tests are reported: 95% confidence intervals (CI), t-values, degrees of freedom (df), and p-values. Significance annotations reflect meaningful differences between styles (* = $p < 0.05$; ** = $p < 0.01$; *** = $p < 0.001$).

Speaker	Mean Diff	95% CI	t-value	df	p-value
BF	35.66	[10.91, 60.41]	3.21	10	0.009**
BX	50.89	[27.59, 74.20]	4.56	20	<0.001***
DW	21.48	[7.52, 35.44]	3.18	23	0.004**
TW	16.42	[1.28, 31.56]	2.24	23	0.03*

³In the published versions of the current experiment and Experiment 2, the acoustic analyses solely consisted of paired t-tests (Aoki & Zellou, 2025a, b). However, taking Xin Xie and Chigusa Kurumada’s post-publication feedback into account, the revised version of Experiment 2 additionally visualizes the distributions of the acoustic cues (see Figures 4.2 and 4.3 below). This extra step was not taken here – since the current paradigm (speech-transcription-in-noise) is not focused on any one acoustic cue, showing cue distributions would not be clearly informative for interpreting the results. However, all of the acoustic data are open-access, and can be examined here: <https://doi.org/10.17605/osf.io/n8hws>

Table 3.2: Results of the speaking rate analysis for each speaker. The “Mean Diff” column subtracts the average value for hard-of-hearing-directed speech by the average value for casual speech. The full output of paired t-tests are reported: 95% confidence intervals (CI), t-values, degrees of freedom (df), and p-values. Significance annotations reflect meaningful differences between styles (* = $p < 0.05$; ** = $p < 0.01$; *** = $p < 0.001$).

Speaker	Mean Diff	95% CI	t-value	df	p-value
BF	-1.83	[-1.97, -1.68]	-25.12	77	<0.001***
BX	-1.53	[-1.70, -1.37]	-18.99	77	<0.001***
DW	-0.90	[-1.01, -0.79]	-15.94	77	<0.001***
TW	-1.01	[-1.10, -0.91]	-21.20	77	<0.001***

3.2.2 Participants

The current experiment has a fully between-participant design (see Section 3.2.3), similar to many prior studies on L2-accent adaptation (e.g., Baese-Berk et al., 2013; Bradlow & Bent, 2008; Xie et al., 2021). Given the lack of within-participant variables, a sufficiently large sample size is especially critical to ensure that any differences between conditions are not merely due to listener idiosyncrasies (Brysbaert & Stevens, 2018).

The target sample size was 800 participants ($n = 80$ across ten conditions; see Section 3.2.4), following the recommendation from Xie et al. (2021), who conducted a very similar speech-transcription-in-noise experiment. After running various simulations, Xie et al. (2021) achieved 45.9-96.4% power with 80 listeners in each condition – this was deemed acceptable, since power was considerably higher than the 21-38.4% range of the original Bradlow and Bent (2008) study.

All 909 participants were recruited from Prolific and paid approximately \$9 per hour (\$1.40 if placed in the shorter condition without an exposure phase and \$2.50 if placed in any other group; see Section 3.2.4 for details). The demographic filters on Prolific were used

to narrow down the participant pool to individuals who were born and living in the United States, between 18 and 35 years old, and whose self-reported first language was English (these specific criteria were selected to align with prior speech-transcription-in-noise studies; e.g., Experiment 3 of McLaughlin, 2022).

Certain participants ($n = 94$) were omitted from the analysis (the 10.3% exclusion rate is comparable to Xie et al. (2021), who reported 6.7% and 15.6% exclusion rates for their first and second experiments, respectively). Listeners were removed based on the following self-reported criteria: (i) audio, technical, and/or hearing difficulties ($n = 31$); (ii) English not being their dominant language ($n = 7$); (iii) childhood and/or daily exposure to any Chinese language or dialect (Mandarin, Cantonese, Mien, Hokkien; $n = 49$); (iv) working proficiency or fluency in Mandarin ($n = 3$); (v) being older than 35 years old (i.e., a mismatch from their official profile on Prolific; $n = 3$); (vi) writing non-words or “don’t know” on the majority of trials in the test phase of the transcription task ($n = 1$).

Slightly exceeding the original sample size target, the final analysis included data from 815 participants (445 women, 340 men, 30 non-binary; mean age = 28.01 years, $sd = 4.54$; self-reported ethnicity: Asian = 70, Black = 107, Latino = 49, Mixed = 124, Native American or Alaska Native = 4, White = 461).

Geographic analysis

The only geographic filter on Prolific was “individuals who were born and living in the United States”, but this is quite vague – where exactly do the participants reside within the United States? For example, are all of the participants concentrated in urban areas of the Midwest, or are they spread out more evenly across the country? Providing geographic context is relevant because participants living in urban areas might have more experience with linguistic variation (Gu et al., 2026), which could influence the role of talker variability on cross-talker generalization (see Section 2.3.3 above).

Unfortunately, detailed information about participants’ residence and background was not collected, other than what was mentioned in the previous section. This section is entirely a post-publication analysis to make up for the lack of demographic information.

This study was conducted through Qualtrics (see Section 3.2.3 below), which adds latitude and longitude information to raw data files. The documentation states the following: “... the location is an approximation determined by comparing the participant’s IP address to a location database. Inside the United States, this data is typically accurate to the city level” (Qualtrics, n.d.); “city level” is later clarified as equivalent to “zip code” (i.e., someone in a rural area will be placed in the correct zip code, not moved to what would typically be thought of as a “city”). Even though it would have been preferable to ask for more detailed information within the original experiment, all is not completely lost – the latitude and longitude data from Qualtrics can still be helpful for making inferences about the participants.

Two important points must be stated before moving on to the analysis. First, the “[l]atitude and longitude data values are based on city/zip code and cannot be used to identify a specific location or individual” (Qualtrics, n.d.). This implies that publishing geographic data does not violate ethical principles, since the participants’ locations are still anonymized. Second, since the data only reflect where the experiment was completed, an underlying assumption is that most listeners were taking Prolific studies at or near their homes.

The latitude and longitude coordinates for each participant were fed into the *reverse_geocode* function from the *tidygeocode* package in R (Cambon et al., 2025), outputting a list of zip codes (the ‘osm’ or “Open Street Map” method was used). These zip codes were then inputted into the *get_ruca* function from the *rurality* package in R (Wimpy, 2026), resulting in a list of Rural-Urban Commuting Area (RUCA) Codes (only the primary codes were used here, not the secondary codes). Developed by the United States government, RUCA codes represent “a detailed and flexible scheme for delineating sub-county rural and urban areas”; they range from 1 to 10, where lower values indicate more urban areas (Dobis

& Sanders, 2026). Overall, these RUCA codes are not a perfect measure, but they at least provide a clue about whether the participants are residing in more urban (and thus more linguistically diverse) environments or in more rural locales.

Overlaid on a map from the *maps* package in R (Becker et al., 2025), Figure 3.1 depicts the approximate location of residence for the participants, color-coded by RUCA value.⁴ Most RUCA values are purple (i.e., have lower values), suggesting that most participants are living in urban, diverse areas of the United States.⁵

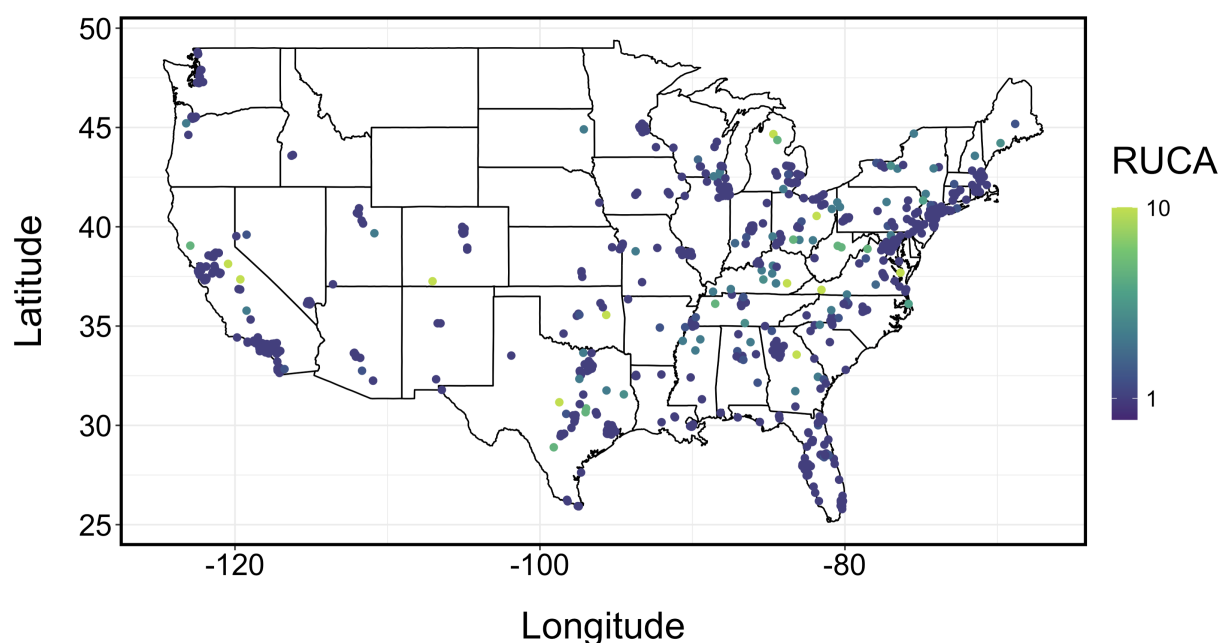


Figure 3.1: A map of the contiguous United States, where each dot represents the estimated location of residence for each participant in Experiment 1. The dots are color-coded by their Rural-Urban Commuting Area (RUCA) code, where lower numbers (more purple) indicate more urban locales.

⁴Figure 3.1 depicts 785 listeners (96.32% of the total in Experiment 1, $n = 30$ excluded). Two participants living in Hawaii were excluded for formatting purposes. For unknown reasons, the *get_ruca* function from the *rurality* package did not work for 28 zip codes.

⁵From Figure 3.1, one can see large clusters of participants around major cities, such as Chicago, San Francisco, and New York City. It is also clear, however, that the RUCA coding system assigns a value of “1” to many cities of different sizes, erasing variation (e.g., workers in Boise, Idaho and New York City are both indicated with purple dots, which misleadingly makes it seem as if they are equally “urban”). An RUCA code of “1” is technically defined as “Metropolitan core: primary flow is within an urban area of 50,000 or more people” (Dobis & Sanders, 2026), which lumps many locales together – this highlights that, although RUCA codes can be helpful for providing context, they should be supplemented by other measures in future work.

Summary statistics corroborate Figure 3.1 – most RUCA codes were very low (mean = 1.445), with 87.04% of participants having a score of 1. Table 3.3 shows that the average RUCA code across the ten between-participant conditions did not differ drastically, with all values being between 1 and 2 (indicating an urban environment).

Table 3.3: Mean RUCA value for each of the ten between-participant conditions in Experiment 1, where the variables “Exposure Style”, “Apparent Exposure Talker Number”, and “Test Style” are abbreviated as “Exposure”, “Variability”, and “Test”, respectively. The term “hard-of-hearing-directed speech” is abbreviated as “HOH-DS”.

Exposure	Variability	Test	Mean RUCA
Casual	Single	Casual	1.49
Casual	Multiple Apparent	Casual	1.61
Casual	Single	HOH-DS	1.54
Casual	Multiple Apparent	HOH-DS	1.29
HOH-DS	Single	Casual	1.33
HOH-DS	Multiple Apparent	Casual	1.38
HOH-DS	Single	HOH-DS	1.37
HOH-DS	Multiple Apparent	HOH-DS	1.47
No Exposure	N/A	Casual	1.41
No Exposure	N/A	HOH-DS	1.54

A critical implication falls out from this post-hoc analysis – recruiting participants from Prolific may result in a comparatively more representative sample than a university participant pool (Behrend et al., 2011), but still does not reflect a fully representative sample. For example, looking at Figure 3.1, large swaths of the United States are barely or not represented at all, particularly the Rocky Mountain Area (even though the original filter was quite broad, asking for anyone “born and living in the United States”). The results from high-powered experiments on crowdsourcing platforms must therefore still be interpreted with caution.

3.2.3 Procedure

The entire study was completed through an online, self-paced Qualtrics survey. After informed consent, participants received instructions to wear headphones and to move to a quiet area with little background noise (explicit headphone and background noise checks were not built into the survey). A sound calibration procedure ensued, followed by a perceptual experiment and a demographic questionnaire. The sound calibration procedure and the main experiment are described in separate subsections below.

Pre-experiment

The sound calibration procedure was identical to that reported in Section 2.1 of Zellou et al. (2023). Participants listened to a single sentence (“She asked about the host”) and were asked to adjust the volume of their computer to a comfortable level, only if they felt it was necessary. They were asked not to alter their volume for the remainder of the experiment.

To confirm that participants had listened to the sentence, a mandatory, forced-choice question was included as an attention check. Listeners were tasked with identifying the sentence from among three options (“She asked about the host”, “She asked about the toast”, “She asked about the “coast”). Proceeding to the next phase of the Qualtrics survey was blocked until the correct answer was provided. The question could be taken an unlimited number of times and did not function as a screener. All participants eventually gave the correct response and were not removed if they initially answered the question incorrectly.

The stimulus was amplitude-normalized to 60 dB SPL (a similar amplitude as the actual experimental stimuli), not embedded in noise, and produced by a male, text-to-speech voice (“Matthew” from Amazon Polly; Amazon Web Services, n.d.). There was no specific reason to use this particular voice, other than following past work (e.g., Zellou et al., 2023). The stimulus could be played up to five times. Replaying the stimulus was not required, but

included as an option just in case participants wanted to adjust their volume level and/or had answered the forced-choice question incorrectly and wanted to hear the sentence again.

After successfully answering the multiple-choice question, listeners were allowed to move on to the perceptual experiment.

Experiment

The main experiment was a speech-transcription-in-noise task. The 78 low-predictability sentences described in Section 3.2.1 were evenly divided across a 39-trial exposure phase and a 39-trial test phase, with each sentence only ever appearing in one of the phases (e.g., “Mr. Black knew about the pad” was always heard in the exposure phase, never in the test phase). Each trial auditorily presented a single sentence masked by noise, and listeners were asked to type the final word into a text box. All sentences were presented in a pseudo-randomized order and only played once with no feedback provided.

Participants were randomly assigned to one of five exposure conditions and one of two test conditions (5 exposure x 2 test = 10 possible conditions overall). In terms of exposure, listeners either heard: (i) no exposure; (ii) a single talker producing casual speech; (iii) a single talker producing hard-of-hearing-directed speech; (iv) multiple (apparent) talkers producing casual speech; (v) multiple (apparent) talkers producing hard-of-hearing-directed speech. There were also two test conditions: (i) a single talker producing casual speech; (ii) a single talker producing hard-of-hearing-directed speech.

The exposure stimuli always originated from a single Mandarin-accented English talker – listeners only differed in the number of *apparent* talkers heard in exposure. Whereas the single talker conditions just presented the veridical (male) stimuli, the multiple apparent talker conditions presented three seemingly different talkers created through the “Change Gender” Praat function (see Section 3.2.1 for details). To reinforce the experimental manipulation, listeners who heard multiple apparent talkers were explicitly told in the exposure phase

instructions that they would be hearing “three speakers”. Sentence content and apparent talker assignment were evenly counterbalanced.

All participants completed a test phase, in which a single (novel) Mandarin-accented English talker produced either casual or hard-of-hearing-directed speech. Only the original productions were presented, so apparent talker variability was not manipulated. As in Xie et al. (2021), the four speakers were counterbalanced across exposure and test.

3.2.4 Statistical analysis

The exposure and test phase responses were cleaned by removing all spaces and punctuation, converting all characters to lowercase, and deleting any non-final keywords (e.g., when an entire sentence was transcribed instead of just the final keyword). Accuracy was then coded binomially as either correct (=1) or incorrect (=0). Similar to prior work (Aoki & Zellou, 2023c; Bent & Bradlow, 2003), a strict standard was employed, with responses only counted as correct if they contained all and only the appropriate affixes (e.g., if “cake” was the keyword, “cakes” was given a 0).⁶ However, both obvious spelling errors (e.g., “sleaves” for “sleeves”) and homophones (e.g., “greece” for “grease”) were considered correct.

The statistical analysis was conducted with Bayesian mixed-effects logistic regression in R (R Core Team, 2021) through the *brms* package (Bürkner, 2017) and Stan (Stan Development Team, 2023).⁷ The prior distributions were identical to the model described in Chapter

⁶In casual speech, one of the speakers accidentally produced an extra affix on the final keyword of one of the sentences (“We’ve spoken about the *trucks*” instead of “We’ve spoken about the *truck*”). Only responses of “trucks” were considered correct for this particular stimulus, even though the original keyword was “truck.”

⁷To analyze categorical data such as binary transcription accuracy, it is now widely acknowledged that logistic regression is a better choice than ANOVA, since the assumptions of the latter are violated by categorical data (Jaeger, 2008). However, among modern-day papers in linguistics, both frequentist and Bayesian approaches are still very commonly implemented (Sonderegger & Sóskuthy, 2025). Going Bayesian in this study (and throughout the dissertation) is more of a personal choice – none of the overall model interpretations in this dissertation really change, regardless of whether a frequentist or Bayesian model is adopted. In general, however, Bayesian models have many advantages relative to frequentist models (e.g., the ability to include information about prior distributions, the output of credible intervals rather than just point estimates for random intercepts and slopes, etc.; see Barreda and Silbert (2023) for more information).

10.5.2 of Barreda and Silbert (2023)), except that the “cor” parameter was removed. Several different models were fitted, with each answering a distinct question – the syntax for each model is described below in Section 3.3.

Before unpacking the results, it should be noted that the interpretation of Bayesian models is somewhat controversial. Certain papers take the output of Bayesian models (credible intervals) and make a differentiation between intervals that cross zero or do not cross zero (e.g., only bolding the latter, implying that there is evidence for an effect when an interval does not cross zero; Cohn et al., 2023). However, others have noted that this is problematic because zero is a somewhat arbitrary cutoff (in a practical sense, is there really a difference between $[-0.001, 5]$ and $[0.001, 5]$, even though one includes zero and the other does not?; Kruschke, 2018). Others implement Bayes Factors to draw conclusions from their models, but recent meta-analyses indicate that almost all papers in the social sciences misinterpret Bayes Factors (Tendeiro et al., 2024). In general, the approach to statistical modeling and interpretation can vary greatly from one study to the next (Coretta et al., 2023), which makes it challenging to determine how data should be analyzed and interpreted.

This study will follow Cohn et al. (2023), interpreting credible intervals that do not cross zero as evidence for an effect (and characterizing intervals that barely cross zero as “marginal”). There are two reasons for this choice: (i) this dissertation is mostly based on already-published work, which makes it awkward to radically change the analysis at this point; (ii) since the models for the other experiment in this dissertation (Experiment 2) contain a few dozen fixed effects and interactions, bolding the credible intervals that do not cross zero is preferable for readability. All of this said, the potential issues with this approach are noted in this section for the record.

3.3 Results

3.3.1 Exposure phase

Although generalization is of primary interest (i.e., differences in test phase accuracy), it is essential to affirm that the exposure phase manipulations were perceived as intended. One must: (i) establish that casual and hard-of-hearing-directed speech differed perceptually; (ii) verify that listeners who heard multiple apparent talkers actually perceived different talkers.

Both concerns can be addressed by analyzing the exposure phase. Accuracy should be higher for hard-of-hearing-directed speech (Picheny et al., 1985); moreover, if the apparent talker manipulation is effective, then it should replicate work on veridical talker variability (i.e, lower intelligibility when listening to multiple apparent talkers; Mullennix et al., 1989).⁸

A statistical model (depicted in Equation 3.1 with R syntax) was fitted with sum-coded fixed effects of Exposure Style (between participants; casual, hard-of-hearing-directed) and Apparent Exposure Talker Number (between participants; single, multiple), as well as their interaction. By-listener, by-talker, and by-sentence random intercepts were included.⁹

$$\begin{aligned} \text{Accuracy} \sim & \text{Exposure Style} * \text{Apparent Exposure Talker Number} + \\ & (1 \mid \text{Listener}) + (1 \mid \text{Speaker}) + (1 \mid \text{Sentence}) \end{aligned} \tag{3.1}$$

⁸The predicted effect of talker variability differs across exposure and test: “greater variability may initially hinder learning [lower accuracy in exposure], but typically leads to an improved ability to generalize learning to new contexts [greater accuracy for a novel test talker]” (Raviv et al., 2022, p. 463). Why talker variability might “initially hinder learning” is outside the scope of this work – for a review, see Luthra (2024).

⁹Determining an appropriate random effects structure remains a major area of uncertainty in mixed-effects modeling (Meteyard & Davies, 2020). In particular, there are two predominant philosophies regarding random slopes: maximal and data-driven. The maximal approach advocates for “add[ing] as many random slopes as possible”, while the data-driven perspective recommends only “add[ing] random slopes that significantly improve the model” (Sonderegger, 2023, p. 264). Deciding on a model is made challenging by the following trade-off – inserting random slopes tends to decrease the Type I error rate (Barr et al., 2013), while also increasing the Type II error rate (Matuschek et al., 2017). Based on feedback during peer-review, a forward-selection, data-driven procedure was used to evaluate the effect of random slopes on model fit (Matuschek et al. (2017)). To make a long story short, it was determined that adding random slopes was not really necessary for any of the models in this study – see Section 2.5.1 of the original paper for details (Aoki & Zellou, 2025b).

Figure 3.2 visualizes the exposure phase results (note that the abbreviation “HOH-DS” stands for “hard-of-hearing-directed speech” in all figures of this chapter). Aligning with the literature on speaking style perception, hard-of-hearing-directed speech was transcribed more accurately than casual speech [β : -0.29, SE: 0.02, 95% highest density interval (HDI) = (-0.33, -0.24)]. There was also a comprehension detriment for listeners who heard multiple apparent talkers compared to listeners who heard a single talker [β : 0.19, SE: 0.02, 95% HDI = (0.14, 0.23)], as expected. No evidence of an interaction was observed [β : -0.02, SE: 0.02, 95% HDI = (-0.07, 0.03)], which implies that the effect of Exposure Style did not differ based on the number of apparent exposure talkers.

In summary, the exposure phase results are highly consistent with prior work and suggest that both the speaking style and apparent talker variability manipulations were successful.

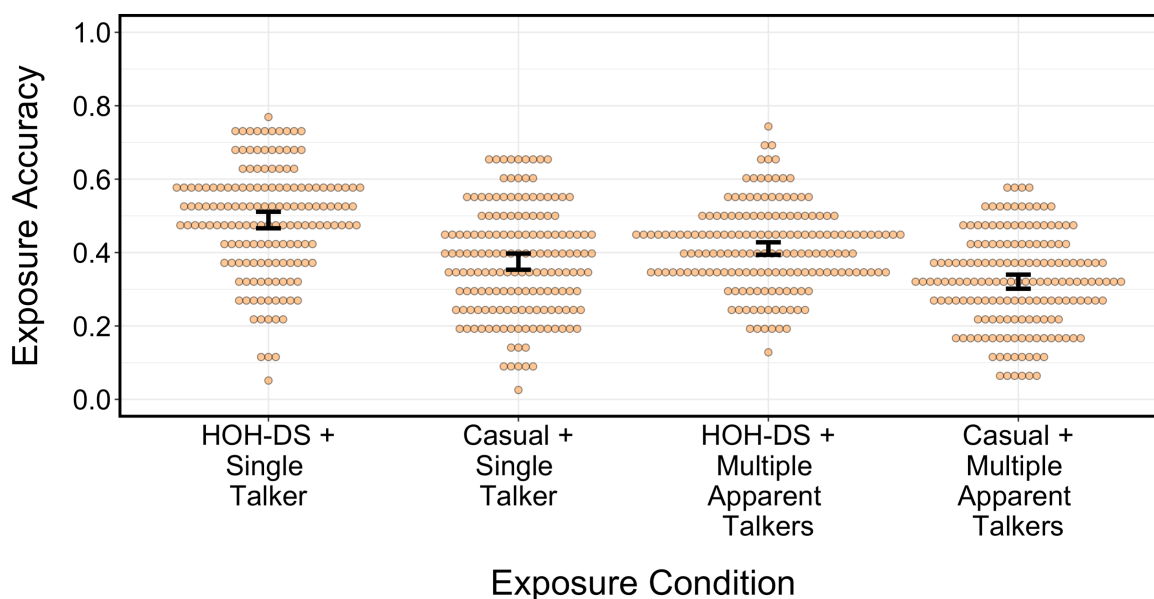


Figure 3.2: Exposure phase transcription accuracy across conditions. Individual points reflect by-participant mean performance. Error bars are 95% confidence intervals bootstrapped over by-participant means.

3.3.2 Test phase

The main goal is to investigate generalization, which is being operationalized as increased test phase accuracy relative to the no-exposure control condition (see Bradlow et al., 2023). As in Cummings and Theodore (2023), the test phase analysis was divided into two stages. The initial stage assessed whether exposure facilitated generalization at all, while the second stage investigated the relative degree of generalization among the critical conditions.

Did generalization occur at all?: Comparing the control and critical conditions

This question was evaluated with eight separate models, which compared each of the eight critical conditions to a no-exposure control condition. The comparisons were always within speaking styles (e.g., listeners tested on casual speech were only compared to participants who were similarly tested on casual speech). Each model, whose syntax is shown in Equation 3.2, contained a treatment-coded fixed effect of Condition (between participants; two levels, reference = no-exposure control) and the same random effects structure as Equation 3.1.

$$\text{Accuracy} \sim \text{Condition} + (1 \mid \text{Listener}) + (1 \mid \text{Speaker}) + (1 \mid \text{Sentence}) \quad (3.2)$$

The test phase results are divided into two images for clarity, with Figures 3.3 and 3.4 showing performance on casual speech and hard-of-hearing-directed speech, respectively. Among listeners tested on casual speech, all four conditions with exposure facilitated cross-talker generalization. More specifically, compared to the no-exposure control, test phase accuracy was greater among listeners exposed to: (i) hard-of-hearing-directed speech and a single apparent talker [β : 0.28, SE: 0.08, 95% HDI = (0.12, 0.45)]; (ii) casual speech and a single apparent talker [β : 0.22, SE: 0.09, 95% HDI = (0.03, 0.40)]; (iii) hard-of-hearing-directed speech and multiple apparent talkers [β : 0.20, SE: 0.09, 95% HDI = (0.03, 0.38)]; (iv) casual speech and multiple apparent talkers [β : 0.41, SE: 0.08, 95% HDI = (0.25, 0.57)].

By contrast, there was no evidence of generalization for listeners tested on hard-of-hearing-directed speech. Relative to the no-exposure control, no robust difference in test phase performance emerged for any condition: (i) hard-of-hearing-directed speech and a single apparent talker [β : 0.03, SE: 0.09, 95% HDI = (-0.14, 0.20)]; (ii) casual speech and a single apparent talker [β : 0.02, SE: 0.07, 95% HDI = (-0.13, 0.17)]; (iii) hard-of-hearing-directed speech and multiple apparent talkers [β : 0.10, SE: 0.08, 95% HDI = (-0.07, 0.26)]; (iv) casual speech and multiple apparent talkers [β : 0.04, SE: 0.08, 95% HDI = (-0.11, 0.19)].

In summary, there was evidence of generalization for listeners tested on casual speech (Figure 3.3), but not for listeners tested on hard-of-hearing-directed speech (Figure 3.4).

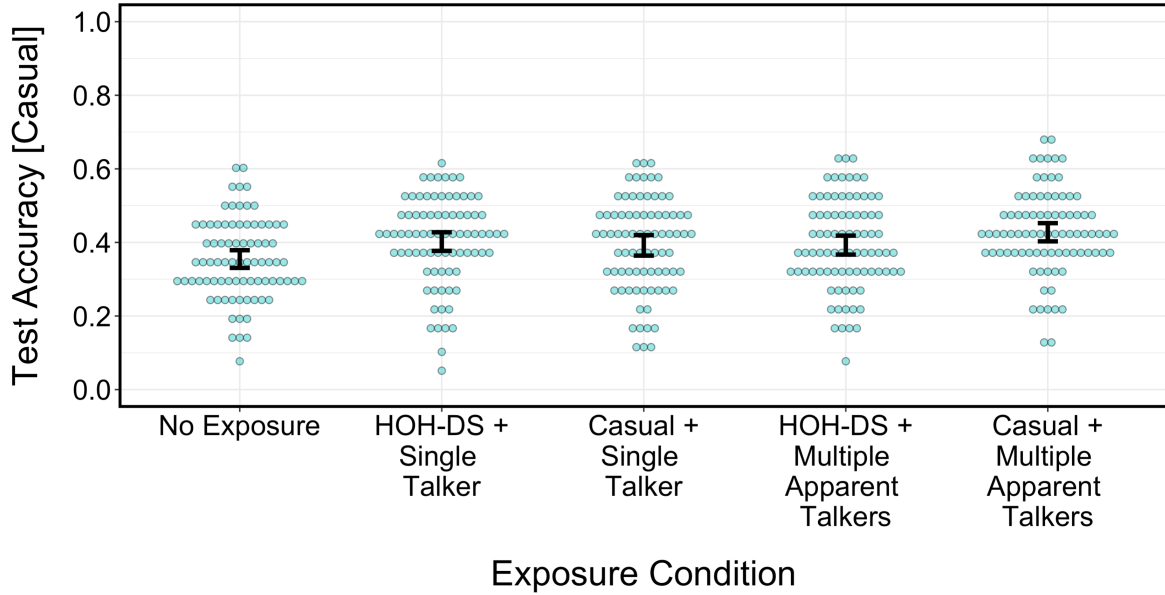


Figure 3.3: Test phase transcription accuracy across conditions, among participants tested on casual speech. Individual points reflect by-participant mean performance. Error bars are 95% confidence intervals bootstrapped over by-participant means.

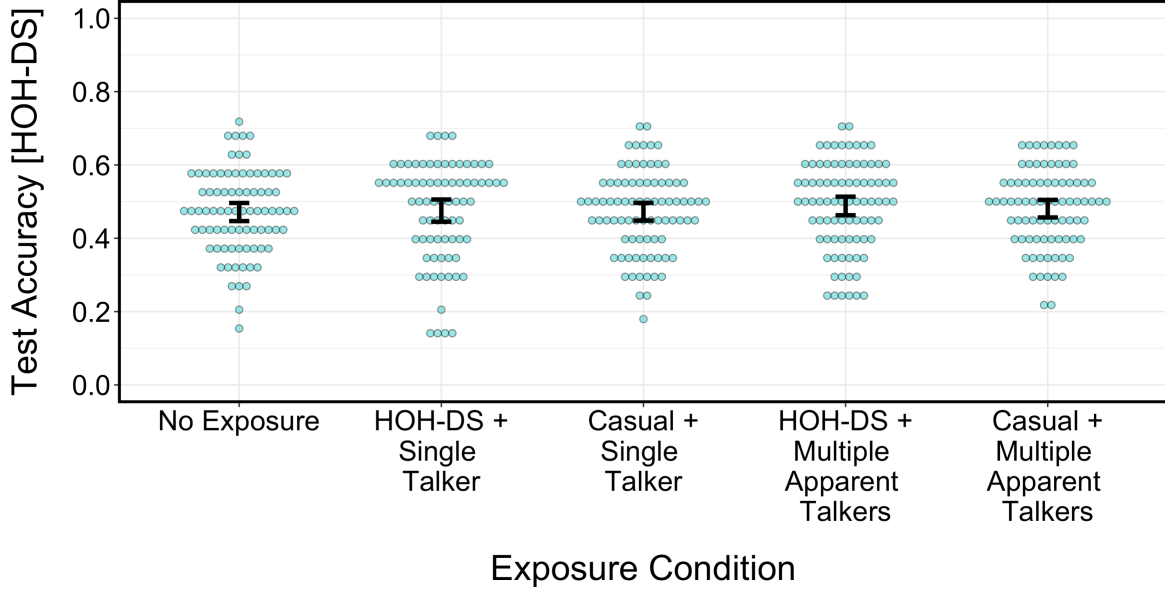


Figure 3.4: Average test phase transcription accuracy across conditions, among participants tested on hard-of-hearing-directed speech. Individual points reflect by-participant mean performance. Error bars are 95% confidence intervals bootstrapped over by-listener means.

Investigating the relative degree of generalization across critical conditions

Since no generalization took place among listeners tested on hard-of-hearing-directed speech, the initial statistical model in this section only includes results for participants tested on casual speech (i.e., the four right-most conditions in Figure 3.3). The model formula was identical to Equation 3.1.

There were no main effects of Exposure Style [β : 0.03, SE: 0.03, 95% HDI = (-0.03, 0.10)] or Apparent Exposure Talker Number [β : -0.03, SE: 0.03, 95% HDI = (-0.09, 0.03)]. However, a marginal interaction emerged [β : -0.06, SE: 0.03, 95% HDI = (-0.13, -0.002)].

The interaction term was interpreted via the *hypothesis* function from *brms* (Bürkner, 2017). The first hypothesis revealed no effect of apparent talker variability for listeners exposed to hard-of-hearing-directed speech [β : 0.03, SE: 0.04, 95% HDI = (-0.05, 0.12)]. However, exposure to multiple apparent talkers led to an increase in test phase accuracy for listeners exposed to casual speech [β : -0.09, SE: 0.04, 95% HDI = (-0.18, -0.008)]. In other

words, among listeners tested on casual speech (Figure 3.3), the No Exposure condition had the lowest accuracy, the “Casual + Multiple Apparent Talkers” condition had the highest accuracy, and the other three exposure conditions had an equally intermediate accuracy.

This interaction is quite weak, and it remains unclear whether the effect would remain in an aggregated analysis that combines all critical conditions (i.e., the four right-most conditions in Figure 3.3 and the four-most conditions in Figure 3.4). To address this point, a final statistical model added Test Style (between participants; casual, hard-of-hearing-directed speech) as a sum-coded fixed effect to Equation 3.1, along with all possible interactions (see Equation 3.3).

$$\begin{aligned} \text{Accuracy} \sim & \text{Exposure Style} * \text{Apparent Exposure Talker Number} * \text{Test Style} + \\ & (1 \mid \text{Listener}) + (1 \mid \text{Speaker}) + (1 \mid \text{Sentence}) \end{aligned} \quad (3.3)$$

In theory, there should be a three-way interaction – Exposure Style and Apparent Exposure Talker Number should jointly influence test phase accuracy, but only for listeners tested on casual speech.

The full summary statistics of this final model are displayed in Table 3.4 for clarity. A strong effect of Test Style was observed [β : -0.21, SE: 0.02, 95% HDI = (-0.25, -0.17)], indicating that casual speech was transcribed less accurately overall than hard-of-hearing-directed speech. Only a marginal three-way interaction emerged between Exposure Style, Test Style, and Apparent Exposure Talker Number [β : -0.04, SE: 0.02, 95% HDI = (-0.08, 0.007)]. Output from the *hypothesis* function showed: (i) marginal evidence of a two-way interaction between Exposure Style and Apparent Exposure Talker Number for listeners tested on casual speech [β : -0.06, SE: 0.03, 95% HDI = (-0.12, 0.0003)]; (ii) no evidence for such an interaction for listeners tested on hard-of-hearing-directed speech [β : 0.01, SE: 0.03, 95% HDI = (-0.04, 0.07)].

Table 3.4: Summary statistics for the final model of the test phase analysis, which investigated the relative degree of generalization across all critical conditions (i.e., a comparison of test phase accuracy in the four right-most conditions of Figure 3.3 and the four right-most conditions of Figure 3.4). The predictors “Exposure Style”, “Apparent Exposure Talker Number”, and “Test Style” are abbreviated as “Exposure”, “Variability”, and “Test”, respectively. Credible intervals that do not cross zero are in bold.

Effect	Estimate	Est. Error	l-95% CI	u-95% CI
Intercept	-0.34	0.48	-1.30	0.60
Exposure	0.01	0.02	-0.04	0.05
Variability	-0.02	0.02	-0.07	0.02
Test	-0.21	0.02	-0.25	-0.17
Exposure : Variability	-0.02	0.02	-0.07	0.02
Exposure : Test	0.03	0.02	-0.01	0.07
Variability : Test	-0.004	0.02	-0.05	0.04
Exposure : Variability : Test	-0.04	0.02	-0.08	0.01

3.4 Discussion

Leveraging the speech-transcription-in-noise paradigm, Experiment 1 exposed listeners to either a single Mandarin-accented English talker or to multiple (apparent) Mandarin-accented English talkers, and then presented a novel test talker from the same language background. The aim was to examine effects of talker variability and acoustic similarity on generalization (i.e., whether listeners would transfer what they learned from the exposure phase to the test phase). As predicted by a *sufficient similarity account*, is single-talker exposure enough for generalization, as long as the exposure and test talkers are acoustically similar (Xie et al., 2021; Xie & Myers, 2017)? Or is multiple-talker exposure helpful/necessary for generalization, in alignment with a *numerosity account* (Bradlow & Bent, 2008)?

The two main findings are consistent with both sufficient similarity and numerosity. First, exposure to Mandarin-accented English facilitated generalization (i.e., led to enhanced comprehension of the novel test talker), but the effects were mediated by speaking style (all listener groups tested on casual speech showed evidence of generalization, in stark contrast to those tested on hard-of-hearing-directed speech). Second, exposure to multiple apparent talkers fostered greater generalization beyond single-talker exposure, but the effect size was fairly weak and only occurred for participants exposed to and tested on casual speech (as opposed to listeners who heard a mismatch in speaking styles across exposure and test).

On the one hand, buttressing a sufficient similarity account, all listener groups tested on casual speech generalized regardless of the number of perceived exposure talkers (since the exposure and test talkers had the same language background, they were presumably similar acoustically, thus facilitating generalization). Meanwhile, supporting both numerosity and sufficient similarity, multiple apparent talker exposure did boost generalization, but only given acoustic similarity in style.¹⁰

Experiment 1 also provides evidence for the *linguistic experience hypothesis*. As theorized in Section 2.3.3, multiple-talker exposure should only be required for generalization when listeners hear variation that they have never (or rarely) encountered. Despite self-reporting little experience with Mandarin-accented English, the participants likely had at least some experience with L2-accents prior to the experiment (ShareAmerica, 2023), especially since they all predominantly live in urban areas (see Section 3.2.2). L2-accents are easily recognized as such (Atagi & Bent, 2017), and listeners presumably know that L2-accents are potentially generalizable. The linguistic experience hypothesis predicts that the bar for generalization should be fairly low in this context – as expected, single-talker exposure was sufficient for generalization, with the benefit of multiple-talker exposure being comparatively weak.

¹⁰Although multiple-talker exposure generally increases both the number of talkers and the amount of acoustic variability relative to single-talker exposure (Raviv et al., 2022), the *simulation* of talker variability kept changes in acoustic variability to a minimum. This suggests that the numerosity benefit is truly due to the greater number of (perceived) talkers, not to incidental acoustic properties. For details about apparent talker variability, see Section 3.1.1.

The rest of the general discussion is organized as follows. Section 3.4.1 explores potential mechanisms of the (small) numerosity effect, ruling out an explanation based on cognitive difficulty. Section 3.4.2 provides some possible explanations for why listeners tested on hard-of-hearing-directed speech unexpectedly showed no generalization, contra listeners tested on casual speech. Finally, Section 3.4.3 serves as a transition to the next experiment.

3.4.1 Ruling out a cognitive difficulty account

The present study has focused on numerosity (an increase in the number of talkers) to explain the benefit of apparent talker variability. However, an entirely different interpretation might also be possible – what if the results of the current experiment are mediated by cognitive difficulty? Listeners exposed to casual speech and multiple apparent talkers showed the lowest transcription accuracy in the exposure phase (Figure 3.2) – in theory, being assigned to a more perceptually challenging task may have engendered increased attention to the exposure stimuli and thus led to greater generalization at test (Tzeng et al., 2024).¹¹

Recent work, however, provides evidence against this explanation. Employing the speech-transcription-in-noise paradigm, Bradlow et al. (2023) found that transcription accuracy for a novel L2-accented test talker was generally the lowest following exposure to a low-intelligibility Turkish-accented talker (as opposed to more intelligible Brazilian Portuguese-, Farsi-, and Spanish-accented talkers). If anything, reducing intelligibility is expected to hinder adaptation, which would not support a cognitive difficulty mechanism.

The ideal adapter framework more plausibly accounts for effects of apparent talker variability (Kleinschmidt & Jaeger, 2015; see Section 2.1.4 above for a review). Repeated exposure to systematic acoustic properties is thought to facilitate the development of generative mental models, which help listeners make predictions about speech production (e.g., given prior exposure to word-final devoicing in Mandarin-accented English, one can later predict that

¹¹Thank you to Reviewer 1 from *Language and Speech* for this point.

“overload” may have been meant as “overload”; Xie & Myers, 2017). If listeners apply a generative model from an exposure phase to a novel test talker (i.e., engage in generalization) and if the talkers are acoustically similar, then test phase comprehension should increase, since listeners are more accurately predicting the test talker’s intended productions.

Listeners exposed to casual speech and multiple apparent talkers exhibited the largest generalization effect. This potentially occurred for two reasons: (i) there were two types of acoustic similarity across exposure and test (language background and speaking style); (ii) the apparent perception of multiple, acoustically similar talkers in exposure provided even greater certainty that the acoustic properties of the exposure talkers were not just idiosyncrasies. The latter point suggests that, even in cases where single-talker exposure is sufficient for generalization, multiple-talker exposure can still be helpful.

3.4.2 Why was no generalization observed among listeners tested on hard-of-hearing-directed speech?

Appealing to the ideal adapter framework only suffices for listeners tested on casual speech (i.e., when prior exposure was helpful in comprehending a novel L2-accented talker). Why was test phase accuracy the same for participants tested on hard-of-hearing-directed speech, regardless of whether or not they received prior exposure?

Perhaps there is some special property of hard-of-hearing-directed speech that blocks generalization. Hard-of-hearing-directed speech is acoustically exaggerated (see Tables 3.1 and 3.2), increasing intelligibility compared to casual speech (see Figures 3.3 and 3.4). These acoustic and perceptual enhancements might allow listeners in the no-exposure control condition to adapt more rapidly to a novel talker, rendering prior exposure unnecessary for successful adaptation. The null result in the test phase could therefore be tentatively interpreted as one of the many benefits conferred by hyper-articulation (Smiljanić, 2021).

Looking at Figure 3.4, it may seem counterintuitive to treat 50% accuracy as a ceiling effect showing “successful adaptation”. This conclusion becomes more plausible, however, when the difficulty of the experiment is taken into account. Participants were not only asked to transcribe the final keyword of L2-accented sentences in noise (a demanding task in itself), but also heard each sentence just once on each trial, and were presented with challenging stimuli (low-predictability sentences such as “Miss Brown considered the coast”; Kalikow et al., 1977). Even experiments with significantly less taxing procedures do not necessarily lead to perfect results – for example, in the absence of background noise, participants in Kim et al. (2025) show approximately 90-95% accuracy when transcribing simple sentences produced by L1-accented speakers (e.g., “Somebody stole the money”; Bradlow, n.d.). A ceiling effect might therefore be expected at a fairly low accuracy for a more difficult task, even if listeners are exposed to hyper-articulated, hard-of-hearing-directed speech.

At the same time, the null effect observed for hard-of-hearing-directed speech might not involve any particular speaking style per se.¹² Bradlow et al. (2023) point to the possible role of intelligibility on adaptation: “lexical knowledge may not be available [in a low-intelligibility context] to guide the mapping of phonetic variation to linguistically meaningful categories, and consequently, adaptation to L2 speech may be constrained for both talker-specific and talker/accent-general adaptation” (p. 1610-1611). While Bradlow et al. (2023) are careful to note the “equivocal” nature of the literature (p. 1610), low stimulus intelligibility could potentially account for both the null effects and the small effect sizes observed in the current study. The average transcription accuracy across all exposure and test conditions in the present experiment is fairly low, regardless of speaking style (32-49%; see Figures 3.3 and 3.4). This is especially striking in comparison to other work – across test conditions, Table I of Bradlow et al. (2023) and Figure 3 of Xie et al. (2021) report an average accuracy of 55-89% and around 75-90%, respectively (this stark contrast likely stems from methodological differences, e.g., the use of low- versus high-predictability sentences). The low overall intelligibility for all

¹²Thank you to Dr. Meghan Clayards for pointing out this possibility.

speaking styles in the current study may have impeded adaptation, leading to both the weak effect of apparent talker variability and the lack of evidence of generalization for listeners tested on hard-of-hearing-directed speech.

In short, the null effect in this study might be driven by hard-of-hearing-directed speech itself or by low overall intelligibility. The former explanation would predict that, regardless of stimulus intelligibility, testing listeners on hard-of-hearing-directed speech should always show a null effect of exposure condition, as observed in the current work. By contrast, the latter explanation should see the null effect dissipate if listeners are presented with more easily transcribed and more intelligible stimuli.

3.4.3 Conclusion and next steps

Experiment 1 focused on a very specific type of sociolinguistic context, in line with prior work. L1-English participants from the United States were recruited from a crowdsourcing platform and completed an L2-accent adaptation experiment. It is reassuring that the results are largely consistent with the literature (Xie et al., 2021; Xie & Myers, 2017) – single-talker exposure was sufficient for generalization (and even though there was a small benefit of multiple (apparent) talker exposure, it was still mediated by acoustic similarity).

However, this finding is largely expected by the linguistic experience hypothesis. Multiple-talker exposure is only predicted to be necessary for generalization when listeners adapt to variation that is extremely *unfamiliar* to them. Given the linguistic diversity of the United States (ShareAmerica, 2023), L2-accented speech may not have been sufficiently unfamiliar to listeners for multiple-talker exposure to have a strong effect.

If we push the envelope and present an entirely new type of test case (one that listeners have definitely never experienced before and that has not been examined in the literature), how might talker variability mediate generalization?

Chapter 4

Experiment 2: Adapting to a Novel Phonetic Shift

Departing from L2-accent adaptation, Experiment 2 focuses on a novel phonetic shift embedded within an otherwise canonical L1-English accent. Participants were exposed to synthetic voices, where the female speaker(s) produced /p/ as [b] (e.g., “paint” as *baint*) and the male speaker(s) remained unshifted. The subsequent test phase presented a single novel female and a single novel male speaker, examining whether listeners would generalize the /p/ to [b] shift from the female talker(s) in exposure to the novel female talker at test.

The linguistic experience hypothesis predicts that multiple-talker exposure should become helpful (or perhaps even necessary) for generalization when listeners encounter *unfamiliar* phonetic variation. Unlike L2-accented speech in Experiment 1, a completely unfamiliar phonetic shift heard in only one talker might be treated as an ungeneralizable idiosyncrasy, regardless of whether the exposure and novel talkers are sufficiently similar acoustically. But if listeners hear more than one female talker who says /p/ as [b], they might realize that the phonetic shift is broadly applicable to female speakers (not just to one individual), thus triggering cross-talker generalization.

As noted in Point 6 of the Declarations above, this experiment is based on a published paper in *Glossa Psycholinguistics* (Aoki & Zellou, 2025a). However, due to the absence of a word limit, much greater detail is now provided, particularly in the Introduction.

The introduction of this chapter opens with Section 4.1.1, which justifies the current test case. Section 4.1.2 then briefly summarizes the experimental design, and Section 4.1.3 outlines the predictions.

4.1 Introduction

4.1.1 Developing a suitable test case

Experiment 1 was straightforward for the most part – a very well-studied test case was examined (Mandarin-accented English) using a very well-established adaptation paradigm (speech-transcription-in-noise). If the experimental design is largely consistent with previous studies, then there is less pressure to motivate one’s methodological choices.

By contrast, as far as the author is aware, no prior adaptation study has investigated a gendered /p/ to [b] shift produced by synthetic voices. To demonstrate that this test case has not been constructed in a random fashion, this section provides a two-part justification, with both parts addressing a potential question that a skeptical reader might have.

Question 1: Why present a novel phonetic shift instead of L2-accented English?

As an initial question, one might wonder why it is necessary to change the test case at all from Experiment 1. Given the objective of this dissertation (to examine how listener familiarity mediates effects of talker variability on generalization), the path of least resistance would be

to keep the same test case (Mandarin-accented English) and only make slight adjustments to the experimental design.

The most straightforward approach would have been to include an individual-differences analysis in Experiment 1. Participants could have filled out a language experience survey at the end of the study (e.g., “Rate how familiar [you are] with... Chinese-accented English on a scale of 0 (Not familiar at all) to 6 (Very familiar)”); Shin et al., 2021, p. 590). The responses could have been quantified as a measure of “accent experience” and added to the statistical models. Individuals with less accent experience should have, in theory, shown a greater numerosity effect (i.e., a more robust benefit of talker variability on generalization).

While not denying the value of individual-differences research (Yu & Zellou, 2019), measuring individual differences in language experience can be a risky endeavor. There is relatively little consensus about how accent/dialect “familiarity” should be evaluated (for discussion, see Miao & Kang, 2023, p. 1-4). Although Likert-scale ratings and other measures of language experience can be helpful, they are also potentially influenced by memory biases (Gu et al., 2026), and their interpretation is often subjective (e.g., the meaning of “familiar” might differ across raters). Overall, the output of individual-differences analyses can be ambiguous, leading to difficulty when interpreting results.

Taking a different angle, why did Experiment 1 not just compare two groups, one familiar with L2-accented speech and another unfamiliar with L2-accented speech? The issue with this proposal involves statistical power. As noted in the Introduction (Section 2.4.1), the goal of this work is to test the linguistic experience hypothesis with a more extreme type of test case (i.e., one where listeners are *completely* unfamiliar, not just somewhat unfamiliar). Although it might be possible to recruit individuals from small-scale societies with no previous exposure to L2-accented speech, finding a sufficient number of participants would likely pose a significant logistical challenge (all of the experiments in this dissertation require several hundred listeners to align with modern-day standards for statistical power; Brysbaert & Stevens, 2018).

In general, addressing the theoretical questions of this work becomes practically difficult if all of the experiments focus on L2-accent adaptation. The current study therefore abandons L2-accented speech, instead exposing listeners to a novel phonetic shift in L1-English.

There are several advantages to this test case. First, the phonetics and phonology of American English are extremely well-studied (Chomsky & Halle, 1968; Clopper et al., 2005), which makes it fairly clear which types of variation are attested and unattested (as discussed below, it seems highly unlikely that any L1-English listener has heard a gendered /p/ to [b] shift produced by synthetic voices). Second, it is much more time-effective to recruit L1-English listeners from an online crowdsourcing platform, without worrying about finding individuals who have never been exposed to L2-accented English. Finally, the potential ambiguity of an individual-differences analysis can be avoided.

It should be made clear that the “advantages” of a novel phonetic shift over L2-accented English only involve concerns about feasibility. All of the approaches described above are valid in principle, and it is left to future work to determine whether the linguistic experience hypothesis holds up in other contexts (e.g., if a robust method emerges for measuring individual differences in accent familiarity, then it would be well-worth using this measure to evaluate the linguistic experience hypothesis for L2-accented speech).

Question 2: Why a synthetic gendered shift from /p/ to [b]?

There are essentially an infinite number of unattested phonetic shifts in English. Why did the current study focus on a /p/ to [b] shift, why is the shift correlated with binary gender (with the female speakers being shifted, but not the male speakers), and why are synthetic voices used instead of human voices?

Before motivating these specific components, the first step is to define the qualities of an ideal test case. Besides *novelty*, two other important characteristics are considered below: *salience* and *plausibility*.

Salience is important for easing learning – otherwise, considering the difficulty of adapting to a novel phonetic shift in a short time frame, listeners might not show any successful generalization at all, even after exposure to multiple talkers. The phonetic shift should be readily noticed and recognizable as novel (e.g., subtle changes in nasal coarticulation below the level of consciousness would not be appropriate for the present purposes; Coetzee et al., 2022). The shifted exposure talker(s) and the novel test talker should also be similar in some clearly identifiable way – if not, then listeners may not be able to mentally group the talkers together, which would block generalization (e.g., without any visual cues, ethnic identification from speech is generally too inconsistent for ethnicity to be implemented as a reliable social cue in the current study; Newman & Wu, 2011). As a final step towards increasing salience, it would also be preferable to introduce explicit *co-variation* in the exposure phase. If the phonetic shift is correlated with a social group in exposure (e.g., binary gender), and listeners hear at least one representative from two categories (shifted female speakers *and* unshifted male speakers), then it should telegraph to listeners how they should generalize (e.g., the shift would be more unambiguously associated with *gender* and not some other social category that is incidentally shared between the shifted exposure talkers).

In addition to being novel and salient, the test case should be (at least somewhat) plausible. As a rather flippant example, it would probably be strange to co-vary a /p/ to [z] shift with shirt color (e.g., to expose listeners to red-shirted speakers who produce “paint” as *zaint* and to unshifted purple-shirted speakers). A change from /p/ to [z] is very unexpected, since the two sounds are quite different from an articulatory and acoustic/auditory perspective; moreover, mutable characteristics like shirt color are highly unlikely to be associated with accent/dialect differences (Kleinschmidt, 2019).¹ The ideal test case is one that sits in a

¹Note that a correlation between a /p/ to [z] shift and shirt color is not *impossible*. We know that phonetic variants can be quite distinct articulatorily (e.g., the alveolar [t] and the glottal [ʔ] are in free variation in English; Sumner & Samuel, 2005). One can also conceive of a scenario where shirt color might be linguistically meaningful (e.g., imagine being at an NCAA sports game where red-shirted California English speakers from Stanford are cheering against purple-shirted Midwest English speakers from Northwestern). However, these two situations are not ideal for the present purposes because they are either too improbable or require the establishment of extensive social context.

“Goldilocks” zone of being just outside the range of variation of English (and is thus relatively plausible), while still being clearly novel and salient.

In acknowledgment that the “perfect” test case probably does not exist, the three components of the current test case (binary gender, /p/ to [b] shift, synthetic voices) reflect a best attempt at simultaneously satisfying all three characteristics described above (novelty, plausibility, salience). Each component is justified below.

1. Binary gender: As one of the most highly recognizable socio-indexical cues (Barreda & Predeck, 2024; Eckert, 1989), binary gender is intended to satisfy the requirements of salience and plausibility. Male and female voices are often acoustically distinct (Barreda, 2021), and listeners can leverage the (on average) lower f0 and formant frequencies of males to identify speaker gender with high confidence and near-ceiling accuracy (Hillenbrand & Clark, 2009). Moreover, unlike shirt color, listeners also know (consciously or not) that gender can be relevant for speech perception in certain cases (e.g., an ambiguous fricative between /s/ and /ʃ/ is more likely to be categorized as /s/ if the talker is male; Yu, 2010).
2. A /p/ to [b] shift: On the one hand, since [p] and [b] only differ in voicing (and are thus very similar articulatorily), a shift from /p/ to [b] is not as unrealistic as other possible phonetic shifts (e.g., the previous example of a /p/ to [z] shift).

However, such an extreme shift from /p/ to [b] is both salient and novel, especially in the absence of any other pronounced phonetic adjustments. Although entirely remapped sounds can be observed in otherwise canonical L1-accents (e.g., unexpected sibilant productions or “lisps”; Lockenvitz et al., 2020), they are usually characterized as talker-specific idiosyncrasies, not talker-general properties that correlate with social categories like gender.

Another novel aspect of the current test case is the co-variation between binary gender and the /p/ to [b] shift, with female speakers being shifted and male speakers remaining

unshifted. In general, compared to fricatives and vowels, male and female speakers in American English do not differ as much in their production of stop consonants like [p] and [b] (Kleinschmidt, 2019; Morris et al., 2008). Among studies that do find gender effects (e.g., Robb et al., 2005; Swartz, 1992), the result is almost always the exact opposite of the co-variation constructed for this study – male speakers usually produce voiceless consonants with a shorter Voice Onset Time (VOT) than female speakers (i.e., it is *male* speakers who produce /p/ as a more acoustically [b]-like sound).

3. Synthetic voices: To compensate for the peculiarity of the /p/ to [b] shift, the usage of synthetic voices is meant to shift the balance by increasing plausibility.

In general, we now live in a digital era, with generative artificial intelligence (AI) platforms receiving billions of visits per month (Liu & Wang, 2026), and millions of people regularly interacting with voice-activated assistants such as Amazon’s Alexa and Siri (Ammari et al., 2019; Keaton & Zellou, 2026). Despite the utility of such technologies (or at least, their perceived utility), they can also cause frustration for users, behaving in highly unexpected ways. Voice assistants can phonetically glitch, replacing one sound with another unrelated sound (e.g., see Figure 4.1 for a Siri user griping about a /t/ to [ɹ] shift); generative AI can output nonsensical strings of text (e.g., Google Gemini’s recommendation to glue cheese onto pizza; Sarkar, 2025); and digital devices can act unpredictably (e.g., Amazon’s Alexa randomly starting to talk without any command from the user; Lutz & Newlands, 2025).

In the context of synthetic speech, it might not be so odd to introduce a random phonetic shift into a voice with an otherwise canonical L1-English accent. The association between synthetic voices and technology might lead listeners to expect the unexpected – a gender-mediated /p/ to [b] shift could be framed as a kind of “glitch” when produced by synthetic voices (e.g., perhaps the programmers made an error with only the female voices, but not the male voices).

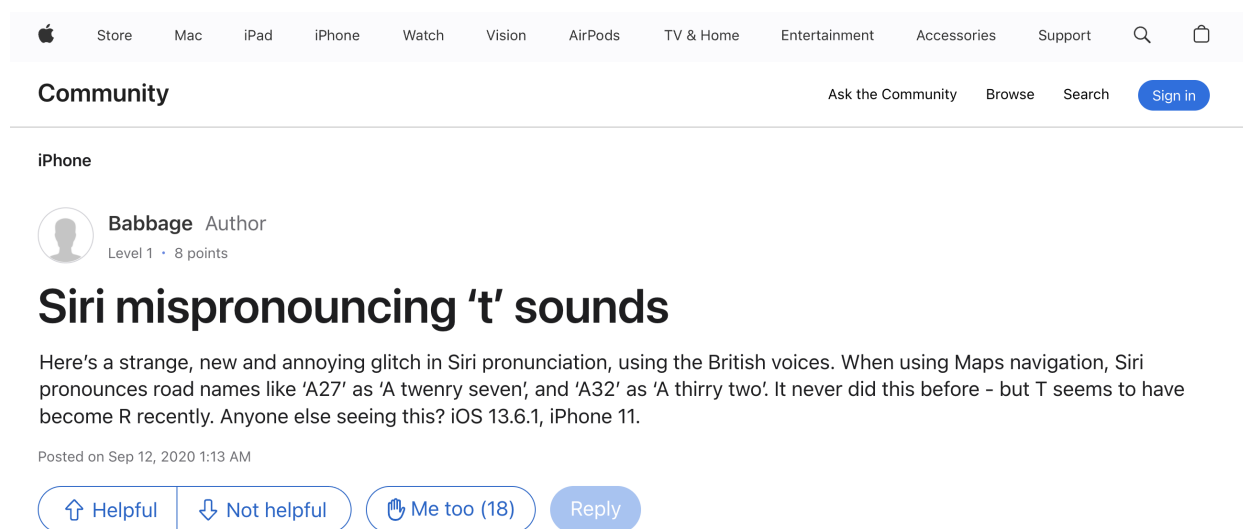


Figure 4.1: A user complaining about the pronunciation of British Siri on an online forum (dated September 12, 2020), with agreement from 18 other users. Siri is apparently producing a “t” to “r” shift (e.g., “twenry” instead of “twenty”), randomly replacing one sound with another: <https://discussions.apple.com/thread/251785651> (accessed May 6, 2026).

4.1.2 Overview of the study design

Now that the test case has been motivated, this section briefly reviews the procedure, so that the predictions in the following section are properly contextualized.

The design mirrored Zellou et al. (2023). All participants completed an exposure phase followed by a test phase. Listeners were first exposed to semantically predictable sentences and asked to identify sentence-final keywords (e.g., “The wall needed a new coat of *paint*”). Participants then categorized stimuli along a “buy-pie” continuum at test.

The test phase always presented a novel synthetic female and a novel synthetic male speaker in separate blocks. The exposure phase contained two between-participant manipulations. First, listeners were either placed in a critical condition with a gender-mediated shift (female = /p/ to [b], male = canonical /p/) or a control condition with no shift. Second, listeners were either exposed to a single talker per gender or to multiple talkers per gender.²

²In the original *Glossa Psycholinguistics* article (Aoki & Zellou, 2025a), the current study was reported as two separate experiments that were identical except for the number of exposure talkers. These two experiments have now been combined into the same chapter for concision.

The key theoretical question is whether the experimental manipulations induce cross-talker generalization. When do listeners apply the /p/ to [b] shift from the exposure female talker(s) to the novel female talker?

4.1.3 Predictions

In the test phase, listeners were asked to make a binary judgment, categorizing stimuli along a “buy-pie” continuum as either “buy” or “pie”. Generalization should specifically be realized as a greater proportion of “pie” responses for the novel female test talker than for the novel male test talker (if a novel speaker is expected to produce a /p/ to [b] shift, then more acoustically [b]-like tokens should be categorized as /p/; Sumner, 2011). An increase in “pie” responses is only expected among listeners exposed to the phonetic shift, not among listeners placed in the control condition with no shift.

The current experiment adjudicates between three possible theoretical accounts: (i) talker normalization; (ii) sufficient similarity; (iii) numerosity. The predictions of each account are summarized in Table 4.1 and further explained below.

Table 4.1: Summary of predictions for each theoretical account in Experiment 2. Each prediction (“No difference” or “More ‘pie’ response”) compares the novel female talker to the novel male talker in the test phase of the critical condition. (In the control condition, neither exposure talker produces the /p/ to [b] phonetic shift, so “buy-pie” categorization is expected to be the same for both test talkers).

Theoretical Account	Single Talker per Gender	Multiple Talkers per Gender
Talker Normalization	No difference	No difference
Sufficient Similarity	More “pie” response	More “pie” response
Numerosity	No difference	More “pie” response

Given that the current phonetic shift is highly inconsistent with the long-term distributional statistics of English and has not been examined in prior research (to the best knowledge of the author), an open question is whether listeners will adapt or generalize at all. Therefore, it is essential to begin with some kind of null hypothesis, before thinking about potential effects of talker variability.

A *talker normalization account* proposes that listeners will “normalize”, or disregard, speaker gender differences in the exposure phase, resulting in similar responses for the novel female and male test talkers across all conditions (Liberman & Mattingly, 1985).³ Since the phonetic shift involves stop consonants, and stop consonant production does not strongly co-vary with gender in American English (Kleinschmidt, 2019), participants might ignore gender altogether (cf. Experiment 1 of Kraljic & Samuel, 2007). All of the critical exposure conditions contain an equal number of shifted and prototypical tokens (cf. Experiment 2 of Tzeng et al., 2021). If talker gender is disregarded and the experiences hearing shifted and prototypical tokens are all weighted equally, then test phase categorization should be the same across all conditions. In other words, according to a talker normalization account, no condition should exhibit a meaningful difference in the proportion of “pie” responses between the novel female and male talkers.

Unlike a talker normalization account, sufficient similarity and numerosity accounts both assume that speaker gender will impact adaptation. However, the latter two accounts diverge in their predictions about how the number of exposure talkers might affect the results. A *sufficient similarity account* claims that generalization should be triggered for any test talker who is sufficiently similar to the phonetically shifted exposure talkers, regardless of how many exposure talkers are heard (Xie et al., 2021; Xie & Myers, 2017). In the critical conditions, the /p/ to [b] phonetic shift is always associated with female speech in the exposure phase –

³Note that the term “normalization” has been employed in many different ways in the speech perception literature. Here, “normalization” is intended to hearken back to more traditional theories, which claimed that “variants of a particular type of linguistic unit are *normalized* [emphasis added] to arrive at an abstract, prototypical representation” (Nygaard, 2005, p. 391). More modern theories of normalization do not necessarily assume that social information is stripped away from the speech signal (Barreda, 2020).

thus, listeners are always expected to generalize the shift to the novel female test talker, given similarity in social category (gender) and in acoustics (e.g., higher f0 and formant frequencies; Hillenbrand & Clark, 2009).

A *numerosity account* posits that the number of exposure talkers should have an impact on the results; according to a stringent version, generalization should only occur when listeners hear multiple talkers per gender (Bradlow & Bent, 2008). Given their lack of experience with the particular /p/ to [b] shift under investigation, participants might treat the shift as a talker-specific idiosyncrasy when produced by only one talker, thus blocking generalization and leading to no difference in “pie” response between the novel test talkers. Listeners should only generalize if the shift is produced by a group of similar talkers (e.g., two female speakers), since this would rule out the shift as a talker-specific idiosyncrasy. Generalization should specifically occur towards the novel female test talker, who aligns more closely with the phonetically shifted exposure speakers in terms of social and acoustic similarity.

4.2 Methods

4.2.1 Exposure stimuli

Development

The exposure phase contained 48 semantically predictable sentences that all ended in a monosyllabic, phrase-final keyword (see Table A.2). Although semantic predictability was not normed, the procedure made the intended keyword very clear (see Section 4.2.4).

The sentences were divided up into 16 filler and 32 critical stimuli (the ratio of one-third filler to two-thirds critical follows prior work; Lai and Tamminga 2024). The 16 filler sentences

did not have any words with /b/ or /p/. The majority were taken directly from Zellou et al. (2023) (n = 11; e.g., “My wool socks are warm and thick”), while the others were slightly adapted to remove any instances of /b/ or /p/ (n = 5; e.g., “*Paula* had no shovel so she could not dig” became “*Lisa* had no shovel so she could not dig”).

The 32 critical sentences were also designed to avoid /b/ and /p/, except for the target words. Each of the critical target words contained one instance of /p/ (all word-initial), did not contain /b/, and were not part of a /p/-/b/ minimal pair (e.g., paint/*baint; “The wall needed a new coat of *paint*”). All of the critical stimuli were original sentences created by the author (since Zellou et al. (2023) exposed listeners to a different phonetic shift, a new set of critical sentences had to be constructed).

All stimuli were automatically generated in Amazon Polly (Amazon Web Services, n.d.). There were three reasons for using Amazon Polly, out of the numerous text-to-speech (TTS) systems that are now available (Alonso Martin et al., 2020). First, it is a cost-effective platform that outputs highly intelligible, state-of-the-art TTS voices (Hosier et al., 2020). Second, TTS voices from Amazon offer a high degree of control over the acoustic properties of each stimulus (e.g., every time “paint” is generated with the “Joanna” voice, the Voice Onset Time of the initial /p/ is always around 53 seconds). Given the potential effect of stimulus acoustics on adaptation (Lai & Tamminga, 2024; Xie & Myers, 2017), using controlled synthetic speech should make the results more reliable and replicable with the same methodology. Third, unlike synthetic voices that are very difficult to distinguish from human voices (e.g., voice clones from ElevenLabs; Bakkouche et al., 2025), many Amazon voices still sound somewhat robotic (Lilley et al., 2025; see Section 4.1.1 for why more synthetic-sounding speech is actually preferred for the current study).

Amazon Polly has four engines that each generate voices slightly differently (from most to least advanced: generative, long-form, neural, standard). Neural TTS was selected because it offers the widest array of voices (13 total US-English voices as of March 23rd, 2026, versus 7

each in generative and standard TTS and 4 in long-form TTS; Amazon Web Services, 2026, p. 199-217). An added benefit of neural TTS is that it does *not* implement the most advanced technology, outputting voices that are not quite as human-like as other engines.

The exposure stimuli were generated in six adult voices (three female: Joanna, Ruth, Salli; three male: Joey, Matthew, Stephen) in the “US-English” setting (at the time of writing, Amazon Polly does not provide any details about the specific dialectal background of the voices).⁴ For each of the three male voices, 48 sentences (16 filler and 32 critical) were typed directly into the Amazon Web Services console and downloaded (3 voices * 48 sentences = 144 total). For each of the three female voices, 80 sentences (16 filler, 32 unshifted critical, 32 shifted critical) were created in the same fashion (3 voices * 80 sentences = 240 total). The two versions of the critical stimuli only differed in the /p/-initial target words – in the shifted version, these words were altered orthographically to begin with “b” (e.g., “paint” was typed as “baint”). The downloaded sentences were converted from .mp3 to .wav in the command line using FFmpeg (Tomar, 2006) and set to a presentation level of 60 dB SPL in Praat (Boersma & Weenink, 2023).

The author listened to each shifted sentence from each female speaker to verify that the keywords were produced as intended (relying on the author’s impressionistic judgments as an L1-English speaker). There were generally no issues; however, since the shifted keywords are essentially nonwords (and thus highly unlikely to appear in Amazon’s training data), a handful of stimuli were not produced appropriately (e.g., “peace” was typed as “beace”, but pronounced as “be ace” [bi ɛis]). In these instances, the sentence was retyped using Speech Synthesis Markup Language and X-SAMPA, following Amazon Polly’s documentation (Amazon Web Services, 2026, p. 240-242). This second step resolved all remaining pronunciation issues, based on the author’s own auditory evaluation of the stimuli.

⁴As noted in the Declarations above, the auditory stimuli cannot be published by the author for proprietary reasons. However, as of March 23rd, 2026, an Amazon Polly demo is available online that plays samples of all six voices in both standard and neural TTS: https://www.youtube.com/watch?v=nUX-6_aRvGE. The video also shows some of the functionalities of the AWS console, including the selection of different voices.

Before moving on to the acoustic analysis, a mistake should be mentioned – one of the sentences erroneously contains a /b/-initial word (“The soccer team won the match *by* one point”; Row 37 of Table A.2). There are two reasons why this is considered a relatively minor issue: (i) “by” is a non-content word representing only 0.2% of the words listeners heard in the exposure phase (1/401); (ii) hearing a single canonical /b/ does not reverse the mapping from /p/ to [b] in the shifted speakers. (Including a non-target word with a canonical /p/ would have been a much more grave error (e.g., “My wool *pants* are warm and thick”), since that would have contradicted the /p/ to [b] shift and potentially blocked adaptation; Tzeng et al., 2021). Nevertheless, the intention was to avoid any instances of /b/ in the exposure stimuli, and the unintended exposure stimulus is noted here for transparency.

Acoustic analysis

Figure 4.2 depicts the VOT distributions of both unshifted and shifted /p/ for each speaker.⁵ According to Bonferroni-corrected paired t-tests, the VOT of the shifted /p/ to [b] stimuli for the female speakers was lower than the VOT of the unshifted /p/ stimuli across all speakers (all $p < 0.001$; compare the distributions with solid lines in the right panel of Figure 4.2 to the other distributions across both panels). Moreover, as can be seen in the left panel of Figure 4.2, the VOT of the unshifted /p/-initial stimuli was slightly higher for each female speaker compared to each male speaker ($p < 0.001$ for nearly all comparisons; the only exception was the comparison between Joanna and Matthew, where $p = 0.01$). This relatively small but meaningful gender difference in the unshifted stimuli is consistent with prior work (Robb et al., 2005; Swartz, 1992).⁶

⁵The original article did analyze the acoustics of the exposure stimuli, but only showed the mean VOT and f0 for each speaker (Tables 2 and 3 in Aoki and Zellou (2025a)). However, as Xin Xie and Chigusa Kurumada rightly pointed out to the author post-publication, visualizing the *variance* of acoustic cues provides critical context for adaptation experiments – even if two speakers have an identical mean VOT, the variance or shape of the VOT distributions can influence perceptual learning (Baese-Berk, 2019; Clayards et al., 2008).

⁶Unlike Tables 3.1 and 3.2 in Experiment 1, the full output of the paired t-tests in Experiment 2 is not shown here for concision (i.e., only p -values are reported, not the mean differences, confidence intervals, t -values, or degrees of freedom). The comprehensive analysis can be found on the Open Science Framework: <https://doi.org/10.17605/osf.io/n8hws>

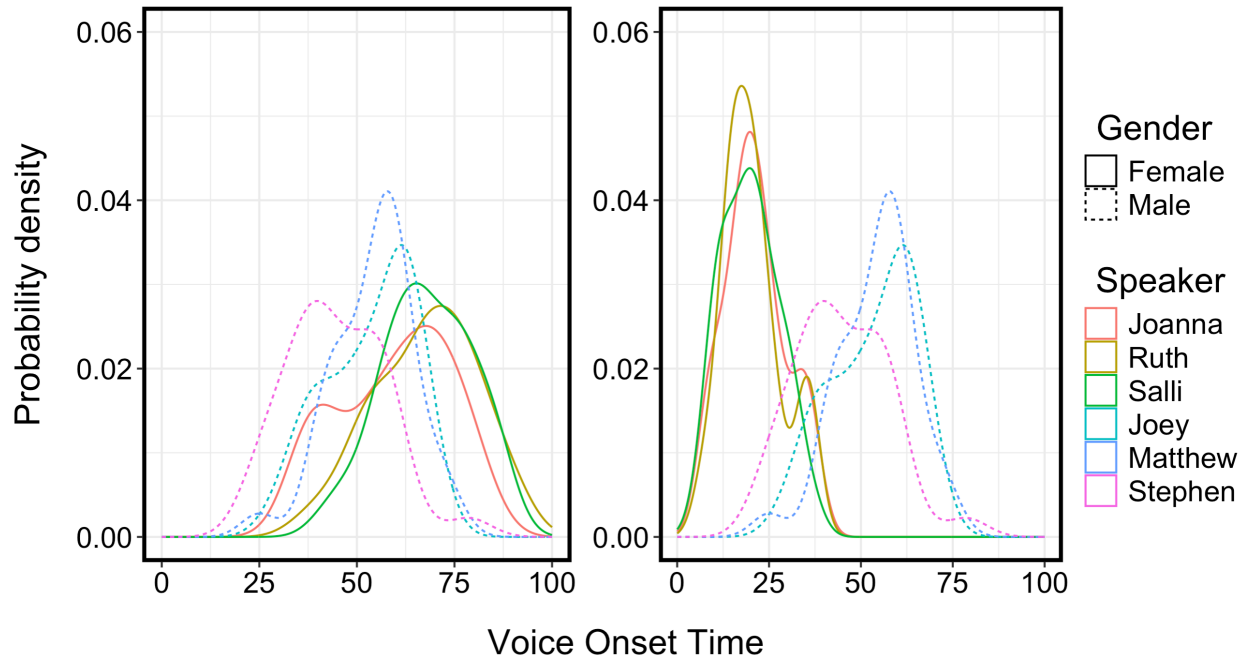


Figure 4.2: Distribution of Voice Onset Time for word-initial /p/ in the exposure stimuli. The left panel shows the unshifted female tokens, while the right panel shows the (/p/ to [b]) shifted female tokens. Both panels show identical distributions for the unshifted male tokens.

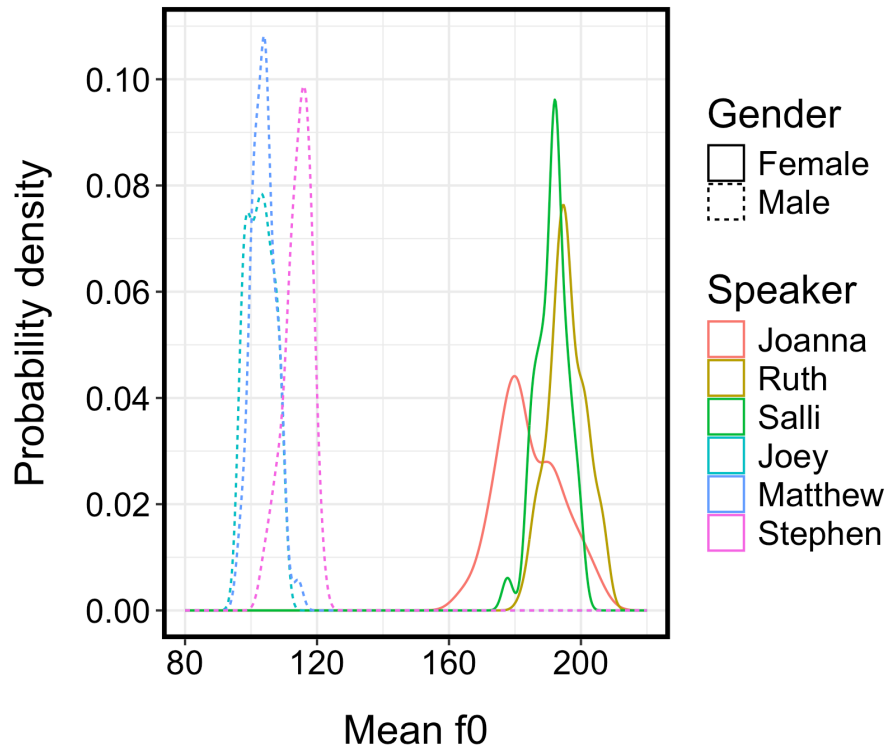


Figure 4.3: Distribution of sentence-level mean f0 in the exposure stimuli.

A custom-made Praat script measured the mean f0 over the entire duration of each exposure stimulus (Cohn, 2019) – the distributions are visualized in Figure 4.3 for each speaker (only the unshifted female stimuli were included in this analysis). Bonferroni-corrected paired t-tests confirm that, as expected (Titze, 1989), all three female speakers have a higher mean f0 than each of the male speakers (all $p < 0.001$).

4.2.2 Test stimuli

Development and norming

The test stimuli initially consisted of a 9-step “buy-pie” continuum for each of the six voices. However, with these initial continua, a norming study (discussed later on in this section) revealed a greater overall “pie” response for the female speakers compared to the male speakers. Since successful generalization is also indicated by a greater “pie” response for female than male speakers (see Section 4.1.3), the responses at each continuum step should ideally be matched by speaker gender to facilitate model interpretation later (a gender difference in an actual experiment would be attributed to an adaptation effect, rather than an inherent property of the stimuli).

The actual test stimuli in the final experiment consisted of 5-step “buy-pie” continua that were, in terms of their perceptual categorization, aligned as closely as possible by gender. The rest of this section details the procedure of the norming study and explains how the results were used to create the final 5-step continua.

For the norming study, the words “buy” and “pie” were separately typed into the Amazon Web Services console and downloaded for each of the six voices used in the current experiment (female: Joanna, Ruth, Salli; male: Joey, Matthew, Stephen). Similar to the exposure stimuli (see Section 4.2.1), all productions of “buy” and “pie” were generated with neural TTS,

converted from .mp3 to .wav files (Tomar, 2006), and set to 60 dB SPL (Boersma & Weenink, 2023). A 9-step “buy-pie” continuum was created in Praat for each voice, using the original “buy” and “pie” stimuli as endpoints (Winn, 2022). To create more naturalistic stimuli, both VOT and f0 were increased as the stimuli became more “pie”-like (Clayards, 2017).

The 9-step continua were presented to 120 listeners on Prolific through a self-paced, online Qualtrics survey. The demographic filters were used to narrow down the participant pool to individuals who were living in the United States, between 18 and 35 years old (inclusive), and whose first language was English. Participants were paid \$0.80 (approximately \$9 per hour) and provided informed consent at the outset of the study.

After informed consent, participants completed a categorization task in which they listened to a stimulus on each trial and stated whether they heard “buy” or “pie”. Each listener was randomly assigned to one of the six speakers and categorized each continuum step five times in a randomized order (9 steps * 5 repetitions = 45 total trials). Data from five individuals were removed who either self-reported a hearing difficulty ($n = 2$), self-reported being older than 35 years old ($n = 1$), or whose first language was not solely English ($n = 2$). The final analysis included 115 participants (gender: 53 women, 60 men, 2 non-binary; mean age: 29.27 years, sd age: 4.12; ethnicity: Asian = 27, Black = 15, Latino = 3, Mixed = 11, White = 59).

The aggregated results by speaker are shown in Figure 4.4 and by speaker gender in the left panel of Figure 4.5. Impressionistically, the left panel of Figure 4.5 indicates that while the results for Steps 1-3 and Steps 7-9 are evenly matched by speaker gender, Steps 4-6 induced a greater “pie” response for the female speakers than for the male speakers overall. To match the overall proportion of “pie” responses by gender as closely as possible, five specific steps were selected for each speaker (Joanna, Ruth, Stephen: 1, 4, 5, 6, 9; Salli: 1, 3, 4, 6, 9; Joey: 1, 3, 5, 6, 9; Matthew: 1, 4, 6, 7, 9). The aggregated 5-step continua are shown by speaker gender in the right panel of Figure 4.5.

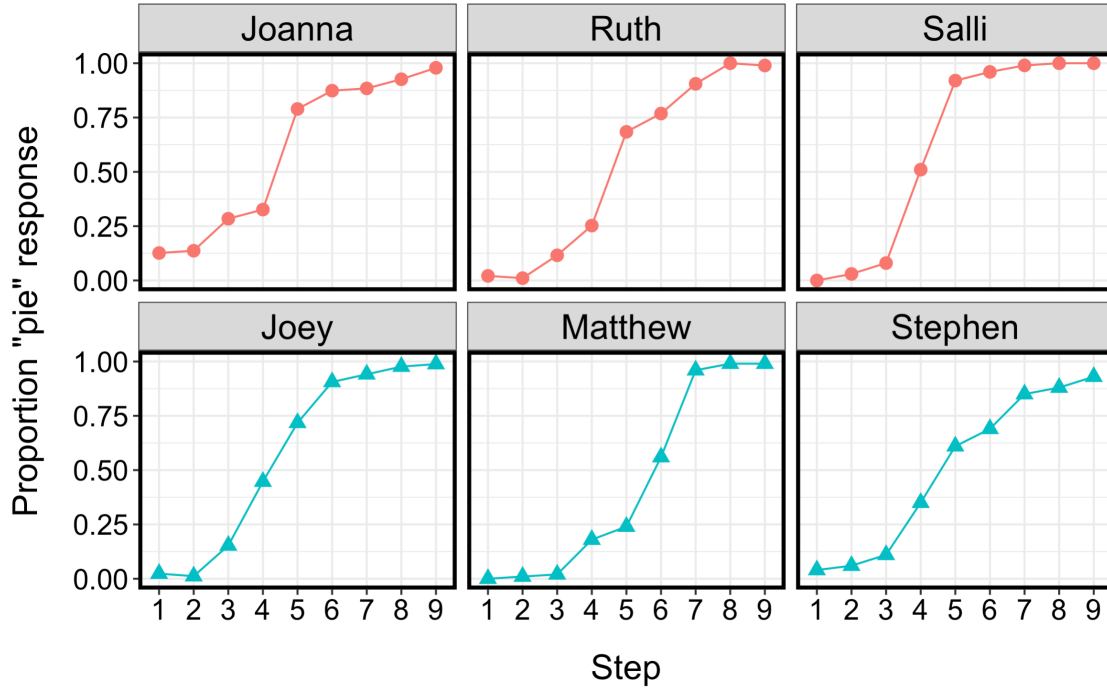


Figure 4.4: Aggregated results by Step and Speaker in the norming experiment (female = top row, red; male = bottom row, blue).

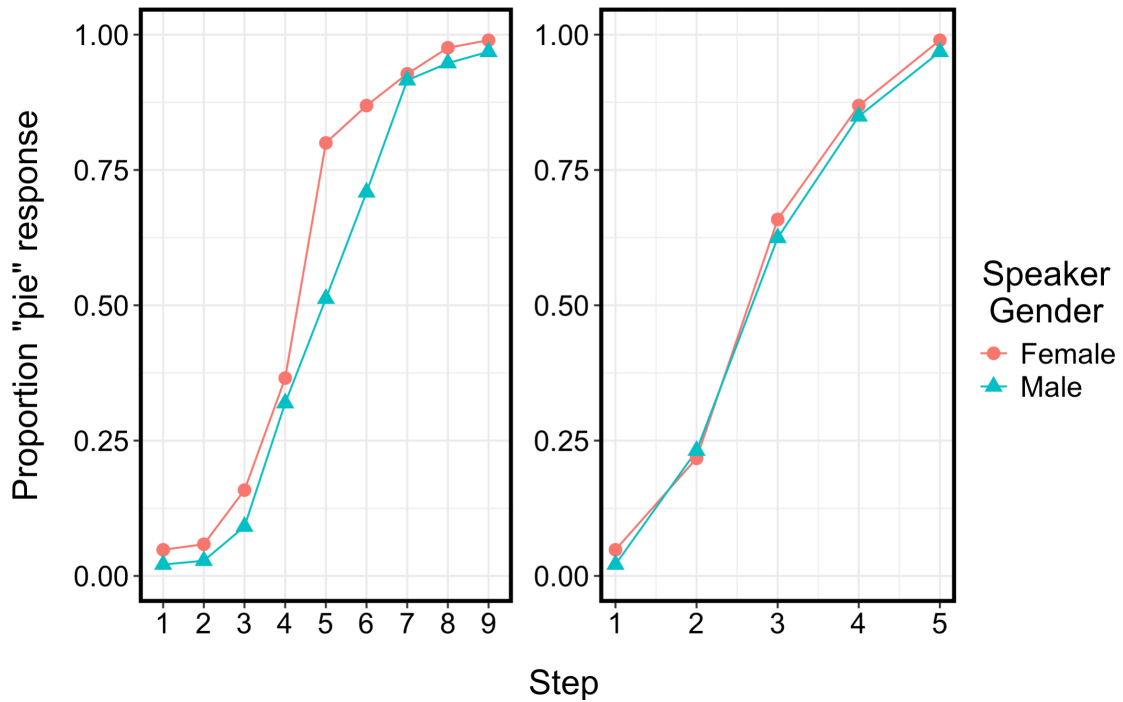


Figure 4.5: Aggregated results by Step and Speaker Gender (female = red; male = blue) for the original 9-step continuum (left panel) and the adjusted 5-step continuum used in Experiment 2 (right panel).

To verify the impressionistic observation that the 5-step continua in Figure 4.5 are gender-matched, a Bayesian mixed-effects logistic regression model was fitted (“pie” = 1, “buy” = 0) through the *brms* package (Bürkner, 2017) and Stan (Stan Development Team, 2023) in R (R Core Team, 2021) using identical priors as Section 10.5.2 of (Barreda & Silbert, 2023). As shown in Equation 4.1, the model had fixed effects of Step (within-participants; numeric, scaled, and centered), Speaker Gender (between-participants; sum-coded; Female, Male) and their interaction, as well as by-speaker and by-listener random intercepts and a by-listener random slope for Step.

$$\text{Resp.} \sim \text{Step} * \text{Speaker Gender} + (1 + \text{Step} \mid \text{Listener}) + (1 \mid \text{Speaker}) \quad (4.1)$$

The model revealed an effect of Step [β : 4.32, SE: 0.29, 95% highest density interval (HDI) = (3.79, 4.91)], indicating that “pie” responses increased with continuum step. There was no main effect of Speaker Gender [β : 0.05, SE: 0.54, 95% HDI = (−1.02, 1.09)] and no interaction [β : 0.01, SE: 0.22, 95% HDI = (−0.43, 0.44)]. Given the lack of gender effect, the 5-step continua were deemed suitable for the test phase of the current experiment.

Acoustic analysis

The VOT and onset f0 of the final stimuli are shown in Table 4.2 and Table 4.3, respectively. In alignment with prior work (Clayards, 2017), VOT and f0 both increase gradually from Step 1 (most “buy”-like) to Step 5 (most “pie”-like) for all speakers.

Bonferroni-corrected paired t-tests were used for the acoustic analysis. Akin to the exposure stimuli, all three female speakers had a higher mean f0 than each of the male speakers (all $p < 0.001$). No significant VOT differences were observed for the test phase continua between any of the speakers after the Bonferroni correction (all $p > 0.006$).

Table 4.2: Voice Onset Time (ms) of at each step along the final version of the continua.

Step	Joanna	Ruth	Salli	Joey	Matthew	Stephen
1	9.81	14.88	10.66	14.65	8.34	12.10
2	31.37	30.51	24.21	24.50	28.77	29.26
3	35.10	38.77	30.78	32.54	43.27	33.83
4	44.49	45.92	46.78	44.65	51.94	43.84
5	67.77	68.83	71.31	64.23	66.69	68.22

Table 4.3: Onset f0 (Hz) of each step along the final version of the continua.

Step	Joanna	Ruth	Salli	Joey	Matthew	Stephen
1	173.92	227.11	199.62	104.71	135.59	139.48
2	183.23	235.08	206.43	110.18	142.07	148.73
3	186.25	238.66	209.25	116.08	148.46	148.89
4	188.10	241.63	214.87	119.37	150.30	153.37
5	195.90	251.62	222.93	126.57	154.38	164.16

4.2.3 Participants

This study has a (mostly) between-participant design (see Section 4.2.4), so several hundred listeners are needed to achieve sufficient power (Brysbaert & Stevens, 2018). However, a large sample size is arguably even more crucial than in Experiment 1 – since participants are now being exposed to a novel phonetic shift, adaptation might become challenging, resulting in a small effect size (and therefore, an *extremely* large number of listeners to statistically detect an effect). The present experiment aimed for around 100 listeners per condition, doubling a comparable adaptation study with 80% power (Cummings & Theodore, 2023).

Using identical demographic filters as the norming study described above (Section 4.2.2), 839 crowdsourced workers were recruited from Prolific, none of whom had completed the norming study. All participants were compensated with \$1.80 (around \$9 per hour) and provided informed consent before the experiment. Anyone who self-reported a hearing difficulty ($n = 16$), whose strongest self-reported language was not solely English ($n = 23$), or who self-reported being older than 35-years-old ($n = 1$) was omitted from the analysis. After these initial removals, exposure phase accuracy was calculated, and anyone whose accuracy was more than three standard deviations below the mean was taken out of the data set ($n = 4$). The final exclusion criterion was an atypical response in the test phase. As discussed in Section 4.2.4, the test phase consisted of a categorization task, where listeners heard two blocks of a 5-step “buy-pie” continuum. For each participant, the responses for both blocks were combined, and a “difference score” was calculated by subtracting the percentage of “pie” responses at Step 5 (expected to be near 100%) by the percentage of “pie” responses at Step 1 (expected to be near 0%). Anyone whose difference score was more than three standard deviations below the mean was excluded ($n = 24$). The average difference score after exclusions was 96.74%.

The analysis included responses from 771 individuals (gender: 422 women, 330 men, 19 non-binary; mean age = 28.15 years, sd age = 4.51; ethnicity: Asian = 89, Black = 125, Latino = 29, Mixed = 99, Native American or Alaska Native = 3, Native Hawaiian or Pacific Islander = 3, White = 423). As there were three between-participant variables with two levels each (eight cells; Section 4.2.4), the sample size goal was met (~96 listeners per cell).⁷

Note that the overall number of participants is marginally different from the original version of this work ($n = 769$ total in Aoki and Zellou (2025)). The reason is that, since the two experiments in the original study have now been combined into one large experiment for

⁷Unlike Experiment 1, a geographic analysis is not quite as relevant here (i.e., a visualization of the participants’ approximate location of residence and urban/rural orientation; see Figure 3.1). Since this experiment presents a completely novel phonetic shift (as opposed to naturalistic L2-accented speech), all listeners are equally unfamiliar with the variant at hand, rendering a geographic analysis less informative.

concision, the cutoffs for exclusion have ever-so-slightly shifted. Adding two extra participants does not meaningfully adjust any of the overall conclusions in the current study.

4.2.4 Procedure

The study was conducted through a self-paced, online Qualtrics survey. After informed consent, participants were asked to wear headphones and to take the study in a quiet room with no background noise. These instructions were followed by a sound calibration procedure (identical to Experiment 1; see “Pre-experiment” within Section 3.2.3).⁸ Participants then completed an exposure phase, a test phase, and a demographic questionnaire.

The exposure phase consisted of 48 trials presented in a pseudo-randomized order (16 filler, 32 critical; see Section 4.2.1). Two between-participant variables were manipulated in exposure: Talker Variability and Shift Type. In terms of Talker Variability, listeners either heard a single talker per gender (one female, one male) or multiple talkers per gender (two female, two male); the instructions explicitly stated the number and gender of the speakers that would be heard. The exposure stimuli were evenly divided across the talkers, with sentence content and talker being evenly counterbalanced.

Participants were also assigned to either a Female Shifted or No Shift condition, which varied the production of critical items in exposure (the fillers remained the same across conditions). All exposure speakers in the No Shift condition produced a canonical /p/ across all critical items. In the Female Shifted condition, the female speaker(s) produced /p/ as [b] (e.g., “paw” as *baw*), while the male speaker(s) produced a canonical /p/. The relative amount of exposure to /p/-shifted words remained the same across listeners in the Female Shifted condition, regardless of the number of the exposure talkers (the /p/-shifted stimuli were either evenly distributed across two female talkers or just produced by one female talker).

⁸Note that this sound calibration procedure did not likely interfere with the effects in the experiment, since /p/ was never presented (the “Matthew” voice from Amazon Polly said “She asked about the host”).

The procedure of the exposure phase adhered to Zellou et al. (2023). Listeners heard a semantically predictable sentence on each trial (e.g., “The dog had a furry paw”) and were asked to identify the final keyword from two options: (i) the target (“paw”); (ii) a competitor (“pawn”).⁹ To promote the mapping from /p/ to [b] in the Female Shifted condition, all target and competitor items for the critical stimuli were always real words with initial /p/, both phonologically and orthographically (see Table A.2 for the competitors). The targets and competitors were randomly placed on either the left or right side of the screen.

The exposure phase was followed by a test phase. There were two test blocks (one with a novel female speaker and the other with a novel male speaker); as a third between-participant variable, block order was evenly counterbalanced, with participants either hearing the novel female speaker first (Female → Male) or the novel male speaker first (Male → Female). On each trial, listeners heard a single stimulus from a “buy”-“pie” continuum were asked whether they heard “buy” or “pie”. The test stimuli came from a 5-step buy-pie continuum (see Section 4.2.2), with each step presented nine times per block in a pseudo-randomized order (45 trials per block, 90 trials total).

Importantly, avoiding the confound in Bradlow and Bent (2008), the three female and three male voices were always evenly counterbalanced as either an exposure or test talker (see Section 2.3.2 for discussion about talker counterbalancing).

Before delving into the statistical analysis, two questions should be clarified about the procedure. First, why does the exposure phase involve a keyword identification task instead of a more conventional lexical decision task? Second, why is talker variability not being simulated, in contrast to Experiment 1?

⁹A custom-made JavaScript function forced participants to listen to the entire sentence before proceeding to the next trial. This prevented the potential selection of the correct option *before* actually hearing the final keyword (e.g., the fragment “The dog had...” seems much more likely to be associated with “paw” than “pawn”; certain participants might just hear “The dog had...” and correctly select “paw” without ever listening to the full sentence; Altmann & Kamide, 1999). Not including this JavaScript function could have been deeply problematic – if listeners in the Female Shifted condition skipped to the next trials without hearing the final keywords, then they would never actually be exposed to the phonetic shift.

Rationale for presenting a keyword identification task in exposure

Lexically-guided perceptual learning can only take place if the critical items in the exposure phase are processed as real words (Babel et al., 2019; Norris et al., 2003). An auditory lexical decision task is usually appropriate, since the subtle phonetic shifts in most adaptation studies are not enough to alter lexical judgments (e.g., “dino?aur” is still considered a real word, even if “?” is replaced by an ambiguous fricative between /s/ and /ʃ/; Kraljic & Samuel, 2007).

However, when a more radical phonetic shift is presented (e.g., the /p/ to [b] shift in the current study), a change in paradigm is vital for listeners to store the shifted stimuli as real words (Charoy & Samuel, 2023; Sumner, 2011). Trying to implement a lexical decision task would make it impossible to differentiate nonwords from words with entirely remapped sounds, especially if they are presented in isolation (e.g., “flirp” vs. “brint”).

The current task (identifying keywords at the end of semantically predictable sentences) is suitable because semantic context strongly biases a “real word” interpretation despite the presence of an extreme phonetic shift. For example, even if /p/ is realized as [b] in “The novel is now in *brint*”, listeners can infer from the prior word “novel” that *brint* must be a shifted version of “print” (Zellou et al., 2023). The intended word becomes even more transparent when listeners are given a binary choice, and the other option is clearly different semantically and phonologically (e.g., between “print” and “prize”, “print” is the only reasonable completion of “The novel is now in. . .”).

The design of the exposure phase also has two other incidental advantages. First, compared to isolated words, sentence-length stimuli provide even more acoustic cues to speaker gender through suprasegmental information (Holliday, 2021). Exposure to full sentences could help listeners clearly differentiate the female and male speakers, which is particularly important when focusing on a gender-mediated phonetic shift.¹⁰ Second, the task itself is extremely straightforward, as it is self-paced and has a very clear correct answer on each trial. Given

¹⁰Thank you to Meredith Tamminga for this point.

that high cognitive load can block perceptual learning (Zhang & Samuel, 2014), and given that listeners have already been burdened with the challenge of adapting to a completely novel phonetic shift, reducing task difficulty could be critical for observing robust generalization.

Rationale for *not* presenting apparent talker variability

Given the extensive discussion of *apparent* talker variability in Experiment 1, one might wonder why talker variability is not being simulated in the present study. There are two reasons why this is so – but for full disclosure, Experiment 1 was conducted and published *after* Experiment 2, meaning that all justifications in this section are necessarily post-hoc.

The first reason is that the method for simulating talker variability is highly impractical in the current context. Experiment 1 relied on the Change Gender function in Praat to take a single male talker and create two novel “female” talkers through the manipulation of f0 and formants. However, the design of this study requires the creation of *three* novel talkers – since the “multiple talkers per gender” condition presents two different female and two different male speakers in exposure, it would be necessary to take a single male talker and generate both two novel “female” talkers *and* a novel male talker. In the author’s personal experience, the latter is extremely challenging because one has to ensure that the novel male talker simultaneously: (i) sounds different enough from the original to be perceived as a new speaker; (ii) still sounds similar enough to the original to be robustly categorized as “male”; (iii) still sounds naturalistic. These issues are not necessarily surprising because the goal of the Change Gender function is really to *change* gender, not to keep gender the same and only adjust apparent talker identity.

The other reason for not simulating talker variability is that it is no longer theoretically important. In short, apparent talker variability was introduced to control for potential effects of cue weighting in L2-accented stimuli (e.g., the greater reliance of Mandarin-accented English speakers on burst and closure duration than vowel duration in word-final voicing;

see Section 3.1.1 for details). The current study, however, presents a very different type of variation (replacing one sound with another, not adjusting cue weights), so it becomes unclear how apparent talker variability would be productive for teasing apart different theories.

4.2.5 Statistical analysis

Accuracy in the exposure phase was essentially at ceiling (99.74% aggregated across all conditions). Due to this lack of variation, the exposure phase results were not analyzed statistically (attempting to do so would likely result in highly unreliable model estimates due to the problem of complete separation in logistic regression; Agresti, 2002).

The test phase responses were coded binomially as “pie” (= 1) or “buy” (= 0) and analyzed with Bayesian mixed-effects logistic regression in R (R Core Team, 2021) using *brms* (Bürkner, 2017) and Stan (Stan Development Team, 2023). The priors were identical to the norming study described in Section 4.2.2. Similar to Experiment 1, effects are interpreted as “meaningful” if credible intervals do not cross zero (see Section 3.2.4 above for discussion about Bayesian model interpretation).

The published version of this chapter framed the current study as two separate experiments with separate analyses (“Experiment 1” was titled “Exposure to a single talker per gender”; “Experiment 2” was titled “Exposure to multiple talkers per gender”; Aoki & Zellou, 2025a). This is an extremely common practice in the adaptation literature – the levels of the key variable of interest are divided into multiple sub-experiments and analyzed on their own.¹¹ For example, in Baese-Berk et al. (2013), performance on the two transcription tasks (one with a Slovakian-accented talker and another with a Mandarin-accented talker) was assessed in two different models, even though both shared an identical structure (Kraljic and Samuel (2007), Zellou et al. (2023), and numerous other papers have the same type of analysis).

¹¹The author is not aware of any explicit justification for this practice. Having multiple sub-experiments is presumably intended to make journal articles easier to read, since it breaks apart potentially complicated experimental designs into more manageable chunks.

Following prior work, the original paper took the most important variable (talker variability) and divided the study into two sub-experiments (single talker per gender and multiple talkers per gender). The models had the exact same structure (see Equation 4.2 below). There were main effects of Step (within-participants; scaled and centered), Speaker Gender (within-participants; Female, Male), Block Order (between-participants; Female $\rightarrow$ Male, Male $\rightarrow$ Female), and Shift Type (between-participants; Female Shifted, No Shift). Step was treated as a numeric variable, while the other main effects were sum-coded, categorical variables. All possible interactions were included. The random effects structure consisted of by-listener and by-speaker random intercepts, as well as by-listener random slopes for Step, Speaker Gender, and their interaction.

$$\begin{aligned} \text{Resp.} \sim & \text{Step} * \text{Speaker Gender} * \text{Block Order} * \text{Shift Type} + \\ & (1 + \text{Step} * \text{Speaker Gender} \mid \text{Listener}) + (1 \mid \text{Speaker}) \end{aligned} \quad (4.2)$$

The issue with a two-part analysis is that interactions between Talker Variability and other predictors are never explored. As explained below, the main interest is really in the three-way interaction between Talker Variability, Speaker Gender, and Shift Type, not necessarily in separately modeled two-way interactions between Speaker Gender and Shift Type.¹²

To remain faithful to the original article while also integrating feedback post-publication, this dissertation: (i) maintains the two-part analysis described above; (ii) adds a combined model that is identical to Equation 4.2, but includes Talker Variability as a predictor (between participants; Single Talker Per Gender, Multiple Talkers Per Gender), along with all possible interactions (Equation 4.3).

$$\begin{aligned} \text{Resp.} \sim & \text{Step} * \text{Speaker Gender} * \text{Block Order} * \text{Shift Type} * \text{Talker Variability} + \\ & (1 + \text{Step} * \text{Speaker Gender} \mid \text{Listener}) + (1 \mid \text{Speaker}) \end{aligned} \quad (4.3)$$

¹²Thank you to Xin Xie and Chigusa Kurumada for this point.

4.3 Main results

This section is divided into three subsections below: (i) “Exposure to a single talker per gender” (Section 4.3.1); (ii) “Exposure to multiple talkers per gender” (Section 4.3.2); (iii) “Combined analysis” (Section 4.3.3). Since all of the models are fairly complex (each contains over a dozen fixed effects), the model output is provided in Appendix B for readability.

Only the credible intervals that are theoretically relevant and/or do not cross zero are reported for clarity. As discussed in Section 4.1.3, generalization is indicated by a greater “pie” response for the novel female talker relative to the novel male talker in the Female Shifted condition only (there should be no difference between the novel talkers in the No Shift control condition). This implies that the “theoretically relevant” effects are any interactions involving Speaker Gender and Shift Type (both two-way and higher order effects).

One last point – all of the models were refitted, implying that the estimates and intervals are not exactly as they were originally (Aoki & Zellou, 2025a). Rerunning the models was necessary because the number of participants being analyzed is now slightly different (since listeners in the single talker and multiple talkers per gender conditions are now being combined into one large study, the exact cutoffs for exclusion based on the exposure and test phases are now slightly different as well; see Section 4.2.3 for details). These marginal differences have no effect on the overall conclusions of this study, but are noted here for the record.

4.3.1 Exposure to a single talker per gender

The results are visualized for the “single talker per gender” condition in Figure 4.6; the model output is shown in Table B.1 in Appendix B.¹³ The consistent effect of Step [β : 4.91, SE: 0.12, 95% HDI = (4.69, 5.14)] demonstrated that as continuum step increased, listeners

¹³Figures 4.6-4.8 are essentially identical to Figures 1-3 in the original paper (Aoki & Zellou, 2025a), except that they now show bootstrapped confidence intervals (the previous figures only showed point estimates). Thank you to Chigusa Kurumada for providing the code to create these intervals in *ggplot* graphs!

selected “pie” more often. There were no other robust main effects or interactions. Notably, no interactions between Speaker Gender and Shift Type surfaced, implying that the exposure conditions (Female Shifted and No Shift) did not differ in “pie” responses by gender – this indicates a lack of generalization among listeners exposed to a single talker per gender.

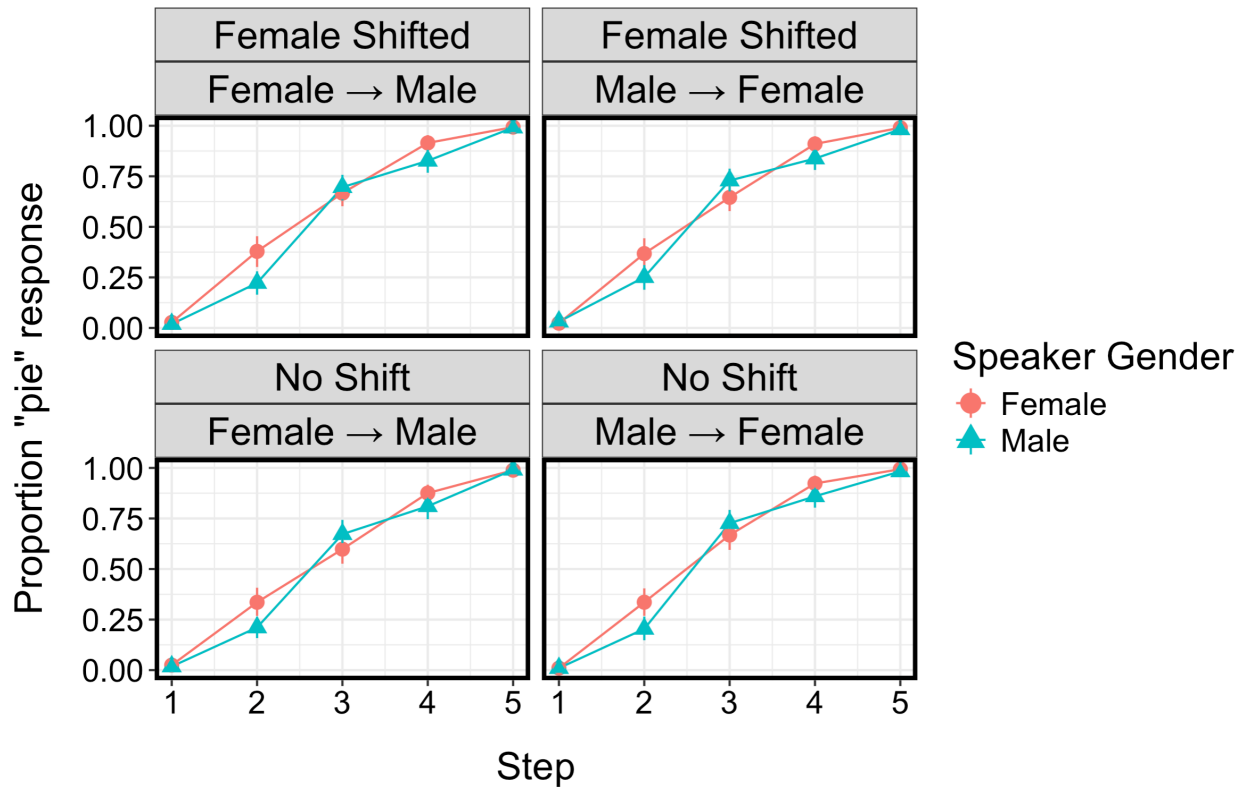


Figure 4.6: Aggregated results in the “single talker per gender” exposure condition by Step, Speaker Gender (female = red, male = blue), Block Order (Female → Male, Male → Female), and Shift Type (Female Shifted, No Shift). Error bars are 95% confidence intervals bootstrapped over by-participant means.

4.3.2 Exposure to multiple talkers per gender

The results are visualized for the “multiple talkers per gender” condition in Figure 4.7; the model output is shown in Table B.2 in Appendix B. A main effect of Step emerged, indicating that listeners provided more “pie” responses as step number increased [β : 4.60, SE: 0.11, 95% HDI = (4.39, 4.82)]. There was also a main effect of Block Order, such that on average,

listeners who heard the female block first made fewer “pie” responses [β : -0.30, SE: 0.11, 95% HDI = (-0.51, -0.09)]. A three-way interaction between Step, Speaker Gender, and Block Order [β : -0.16, SE: 0.07, 95% HDI = (-0.30, -0.03)] indicated: (i) a steeper slope of Step for the novel male speaker compared to the novel female speaker in the Female $\rightarrow$ Male block order [β : -0.31, SE: 0.11, 95% HDI = (-0.52, -0.09)]; (ii) no difference in the slope of Step by gender in the Male $\rightarrow$ Female block order [β : 0.02, SE: 0.11, 95% HDI = (-0.20, 0.24)].

Critically, there was a three-way interaction between Speaker Gender, Block Order, and Shift Type [β : 0.17, SE: 0.07, 95% HDI = (0.03, 0.33)]. Using the *hypothesis* function from *brms*, the two-way interaction between Speaker Gender and Shift Type was examined for each block order. An interaction with a positive coefficient was found for listeners who heard the female block first [β : 0.26, SE: 0.11, 95% HDI = (0.05, 0.47)], but not for listeners who heard the male block first [β : -0.09, SE: 0.10, 95% HDI = (-0.29, 0.11)]. In other words, relative to the No Shift control condition, participants in the Female Shifted exposure condition provided more “pie” responses for the female speaker than for the male speaker, but only if they heard the female block first (i.e., Female $\rightarrow$ Male, not Male $\rightarrow$ Female).

This result is visually evident – looking just at the top-left panel of Figure 4.7, it becomes clear how listeners assigned to the Female Shifted exposure condition and Female $\rightarrow$ Male test block order provided a greater “pie” response for the female speaker (red line) relative to the male speaker (blue line). By contrast, in the other three panels, there seems to be no appreciable difference in categorization between the female and male speakers.¹⁴

¹⁴When the top-left panel of Figure 4.7 is compared to the other three panels, the adaptation effect might appear more nuanced – the three-way interaction seems to be driven by a reduced “pie” response for the novel male speaker, rather than a greater “pie” response for the novel female speaker (thank you to Reviewer 2 from *Glossa Psycholinguistics* for this point). This was investigated through an additional *hypothesis* examining the two-way interaction between Block Order and Shift Type for each gender. There is ultimately no obvious evidence to support this impressionistic observation, as there were no robust interactions found for either the male test speaker [β : -0.21, SE: 0.14, 95% HDI = (-0.50, 0.06)], or the female test speaker [β : 0.13, SE: 0.12, 95% HDI = (-0.11, 0.38)]. That said, there is a marginal interaction for the male test speaker, which accounts for why the effect of Block Order in the main model has a negative coefficient.

In summary, there was evidence of generalization in the “multiple talkers per gender” condition (unlike the “single talker per gender” condition), but the effect was modulated by test block order.

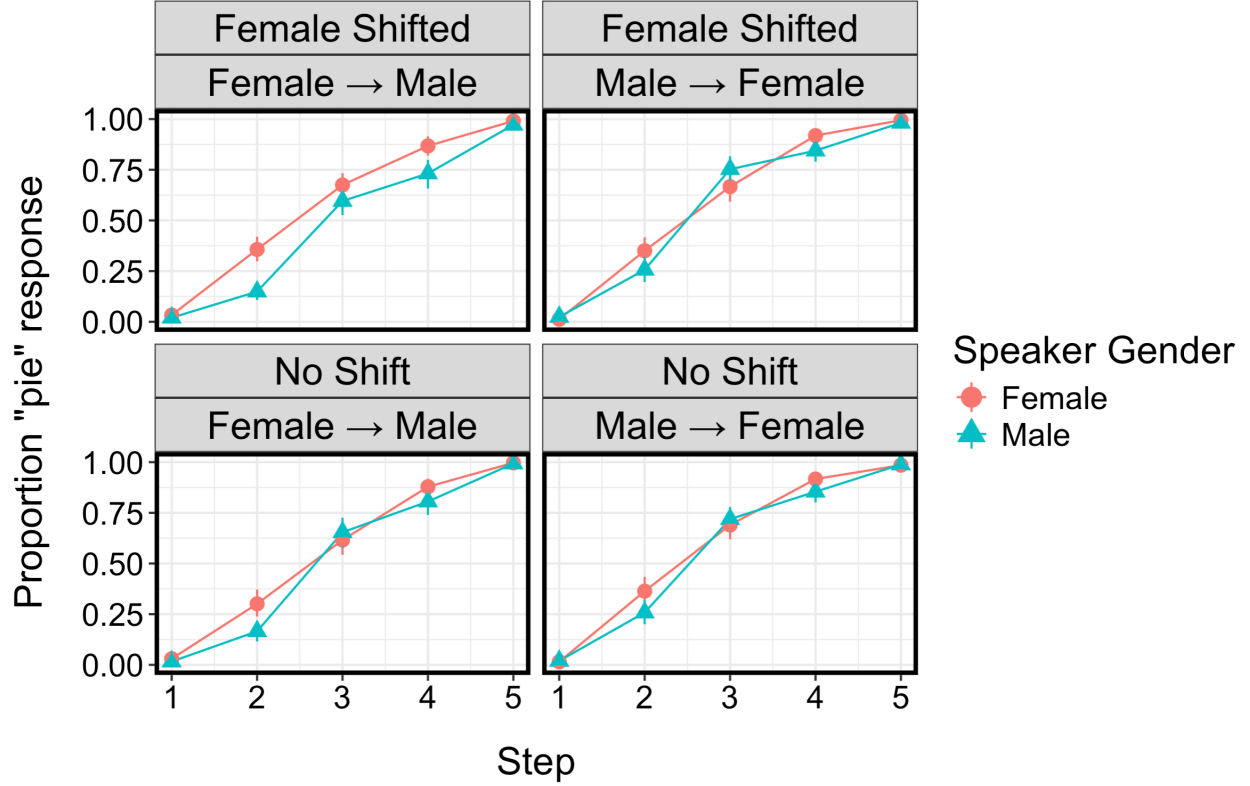


Figure 4.7: Aggregated results in the “multiple talkers per gender” exposure condition by Step, Speaker Gender (female = red, male = blue), Block Order (Female → Male, Male → Female), and Shift Type (Female Shifted, No Shift). Error bars are 95% confidence intervals bootstrapped over by-participant means.

4.3.3 Combined analysis

This subsection is where this dissertation departs from the original article (Aoki & Zellou, 2025a). Table B.3 in Appendix B exhibits the results of the “combined” analysis (i.e., that compares the “single talker per gender” and “multiple talkers per gender” conditions in the same model). Only the interactions between Speaker Gender and Shift Type will be discussed here, as they are the most pertinent theoretically.

A three-way interaction was found in the combined model between Speaker Gender, Block Order, and Shift Type [β : 0.11, SE: 0.05, 95% HDI = (0.01, 0.22)] – the interpretation is the same as the previous subsection (Section 4.3.2) and is not repeated here for concision.

When the data were modeled in two separate stages, the critical three-way interaction only appeared in the “multiple talkers per gender” condition, but not in the “single talker per gender” condition. Is this effect robust enough to appear in the combined model as a four-way interaction between Speaker Gender, Block Order, Shift Type, and Talker Variability? The answer is no – the credible interval for the four-way interaction crosses zero, meaning that the effect is not very robust [β : 0.06, SE: 0.05, 95% HDI = (-0.04, 0.17)].

4.4 Interim discussion

Taking the results from all of Section 4.3 together (i.e., both the separated analyses and the combined analysis), two findings emerge. First, there was evidence of generalization – listeners who were exposed to a gender-mediated shift (female = /p/ as [b], male = canonical /p/) provided more “pie” responses for the novel female test talker than for the novel male test talker, as long as they were in the “Female → Male” test block order (i.e., they heard the novel female talker first). Second, the effect only seemed to appear in the “multiple talkers per gender” exposure condition (not the “single talker per gender” exposure condition), but the effect size is incredibly weak.¹⁵

The current study very tentatively supports a numerosity account over a sufficient similarity account. It seems that multiple-talker exposure may be necessary for generalization when listeners are presented with a completely unattested variant. Given their unfamiliarity with a gendered /p/ to [b] shift, listeners were unwilling to generalize from a single female talker

¹⁵Although reporting separated analyses is not necessarily an unacceptable practice, the author realizes upon reflection that the published article in *Glossa Psycholinguistics* should have included a combined model for context. Thank you again to Xin Xie and Chigusa Kurumada for their suggestion to fit and add this extra model to this dissertation.

to a novel female talker due to their uncertainty about whether the shift was idiosyncratic or generalizable. This ambiguity makes it important to hear multiple types of evidence for generalization (i.e., hearing the shift in multiple female talkers *and* acoustic similarity between the female talkers across the exposure and test phases).

There are two caveats of the analysis and results so far. As alluded to above, generalization was somewhat short-lived, with a categorization shift only occurring among listeners assigned to the “Female $\rightarrow$ Male” block order, not among those placed in the “Male $\rightarrow$ Female” order (i.e., when there is a delay in presenting the novel female talker). There are at least four different, non-mutually exclusive explanations for this order effect, which are unpacked below in the main discussion (Section 4.6.1).

A second caveat is that only the aggregated results have been shown so far, not the results by individual test talker (as mentioned in Section 4.2.4, the three female and three male talkers were fully counterbalanced so that each was presented as a test talker equally often). There is a minor possibility that the results genuinely support a sufficient similarity account (rather than a numerosity account), but that the aggregated analysis is masking this finding. The next section presents a post-hoc analysis that investigates this idea in greater depth.

4.5 Post-hoc analysis

A strong version of a numerosity account should ideally be robust to talker counterbalancing. More specifically, generalization should occur for zero exposure-test talker pairs in the “single talker per gender” condition and for all possible talker combinations in the “multiple talkers per gender” condition. This is depicted by the hypothetical set of results in Table 4.4.

Table 4.4: Hypothetical set of results supporting a strong version of a numerosity account (i.e., where multiple-talker exposure is *necessary* for generalization). The column and row labels stand for the possible exposure and test talkers, respectively. The values within each cell refer to presence (‘Yes’) or absence (‘No’) of generalization in this scenario. Dashes reflect the lack of talker-specific conditions in the current study (i.e., the exposure talkers were never presented in the test phase).

	Joanna	Ruth	Salli	Joanna + Ruth	Joanna + Salli	Ruth + Salli
Joanna	—	No	No	—	—	Yes
Ruth	No	—	No	—	Yes	—
Salli	No	No	—	Yes	—	—

Although the output of the aggregated models very tentatively support a numerosity account (see Sections 4.3-4.4), the results could also be influenced by sufficient similarity. This latter possibility might arise, for example, if two of the female talkers are similar to each other and both are different from the third female talker.

To more clearly illustrate this idea, Table 4.5 exhibits a hypothetical scenario in which Ruth and Salli are similar to each other, but both differ from Joanna. A sufficient similarity account predicts that the degree of similarity between exposure and test talkers is what mediates cross-talker generalization, not the number of exposure talkers (Xie et al., 2021; Xie & Myers, 2017) – if so, then two of six exposure-test talker combinations should show generalization in the “single talker per gender” condition (Ruth $\rightarrow$ Salli, Salli $\rightarrow$ Ruth). Moreover, given the same similarity relations, two of three exposure-test combinations in the “multiple talkers per gender” condition should also result in generalization (Joanna + Ruth $\rightarrow$ Salli, Joanna + Salli $\rightarrow$ Ruth), since at least one exposure talker is similar to the test talker. In an analysis that glosses over the individual exposure-test groupings (Section 4.3), an aggregated generalization effect could be shown with exposure to multiple talkers per gender, since the majority of participants (two-thirds) would hear at least one exposure talker who is similar to the test talker. By contrast, among participants exposed to a single

talker per gender, only a minority (one-third) would be presented with sufficiently similar exposure and test talkers, leading to a null aggregate effect.

Table 4.5: Hypothetical set of results that support a sufficient similarity account *and* still lead to the aggregated results in Section 4.3. The layout of the table is identical to Table 4.4.

	Joanna	Ruth	Salli	Joanna + Ruth	Joanna + Salli	Ruth + Salli
Joanna	—	No	No	—	—	No
Ruth	No	—	Yes	—	Yes	—
Salli	No	Yes	—	Yes	—	—

Although the cursory acoustic analysis in Sections 4.2.1 and 4.2.2 confirmed that all of the female speakers have a higher f0 than the male speakers, certain female speakers might still be insufficiently similar to each other in additional variables besides f0. A separate post-hoc analysis is therefore needed to ascertain whether the overall results are truly driven by numerosity or by sufficient similarity in disguise.

Since speech contrasts are highly multidimensional (Schertz & Clare, 2020) and since there are individual differences among listeners in cue weighting (Kapnoula et al., 2017), it is difficult in practice to determine which specific acoustic cues are being leveraged in judgments of cross-talker “similarity”. The current post-hoc analysis avoids this issue by adopting a perceptual approach – two individuals are assumed to be sufficiently similar to each other if exposure to one talker leads to cross-talker generalization for the other.¹⁶

¹⁶One possible question the reader might be thinking is why a by-talker post-hoc analysis is being included for Experiment 2, but not for Experiment 1, which also evenly counterbalanced the talkers across the exposure and test phases. The first reason involves a methodological difference – talker variability was *simulated* in Experiment 1. This implies that, unlike the current study, the acoustic similarity relations in the single-talker and “multiple”-talker exposure conditions were not substantively different, nullifying the concerns laid out in this section (each of the “different” talkers in the “multiple”-talker conditions were, in theory, equally similar to each other and to the other talkers). The second reason has to do with statistical power. In the critical conditions of Experiment 1, there were 12 possible counterbalancing combinations – one of the four talkers was randomly chosen for the exposure phase and then one of the remaining three talkers was presented at test. Since there were around 80 participants per condition, 12 combinations would result in only 6-7 listeners per post-hoc model (such dearth of statistical power contrasts with the post-hoc models in Experiment 2; see the following footnote).

4.5.1 Statistical analysis and results

Section 4.3 showed tentative evidence of generalization in the “multiple talkers per gender” condition, but only among listeners who heard the novel female first. Given this order effect, the scope of this post-hoc analysis is narrowed to the “Female $\rightarrow$ Male” block order. Since there was no strong evidence of Step interacting with both Speaker Gender and Shift Type, Step was omitted as a predictor from the post-hoc models.¹⁷

As depicted in Equation 4.4, all post-hoc models contained fixed effects of Speaker Gender (within-participants; Female, Male), Shift Type (between-participants; Female Shifted, No Shift), and their interaction, by-speaker and by-listener random intercepts, and by-listener random slopes for Speaker Gender. A separate model was fitted for each of the nine possible combinations of female talkers (i.e., each non-empty cell in Tables 4.4 and 4.5).¹⁸

$$\begin{aligned} \text{Resp.} \sim & \text{Speaker Gender} * \text{Shift Type} + \\ & (1 + \text{Speaker Gender} \mid \text{Listener}) + (1 \mid \text{Speaker}) \end{aligned} \tag{4.4}$$

If cross-talker generalization has occurred, then the interaction between Speaker Gender and Shift Type should have a positive coefficient (i.e., a greater “pie” response for the novel female talker than the novel male talker, but only in the Female Shifted exposure condition).

¹⁷Although this post-hoc analysis decreases power, the number of participants in each model (Single Talker Per Gender: $n \approx 33$; Multiple Talkers Per Gender: $n \approx 62$) is still comparable to prior work with 80% power (e.g., $n = 50$ in Cummings and Theodore (2023)).

¹⁸Xin Xie and Chigusa Kurumada made an important point post-publication – there were also three male talkers who were evenly counterbalanced across exposure and test, and a post-hoc analysis that just focuses on the the possible combinations of female talkers ignores the possible role that the male talkers might be having on adaptation. Although it would have been ideal to include all of the possible counterbalancing combinations (for both the female and male talkers), that would have decreased power significantly (e.g., there end up being 36 possible combinations in the “single talker per gender” condition, instead of just the 6 possible combinations shown in Tables 4.4 and 4.5). A more thorough post-hoc analysis with all possible combinations of exposure and test talkers is left for future experiments with a larger number of total participants.

Table 4.6 shows the credible intervals for the interaction terms in each post-hoc model.¹⁹ Among listeners exposed to multiple talkers per gender, robust interactions were present for two of three combinations (Joanna + Salli $\rightarrow$ Ruth, Joanna + Ruth $\rightarrow$ Salli). No exposure-test talker pair facilitated generalization in the “single talker per gender” condition.

These results imply that, consistent with a numerosity account, multiple-talker exposure seems truly necessary for generalization in the current experiment. The tentative nature of the talker variability effect in Section 4.3.3 is likely due to the absence of generalization in the “Ruth + Salli $\rightarrow$ Joanna” group.

Table 4.6: Summary statistics for the interaction between Speaker Gender and Shift Type in each post-hoc model. Credible intervals that do not cross zero are in bold.

Exposure Talker(s)	Test Talker	Estimate	Est. Error	l-95% CI	u-95% CI
Joanna	Ruth	0.04	0.07	-0.11	0.18
Joanna	Salli	-0.01	0.06	-0.13	0.12
Ruth	Joanna	-0.05	0.07	-0.19	0.08
Ruth	Salli	0.06	0.05	-0.03	0.15
Salli	Joanna	-0.02	0.07	-0.16	0.11
Salli	Ruth	0.03	0.07	-0.10	0.17
Joanna + Salli	Ruth	0.14	0.04	0.06	0.22
Joanna + Ruth	Salli	0.10	0.05	0.01	0.19
Ruth + Salli	Joanna	-0.08	0.05	-0.19	0.02

¹⁹The main effects of Speaker Gender and Shift Type are not directly relevant for the current research question and are thus omitted for clarity. However, all of the statistical models are open-access and available here: <https://doi.org/10.17605/osf.io/n8hws>

4.5.2 Refitting the combined model

As a review, the analysis in Section 4.3 reported one model that only looked at the “single talker per gender” condition (Section 4.3.1), one model that only looked at the “multiple talkers per gender” condition (Section 4.3.2), and then a combined model that looked at all of the data (Section 4.3.3). The author had originally only tried fitting the first two models, and in the published version of this chapter (Aoki & Zellou, 2025a), the presence of a three-way interaction between Speaker Gender, Block Order, and Shift Type in the “multiple talkers per gender” model (but not the “single talker per gender” model) was used to claim that multiple-talker exposure was necessary for generalization. However, upon receiving feedback from Xin Xie and Chigusa Kurumada post-publication, the author fitted the extra combined model and found relatively little evidence for a *four*-way interaction between Speaker Gender, Block Order, Shift Type, and Talker Variability.

The post-hoc analysis in Section 4.5.1 (which was reported in Aoki and Zellou (2025a)) provides context about why this four-way interaction may have been so weak – as shown in Table 4.6, generalization in the “multiple talkers per gender” condition only occurred when the test talker was Ruth or Salli, not Joanna. An immediate question is whether a “combined” model would show a robust four-way interaction if Joanna is not included as the test talker.

A model with the same formula as Equation 4.3 was fitted that included all participants in Experiment 2 except for those who heard Joanna in the test phase (the full output is shown in Table B.4 in Appendix B). As expected, there was a four-way interaction Speaker Gender, Block Order, Shift Type, and Talker Variability [β : 0.13, SE: 0.06, 95% HDI = (0.02, 0.24)], with the effect being much more robust for the “multiple talkers per gender” condition [β : 0.36, SE: 0.08, 95% HDI = (0.20, 0.52)] than for the “single talker per gender” condition [β : 0.10, SE: 0.08, 95% HDI = (-0.05, 0.25)].

4.5.3 Interim discussion

The results of the post-hoc analysis are summarized in Table 4.7, which shows clear support for a numerosity account over a sufficient similarity account. Unlike the “multiple talkers per gender condition”, no strong evidence of generalization was observed between any of the exposure-test talker combinations in the “single talker per gender” condition. A sufficient similarity account would attribute this set of null effects to a lack of sufficient similarity between all of the female talkers (i.e., the “Joanna” voice is dissimilar to both the “Ruth” and “Salli” voices, and the “Ruth” voice is dissimilar to the “Salli” voice). According to a sufficient similarity account, generalization should be blocked if the exposure phase presents two female talkers who are both dissimilar to the test talker (i.e., all of the exposure-test talker combinations in the “multiple talkers per gender” condition should show a null effect; Xie & Myers, 2017; Xie et al., 2021).

Table 4.7: Actual results of the post-hoc analysis. The structure of this table is the same as Tables 4.4 and 4.5, such that the column labels represent the exposure talkers and the row labels represent the test talkers. ‘Yes’ and ‘No’ refer to the presence or absence of generalization, respectively.

	Joanna	Ruth	Salli	Joanna + Ruth	Joanna + Salli	Ruth + Salli
Joanna	—	No	No	—	—	No
Ruth	No	—	No	—	Yes	—
Salli	No	No	—	Yes	—	—

However, this prediction is not borne out. Two of the possible exposure-test talker combinations in the “multiple talkers per gender” condition found robust evidence of cross-talk generalization (Joanna + Salli $\rightarrow$ Ruth, Joanna + Ruth $\rightarrow$ Salli). This suggests that, as predicted by a numerosity account (Bradlow & Bent, 2008), exposure to multiple talkers per gender is necessary to generalize an unfamiliar phonetic shift to a novel talker.

Why was no generalization observed for the participants who were exposed to Ruth and Salli and tested on Joanna? One tentative possibility is that effects of acoustic similarity are at play. Even though Joanna has a higher f_0 than all three male speakers in the current study (see Figure 4.3 and Table 4.3), t-tests demonstrate that the test stimuli of Joanna have a lower f_0 than both the exposure stimuli of Ruth (mean difference = 22.88 Hz, $t = 23.953$, $p < 0.001$) and the exposure stimuli of Salli (mean difference = 18.37 Hz, $t = 20.585$, $p < 0.001$). The exposure stimuli of Ruth and Salli, meanwhile, are more similar to each other acoustically.²⁰ Perhaps Joanna was not sufficiently similar acoustically to Ruth and Salli, which blocked generalization among listeners in the “Ruth + Salli $\rightarrow$ Joanna” condition.

A sufficient similarity account, however, is not entirely satisfactory for two reasons. First, assuming that Ruth and Salli are more similar to each other, it becomes unclear why there was little to no evidence of generalization among listeners who were exposed to Ruth and tested on Salli (Ruth $\rightarrow$ Salli) or exposed to Salli and tested on Ruth (Salli $\rightarrow$ Ruth).

Furthermore, contra the current findings, a sufficient similarity account would predict generalization to be *weaker* among listeners who were either exposed to Joanna and Salli and tested on Ruth (Joanna + Salli $\rightarrow$ Ruth) or exposed to Joanna and Ruth and tested on Salli (Joanna + Ruth $\rightarrow$ Salli). The shifted stimuli in the critical conditions ($n = 16$) were either produced by a single female exposure talker or evenly distributed across two female exposure talkers ($n = 8$ each) – if Ruth and Salli are truly the only two sufficiently similar female speakers, then listeners in the “Joanna + Salli $\rightarrow$ Ruth” and “Joanna + Ruth $\rightarrow$ Salli” conditions were exposed to fewer sufficiently similar tokens than the listeners in the “Salli $\rightarrow$ Ruth” and “Ruth $\rightarrow$ Salli” conditions ($n = 8$ versus $n = 16$). Assuming that reduced exposure leads to *less* robust adaptation (Cummings & Theodore, 2023), generalization should have been less strongly facilitated in the “Joanna + Salli $\rightarrow$ Ruth” and “Joanna + Ruth $\rightarrow$ Salli” conditions compared to the “Salli $\rightarrow$ Ruth” and “Ruth $\rightarrow$ Salli” conditions

²⁰Although Salli does have a greater f_0 than Ruth (mean difference = 4.50 Hz, $t = 5.58$, $p < 0.001$), the mean difference is reduced compared to the pairwise comparisons between Joanna/Ruth and Joanna/Salli.

(the exact opposite of the pattern depicted in Table 4.7). Overall, the acoustic differences between Joanna and Ruth/Salli do not appear to adequately explain the current results.

An alternative explanation could involve the properties of the test continua themselves, which are displayed in Figure 4.8. Although efforts were made to match the continua as closely as possible (see Section 4.2.2), Figure 4.8 shows that listeners in the “No Shift” condition (where no phonetic shift was presented in the exposure phase) were already heavily biased towards a “pie” response for Joanna compared to Ruth and Salli (Joanna: 70.4%; Ruth: 51.6%; Salli: 51.5%).²¹ The presence of generalization is also indicated by a greater proportion of “pie” responses, so it is possible that there was a ceiling effect – listeners in the “Female Shifted” condition could not provide a greater “pie” response than the No Shift condition, resulting in a null effect. The test phase stimuli for Ruth and Salli are not nearly as biased towards a “pie” response, so participants tested on Ruth and Salli were able to show meaningful generalization.

Although it is true that there are certain acoustic disparities among the female speakers (e.g., Joanna has a lower f_0 than Ruth and Salli), the differences in f_0 between the female and male speakers are much starker (see Figure 4.3). It therefore seems more probable for listeners to group Joanna with one of the other female speakers than with the male speakers, based on (relative) acoustic and social similarity (i.e., listeners in the “multiple talkers per gender” condition likely placed Joanna within the same mental model as Salli/Ruth). In accordance with a numerosity account, listeners attuned to the number of talkers with the /p/ to [b] shift and, given that the shift was heard in more than one exposure speaker, participants then generalized the shift to the most similar test talker (i.e., the novel female speaker).

²¹As noted in Section 4.2.2, a norming study was conducted on the test stimuli, and efforts were made to match the continua by gender. Although these efforts were successful overall, since there was no effect of Speaker Gender in the “No Shift” condition (see Figures 4.6 and 4.7), the source of the “pie” response bias for Joanna remains unclear. It should be noted that the norming study had fewer participants than the main study ($n = 120$ versus $n = 771$) and a between-participant design, with listeners categorizing stimuli from only one of the six speakers (~20 listeners per speaker). The lower power of the norming study could explain why the “buy”-“pie” continuum of Joanna seemed more evenly matched to the other female speakers in norming, but not the actual experiment.

Taken together, the results of the post-hoc analysis are more consistent with a numerosity account than a sufficient similarity account. A numerosity account more satisfactorily rationalizes why cross-talker generalization solely occurs in the “multiple-talkers per gender” condition and never in the “single-talker per gender” condition. The absence of generalization in one of the three possible exposure-test talker combinations in the “multiple-talkers per gender” condition seems to reflect idiosyncratic variation in the test phase continua, not evidence against a numerosity account.

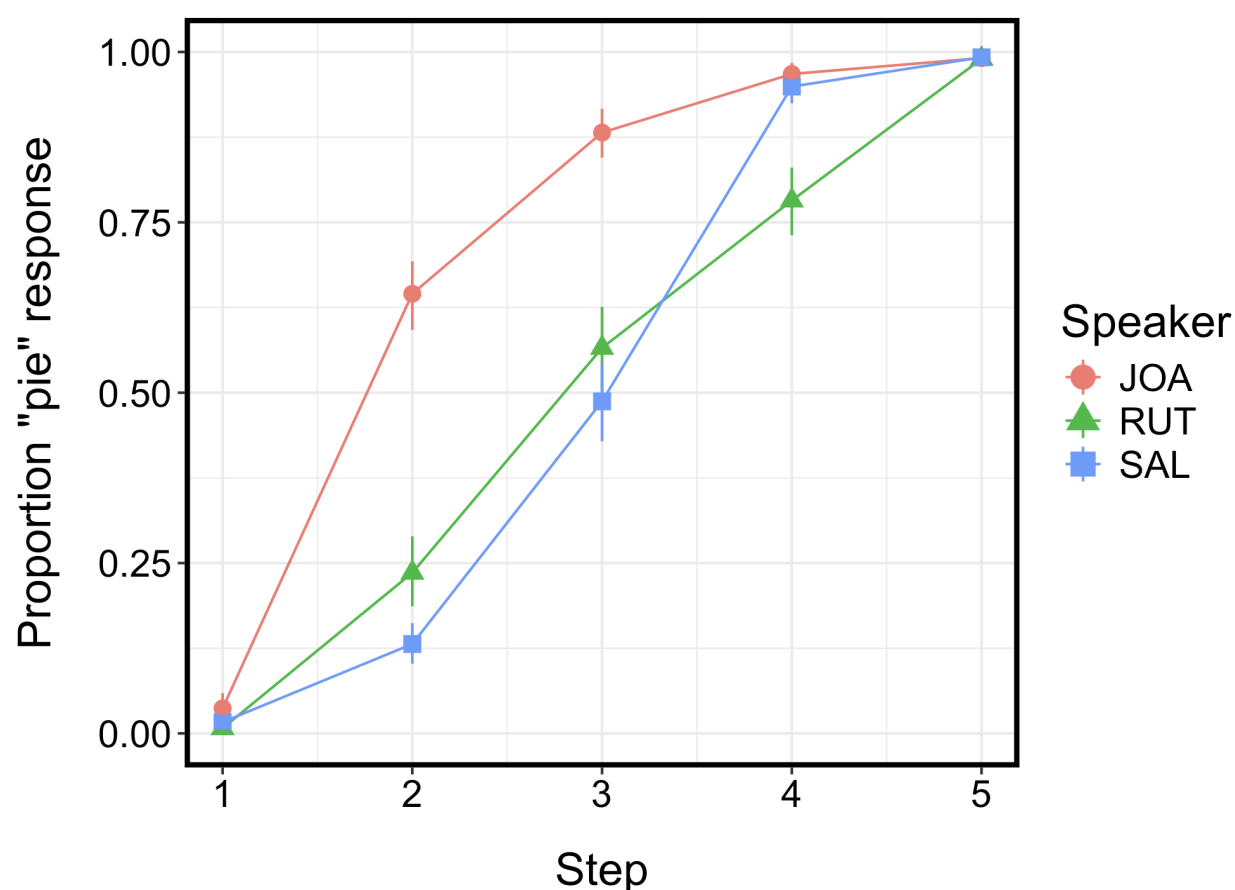


Figure 4.8: Aggregated results by Step and Speaker (JOA = Joanna; RUT = Ruth; SAL = Salli) in Experiment 2 for all participants assigned to the No Shift condition (i.e., no /p/ to [b] phonetic shift in the exposure phase). Error bars are 95% confidence intervals bootstrapped over by-participant means.

4.6 Discussion

The current study first exposed listeners to an unattested variant in L1-English: a gender-mediated phonetic shift, where females produce /p/ as [b] (e.g., “paint” as *baint*) and males produce a canonical /p/. A subsequent test phase presented a novel female talker and a novel male talker. The goal was to investigate the role of talker variability and acoustic similarity on cross-talkers generalization. For listeners to apply the phonetic shift from exposure to test, would it be adequate to just hear a single female exposure talker replace /p/ with [b], as expected by a *sufficient similarity account* (Xie et al., 2021; Xie & Myers, 2017)? Or, in accordance with a *numerosity account* (Bradlow & Bent, 2008), would it be helpful/necessary for generalization to hear the phonetic shift in multiple female exposure talkers?

A series of analyses provided clear support for numerosity over sufficient similarity. Robust generalization was only observed among listeners exposed to multiple talkers per gender, not among listeners exposed to a single talker per gender. Importantly, this finding seems to differ dramatically from Experiment 1, which found single-talkers exposure to be sufficient for generalization to Mandarin-accented English.

Xie et al. (2021) appealed to a methodological explanation to account for differing effects of talker variability across studies (i.e., why Bradlow and Bent (2008) found that multiple-talkers exposure was necessary for generalization, but not Xie and Myers (2017) or Xie et al. (2021)). In particular, there were three procedural issues with Bradlow and Bent (2008): (i) low statistical power ($n = 87$ across five between-participant conditions); (ii) the use of different exposure talkers in the single- and multiple-talkers conditions; (iii) the presentation of only one test talker for all participants. However, these factors cannot explain the current findings because a large number of listeners were included in the analysis ($n = 771$), the exposure talkers were drawn from the same pool for all conditions, and test talker identity was evenly counterbalanced. Therefore, the presence of generalization in only the “multiple talkers per gender” condition cannot merely be attributed to confounds or design flaws.

The *linguistic experience hypothesis* outlined in Section 2.3.3 more effectively unites the findings across experiments. In contrast to L2-accented English, which the listeners in Experiment 1 had likely experienced before, the test case in the current study was designed to be completely unattested in L1-English (see Section 4.1.1 above). The linguistic experience hypothesis predicts that in situations of less familiarity, more evidence should be needed for generalization to convince listeners that variation is truly generalizable and not just an idiosyncrasy. This prediction is supported by the studies in this dissertation – whereas listeners in Experiment 1 relied on only one type of evidence to generalize an L2-accent (acoustic similarity), listeners in Experiment 2 depended on two types of evidence to generalize an unusual /p/ to [b] shift (acoustic similarity *and* the number of exposure talkers).

There is only one aspect of Experiment 2 left to explain. Among listeners exposed to multiple talkers per gender, why was generalization only observed in the “Female → Male” order, not in the “Male → Female” order (i.e., only when the novel female talker was presented first in the test phase, not when the novel male talker was presented first)? Section 4.6.1 provides four possible explanations for this order effect.

4.6.1 Accounting for the order effect

A robust numerosity effect only occurred when listeners engaged in generalization immediately after the exposure phase (the “Female → Male” order); by contrast, generalization was blocked in the “Male → Female” order, when there was a delay in presenting the novel female talker. The “delay” was extremely short (just 45 trials of hearing a novel male talker), which suggests that generalization was short-lived. There could be at least four (non-mutually exclusive) explanations for such ephemerality.

One involves the novelty of the test case. The synthetic gendered /p/ to [b] shift in the exposure phase was designed to be completely unattested in L1-English (see Section 4.1.1), which could have increased the difficulty of learning and resulted in a more transitory

effect. Generalization might have been more prominent had the test case aligned more closely English norms (e.g., if the *male* speakers produced /p/ as [b] instead of the female speakers; Robb et al., 2005).

A second possibility is that there was not enough exposure to prolong generalization. Listeners in the critical conditions only heard 16 phonetically shifted stimuli from only two female talkers during a several-minute-long exposure phase. Given the peculiarity of the shift, greater exposure might have been needed for memory retention. A more stable effect might be observed if listeners were presented with more stimuli (Cummings & Theodore, 2023), heard more than two shifted talkers (Bradlow & Bent, 2008), and/or completed several iterations of this task over multiple sessions (Xie & Kurumada, 2024).²²

Yet a third consideration is the structure of the exposure phase. Listeners heard a mixture of filler and critical stimuli from two female and two male talkers in a randomized order – perhaps presenting the exposure sentences in a blocked order (either by talker or by gender) would have made it easier to keep track of the speakers and to learn that only the female talkers were phonetically shifted. This account accords with work showing detrimental effects of talker-switching on speech perception (Magnuson et al., 2021), and with the intuition that blocked training may be more helpful at the incipient stages of learning (for discussion, see p. 476 of Raviv et al., 2022).²³

²²Although increasing the number of critical stimuli should generally facilitate adaptation (Cummings & Theodore, 2023), too many critical stimuli could lead to a negative after-effect, resulting in diminished adaptation (for details, see Kleinschmidt & Jaeger, 2016, p. 683).

²³The effect of variability on speech perception is highly complex (Quam & Creel, 2021), and in contrast to this proposal, Tzeng et al. (2016) found that a randomized exposure stimulus order facilitated cross-talkers generalization, as opposed to a blocked order. However, the experimental design of Tzeng et al. (2016) diverges significantly from the current study, as they focus on a different phenomenon (adaptation towards Spanish-accented English), use a different task (sentence-transcription-in-noise), and implement a different method of measuring generalization (transcription accuracy of a novel test talker). It is therefore unclear whether the results in Tzeng et al. (2016) should be expected in the current study. In general, future work examining the effect of training structure on adaptation is warranted.

The three points above are already published in Aoki and Zellou (2025a), but a fourth option has arisen, now that the acoustic cue distributions have been visualized for each talker (as opposed to just reporting talker-specific means in a table). The right panel of Figure 4.2 compares the distributions of Voice Onset Time (VOT) between the unshifted male speakers and the /p/ to [b] shifted female speakers. Although the unshifted male speakers have a higher *mean* VOT than the shifted female speakers (as expected), the distributions slightly overlap (i.e., certain male and female stimuli unintentionally had similar VOTs). Extensive work demonstrates that learning and adaptation are weakened when distributions are overlapping (see Kleinschmidt, 2018), which suggests that generalization would have been stronger had all of the male stimuli exhibited a higher VOT than the female stimuli.

The previous paragraph highlights a lesson learned – the importance of looking at the *variance* of acoustic cue distributions (as opposed to just the means) before running a perceptual experiment. Although the VOT issue does not necessarily alter the overall conclusions of this study, the results may have been cleaner if the author had adjusted the stimuli so that the female and male VOT distributions were non-overlapping.²⁴

²⁴Thank you again to Xin Xie and Chigusa Kurumada, not just for their advice about acoustic cue distributions, but for all of their helpful feedback on this work.

Chapter 5

General Discussion

Imagine that you strike up a conversation with the speaker in the Prologue. Your interlocutor seems to have a canonical L1 American English accent – you have no reason to suspect that they are not from the United States. But then, you gradually begin to notice certain unexpected phonetic shifts (e.g., their /w/ is pronounced as [v], as in “You’re *velcome*”), and you later learn that the speaker’s first language is actually Farsi. In this situation, would you treat the /w/ to [v] shift as an idiosyncrasy, or would you assume the shift to be generalizable to all Farsi speakers? This dilemma highlights a longstanding question in cognitive science: When listeners are exposed to a particular accent or phonetic variant, under what conditions do they transfer (or *generalize*) what they have learned to a novel talker?

There are two prominent theories of cross-talker generalization that seem to oppose each other: numerosity and sufficient similarity. A *numerosity account* emphasizes the importance of multiple-talker exposure (Bradlow & Bent, 2008); meanwhile, a *sufficient similarity account* claims that acoustic similarity between the exposure and novel talkers is what mediates generalization (Xie et al., 2021; Xie & Myers, 2017). The latter account aligns with a trend to focus on acoustic explanations for perceptual adaptation, particularly through quantitative modeling (e.g., Feldman et al., 2009; McMurray & Jongman, 2011; Xie et al., 2023, etc.).

This dissertation unites the two accounts through the following observation – multiple-talker exposure only affects generalization when the exposure and novel talkers are acoustically similar (e.g. when they share the same language background; Bradlow & Bent, 2008). This suggests that, rather than being competing perspectives, numerosity and sufficient similarity are jointly positioned along a continuum. A “sufficient similarity” account is observed when listeners only count on one type of evidence to generalize (acoustic similarity), while a “numerosity” account is supported when listeners rely on two types of evidence (acoustic similarity *and* multiple exposure talkers).

This insight lends itself to a novel *linguistic experience hypothesis* (see Section 2.3.3 for details). In brief, the theory predicts an inverse relationship between the listener’s sociolinguistic experience and the amount of evidence needed for generalization. When listeners possess less experience with variation, they tend to require more evidence to convince them that such variation is truly generalizable and not just an idiosyncrasy; meanwhile, more experience tends to be associated with less evidence needed for generalization.

Two experiments tested the linguistic experience hypothesis as a proof of concept. Experiment 1 focused on a situation where listeners were relatively familiar with the test case at hand (L1-English listeners living in urban areas of the United States were exposed to L2-accented English). Experiment 2 then presented L1-English listeners with variation that was designed to be completely unfamiliar to them (a synthetic, gendered /p/ to [b] shift). Consistent with the linguistic experience hypothesis, less evidence was needed for generalization in Experiment 1 (just acoustic similarity); meanwhile, more evidence was required for generalization in Experiment 2 (both acoustic similarity *and* exposure to multiple talkers). Altogether, this body of work makes two points: (i) sufficient similarity and numerosity are mediated by social context, rather than being contradictory accounts; (ii) while not discrediting the importance of acoustic cues, the listener’s sociolinguistic experience should have an equal status in theories of adaptation and generalization.

The rest of the general discussion is structured as follows. Section 5.1 first advocates for a greater integration of sociolinguistic perspectives into theories of cognitive science. Section 5.2 highlights three concrete steps for future work, and Section 5.3 concludes the dissertation.

5.1 The utility of sociolinguistic experience in theories of cognitive science

Any field (including cognitive science and linguistics) intends to make very broad generalizations about some phenomenon that span across time and space. For example, statements like the following are prevalent in the literature:

Language is estimated to have emerged in humans between 100,000 and 1,000,000 years ago. The functions of language and the causal drivers in its origins have long been fiercely debated across diverse fields, including philosophy, linguistics, cognitive science, evolutionary biology and anthropology. Two broad hypotheses have dominated this discussion. . . One proposal is that language primarily serves a communicative function. . . Another proposal is that language mediates thinking and cognition. . . (Fedorenko et al., 2024, p. 575)

The perennial issue is that, despite our aim to make general conclusions, we only have access to limited data. Any individual study must necessarily be focused on a narrow test case; but more crucially, certain test cases are impossible to investigate in a practical sense due to ethical concerns, funding, or other reasons (e.g., as of yet, we can't time travel back to 100,000 to 1,000,000 years ago to test theories about the function of language). This results in situations where there is necessarily a mismatch between the breadth of claims that are made and the data that are presented. Although it is probably impossible to completely nullify this issue, the goal of scientific inquiry is to minimize the gap between theory and data as much as possible.

There can be no doubt that our understanding of adaptation has grown significantly over time, but as discussed in Section 2.3.5, many studies often reuse the same test cases. For example, a disproportionate number of experiments specifically investigate adaptation of L1-English undergraduates to Mandarin-accented English undergraduates at large public universities in the United States (note that this includes the author’s own work; Aoki & Zellou, 2023c). This is not to fault the researchers – conducting experiments is challenging enough, and this particular listener-speaker pairing is arguably one of the easiest to recruit. The problem is that this situation triggers the following chain of events: (i) similar or identical test cases lead to the same result; (ii) evidence accumulates for one theoretical conclusion over another; (iii) the greater amount of evidence for one conclusion increases the perception of robustness and generality, even though the conclusion may be an artefact of the particular test case being employed. This chain of events could be avoided if we work with more varied test cases, but how can we find test cases that are realistic for researchers to implement?

Part of the goal of the Prologue is to highlight the potential utility of sociolinguistic experience for theoretical development. In particular, the L1-Farsi, L2-English speaker in the Prologue inspired the synthetic, gendered /p/ to [b] shift in Experiment 2 (i.e., an instance where someone who sounds like an L1-English speaker unexpectedly replaces one segment with another). By implementing a completely different type of test case in this dissertation, a different theoretical conclusion was reached from prior work, planting the seed for the linguistic experience hypothesis. Overall, this suggests that novel and feasible test cases are already available – but in order to find these test cases, it is necessary to draw upon more diverse sociolinguistic experiences.

None of this is to say that we shouldn’t reuse test cases, or that every experiment must present a new test case. But if we continue working within existing paradigms and supplement what we already have with something new, then a stronger and more robust understanding of language and cognition can be achieved.

This dissertation makes a very small, incremental step through two experiments, but must be accompanied by future work. The following section highlights multiple possible ideas for how the linguistic experience hypothesis can be tested and how a greater sociolinguistic perspective can be integrated into work on adaptation.

5.2 Future directions and limitations

5.2.1 Refining the linguistic experience hypothesis

This dissertation tested the linguistic experience hypothesis in a fairly extreme fashion – a binary comparison was made between a social context where listeners had some familiarity and a social context where listeners had no familiarity. Perhaps unsurprisingly, there was an extreme disparity in results, where multiple-talker exposure was necessary for generalization in the latter scenario but not the former.

Note that this dichotomy was more intended as a proof of concept. The hypothesis is framed in a fairly moderate way, which lends itself to more gradient predictions (see Section 2.3.3 for a full explanation of the wording):

The Linguistic Experience Hypothesis: The less experience a listener has with variation, the more evidence will likely be helpful/needed for generalization, above and beyond acoustic similarity.

The linguistic experience hypothesis cannot be set in stone – the specific wording may need to be updated given greater empirical evidence. The three subsections below highlight various ways of refining the linguistic experience hypothesis, while still maintaining the importance of sociolinguistic experience on adaptation.

Establishing an evidence hierarchy

An *evidence hierarchy* is an implicit idea within the wording of the linguistic experience hypothesis – acoustic similarity is assumed to be ranked the highest, with other potentially relevant cues only having a weaker effect (e.g., multiple-talker exposure). Although this dissertation has focused on acoustic similarity and talker variability, other types of important evidence could affect generalization, such as visual similarity. What might the evidence hierarchy look like after acoustic similarity?

On the one hand, everything besides acoustic similarity might be weighted equally, which predicts that generalization should increase as the number of types of evidence increases. For example, in Experiment 2, generalization could have been robust in the single-talker exposure condition had the synthetic female exposure and test talkers been acoustically similar *and* paired with matching pictures (e.g., if they were all shown with the same images of robots; Tuttösi et al., 2023). There could alternatively be a more complex sub-hierarchy where multiple-talker exposure is ranked second and visual similarity is third (this would predict no generalization in the single-talker exposure condition of Experiment 2, even if visual similarity was present). Ultimately, the more types of evidence are integrated into adaptation experiments, the greater ecological validity can be achieved.

The quality of evidence

As a second area of further research, one might wonder about the *quality* of evidence. Is all “evidence” of the same type treated equally by the listener?¹

Gradiance in similarity represents one way of thinking about the “quality” of evidence. Let’s say that there are two single-talker exposure conditions – one presents a single talker who is strongly similar acoustically to the novel talker, while the other presents a single talker

¹Thank you to Fernanda Ferreira for this broad point and for the two specific ideas in this subsection.

who is weakly similar acoustically to the novel talker. The former condition represents higher quality evidence, which should facilitate generalization. But what if listeners heard many talkers in the “weak acoustic similarity” condition – would the perceptual results become equivalent to hearing a single talker in the “strong acoustic similarity” condition? If listeners value the quality of evidence over the quantity of evidence, then the presence of strong acoustic similarity should always supersede a “weak similarity” condition, regardless of the number of talkers; however, as an alternative hypothesis, a greater quantity of evidence might be able to overcome a lack of quality.

Another interesting idea could be to manipulate the trustworthiness of the speaker. Prior work has demonstrated that word learning and adaptation are facilitated when the speaker plays a more socially dominant role (e.g., a teacher, as opposed to a student; Hayes-Harb et al., 2022). If listeners are sensitive to social dominance effects, then generalization should be facilitated when the exposure talkers are both teachers rather than students.

The relationship between the amount and the type of evidence

Finally, one might wonder about the relationship between the *amount* of evidence and the *type* of evidence in generalization. Prior research has established that, despite some caveats (Kleinschmidt & Jaeger, 2016, p. 683), exposure to more tokens tends to facilitate adaptation (e.g., shifts in phonetic categorization are stronger if listeners hear an ambiguous fricative 20 times in exposure versus just 4 times; Cummings & Theodore, 2023). The linguistic experience hypothesis suggests that more types of evidence should also enhance generalization, but how might the amount and type of evidence interact?

One might envision an experiment where all listeners hear two types of evidence (e.g., acoustic similarity and talker variability), but the relative amount of evidence differs in a between-participant design. For example, one group might hear ten talkers produce three shifted tokens each; another might hear five talkers produce six tokens each. Everyone would

hear the same number of tokens ($n = 30$ total), and two types of evidence, but the number of talkers would differ. Would listeners generalize to the same degree given the same number of tokens, or would they show stronger generalization in the first condition, given the greater number of exposure talkers?

As a second variation of this broad idea, one might place all listeners in a multiple-talker exposure condition where the number of talkers is the same, but the number of relevant tokens is not evenly distributed across the talkers (e.g., unlike Experiment 2 of this dissertation, a /p/ to [b] shift might be heard 12 times in one female exposure talker and only 4 times in another female exposure talker). This situation could lead to at least two possible outcomes: (i) the distribution of tokens might make no difference, with listeners only caring about the presence or absence of evidence in a binary fashion (i.e., as long as acoustic similarity and multiple-talker exposure are both present, generalization will occur to the same degree); (ii) an unbalanced cross-talker distribution of tokens might weaken generalization, since it would become less apparent whether a particular variant is generalizable to other speakers.

Overall, the ideas in this section represent a small fraction of the ways that the linguistic experience hypothesis could be refined and improved upon through future research.

5.2.2 Other ways of testing the linguistic experience hypothesis: Integrating typology and linguistic diversity

In addition to refining the linguistic experience hypothesis, another important step is to vary the types of test cases in adaptation experiments in a way that modulates listener experience.

One approach is to recruit participants in language immersion programs (Chang, 2012). An L2-accent adaptation experiment might be administered before and after a lengthy stay in a foreign country through a language learning program. If greater experience reduces the amount of evidence needed for generalization, then effects of talker variability should

weaken over time; this implies that any benefit of multiple-talker exposure should diminish from pre- to post-immersion, after listeners have been extensively exposed to phonological patterns not present in their first language. The relationship between the participants' first language and their immersion country language might also affect the results – if the two are quite typologically different (e.g., imagine a Japanese speaker learning English in the United States), then talker variability might be much more helpful in the pre-immersion session.

Insights from phonological typology should also help find novel test cases, given the extensive documentation on which sound patterns are more or less frequent (and thus more or less likely to have been experienced by listeners; Gordon, 2016). For instance, in reinterpreting Xie and Myers (2017), Section 2.3.4 cited the prevalence of word-final devoicing as a key reason for the lack of effect of talker variability; in a converse fashion, exposure to word-final *voicing* should increase the helpfulness of multiple-talker exposure (or other types of evidence). As another example, cross-linguistically rare phonological patterns in languages like Tashlhiyt Berber should also require more evidence for generalization (e.g., vowelless words, words that violate the Sonority Sequencing Principle; Zellou et al., 2022, 2024).

The ideas in this section highlight the potential usefulness of the linguistic experience hypothesis for facilitating language learning and accent adaptation in everyday life. More broadly, the hypothesis suggests that diversity and the multilingual experience are helpful for speech perception in the long run – the more exposure one has to variation, the less evidence one needs to generalize, thus facilitating more rapid adaptation to novel talkers.

5.2.3 Individual differences and theories of sound change

An issue with many adaptation studies (including the current set of experiments) is the absence of within-participant variables. This is sometimes necessary – listeners cannot simultaneously be exposed to a single talker *and* to multiple talkers within the same session (there could be separate blocks, but that would defeat the purpose of a single-talker exposure

condition). A between-participant design, however, can obscure individual differences (e.g., most listeners might require multiple-talker exposure for generalization in a certain condition, but there might be a minority who generalize with acoustic similarity alone). Integrating an individual differences approach with the linguistic experience hypothesis could help advance linguistic theory, particularly theories of sound change.

A big-picture question in the sound change literature is the *actuation problem*. “Why [does] a linguistic change occur in a particular language at a particular time and space” (Yu, 2023, p. 215)? Solutions to the actuation problem are varied and complex, but in sociolinguistics, individual differences in social network are thought to play a crucial role:

... the most advanced speakers [in a sound change in progress] are the persons with the largest number of local contacts within the neighborhood, yet who have at the same time the highest proportion of their acquaintances outside the neighborhood. (Labov, 1980, p. 261)

One way of interpreting this idea is that the actuators of change are those with the most linguistic experience (i.e., who have the most diverse social network within a community).

A second line of work within phonetics has claimed that sound changes are driven by the listener (essentially, listeners misattribute phonetic coarticulatory variation to phonological variation; Beddor, 2009). Taking this theory a step further, Tamminga et al. (2020) highlight a potentially crucial role of generalization – a sound change can only propagate if listeners generalize misattributions to different talkers.

Building upon Tamminga et al. (2020), the linguistic experience hypothesis offers a way of linking both sociolinguistic and phonetic theory – individuals with greater exposure to variation through their social networks are *also* the same individuals who will more readily generalize (i.e., they require less evidence for generalization). By situating generalization as a key ingredient in the socio-cognitive mechanisms of sound change, this proposal helps to clarify *why* “people with the largest number of contacts” might be the actuators of change.

Interestingly, the insights derived from the linguistic experience hypothesis additionally align well with linguistic anthropology. Eckert (1989) notes the following about the relationship between speaker sex/gender and sound change:

...sex does not have a uniform effect on variables or even on variables that represent sound change in progress. This is because sex is not directly related to linguistic behavior but reflects complex social practice... the significance of gender in variation cannot be reduced to notions of male or female speech as “more or less conservative”. (p. 245)

In other words, when an ethnographer or sociolinguistic fieldworker finds that women are more “linguistically innovative” than men, such differences do not involve categories in a vacuum (i.e., it is not the quality of being a “woman” in and of itself that mediates change). Rather, linguistic variation reflects one’s experiences, which happen to correlate with macro-social categories such as gender, age, etc.

Returning to the very beginning of this subsection, one way of turning all of this into a testable set of hypotheses is to investigate individual differences in cross-talker generalization through a *within*-participant design. Let’s say that listeners complete two generalization experiments that are sufficiently far apart in time for effects to fully decay – one session exposes listeners to single talker, while the other presents multiple talkers in exposure. According to the proposal described above, there should be a correlation with one’s social network (as measured by a social network analysis; Kutlu et al., 2024) and effects of talker variability on generalization (less evidence should be required, implying less of a difference between the single- and multiple-talker exposure conditions).

A broader point made by this section is to highlight the value of a greater sociolinguistic perspective in cognitive science – the more experiments are situated in context, the more connections can be made to other fields (e.g., anthropology, sociolinguistics, phonetics), thus galvanizing new, cross-disciplinary collaborations.

5.3 Conclusion

The adaptation and generalization literature tends to focus heavily on the acoustics of the speech signal. Although there can be no question that speech acoustics play a key role in speech perception, this dissertation claims that the listener's sociolinguistic experience is of equal importance, and must be more deeply embedded within theories of adaptation.

This dissertation proposes a novel theory (the linguistic experience hypothesis) that makes specific, testable predictions about how listener experience mediates cross-talker generalization. While still appreciating the contributions of current work, the present study underscores the need of a more interdisciplinary, sociolinguistic perspective for enriching our theoretical understanding of language and cognition.

References

- Agresti, A. (2002). *Categorical data analysis* (2nd ed.). John Wiley & Sons.
<https://doi.org/10.1002/0471249688>
- Alonso Martin, F., Malfaz, M., Castro-González, Á., Castillo, J. C., & Salichs, M. Á. (2020). Four-features evaluation of text to speech systems for three social robots. *Electronics*, 9(2), 267. <https://doi.org/10.3390/electronics9020267>
- Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, 73(3), 247-264.
[https://doi.org/10.1016/S0010-0277\(99\)00059-1](https://doi.org/10.1016/S0010-0277(99)00059-1)
- Amazon Web Services. (n.d.). *Amazon Polly: AI Voice Generator*. Retrieved April 18, 2024 from: <https://aws.amazon.com/polly/>
- Amazon Web Services. (2026). *Amazon Polly: Developer Guide*. Retrieved April 27, 2026 from: <https://docs.aws.amazon.com/pdfs/polly/latest/dg/polly-dg.pdf>
- Ammari, T., Kaye, J., Tsai, J. Y., Bentley, F. (2019). Music, search, and IoT: How people (really) use voice assistants. *ACM Transactions on Computer-Human Interaction*, 26(3), 1-28, 17. <https://doi.org/10.1145/3311956>
- Aoki, N. B., Cohn, M., & Zellou, G. (2022). The clear speech intelligibility benefit for text-to-speech voices: Effects of speaking style and visual guise. *JASA Express Letters*, 2(4), 045204. <https://doi.org/10.1121/10.0010274>

- Aoki, N. B., & Zellou, G. (2023a). Speakers talk more clearly when they see an East Asian face: Effects of visual guise on speech production. In R. Skarnitzl & J. Volín (Eds.), *Proceedings of the 20th International Congress of Phonetic Sciences* (pp. 2294-2298). Guarant International. https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2023/full_papers/45.pdf
- Aoki, N. B., & Zellou, G. (2023b). When clear speech does not enhance memory: Effects of speaking style, voice naturalness, and listener age. *Proceedings of Meetings on Acoustics*, 51(1), 060002. <https://doi.org/10.1121/2.0001766>
- Aoki, N. B., & Zellou, G. (2023c). Visual information affects adaptation to novel talkers: Ethnicity-specific and ethnicity-independent learning of L2-accented speech. *The Journal of the Acoustical Society of America*, 154(4), 2290-2304. <https://doi.org/10.1121/10.0021289>
- Aoki, N. B., & Zellou, G. (2024). Being *clear* about clear speech: Intelligibility of hard-of-hearing-directed, non-native-directed, and casual speech for L1- and L2-English listeners. *Journal of Phonetics*, 104, 101328. <https://doi.org/10.1016/j.wocn.2024.101328>
- Aoki, N. B., & Zellou, G. (2025a). When multiple talker exposure is necessary for cross-talker generalization: Insights into the emergence of sociolinguistic perception. *Glossa Psycholinguistics*, 4(1). <https://doi.org/10.5070/g6011.21217>
- Aoki, N. B., & Zellou, G. (2025b). Apparent talker variability and speaking style similarity can enhance comprehension of novel L2-accented talkers. *Language and Speech*. <https://doi.org/10.1177/00238309251390505>
- Apfelbaum, K. S., & McMurray, B. (2011). Using variability to guide dimensional weighting: Associative mechanisms in early word learning. *Cognitive Science*, 35(6), 1105-1138. <https://doi.org/10.1111/j.1551-6709.2011.01181.x>
- Atagi, E., & Bent, T. (2017). Nonnative accent discrimination with words and sentences. *Phonetica*, 74(3), 173-191. <https://doi.org/10.1159/000452956>
- Babel, M., McAuliffe, M., Norton, C., Senior, B., & Vaughn, C. (2019). The Goldilocks zone of perceptual learning. *Phonetica*, 76(2-3), 179-200. <https://doi.org/10.1159/000494929>
- Baese-Berk, M. M. (2019). Interactions between speech perception and production during learning of novel phonemic categories. *Attention, Perception, & Psychophysics*, 81, 981-1005. <https://doi.org/10.3758/s13414-019-01725-4>

- Baese-Berk, M. M., Bradlow, A. R., & Wright, B. A. (2013). Accent-independent adaptation to foreign accented speech. *The Journal of the Acoustical Society of America*, 133(3), EL174-EL180. <https://doi.org/10.1121/1.4789864>
- Baese-Berk, M. M., Levi, S. V., & Van Engen, K. J. (2023). Intelligibility as a measure of speech perception: Current approaches, challenges, and recommendations. *The Journal of the Acoustical Society of America*, 153(1), 68-76. <https://doi.org/10.1121/10.0016806>
- Baese-Berk, M. M., Morrill, T. H. (2015). Speaking rate consistency in native and non-native speakers of English. *The Journal of the Acoustical Society of America*, 138(3), EL223-EL228. <https://doi.org/10.1121/1.4929622>
- Bakkouche, L., McGhee, C., Lau, E., Cooper, S., Luo, X., Rees, M., Alter, K., Post, B., & Schwarz, J. (2025). Finding the human voice in AI: Insights on the perception of AI-voice clones from naturalness and similarity ratings. *Proceedings of Interspeech 2025* (2190-2194). <https://doi.org/10.21437/Interspeech.2025-947>
- Banks, B., Gowen, E., Munro, K. J., & Adank, P. (2015). Cognitive predictors of perceptual adaptation to accented speech. *The Journal of the Acoustical Society of America*, 137(4), 2015-2024. <https://doi.org/10.1121/1.4916265>
- Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3), 255-278. <https://doi.org/10.1016/j.jml.2012.11.001>
- Barreda, S. (2015). *phonTools: Functions for phonetics in R* (Version 0.2-2.1). [R package]. Retrieved from: <https://cran.r-project.org/web/packages/phonTools/phonTools.pdf>
- Barreda, S. (2020). Vowel normalization as perceptual constancy. *Language*, 96(2), 224-254. <https://doi.org/10.1353/lan.2020.0018>
- Barreda, S. (2021). Perceptual validation of vowel normalization methods for variationist research. *Language Variation and Change*, 33(1), 27-53. <https://doi.org/10.1017/S0954394521000016>
- Barreda, S., & Nearey, T. M. (2012). The direct and indirect roles of fundamental frequency in vowel perception. *The Journal of the Acoustical Society of America*, 131(1), 466-477. <https://doi.org/10.1121/1.3662068>

- Barreda, S., & Predeck, K. (2024). Inaccurate but predictable: Vocal-tract length estimation and gender stereotypes in height perception. *Journal of Phonetics*, 102, 101290. <https://doi.org/10.1016/j.wocn.2023.101290>
- Barreda, S., & Silbert, N. (2023). *Bayesian multilevel models for repeated measures data: A conceptual and practical introduction in R*. Routledge. <https://santiagobarreda.com/bmmrmd/>
- Bayley, R., & Holland, C. (2014). Variation in Chicano English: The case of final (z) devoicing. *American Speech*, 89(4), p. 385-407. <https://doi.org/10.1215/00031283-2908200>
- Becker, R. A., Wilks, A. R., Brownrigg, R., Minka, T. P., & Deckmyn, A. (2025). *maps* (Version 3.4.3). [R package]. Retrieved from: <https://cran.r-project.org/web/packages/maps/maps.pdf>
- Beddor, P. S. (2009). A coarticulatory path to sound change. *Language*, 85(4), 785-821. <https://doi.org/10.1353/lan.0.0165>
- Behrend, T. S., Sharek, D. J., Meade, A. W., & Wiebe, E. N. (2011). The viability of crowdsourcing for survey research. *Behavior Research Methods*, 43, 800-813. <https://doi.org/10.3758/s13428-011-0081-0>
- Bent, T., and Bradlow, A. R. (2003). The interlanguage speech intelligibility benefit. *The Journal of the Acoustical Society of America*, 114(3), 1600-1610. <https://doi.org/10.1121/1.1603234>
- Bent, T., Buchwald, A., & Pisoni, D. B. (2009). Perceptual adaptation and intelligibility of multiple talkers for two types of degraded speech. *The Journal of the Acoustical Society of America*, 126(5), 2660-2669. <https://doi.org/10.1121/1.3212930>
- Bent, T., & Holt, R. F. (2017). Representation of speech variability. *WIREs Cognitive Science*, 8(4), e1434. <https://doi.org/10.1002/wcs.1434>
- Blasi, D. E., Henrich, J., Adamou, E., Kemmerer, D., & Majid, A. (2022). Over-reliance on English hinders cognitive science. *Trends in Cognitive Sciences*, 26(12), P1153-1170. <https://doi.org/10.1016/j.tics.2022.09.015>
- Blumstein, S. E., & Stevens, K. N. (1979). Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. *The Journal of the Acoustical Society of America*, 66(4), 1001-1017. <https://doi.org/10.1121/1.383319>

- Boersma, P., & Weenink, D. (2023). *Praat: doing phonetics by computer* (Version 6.3.11) [Computer program]. <https://praat.org/>
- Bradlow, A. R. (n.d.) *ALLSTAR: Archive of L1 and L2 scripted and spontaneous transcripts and recordings*. Retrieved September 14, 2025 from: <https://speechbox.linguistics.northwestern.edu/allstar>
- Bradlow, A. R., Bassard, A. M., & Paller, K. A. (2023). Generalized perceptual adaptation to second-language speech: Variability, similarity, and intelligibility. *The Journal of the Acoustical Society of America*, 154(3), 1601-1613. <https://doi.org/10.1121/10.0020914>
- Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, 106(2), 707-729. <https://doi.org/10.1016/j.cognition.2007.04.005>
- Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., & Tohkura, Y. (1997). Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production. *The Journal of the Acoustical Society of America*, 101(4), 2299–2310. <https://doi.org/10.1121/1.418276>
- Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition*, 1(1), 9. <https://doi.org/10.5334/joc.10>
- Bucholtz, M., Bermudez, N., Fung, V., Edwards, L., & Vargas, R. (2007). Hella Nor Cal or Totally So Cal: The perceptual dialectology of California. *Journal of English Linguistics*, 35(4), 325-352. <https://doi.org/10.1177/0075424207307780>
- Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1), 1-28. <https://doi.org/10.18637/jss.v080.i01>
- Cambon, J., Hernangómez, D., Belanger, C., Possenriede, D., & Hansen, O. (2025). *tidygeocoder* (Version 1.0.6). [R package]. Retrieved from: <https://cran.r-project.org/web/packages/tidygeocoder/tidygeocoder.pdf>
- Campbell-Kibler, K. (2010). Sociolinguistics and perception. *Language and Linguistics Compass*, 4(6), 377–389. <https://doi.org/10.1111/j.1749-818X.2010.00201.x>
- Cebrian, J. (2007). Old sounds in new contrasts: L2 production of the English tense-lax vowel distinction. In J. Trouvain & W. J. Barry (Eds.), *Proceedings of the 16th International Congress of Phonetic Sciences* (pp. 1637-1640). <https://www.icphs2007.de/conference/Papers/1576/1576.pdf>

- Chang, C. B. (2012). Rapid and multifaceted effects of second-language learning on first-language speech production. *Journal of Phonetics*, 40(2), 249-268.
<https://doi.org/10.1016/j.wocn.2011.10.007>
- Chao, K-Y., & Chen, L-M. (2008). A cross-linguistic study of Voice Onset Time in stop consonant productions. *International Journal of Computational Linguistics & Chinese Language Processing*, 13(2), 215-232. <https://aclanthology.org/O08-4005/>
- Charoy, J., & Samuel, A. G. (2023). Bad maps may not always get you lost: Lexically driven perceptual recalibration for substituted phonemes. *Attention, Perception, & Psychophysics*, 85, 2437-2458. <https://doi.org/10.3758/s13414-023-02725-1>
- Chodroff, E., & Wilson, C. (2017). Structure in talker-specific phonetic realization: Covariation of stop consonant VOT in American English. *Journal of Phonetics*, 61, 30-47.
<https://doi.org/10.1016/j.wocn.2017.01.001>
- Chomsky, N., & Halle, M. (1968). *The sound pattern of English*. Harper & Row.
<https://mitpress.mit.edu/9780262031790/the-sound-pattern-of-english/>
- Clayards, M. (2017). Individual talker and token covariation in the production of multiple cues to stop voicing. *Phonetica*, 75(1), 1-23. <https://doi.org/10.1159/000448809>
- Clayards, M. (2018). Differences in cue weights for speech perception are correlated for individuals within and across contrasts. *The Journal of the Acoustical Society of America*, 144(3), EL172-EL177. <https://doi.org/10.1121/1.5052025>
- Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, 108(3), 804-809.
<https://doi.org/10.1016/j.cognition.2008.04.004>
- Clopper, C. G. (2011). Effects of dialect variation on the semantic predictability benefit. *Language and Cognitive Processes*, 27(7-8), 1002-1020.
<https://doi.org/10.1080/01690965.2011.558779>
- Clopper, C. G., & Pisoni, D. B. (2004). Effects of talker variability on perceptual learning of dialects. *Language and Speech*, 47(3), 207-238.
<https://doi.org/10.1177/00238309040470030101>

- Clopper, C. G., Pisoni, D. B., & de Jong, K. (2005). Acoustic characteristics of the vowel systems of six regional varieties of American English. *The Journal of the Acoustical Society of America*, 118(3), 1661-1676. <https://doi.org/10.1121/1.2000774>
- Coetzee, A. W., Beddor, P. S., Styler, W., Tobin, S., Bekker, I., & Wissing, D. (2022). Producing and perceiving socially structured coarticulation: Coarticulatory nasalization in Afrikaans. *Laboratory Phonology*, 13(1). <https://doi.org/10.16995/labphon.6450>
- Cohn, M. (2019). *Get intensity, duration, & mean f0 over 12 points of the vowel* [Praat script]. Retrieved June, 4 2024 from: https://github.com/michellecohn/praat-scripts/blob/master/!get_mean_f0_at_12_intervals_for_all_files_looping_through_subdirectories
- Cohn, M., Barreda, S., & Zellou, G. (2023). Differences in a musician's advantage for speech-in-speech perception based on age and task. *Journal of Speech, Language, and Hearing Research*, 66(2), 545-564. https://doi.org/10.1044/2022_JSLHR-22-00259
- Cohn, M., Pycha, A., & Zellou, G. (2021). Intelligibility of face-masked speech depends on speaking style: Comparing casual, clear, and emotional speech. *Cognition*, 210, 104570. <https://doi.org/10.1016/j.cognition.2020.104570>
- Coretta S., Casillas J. V., Roessig, S., Franke, M., ... Roettger, T. B. (2023). Multidimensional signals and analytic flexibility: Estimating degrees of freedom in human-speech analyses. *Advances in Methods and Practices in Psychological Science*, 6(3). <https://doi.org/10.1177/25152459231162567>
- Coupland, N. (2010). *The handbook of language and globalization*. Blackwell. <https://doi.org/10.1002/9781444324068>
- Cummings, S. N., & Theodore, R. M. (2022). Perceptual learning of multiple talkers: Determinants, characteristics, and limitations. *Attention, Perception, & Psychophysics*, 84, 2335-2359. <https://doi.org/10.3758/s13414-022-02556-6>
- Cummings, S. N., & Theodore, R. M. (2023). Hearing is believing: Lexically guided perceptual learning is graded to reflect the quantity of evidence in speech input. *Cognition*, 235, 105404. <https://doi.org/10.1016/j.cognition.2023.105404>
- Derwing, T. M., & Munro, M. J. (2014). Training native speakers to listen to L2 speech. In J. Levis & A. Moyer (Eds.), *Social Dynamics in Second Language Accent* (pp. 219-236). De Gruyter Mouton. <https://doi.org/10.1515/9781614511762.219>

- DiCanio, C., Chen, W. R., & Shadle, C. (2021). *Spectral moments of fricative spectra script in Praat* (Version 4) [Praat script]. Retrieved September 18, 2025 from: https://www.acsu.buffalo.edu/~cdicanio/scripts/Time_averaging_for_fricatives_4.0.praat
- Dobis, E. A., & Sanders, A. (2026). *Rural Urban Commuting Area Codes – Documentation*. Economic Research Service: U.S. Department of Agriculture. Retrived May 16, 2026 from: <https://www.ers.usda.gov/data-products/rural-urban-commuting-area-codes/documentation>
- D’Onofrio, A. (2015). Persona-based information shapes linguistic perception: Valley Girls and California vowels. *Journal of Sociolinguistics*, 19(2), 241–256. <https://doi.org/10.1111/josl.12115>
- Drager, K. (2010). Sociophonetic variation in speech perception. *Language and Linguistics Compass*, 4(7), 473–480. <https://doi.org/10.1111/j.1749-818X.2010.00210.x>
- Eckert, P. (1989). The whole woman: Sex and gender differences in variation. *Language Variation and Change*, 1(3), 245–267. <https://doi.org/10.1017/S095439450000017X>
- Eckert, P. (2008). Where do ethnolects stop? *International Journal of Bilingualism*, 12(1-2), 25–42. <https://doi.org/10.1177/13670069080120010301>
- Eisner, F., Melinger, A., & Weber, A. (2013). Constraints on the transfer of perceptual learning in accented speech. *Frontiers in Psychology*, 4, 148. <https://doi.org/10.3389/fpsyg.2013.00148>
- Fedorenko, E., Piantadosi, S. T., & Gibson, E. A. F. (2024). Language is primarily a tool for communication rather than thought. *Nature*, 630, 575–586. <https://doi.org/10.1038/s41586-024-07522-w>
- Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009). The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review*, 116(4), 752–782. <https://doi.org/10.1037/a0017196>
- Feng, H., & Wang, L. (2024). Acoustic analysis of English tense and lax vowels: Comparing the production between Mandarin Chinese learners and native English speakers. *The Journal of the Acoustical Society of America*, 155(5), 3071–3089. <https://doi.org/10.1121/10.0025931>

- Ferguson, S. H., & Kewley-Port, D. (2002). Vowel intelligibility in clear and conversational speech for normal-hearing and hearing-impaired listeners. *The Journal of the Acoustical Society of America*, 112(1), 259-271. <https://doi.org/10.1121/1.1482078>
- Flege, J. E. (1984). The detection of French accent by American listeners. *The Journal of the Acoustical Society of America*, 76(3), 692-707. <https://doi.org/10.1121/1.391256>
- Flege, J. E., & Bohn, O-S. (2021). The Revised Speech Learning Model (SLM-r). In R. Wayland (Ed.), *Second Language Speech Learning: Theoretical and Empirical Progress* (pp. 3-83). Cambridge University Press. <https://doi.org/10.1017/9781108886901>
- Flege, J. E., MacKay, I. R. A., & Meador, D. (1999). Native Italian speakers' perception and production of English vowels. *The Journal of the Acoustical Society of America*, 106(5), 2973-2987. <https://doi.org/10.1121/1.428116>
- Godfrey, J. J., Holliman, E. C., & McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1, 517-520. <https://doi.org/10.1109/ICASSP.1992.225858>
- Gordon, M. K. (2016). *Phonological typology*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199669004.001.0001>
- Graddol, D. (2003). The decline of the native speaker. In G. Anderman & M. Rogers (Eds.), *Translation Today: Trends and Perspectives* (pp. 152-167). Multilingual Matters. <https://doi.org/10.21832/9781853596179-013>
- Gu, Y., Cutler, S., Xie, X., & Kurumada, C. (2026). The accent atlas: A geolocation-based assessment of nonnative accent familiarity and linguistic diversity. *Psychonomic Bulletin & Review*, 33, 93. <https://doi.org/10.3758/s13423-025-02768-5>
- Hay, J., Warren, P., and Drager, K. (2006). Factors influencing speech perception in the context of a merger-in-progress. *Journal of Phonetics*, 34(4), 458-484. <https://doi.org/10.1016/j.wocn.2005.10.001>
- Hayes, B. P. (1999). Phonetically driven phonology: The role of Optimality Theory and inductive grounding. In M. Darnell, E. A. Moravcsik, M. Noonan, F. J. Newmeyer, & K. Wheatley (Eds.), *Functionalism and Formalism in Linguistics: Volume I: General Papers* (pp. 243-286). John Benjamins. <https://doi.org/10.1075/slcs.41.13hay>

- Hayes-Harb, R., Barrios, S. L., & Tripp, A. (2022). Whose input matters?: The influence of socially-differentiated input sources in adult Lx phonetic learning. *Journal of Second Language Pronunciation*, 8(3), 363-388. <https://doi.org/10.1075/jslp.21050.hay>
- Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2-3), 61-83. <https://doi.org/10.1017/S0140525X0999152X>
- Hillenbrand, J. M., Clark, M. J. (2009). The role of f0 and formant frequencies in distinguishing the voices of men and women. *Attention, Perception, & Psychophysics*, 71, 1150–1166. <https://doi.org/10.3758/APP.71.5.1150>
- Hillenbrand, J. M., Clark, M. J., & Houde, R. A. (2000). Some effects of duration on vowel recognition. *The Journal of the Acoustical Society of America*, 108(6), 3013-3022. <https://doi.org/10.1121/1.1323463>
- Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America*, 97(5), 3099-3111. <https://doi.org/10.1121/1.411872>
- Holliday, N. (2021). Prosody and sociolinguistic variation in American Englishes. *Annual Review of Linguistics*, 7, 55–68. <https://doi.org/10.1146/annurev-linguistics-031220-093728>
- Hosier, J., Kalfen, J., Sharma, N., & Gurbani, V. K. (2020). A systematic study of open source and commercial text-to-speech (TTS) engines. In P. Sojka, I. Kopeček, K. Pala, & A. Horák (Eds.), *Proceedings of the 23rd International Conference on Text, Speech, and Dialogue*. https://doi.org/10.1007/978-3-030-58323-1_34
- Idemaru, K., Holt, L. L., & Seltman, H. (2012). Individual differences in cue weights are stable across time: The case of Japanese stop lengths. *The Journal of the Acoustical Society of America*, 132(6), 3950-3964. <https://doi.org/10.1121/1.4765076>
- Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, 59(4), 434-446. <https://doi.org/10.1016/j.jml.2007.11.007>
- Jam, B. (2020). Vowel harmony in Persian. *Lingua*, 246, 102905. <https://doi.org/10.1016/j.lingua.2020.102905>

- Jatteau, A., Audibert, N., Vasilescu, I., Lamel, L., & Adda-Decker, M. (2025). Final devoicing before it happens: A large-scale study of word-final obstruents in French. *Laboratory Phonology*, 16(1). <https://doi.org/10.16995/labphon.10855>
- Johnstone, B., Andrus, J., & Danielson, A. E. (2006). Mobility, indexicality, and the enregisterment of “Pittsburghese”. *Journal of English Linguistics*, 34(2), 77-104. <https://doi.org/10.1177/0075424206290692>
- Jongman, A., Wayland, R., & Wong, S. (2000). Acoustic characteristics of English fricatives. *The Journal of the Acoustical Society of America*, 108(3), 1252-1263. <https://doi.org/10.1121/1.1288413>
- Kalikow, D. N., Stevens, K. N., & Elliott, L. L. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *The Journal of the Acoustical Society of America*, 61(5), 1337-1351. <https://doi.org/10.1121/1.381436>
- Kang, O. (2010). Relative salience of suprasegmental features on judgments of L2 comprehensibility and accentedness. *System*, 38(2), 301-315. <https://doi.org/10.1016/j.system.2010.01.005>
- Kapnoula, E. C., Winn, M. B., Kong, E. J., Edwards, J., & McMurray, B. (2017). Evaluating the sources and functions of gradiency in phoneme categorization: An individual differences approach. *Journal of Experimental Psychology: Human Perception and Performance*, 43(9), 1594–1611. <https://doi.org/10.1037/xhp0000410>
- Kato, M., & Baese-Berk, M. M. (2022). Perceptual consequences of native and non-native clear speech. *The Journal of the Acoustical Society of America*, 151(2), 1246-1258. <https://doi.org/10.1121/10.0009403>
- Kato, M., & Baese-Berk, M. M. (2024). The effects of acoustic and semantic enhancements on perception of native and non-native speech. *Language and Speech*, 67(1), 40-71. <https://doi.org/10.1177/00238309231156615>
- Keaton, A. & Zellou, G. (2026). Exploring variation in sociolinguistic evaluation across human and machine talkers. *Linguistics Vanguard*. <https://doi.org/10.1515/lingvan-2025-0101>

- Kim, S-E., Chernyak, B. R., Keshet, J., Goldrick, M., & Bradlow, A. R. (2025). Predicting relative intelligibility from inter-talker distances in a perceptual similarity space for speech. *Psychonomic Bulletin & Review*, 32, 1664-1675.
<https://doi.org/10.3758/s13423-025-02652-2>
- Kleinschmidt, D. F. (2019). Structure in talker variability: How much is there and how much can it help? *Language, Cognition and Neuroscience*, 34(1), 43-68.
<https://doi.org/10.1080/23273798.2018.1500698>
- Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, 122(2), 148-203. <https://doi.org/10.1037/a0038695>
- Kleinschmidt, D. F., & Jaeger, T. F. (2016). Re-examining selective adaptation: Fatiguing feature detectors, or distributional learning? *Psychonomic Bulletin & Review*, 23, 678–691.
<https://doi.org/10.3758/s13423-015-0943-z>
- Kleinschmidt, D. F., Weatherholtz, K., & Jaeger, T. F. (2018). Sociolinguistic perception as inference under uncertainty. *Topics in Cognitive Science*, 10(4), 818-834.
<https://doi.org/10.1111/tops.12331>
- Koops, C., Gentry, E., and Pantos, A. (2008). The effect of perceived speaker age on the perception of PIN and PEN vowels in Houston, Texas. *University of Pennsylvania Working Papers in Linguistics*, 14(2), 93-101.
<https://repository.upenn.edu/handle/20.500.14332/44691>
- Kraljic, T., & Samuel, A. G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology*, 51(2), 141-178.
<https://doi.org/10.1016/j.cogpsych.2005.05.001>
- Kraljic, T., & Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. *Journal of Memory and Language*, 56(1), 1-15. <https://doi.org/10.1016/j.jml.2006.07.010>
- Kruschke, J. K. (2018). Rejecting or accepting parameter values in Bayesian estimation. *Advances in Methods and Practices in Psychological Science*, 1(2), 270-280.
<https://doi.org/10.1177/2515245918771304>
- Kutlu, E., Baxelbaum, K., Sorensen, E., Oleson, J., & McMurray, B. (2024). Linguistic diversity shapes flexible speech perception in school age children. *Scientific Reports*, 14, 28825. <https://doi.org/10.1038/s41598-024-80430-1>

- Labov, W. (1980). The social origins of sound change. In W. Labov (Ed.), *Locating Language in Time and Space* (pp. 251-266). Academic Press.
- Labov, W., Ash, S., & Boberg, C. (2005). *The atlas of North American English: Phonetics, phonology and sound change*. De Gruyter Mouton.
<https://doi.org/10.1515/9783110167467>
- Lai, W., & Tamminga, M. (2024). Phonetics-phonology mapping in the generalization of perceptual learning. *Journal of Phonetics*, 103, 101295.
<https://doi.org/10.1016/j.wocn.2024.101295>
- Liberman, A. M., Cooper, F. S., Shankweiler, D. P. & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, 74(6), 431-461.
<https://doi.org/10.1037/h0020279>
- Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, 54(5), 358-368. <https://doi.org/10.1037/h0044417>
- Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, 21(1), 1-36. [https://doi.org/10.1016/0010-0277\(85\)90021-6](https://doi.org/10.1016/0010-0277(85)90021-6)
- Lilley, K. D., Dossey, E., Cohn, M., Clopper, C. G., Wagner, L., & Zellou, G. (2025). Social evaluation of text-to-speech voices by adults and children. *Speech Communication*, 166, 103163. <https://doi.org/10.1016/j.specom.2024.103163>
- Lippi-Green, R. (1994). Accent, standard language ideology, and discriminatory pretext in the courts. *Language in Society*, 23(2), 163-198. <https://doi.org/10.1017/S0047404500017826>
- Liu, L. (2018). *Understanding talkers: Adaptation, generalization, and causal reasoning during speech perception*. [Doctoral dissertation, University of Rochester].
<http://hdl.handle.net/1802/34302>
- Liu, Y., & Wang, H. (2026). Who on earth is using Generative AI? *World Development*, 107260. <https://doi.org/10.1016/j.worlddev.2025.107260>
- Lockenwitz, S., Tetnowski, J. A., & Oxley, J. (2020). The sociolinguistics of lisping: A review. *Clinical Linguistics & Phonetics*, 34(12), 1169-1184.
<https://doi.org/10.1080/02699206.2020.1788167>

- Luthra, S. (2024). Why are listeners hindered by talker variability? *Psychonomic Bulletin & Review*, 31, 104-121. <https://doi.org/10.3758/s13423-023-02355-6>
- Luthra, S., Mechtenberg, H., & Myers, E. B. (2021). Perceptual learning of multiple talkers requires additional exposure. *Attention, Perception, & Psychophysics*, 83, 2217-2228. <https://doi.org/10.3758/s13414-021-02261-w>
- Lutz, C., & Newlands, G. (2025). Glitch studies and smart speakers. In S. Habscheid, T. Hector, D. Hoffman, & D. Waldecker (Eds.), *Voice Assistants in Private Homes: Media, Data and Language in Interaction and Discourse* (pp. 243-262). Transcript Verlag. <https://doi.org/10.14361/9783839472002-009>
- Magen, H. S. (1998). The perception of foreign-accented speech. *Journal of Phonetics*, 26(4), 381-400. <https://doi.org/10.1006/jpho.1998.0081>
- Magnuson, J. S., & Nusbaum, H. C. (2007). Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. *Journal of Experimental Psychology: Human Perception and Performance*, 33(2), 391-409. <https://doi.org/10.1037/0096-1523.33.2.391>
- Magnuson, J. S., Nusbaum, H. C., Akahane-Yamada, R., & Saltzman, D. (2021). Talker familiarity and the accommodation of talker variability. *Attention, Perception, & Psychophysics*, 83, 1842-1860. <https://doi.org/10.3758/s13414-020-02203-y>
- Martone, M. E., Garcia-Castro, A., & VandenBos, G. R. (2018). Data sharing in psychology. *American Psychologist*, 73(2), 111-125. <https://doi.org/10.1037/amp0000242>
- Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, 94, 305-315. <https://doi.org/10.1016/j.jml.2017.01.001>
- McCloy, D. (2015). *Mix speech with noise* [Praat script]. Retrieved April, 5 2024 from: <https://github.com/drammock/praat-semiauto/blob/master/MixSpeechNoise.praat>
- McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264, 746-768. <https://doi.org/10.1038/264746a0>
- McLaughlin, D. J. (2022). *The effect of social primes on the perception of native- and nonnative-accented speech*. [Doctoral dissertation, Washington University in St. Louis]. <https://doi.org/10.7936/fg5y-cz79>

- McLaughlin, D. J., & Van Engen, K. J. (2023). Exploring effects of social information on talker-independent accent adaptation. *JASA Express Letters*, 3(12), 125201. <https://doi.org/10.1121/10.0022536>
- McMurray, B. (2022). The myth of categorical perception. *The Journal of the Acoustical Society of America*, 152(6), 3819-3842. <https://doi.org/10.1121/10.0016614>
- McMurray, B., & Jongman, A. (2011). What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological Review*, 118(2), 219-246. <https://doi.org/10.1037/a0022325>
- Merritt, B., & Bent, T. (2022). Revisiting the acoustics of speaker gender perception: A gender expansive perspective. *The Journal of the Acoustical Society of America*, 151(1), 484-499. <https://doi.org/10.1121/10.0009282>
- Meteyard, L., & Davies, R. A. I. (2020). Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language*, 112, 104092. <https://doi.org/10.1016/j.jml.2020.104092>
- Miao, Y., & Kang, O. (2023). An empirical approach to measuring accent familiarity: Phonological and correlational analyses. *System*, 116, 103089. <https://doi.org/10.1016/j.system.2023.103089>
- Morris, R. J., McCrea, C. R., & Herring, K. D. (2008). Voice Onset Time differences between adult males and females: Isolated syllables. *Journal of Phonetics*, 36(2), 308-317. <https://doi.org/10.1016/j.wocn.2007.06.003>
- Mullennix, J. W., Pisoni, D. B., & Martin, C. S. (1989). Some effects of talker variability on spoken word recognition. *The Journal of the Acoustical Society of America*, 85(1), 365-378. <https://doi.org/10.1121/1.397688>
- National Library of Medicine. (2024, February 26). *Generative Artificial Intelligence*. Retrieved December 27, 2025 from: <https://www.nlm.gov/guides/data-thesaurus/generative-artificial-intelligence>
- Newman, M., & Wu, A. (2011). “Do you sound Asian when you speak English?”: Racial identification and voice in Chinese and Korean Americans’ English. *American Speech*, 86(2), 152-178. <https://doi.org/10.1215/00031283-1336992>

- Newman, R. S., Clouse, S. A., & Burnham, J. L. (2001). The perceptual consequences of within-talker variability in fricative production. *The Journal of the Acoustical Society of America*, 109(3), 1181-1196. <https://doi.org/10.1121/1.1348009>
- Niedzielski, N. (1999). The effect of social information on the perception of sociolinguistic variables. *Journal of Language and Social Psychology*, 18(1), 62-85. <https://doi.org/10.1177/0261927X99018001005>
- Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, 47(2), 204-238. [https://doi.org/10.1016/S0010-0285\(03\)00006-9](https://doi.org/10.1016/S0010-0285(03)00006-9)
- Nycz, J. (2019). Media and second dialect acquisition. *Annual Review of Applied Linguistics*, 39, 152-160. <https://doi.org/10.1017/S0267190519000060>
- Nygaard, L. C. (2005). Perceptual integration of linguistic and nonlinguistic properties of speech. In D. B. Pisoni & R. E. Remez (Eds.), *The Handbook of Speech Perception* (pp. 390-413). Blackwell. <https://doi.org/10.1002/9780470757024.ch16>
- Piantadosi, S. T. (2024). Modern language models refute Chomsky's approach to language. In E. Gibson & M. Poliak (Eds.), *From Fieldwork to Linguistic Theory: A Tribute to Dan Everett* (pp. 353-414). Language Science Press. <https://lingbuzz.net/lingbuzz/007180>
- Piazza, G., Martin, C. D., & Kalashnikova, M. (2022). The acoustic features and didactic function of foreigner-directed speech: A scoping review. *Journal of Speech, Language, and Hearing Research*, 65(8), 2896-2918. https://doi.org/10.1044/2022_JSLHR-21-00609
- Picheny, M. A., Durlach, N. I., & Braida, L. D. (1985). Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech. *Journal of Speech, Language, and Hearing Research*, 28(1), 96-103. <https://doi.org/10.1044/jshr.2801.96>
- Picheny, M. A., Durlach, N. I., & Braida, L. D. (1986). Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech. *Journal of Speech, Language, and Hearing Research*, 29(4), 434-446. <https://doi.org/10.1044/jshr.2904.434>
- Pisoni, D. B. (1981). Some current theoretical issues in speech perception. *Cognition*, 10(1-3), 249-259. [https://doi.org/10.1016/0010-0277\(81\)90054-8](https://doi.org/10.1016/0010-0277(81)90054-8)
- Pitt, M. A., Johnson, K., Hume, E., Kiesling, S., & Raymond, W. (2005). The Buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability. *Speech Communication*, 45(1), 89-95. <https://doi.org/10.1016/j.specom.2004.09.001>

- Port, R. F., & Leary, A. P. (2005). Against formal phonology. *Language*, 81(4), 927-964. <https://doi.org/10.1353/lan.2005.0195>
- Purnell, T., Salmons, J., & Tepeli, D. (2005). German substrate effects in Wisconsin English: Evidence for final fortition. *American Speech*, 80(2), 135-164. <https://doi.org/10.1215/00031283-80-2-135>
- Qualtrics. (n.d.). *Understanding your dataset*. Retrieved May 16, 2026 from: <https://www.qualtrics.com/support/survey-platform/data-and-analysis-module/data/download-data/understanding-your-dataset/>
- Quam, C., & Creel, S. C. (2021). Impacts of acoustic-phonetic variability on perceptual development for spoken language: A review. *WIREs Cognitive Science*, 12(5), e1558. <https://doi.org/10.1002/wcs.1558>
- R Core Team. (2021). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. <https://www.R-project.org/>
- Raviv, L., Lupyan, G., & Green, S. C. (2022). How variability shapes learning and generalization. *Trends in Cognitive Sciences*, 26(6), 462-483. <https://doi.org/10.1016/j.tics.2022.03.007>
- Reinisch, E. & Holt, L. L. (2014). Lexically guided phonetic retuning of foreign-accented speech and its generalization. *Journal of Experimental Psychology: Human Perception and Performance*, 40(2), 539-555. <https://doi.org/10.1037/a0034409>
- Rickford, J. R., & King, S. (2016). Language and linguistics on trial: Hearing Rachel Jeantel (and other vernacular speakers) in the courtroom and beyond. *Language*, 92(4), 948-988. <https://doi.org/10.1353/lan.2016.0078>
- Robb, M., Gilbert, H., & Lerman, J. (2005). Influence of gender and environmental setting on Voice Onset Time. *Folia Phoniatrica et Logopaedica*, 57(3), 125-133. <https://doi.org/10.1159/000084133>
- Ronge, R., Maier, M. & Rathgeber, B. (2025). Towards a definition of Generative Artificial Intelligence. *Philosophy & Technology*, 38, 31. <https://doi.org/10.1007/s13347-025-00863-y>
- Samuel, A. G. (2020). Psycholinguists should resist the allure of linguistic units as perceptual units. *Journal of Memory and Language*, 111, 104070. <https://doi.org/10.1016/j.jml.2019.104070>

- Samuel, A. G., & Kraljic, T. (2009). Perceptual learning for speech. *Attention, Perception, & Psychophysics*, 71, 1207-1218. <https://doi.org/10.3758/APP.71.6.1207>
- Sancier, M. L., & Fowler, C. A. (1997). Gestural drift in a bilingual speaker of Brazilian Portuguese and English. *Journal of Phonetics*, 25(4), 421-436. <https://doi.org/10.1006/jpho.1997.0051>
- Saniyar, E., & Tunçer, Ü. C. (2022). Continuant devoicing in Turkish: Reanalysis of Turkish word-final devoicing. *Proceedings of the Workshop on Turkic and Languages in Contact with Turkic*, 7, 164-171. <https://doi.org/10.3765/ptu.v7i1.5314>
- Sarkar, A. (2025). Gluing pizza, eating rocks, and counting Rs in strawberry: The discursive social function of stupid AI answers. *Proceedings of the 4th Annual Symposium on Human-Computer Interaction for Work*, Article 25. <https://doi.org/10.1145/3729176.3729189>
- Scarborough, R., & Zellou, G. (2013). Clarity in communication: “Clear” speech authenticity and lexical neighborhood density effects in speech production and perception. *The Journal of the Acoustical Society of America*, 134(5), 3793-3807. <https://doi.org/10.1121/1.4824120>
- Schertz, J., & Clare, E. J. (2020). Phonetic cue weighting in perception and production. *WIREs Cognitive Science*, 11(2), e1521. <https://doi.org/10.1002/wcs.1521>
- ShareAmerica. (2023). *The United States is rich in languages*. Retrieved December 17, 2024 from: <https://share.america.gov/united-states-is-rich-in-languages/>
- Shi, L., Griffiths, T. L., Feldman, N. H., & Sanborn, A. N. (2010). Exemplar models as a mechanism for performing Bayesian inference. *Psychonomic Bulletin & Review*, 17(4), 443-464. <https://doi.org/10.3758/pbr.17.4.443>
- Shin, S.-Y., Lee, S., & Lidster, R. (2021). Examining the effects of different English speech varieties on an L2 academic listening comprehension test at the item level. *Language Testing*, 38(4), 580–601. <https://doi.org/10.1177/0265532220985432>
- Sidasaras, S. K., Alexander, J. E. D., & Nygaard, L. C. (2009). Perceptual learning of systematic variation in Spanish-accented speech. *The Journal of the Acoustical Society of America*, 125(5), 3306-3316. <https://doi.org/10.1121/1.3101452>

- Smiljanić, R. (2021). Clear speech perception: Linguistic and cognitive benefits. In J. S. Pardo, L. C. Nygaard, R. E. Remez, & D. B. Pisoni (Eds.), *The Handbook of Speech Perception, Second Edition* (pp. 177-205). Wiley. <https://doi.org/10.1002/9781119184096.ch7>
- Smiljanić, R., & Bradlow, A. R. (2009). Speaking and hearing clearly: Talker and listener factors in speaking style changes. *Language and Linguistics Compass*, 3(1), 236-264. <https://doi.org/10.1111/j.1749-818X.2008.00112.x>
- Smith, B. L., & Kenney, M. K. (1994). Variability control in speech production tasks performed by adults and children. *The Journal of the Acoustical Society of America*, 96(2), 699-705. <https://doi.org/10.1121/1.410308>
- Sonderegger, M. (2023). *Regression modeling for linguistic data*. MIT Press. Retrieved June 19, 2025 from https://github.com/msonderegger/rmld-v1.1/blob/main/rmld_v1.1.pdf
- Sonderegger, M., & Sóskuthy, M. (2025). Advancements of phonetics in the 21st century: Quantitative data analysis. *Journal of Phonetics*, 111, 101415. <https://doi.org/10.1016/j.wocn.2025.101415>
- Stan Development Team (2023). *Stan Modeling Language Users Guide and Reference Manual* (Version 2.2.1). <https://mc-stan.org>
- Strand, E. A., & Johnson, K. (1996). Gradient and visual speaker normalization in the perception of fricatives. In D. Gibbon (Ed.), *Natural language processing and speech technology: Results of the 3rd KONVENS conference* (pp. 14-26). Mouton.
- Sumner, M. (2011). The role of variation in the perception of accented speech. *Cognition*, 119(1), 131-136. <https://doi.org/10.1016/j.cognition.2010.10.018>
- Sumner, M., & Samuel, A. G. (2005). Perception and representation of regular variation: The case of final /t/. *Journal of Memory and Language*, 52(3), 322-338. <https://doi.org/10.1016/j.jml.2004.11.004>
- Swartz, B. L. (1992). Gender difference in Voice Onset Time. *Perceptual and Motor Skills*, 75(3), 983-992. <https://doi.org/10.2466/pms.1992.75.3.983>
- Tamminga, M., Wilder, R., Lai, W., & Wade, L. (2020). Perceptual learning, talker specificity, and sound change. *Papers in Historical Phonology*, 5, 90-122. <https://doi.org/10.2218/pihph.5.2020.4439>

- Tang, A., Li, K-K., Kwok, K. O., Cao, L., Luong, S. & Tam, W. (2024). The importance of transparency: Declaring the use of generative artificial intelligence (AI) in academic writing. *Journal of Nursing Scholarship*, 56(2), 314–318. <https://doi.org/10.1111/jnu.12938>
- Tendeiro, J. N., Kiers, H. A. L., Hoekstra, R., Wong, T. K., & Morey, R. D. (2024). Diagnosing the misuse of the Bayes Factor in applied research. *Advances in Methods and Practices in Psychological Science*, 7(1), 1-19. <https://doi.org/10.1177/25152459231213371>
- Titze, I. R. (1989). Physiologic and acoustic differences between male and female voices. *The Journal of the Acoustical Society of America*, 85(4), 1699-1707. <https://doi.org/10.1121/1.397959>
- Tomar, S. (2006). Converting video formats with FFmpeg. *Linux Journal*, 2006(146), 1-10. <https://dl.acm.org/doi/abs/10.5555/1134782.1134792>
- Tuttösi, P., Hughson, E., Matsufuji, A., Zhang, C., & Lim, A. (2023). Read the room: Adapting a robot's voice to ambient and social contexts. *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 3998-4005. <https://doi.org/10.1109/IROS55552.2023.10341925>
- Tzeng, C. Y., Alexander, J. E. D., Sidaras, S. K., & Nygaard, L. C. (2016). The role of training structure in perceptual learning of accented speech. *Journal of Experimental Psychology: Human Perception and Performance*, 42(11), 1793–1805. <https://doi.org/10.1037/xhp0000260>
- Tzeng, C. Y., Nygaard, L. C., & Theodore, R. M. (2021). A second chance for a first impression: Sensitivity to cumulative input statistics for lexically guided perceptual learning. *Psychonomic Bulletin & Review*, 28, 1003-1014. <https://doi.org/10.3758/s13423-020-01840-6>
- Tzeng, C. Y., Russell, M. L., & Nygaard, L. C. (2024). Attention modulates perceptual learning of non-native-accented speech. *Attention, Perception, & Psychophysics*, 86, 339–353. <https://doi.org/10.3758/s13414-023-02790-6>
- Uchanski, R. M. (2005). Clear speech. In D. B. Pisoni & R. E. Remez (Eds.), *The Handbook of Speech Perception* (pp. 207-235). Blackwell. <https://doi.org/10.1002/9780470757024.ch9>

- University of Connecticut. (n.d.) *Total University (Excluding Health Center) Enrollment Headcount by Ethnicity, Gender, Percent Minority, and Percent Female: Fall 1990-2024*. Retrieved May 16, 2026 from:
https://bpir.media.uconn.edu/wp-content/uploads/sites/3452/2024/12/enr_eth_ALL_90-24.pdf
- Van Engen, K. J., Chandrasekaran, B., & Smiljanić, R. (2012). Effects of speech clarity on recognition memory for spoken sentences. *PLoS ONE*, 7(9), e43753.
<https://doi.org/10.1371/journal.pone.0043753>
- Vasishth, S., Nicenboim, B., Beckman, M. E., Li, F., & Kong, E. J. (2018). Bayesian data analysis in the phonetic sciences: A tutorial introduction. *Journal of Phonetics*, 71, 147-161. <https://doi.org/10.1016/j.wocn.2018.07.008>
- Vaughn, C. R. (2019). Expectations about the source of a speaker's accent affect accent adaptation. *The Journal of the Acoustical Society of America*, 145(5), 3218-3232.
<https://doi.org/10.1121/1.5108831>
- Verhoeven, J., Hirson, A., & Basavaraj, K. (2011). The devoicing of fricatives in Southern British English. In W-S. Lee & E. Zee (Eds.), *Proceedings of the 17th International Congress of Phonetic Sciences* (pp. 2070-2073).
<https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2011/OnlineProceedings/RegularSession/Verhoeven/Verhoeven.pdf>
- Vonessen, J., Aoki, N. B., Cohn, M., & Zellou, G. (2024). Comparing perception of L1 and L2 English by human listeners and machines: Effect of interlocutor adaptations. *The Journal of the Acoustical Society of America*, 155(5), 3060-3070.
<https://doi.org/10.1121/10.0025930>
- Wang, H., Knight, R-A., Dipper, L., Alderton, R., & Alyahya, R. S. W. (2025). Systematic review: The identification of segmental Mandarin-accented English features. *Speech Communication*, 167, 103168. <https://doi.org/10.1016/j.specom.2024.103168>
- Wang, Y., Spence, M. M., Jongman, A., & Sereno, J. A. (1999). Training American listeners to perceive Mandarin tones. *The Journal of the Acoustical Society of America*, 106(6), 3649-3658. <https://doi.org/10.1121/1.428217>
- Weatherholtz, K., & Jaeger, T. F. (2016). Speech perception and generalization across talkers and accents. *Oxford Research Encyclopedia of Linguistics*.
<https://doi.org/10.1093/acrefore/9780199384655.013.95>

- Wimpy, C. (2026). *rurality* (Version 0.1.1). [R package]. Retrieved from:
<https://cran.r-project.org/web/packages/rurality/rurality.pdf>
- Winn, M. (2019). *Make speech-shaped noise* [Praat script]. Retrieved April 5, 2024 from:
http://www.mattwinn.com/praat/Make_SSN_from_LTAS_selected_sounds.txt
- Winn, M. (2022). *Make Voice Onset Time (VOT)/F0 continuum* (Version 33) [Praat script]. Retrieved April 18, 2024 from:
http://www.mattwinn.com/praat/Make_VOT_Continuum_v33.txt
- World Medical Association. (2025). World Medical Association Declaration of Helsinki: Ethical principles for medical research involving human participants. *Journal of the American Medical Association*, 333(1), 71-74. <https://doi.org/10.1001/jama.2024.21972>
- Xie, X., & Fowler, C. A. (2013). Listening with a foreign-accent: The interlanguage speech intelligibility benefit in Mandarin speakers of English. *Journal of Phonetics*, 41(5), 369–378. <https://doi.org/10.1016/j.wocn.2013.06.003>
- Xie, X., & Jaeger, T. F. (2020). Comparing non-native and native speech: Are L2 productions more variable? *The Journal of the Acoustical Society of America*, 147(5), 3322-3347. <https://doi.org/10.1121/10.0001141>
- Xie, X., Jaeger, T. F., & Kurumada, C. (2023). What we do (not) know about the mechanisms underlying adaptive speech perception: A computational framework and review. *Cortex*, 166, 377–424. <https://doi.org/10.1016/j.cortex.2023.05.003>
- Xie, X., & Kurumada, C. (2024). From first encounters to longitudinal exposure: A repeated exposure-test paradigm for monitoring speech adaptation. *Frontiers in Psychology*, 15, 1383904. <https://doi.org/10.3389/fpsyg.2024.1383904>
- Xie, X., Liu, L., & Jaeger, T. F. (2021). Cross-talker generalization in the perception of nonnative speech: A large-scale replication. *Journal of Experimental Psychology: General*, 150(11), e22-e56. <https://doi.org/10.1037/xge0001039>
- Xie, X., & Myers, E. B. (2017). Learning a talker or learning an accent: Acoustic similarity constrains generalization of foreign accent adaptation to new talkers. *Journal of Memory and Language*, 97, 30-46. <https://doi.org/10.1016/j.jml.2017.07.005>

- Yu, A. C. L. (2010). Perceptual compensation is correlated with individuals' "autistic" traits: Implications for models of sound change. *PLOS One*, 5(8), e11950. <https://doi.org/10.1371/journal.pone.0011950>
- Yu, A. C. L. (2023). The actuation problem. *Annual Review of Linguistics*, 9, 215-231. <https://doi.org/10.1146/annurev-linguistics-031120-101336>
- Yu, A. C. L., & Zellou, G. (2019). Individual differences in language processing: Phonology. *Annual Review of Linguistics*, 5, 131-150. <https://doi.org/10.1146/annurev-linguistics-011516-033815>
- Zellou, G. (2017). Individual differences in the production of nasal coarticulation and perceptual compensation. *Journal of Phonetics*, 61, 13-29. <https://doi.org/10.1016/j.wocn.2016.12.002>
- Zellou, G., Barreda, S., Lahrouchi, M., & Smiljanić, R. (2024). Learning a language with vowelless words. *Cognition*, 251, 105909. <https://doi.org/10.1016/j.cognition.2024.105909>
- Zellou, G., Cohn, M., & Pycha, A. (2023). Listener beliefs and perceptual learning: Differences between device and human guises. *Language*, 99(4), 692-725. <https://doi.org/10.1353/lan.2023.a914191>
- Zellou, G., Lahrouchi, M., & Bensoukas, K. (2022). Clear speech in Tashlhiyt Berber: The perception of typologically uncommon word-initial contrasts by native and naive listeners. *The Journal of the Acoustical Society of America*, 152(6), 3429-3443. <https://doi.org/10.1121/10.0016579>
- Zhang, X., & Samuel, A. G. (2014). Perceptual learning of speech under optimal and adverse conditions. *Journal of Experimental Psychology: Human Perception and Performance*, 40(1), 200-217. <https://doi.org/10.1037/a0033182>

Appendix A: Experimental Materials

Table A.1: Sentences presented in the exposure and test phases of Experiment 1. All of the sentences come from Kalikow et al. (1977).

ID	Phase	Sentence
1	Exposure	The old man discussed the dive.
2	Exposure	Bob heard Paul called about the strips.
3	Exposure	I should have considered the map.
4	Exposure	Miss Brown shouldn't discuss the sand.
5	Exposure	They might have considered the hive.
6	Exposure	David has discussed the dent.
7	Exposure	He can't consider the crib.
8	Exposure	David does not discuss the hug.
9	Exposure	We've been discussing the crates.
10	Exposure	Miss Black knew about the doll.
11	Exposure	She couldn't discuss the pine.
12	Exposure	Miss Black thought about the lap.
13	Exposure	Mr. Black knew about the pad.
14	Exposure	You heard Jane called about the van.
15	Exposure	Tom wants to know about the cake.
16	Exposure	We hear you called about the lock.

ID	Phase	Sentence
17	Exposure	The old man discussed the yell.
18	Exposure	They're glad we heard about the track.
19	Exposure	Sue was interested in the bruise.
20	Exposure	Ruth will consider the herd.
21	Exposure	The girl talked about the gin.
22	Exposure	Paul can't discuss the wax.
23	Exposure	I hope Paul asked about the mate.
24	Exposure	They did not discuss the screen.
25	Exposure	They were interested in the strap.
26	Exposure	I had a problem with the bloom.
27	Exposure	Peter should speak about the mugs.
28	Exposure	The boy would discuss the scab.
29	Exposure	Tom could have thought about the sport.
30	Exposure	You'd been considering the geese.
31	Exposure	Mr. Black considered the fleet.
32	Exposure	I want to know about the crop.
33	Exposure	Betty has talked about the draft.
34	Exposure	Tom discussed the hay.
35	Exposure	Jane was interested in the stamp.
36	Exposure	I had not thought about the growl.
37	Exposure	Paul should know about the net.
38	Exposure	Tom heard Jane called about the booth.
39	Exposure	We can't consider the wheat.
40	Test	We have not discussed the steam.
41	Test	Miss Brown might consider the coast.

ID	Phase	Sentence
42	Test	He hasn't considered the dart.
43	Test	Mr. Smith thinks about the cap.
44	Test	We should have considered the juice.
45	Test	Bob could have known about the spoon.
46	Test	Bill might discuss the foam.
47	Test	Tom could not discuss the barn.
48	Test	Nancy should consider the fist.
49	Test	I've spoken about the pile.
50	Test	Jane has a problem with the coin.
51	Test	Mary could not discuss the tack.
52	Test	They have a problem with the limb.
53	Test	Nancy had considered the sleeves.
54	Test	Bill heard Tom called about the coach.
55	Test	The old man thinks about the mast.
56	Test	Tom will discuss the swan.
57	Test	Ann was interested in the breath.
58	Test	Ruth hopes he heard about the hips.
59	Test	Tom is talking about the fee.
60	Test	Miss Smith considered the scare.
61	Test	David should consider the blame.
62	Test	He's glad you called about the jar.
63	Test	He hopes Tom asked about the bar.
64	Test	Miss Black could have discussed the rope.
65	Test	Ruth's grandmother discussed the broom.
66	Test	I've been considering the crown.

ID	Phase	Sentence
67	Test	I want to speak about the crash.
68	Test	She's glad Jane asked about the drain.
69	Test	He had considered the robe.
70	Test	I haven't discussed the sponge.
71	Test	You cannot have discussed the grease.
72	Test	She wants to talk about the crew.
73	Test	They knew about the fur.
74	Test	We've spoken about the truck.
75	Test	You're glad she called about the bowl.
76	Test	You hope they asked about the vest.
77	Test	You want to talk about the ditch.
78	Test	We hear she called about the drum.

Table A.2: Sentences presented in the exposure phase of Experiment 2. The first 16 sentences are based on Zellou et al. (2023), while the rest were created by the author.

ID	Sentence	Target	Competitor
1	Lisa had no shovel so she could not dig.	dig	sick
2	A large aquarium holds many fish.	fish	switch
3	The shoes were so small that her feet would not fit.	fit	wig
4	Molly received a doll as a gift.	gift	dish
5	She smiled with a cheeky grin.	grin	twin
6	The mother gave her child a kiss.	kiss	fin
7	The man who owned the mansion was rich.	rich	kid
8	Gordon tried to climb over the ridge.	ridge	rim
9	He had many dishes to rinse.	rinse	dense
10	He would not sky dive due to the risk.	risk	desk
11	At the soccer game, Kayla hurt her shin.	shin	thin
12	She came out to the lake for a swim.	swim	swag
13	My wool socks are warm and thick.	thick	fix
14	The storm reduced the large tree into a twig.	twig	fig
15	The kite was lost to the wind.	wind	hint
16	The genie announced he would grant one wish.	wish	kick
17	The wall needed a new coat of paint.	paint	part
18	Instead of eating the entire cake, Joey only ate a part.	part	past
19	Martha focuses on the future without dwelling on the past.	past	pause
20	She started talking after a long pause.	pause	paw
21	The dog had a furry paw.	paw	pawn
22	Chess games start with sixteen pawns.	pawns	peace
23	The minister vowed to achieve world peace.	peace	peel
24	A key ingredient in the dessert is the zest from an orange peel.	peel	perk

ID	Sentence	Target	Competitor
25	Free food was a great perk.	perk	pinch
26	Do not add too much salt to the stew, just add a pinch.	pinch	pine
27	The travelers walked through forests of oak and pine.	pine	pink
28	Flamingos are animals that are pink.	pink	place
29	New York City is a crowded place.	place	plain
30	The dessert was unsatisfactory as the flavor was dull and plain.	plain	plan
31	To save the organization, the CEO had a plan.	plan	plant
32	The seedling grew into a healthy plant.	plant	plate
33	The naughty girl had no veggies on her plate.	plate	plea
34	The crying man made an emotional plea.	plea	plug
35	Charging a cell phone often requires a plug.	plug	plunge
36	With her swimsuit on, Rachel took a plunge.	plunge	point
37	The soccer team won the match by one point.	point	pork
38	The tacos contained chicken and pork.	pork	port
39	A horn announced that the ferry had left the port.	port	pose
40	At the end of the runway, the model struck a pose.	pose	pouch
41	The kangaroos were nursed in the mothers pouch.	pouch	prank
42	Danny was a jokester who was known to love a good prank.	prank	priest
43	After Sunday mass, the churchgoer chatted with the priest.	priest	prince
44	The resident of the castle is a young prince.	prince	print
45	The novel is now in print.	print	prize
46	In the athletic contest, the fastest runner won the grand prize.	prize	pulse
47	The doctor tried to revive the girl who had no pulse.	pulse	purse
48	Sandra always stored her wallet in her purse.	purse	paint

Appendix B: Model Output

Table B.1: Summary statistics for the original ‘single talker per gender’ exposure model in Experiment 2. Credible intervals that do not cross zero are in bold.

Effect	Estimate	Est. Error	l-95% CI	u-95% CI
Intercept	1.48	0.76	-0.15	2.92
Step	4.91	0.12	4.69	5.14
Gender	0.22	0.74	-1.27	1.74
Order	-0.14	0.11	-0.36	0.09
Shift	0.08	0.12	-0.14	0.31
Step : Gender	-0.004	0.09	-0.18	0.17
Step : Order	-0.11	0.10	-0.31	0.09
Gender : Order	-0.02	0.07	-0.16	0.13
Step : Shift	-0.02	0.10	-0.21	0.18
Gender : Shift	-0.01	0.07	-0.16	0.13
Order : Shift	0.08	0.12	-0.16	0.31
Step : Gender : Order	-0.12	0.07	-0.25	0.02
Step : Gender : Shift	0.02	0.07	-0.11	0.15
Step : Order : Shift	0.15	0.10	-0.04	0.35
Gender : Order : Shift	0.05	0.07	-0.09	0.19
Step : Gender : Order : Shift	0.02	0.07	-0.11	0.15

Table B.2: Summary statistics for the original ‘multiple talkers per gender’ exposure model in Experiment 2. Credible intervals that do not cross zero are in bold.

Effect	Estimate	Est. Error	l-95% CI	u-95% CI
Intercept	1.28	0.65	-0.09	2.59
Step	4.60	0.11	4.39	4.82
Gender	0.19	0.64	-1.12	1.50
Order	-0.30	0.11	-0.51	-0.09
Shift	-0.03	0.11	-0.24	0.19
Step : Gender	-0.14	0.09	-0.31	0.02
Step : Order	-0.11	0.09	-0.29	0.06
Gender : Order	0.13	0.07	-0.01	0.28
Step : Shift	-0.13	0.09	-0.30	0.05
Gender : Shift	0.07	0.07	-0.08	0.22
Order : Shift	-0.04	0.11	-0.26	0.17
Step : Gender : Order	-0.16	0.07	-0.30	-0.03
Step : Gender : Shift	-0.01	0.07	-0.15	0.12
Step : Order : Shift	-0.15	0.09	-0.33	0.03
Gender : Order : Shift	0.17	0.07	0.03	0.33
Step : Gender : Order : Shift	0.11	0.07	-0.03	0.25

Table B.3: Summary statistics for the combined model in Experiment 2. Credible intervals that do not cross zero are in bold.

Effect	Estimate	Est. Error	l-95% CI	u-95% CI
Intercept	1.37	0.71	-0.08	2.78
Step	4.75	0.08	4.60	4.91
Gender	0.20	0.71	-1.19	1.69
Order	-0.22	0.08	-0.37	-0.06
Shift	0.03	0.08	-0.12	0.18
Variability	-0.09	0.08	-0.24	0.06
Step : Gender	-0.08	0.06	-0.20	0.04
Step : Order	-0.11	0.07	-0.25	0.02
Gender : Order	0.06	0.05	-0.05	0.16
Step : Shift	-0.08	0.07	-0.21	0.06
Gender : Shift	0.03	0.05	-0.07	0.13
Order : Shift	0.02	0.08	-0.14	0.18
Step : Variability	-0.11	0.07	-0.25	0.02
Gender : Variability	-0.001	0.05	-0.10	0.10
Order : Variability	-0.09	0.08	-0.24	0.07
Shift : Variability	-0.05	0.08	-0.22	0.10
Step : Gender : Order	-0.14	0.05	-0.23	-0.04
Step : Gender : Shift	-0.005	0.05	-0.10	0.09
Step : Order : Shift	0.003	0.07	-0.13	0.14
Gender : Order : Shift	0.11	0.05	0.01	0.22
Step : Gender : Variability	-0.04	0.05	-0.13	0.06
Step : Order : Variability	-0.01	0.07	-0.14	0.12
Gender : Order : Variability	0.07	0.05	-0.03	0.18
Step : Shift : Variability	-0.06	0.07	-0.19	0.08

Effect	Estimate	Est. Error	l-95% CI	u-95% CI
Gender : Shift : Variability	0.04	0.05	-0.06	0.15
Order : Shift : Variability	-0.06	0.08	-0.21	0.09
Step : Gender : Order : Shift	0.06	0.05	-0.03	0.16
Step : Gender : Order : Variability	-0.02	0.05	-0.12	0.07
Step : Gender : Shift : Variability	-0.01	0.05	-0.11	0.08
Step : Order : Shift : Variability	-0.15	0.07	-0.29	-0.02
Gender : Order : Shift : Variability	0.06	0.05	-0.04	0.17
Step : Gender : Order : Shift : Variability	0.05	0.05	-0.05	0.14

Table B.4: Summary statistics for the combined model in Experiment 2 that omits all listeners who heard Joanna at test. Credible intervals that do not cross zero are in bold.

Effect	Estimate	Est. Error	l-95% CI	u-95% CI
Intercept	0.89	0.28	0.36	1.41
Step	4.67	0.09	4.50	4.86
Gender	-0.46	0.26	-1.00	0.11
Order	-0.11	0.09	-0.29	0.08
Shift	0.02	0.09	-0.16	0.21
Variability	0.03	0.09	-0.15	0.22
Step : Gender	-0.17	0.08	-0.32	-0.02
Step : Order	-0.09	0.08	-0.25	0.07
Gender : Order	0.08	0.06	-0.03	0.19
Step : Shift	-0.08	0.08	-0.24	0.08
Gender : Shift	0.09	0.06	-0.02	0.19
Order : Shift	-0.01	0.09	-0.19	0.17
Step : Variability	-0.14	0.08	-0.30	0.02
Gender : Variability	0.01	0.06	-0.11	0.12
Order : Variability	-0.04	0.09	-0.22	0.14
Shift : Variability	-0.11	0.10	-0.30	0.08
Step : Gender : Order	-0.09	0.06	-0.21	0.02
Step : Gender : Shift	-0.03	0.06	-0.14	0.09
Step : Order : Shift	-0.01	0.08	-0.16	0.15
Gender : Order : Shift	0.23	0.06	0.12	0.33
Step : Gender : Variability	-0.02	0.06	-0.13	0.09
Step : Order : Variability	0.01	0.08	-0.15	0.17
Gender : Order : Variability	0.02	0.06	-0.09	0.13
Step : Shift : Variability	-0.23	0.08	-0.39	-0.07

Effect	Estimate	Est. Error	l-95% CI	u-95% CI
Gender : Shift : Variability	0.04	0.06	-0.07	0.15
Order : Shift : Variability	-0.07	0.09	-0.25	0.12
Step : Gender : Order : Shift	0.09	0.06	-0.03	0.20
Step : Gender : Order : Variability	-0.05	0.06	-0.17	0.06
Step : Gender : Shift : Variability	-0.08	0.06	-0.19	0.04
Step : Order : Shift : Variability	-0.20	0.08	-0.36	-0.03
Gender : Order : Shift : Variability	0.13	0.06	0.02	0.24
Step : Gender : Order : Shift : Variability	0.04	0.06	-0.08	0.15